Bayesian Modelling of Multivariate Quantitative Traits Using Seemingly Unrelated Regressions

Claudio J. Verzilli1*, Nigel Stallard,2 and John C. Whittaker1

1Department of Epidemiology and Public Health, Imperial College London, London, United Kingdom
2Medical and Pharmaceutical Statistics Research Unit, The University of Reading, Reading, United Kingdom

We investigate a Bayesian approach to modelling the statistical association between markers at multiple loci and multivariate quantitative traits. In particular, we describe the use of Bayesian Seemingly Unrelated Regressions (SUR) whereby genotypes at the different loci are allowed to have non-simultaneous effects on the phenotypes considered with residuals from each regression assumed correlated. We present results from simulations showing that, under rather general conditions that are likely to hold in real situations, the Bayesian SUR approach has increased probability of selecting the true model compared to univariate analyses. Finally, we apply our methods to data from subjects genotyped for 12 SNPs in the apolipoprotein E (APOE) gene. Phenotypes relate to response to treatment with atorvastatin and include changes in total cholesterol, low-density lipoprotein cholesterol, and triglycerides. Missing genotype data are naturally accommodated in our Bayesian framework by imputing them using a nested haplotype phasing algorithm. Genet. Epidemiol. 28:313–325, 2005.
& 2005 Wiley-Liss, Inc.

Key words: pharmacogenetics; multiple traits; Markov chain Monte Carlo; Bayesian methods

INTRODUCTION

The focus of much recent research in human genetics has been on how to exploit the wealth of information brought about by the numerous genome sequence variation projects, important corollaries of the international Human Genome Project. The availability of ever-improving marker maps offers great promises of successfully employing association approaches to find susceptibility genes for complex traits. As the most abundant source of DNA variation, single nucleotide polymorphisms (SNPs) are arguably the most commonly used genetic marker. Extensive libraries containing hundreds of thousands of SNPs across the human genome are being compiled [International SNP Map Working Group, 2001; Thorisson and Stein, 2003; International HapMap Consortium, 2003] and made available on numerous web-based databases [Brookes, 2001; Hirikawa, 2002; Smigielski et al., 2000; Klein and Altman, 2004]. At the same time, the statistical challenges that the analysis of this large amount of data poses can be formidable. One of the main difficulties relates to the fact that we can expect only small-to-moderate effects of individual genes or interactions thereof on one or more complex traits of interest. In pharmacogenetics studies, for instance, any association between drug response and individual genetic variants might be influenced, among other factors, by variation in gene expression levels, post-translational modification of proteins, and drug dose [McCarthy and Hilfiker, 2000]. Thus, in general, the number of subjects needed to detect as statistically significant any association between gene SNPs and a complex trait at commonly used levels of power can be very large. These difficulties are compounded by the low heterozygosity of SNPs as opposed to, for instance, microsatellite markers, the low minor allele frequencies that lead to data sparseness and the large number of hypotheses tested that need to be adjusted for if the risk of finding false-positive associations is not to be increased.

Another important issue is that any association could be the result of the variants being in tight

© 2005 Wiley-Liss, Inc.
linkage disequilibrium (LD) with a causative, unassayed, polymorphic site rather than being directly involved in the etiological pathway. In such cases, haplotypic or multilocus effects may be non-negligible and an analysis based on haplotypes or models with fully-saturated genotypic effects may have more power to detect associations than a single-point, SNP-based approach [Drysdale et al., 2000; Subrahmanyan et al., 2001].

Statistical methods for the analysis of multiple SNPs in their association to complex traits have received the attention of many authors [Hoh and Ott, 2003]. Nelson et al. [2001] propose the Combinatorial Partitioning Method to find patterns or partitions of multi-locus genotypes that minimise the within-partition variability in the (quantitative) trait of interest while maximising the variability across partitions. They point out, however, that their method is valid only for exploratory purposes and further research is needed to ascertain its power and coverage of nominal levels of significance [Moore et al., 2002]. Focusing on variants at different sites within a small genetic region, Cordell and Clayton [2002] propose a stepwise regression approach to identify the relative importance of genotype effects at polymorphic sites on a binary trait. Their models allow testing of the statistical significance of additive and dominance effects of SNPs at each site as well as all possible two-way, inter-loci, interactions. Furthermore, if the phase of the genotype data is known, haplotype effects can also be tested, accounting for the possibility mentioned earlier that any association might be the results of the typed SNPs being in tight LD with causative, untyped polymorphisms.

The methods cited, as do the majority of those reported in the literature, focus on the association between genetic variants at multiple loci and a single trait. In many cases, however, data on more than one phenotype are collected. This is, however, seldom fully exploited in any analysis. In this article, we propose a Bayesian version of the Seemingly Unrelated Regressions (SUR) method [Zellner, 1962; Denison et al., 2002] which is used to model the association between a set of multivariate quantitative traits and unphased SNP genotype data. In the SUR model, the SNP alleles at the different loci are allowed to have possibly different effects on each of the phenotypes considered, with residuals from each regression assumed to be correlated. Therefore, different sets of SNPs can be selected for each trait. In simulation studies, we show that the Bayesian multivariate SUR approach leads to higher posterior probability being attached to the true generating model as opposed to Bayesian univariate analyses of the same data, especially for small sample sizes. Thus, the SUR model is more successful at detecting weak main and/or interaction effects associated with polymorphisms compared to univariate methods. Furthermore, missing genotype data are naturally accommodated in our Bayesian framework by treating them as augmented data and sampling from their conditional posterior distribution.

We apply our approach to data from subjects genotyped for 12 SNPs in the apolipoprotein (APOE) gene, including the two SNPs determining the common ε2, ε3 and ε4 alleles [Rall et al., 1982]. The aim was to relate the 12 polymorphisms to response to treatment with atorvastatin, a hypolipidemic pharmacological agent, over a 52-week period. Among the outcomes measured were changes in total cholesterol (TC), low-density lipoprotein cholesterol (LDL-C), and triglycerides (TG) since the start of the trial. In a previous study, these have been shown to be associated to the three common alleles in men but not women, with ε2 lowering cholesterol levels and ε4 raising them [Pedro-Botet et al., 2001].

The article is structured as follows. The next section describes the proposed method while introducing the notation used throughout. Results from simulation studies are presented followed by the application to genotype data in the APOE region. We end with a discussion of the advantages and disadvantages of the proposed approach.

METHODS

SEEMINGLY UNRELATED REGRESSIONS MODEL

Before presenting the Bayesian formulation of the Seemingly Unrelated Regressions (SUR) model, we illustrate the parameterisation used for the genotype model and present the generic expression for a system of SUR. Let \( y_1, y_2, \ldots, y_M \) denote measurements on \( M \) continuous phenotypes taken on \( n \) subjects, with \( y_m \) a \( n \times 1 \) vector, \( m=1, \ldots, M \). The same subjects are genotyped at a set of \( L \) SNP loci. Let \( q_l \) and \( Q_l \) indicate the alleles present at locus \( l, l=1, \ldots, L, \) so that the genotype at this locus may be \( q_l q_l \), \( q_l Q_l \) or \( Q_l Q_l \). A linear model can be used to describe the statistical association between
genotypes and phenotypes. A “saturated” model would include distinct terms for the genotypic effects at each locus and all multilocus interaction effects. In the simple case of just two loci, a saturated model contains nine parameters corresponding to the nine possible two-locus genotypes.

An alternative parameterisation of the genotypic effects, standard in quantitative genetics and adopted here, models the additive and dominance effects of the genotypes at the different loci. For generic subject $i$ and locus $l$, the additive and dominance effects are specified by introducing additional variables $x_{ilA}$ and $x_{ilD}$, say, coded $-1, 0, 1$ and $-0.5, 0.5, -0.5$, respectively, for the three possible genotypes $Q_lQ_l, q_lQ_l$, and $q_lq_l$. The coding scheme for additive term $x_{ilA}$ implies that the effect of being homozygous for allele $Q_l$ is twice that of being homozygous for $q_l$ whereas the coefficient for dominance term $x_{ilD}$ measures any deviation from the additive assumption. Thus, for generic trait $m$ and subject $i$, a two-locus model with additive and dominance effects at the first locus and only additive effects at the second plus two-way interactions between these main effects can be written as

$$y_{mi} = \beta_{0m} + \beta_{1m}x_{i1} + \beta_{1m}^a x_{i1} + \beta_{2m}x_{i2} + \beta_{3m}^a x_{i1}x_{i2} + \epsilon_{mi}$$

where $\epsilon_{mi}$ is a zero-mean normally distributed random variable, $\epsilon_{mi} \sim N(0, \sigma_m^2)$. The advantage of this parameterisation in designed experiments is that it leads to orthogonal contrasts so that parameter estimates do not depend on the current complexity of the model allowing for testing of non-nested models. However, as noted by Cordell and Clayton [2002], this property is unlikely to apply to population-based studies. This means that, in the latter case, any model comparison should be carried out between nested models, so that terms describing dominance effects at each locus should be included only if the corresponding additive terms appear in the model and interaction terms can only involve loci with main additive or dominance effects already present in the model (as in expression (1)). We will return to this point later in the article when describing our Bayesian model search strategy.

To simplify the notation, we indicate with $X_m$ the $n \times (K_m+1)$ design matrix coding for all additive, dominance and interaction effects, with $K_m$ representing the number of terms in the model for phenotype $m$ excluding the intercept term, $m=1,\ldots,M$, so that $y_m= X_m \beta_m + \epsilon_m$ with $\beta = (\beta_{0m}, \ldots, \beta_{3m})'$ a $(K_m+1) \times 1$ vector of coefficients and $\epsilon_m = (\epsilon_{m1}, \epsilon_{m2}, \ldots, \epsilon_{mn})'$ a $n \times 1$ vector of zero-mean random error terms.

Then, a system of SUR can be written as [Zellner, 1962]

$$
\begin{pmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_M \\
\end{pmatrix} = \begin{pmatrix}
  X_1 & 0 & \cdots & 0 \\
  0 & X_2 & \cdots & 0 \\
  \vdots & \vdots & \ddots & \vdots \\
  0 & 0 & \cdots & X_M \\
\end{pmatrix} \begin{pmatrix}
  \beta_1 \\
  \beta_2 \\
  \vdots \\
  \beta_M \\
\end{pmatrix} + \begin{pmatrix}
  \epsilon_1 \\
  \epsilon_2 \\
  \vdots \\
  \epsilon_M \\
\end{pmatrix}
$$

We can write (2) more compactly as

$$y = X \beta + \epsilon$$

with $y = (y_1, y_2, \ldots, y_M)$, $\beta = (\beta_1, \beta_2, \ldots, \beta_M)'$, $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_M)'$, and $X$ a $n \times K$ block-diagonal matrix, $N=nM$, $K=\sum_{m=1}^{M} K_m$. The vector of random error terms $\epsilon$ in (2) and (3) is assumed to be zero-mean normally distributed with variance-covariance matrix given by

$$\Sigma = \Sigma_e \otimes I$$

where $I$ is a $n \times n$ unit matrix and $\Sigma_e$ a $M \times M$ matrix with generic entry $\sigma_{mm'} = E(\epsilon_{mi}\epsilon_{m'i})$ for $m=1,\ldots,M$, $i=1,\ldots,n$. In this framework, expression (2) allows for a differential effect of SNP genotypes on phenotypes as well as the possibility that some loci might be associated with some of the quantitative traits modelled but not all of them. Notice that the SUR model is equivalent to the univariate, single-equation approach when $\Sigma$ is diagonal while the standard multivariate regression model is a special case of (2) corresponding to having $X_1 = X_2 = \cdots = X_M$.

In the frequentist setting, there are several ways of estimating parameters in (2); for a recent review see Foschi et al. [2003]. The statistical significance of genotypic effects can then be assessed using, for instance, a stepwise procedure based on likelihood ratios. Our approach is fully Bayesian and uses a Reversible Jump (RJ) algorithm to obtain samples from the conditional posterior distribution of models given the data. The posterior probabilities thus obtained are then used to compare the fit of different models via Bayes factors. There are two main advantages with this approach. First, variable selection using the
Bayesian paradigm allows us to make probabilistic statements about the relative plausibility of different models. There is also some evidence showing that the Bayesian approach achieves better coverage of the nominal levels of significance than a frequentist stepwise selection of important predictors [Viallefont et al., 2001]. Second, the Bayesian framework naturally handles incomplete genotype data by treating missing data as extra parameters and sampling from their conditional posterior distribution. In particular, in the application to the APOE data, we describe a nested haplotype phasing algorithm that is used for this purpose, assuming that genotype data are missing at random in the sense of Rubin [1976]. In this way, the extra uncertainty due to data incompleteness is taken into account, while making a fuller use of the available data.

BAYESIAN SUR

Before proceeding, we need to introduce further notation. We indicate with \( \mathcal{M} = \{ \mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m \} \) the set of all possible models for a given maximum degree of interaction between polymorphic loci. For example, in the case of two loci, a model for the full genotypic effects on trait \( m \) can be written as

\[
y_{mi} = \beta_{0m} + \sum_{l=1}^{2} \sum_{j=d,a} \beta_{lj} x_{li} + \sum_{j=ad} \beta_{jj} x_{j} + e_{mi}.
\]

Next, consider a random vector \( \theta \) that completely specifies which genotype effects characterize each model. That is, \( \theta \) is a random vector of varying dimension and contains the indexes of the columns of the matrix \( X_{full} \) that specify the various nested models, with \( \theta_{\mathcal{M}_0} = 1 \) a vector with all entries one. Then, given the data \( D \), comparison between any two models \( \mathcal{M}_0 \) and \( \mathcal{M}_1 \) can be based on the Bayes factor

\[
BF_{10} = \frac{p(\mathcal{M}_1|D)}{p(\mathcal{M}_0|D)} = \frac{p(\theta_{\mathcal{M}_1}|D)}{p(\theta_{\mathcal{M}_0}|D)} \frac{p(\mathcal{M}_0)}{p(\mathcal{M}_1)}
\]

where, for instance, \( \theta_{\mathcal{M}_0} \) is the value of \( \theta \) that corresponds to a null model \( \mathcal{M}_0 \). In the next subsection, we outline a hybrid sampling strategy used to obtain samples from \( p(\theta_{\mathcal{M}_1}|D) \) which are then used to approximate (5) and discuss the choice of prior distributions for parameters \( \beta, \Sigma \) and \( \theta \). A more detailed description of the algorithm used is given in the Appendix.

PRIOR DISTRIBUTIONS AND POSTERIOR SAMPLING

The joint distribution of the phenotypes, and model parameters can be written as

\[
p(Y, \beta, \Sigma, \theta, X) = p(Y|\beta, \Sigma, \theta, X)p(\beta, \Sigma, \theta)
\]

where, from (3), the likelihood on the right-hand side is

\[
p(Y|\beta, \Sigma, \theta, X) = |\Sigma|^{-\frac{N}{2}}(2\pi)^{-\frac{M}{2}}\exp\left\{ -\frac{1}{2} (Y - X\beta)^\top \Sigma^{-1} (Y - X\beta) \right\}
\]

Following Denison et al. [2002], we assume independent prior distributions for the vector of regression coefficients \( \beta \) and the precision matrix \( \Sigma^{-1} \) with \( \beta \sim N(0, \Omega^{-1}) \), \( \Sigma^{-1} \sim Wi(z, S) \) and partition the joint prior as

\[
p(\beta, \Sigma^{-1}, \theta) = p(\beta|\theta)p(\Sigma^{-1}p(\theta)
\]

where Wi(zS) is a Wishart distribution with parameters \( z \), a scalar, and \( S \) a \( M \times M \) positive-definite matrix. For the reasons mentioned in the previous section, we adopt a model space prior \( p(\theta) \) that imposes certain constraints on the terms entering each model [Chipman, 1996]. Namely, dominance terms are only allowed if the corresponding additive terms are already in the model. Similarly, interactions between loci may be included only if the corresponding main effects are present in the model. Thus, in the case of two loci in expression (4), the model space prior factors as

\[
p(\theta) = p(\theta_0)p(\theta_1)p(\theta_{12})p(\theta_{12})\prod_{l=ad} p(\theta_{l_1l_2}|\theta_{l_1}, \theta_{l_2})
\]

and

\[
p(\theta_{l_1l_2}|\theta_{l_1}, \theta_{l_2}) = \begin{cases} 0.5 & \text{if } (\theta_{l_1}, \theta_{l_2}) = (1, 1) \\ 0 & \text{otherwise} \end{cases}
\]

A similar dependence applies to \( p(\theta_{l_1}|\theta_{l_0}) \) for the dominance terms.

The independent priors assumption (6) makes it easy to sample from the relevant conditional posterior distributions. In particular, a hybrid sampling strategy is used that alternates an RJ step for updating \( \theta \) with draws from full conditional distributions for \( \beta \) and \( \Sigma^{-1} \). Specifically, to update \( \theta \), a move is proposed with equal probability to modify the current model by adding an explanatory variable, deleting an explanatory variable or replacing a term that is currently included in the model with a new one. All these
moves have to satisfy the restriction that the resulting model nests or is nested in the current one for the reasons mentioned above. Any move from $\theta$ to $\theta'$ is then accepted with probability

$$\min \left\{ 1, \frac{p(y | \Sigma, \theta') p(\theta') q(\theta | \theta')} {p(y | \Sigma, \theta) p(\theta) q(\theta | \theta')} \right\}$$

where $p(y | \Sigma, \theta)$ is the partial marginal likelihood (integrating over the vector of coefficients $\beta$) and $q(\cdot | \cdot)$ is a proposal distribution that is non-symmetric due to the constraints imposed on the model space.

**RESULTS**

**SIMULATION STUDIES**

In this section, we present results from simulation studies comparing the Bayesian SUR approach with univariate Bayesian analyses in terms of posterior probabilities of selecting the genotype effects that are known to be non-zero. Since the generating model is known, the performance of the two approaches can thus be compared objectively.

Various scenarios are considered that differ in the size of the generated sample and degree of correlation among residuals. In all cases, the total number of loci considered is fixed at 12 and we assume that we are interested in their potential association with three quantitative traits. SNP genotype data were generated by sampling with replacement from the APOE data set mentioned above and described more thoroughly in the next section. Under the first scenario, data are generated from the SUR model

$$y_1 = -10 + 1.1x_1 + 0.9x_3 + \epsilon_1 \quad (7)$$
$$y_2 = -10 + 0.9x_1 + 0.95x_3 + 0.75x_5 + \epsilon_2 \quad (8)$$
$$y_3 = -10 + 0.9x_4 + 0.8x_8 + 0.8x_{12} + \epsilon_3 \quad (9)$$

so that, for instance, we have an additive effect of locus 1 on $Y_1$ and $Y_2$ and additive and dominance effects of locus 5 on $Y_2$. Residuals are normally distributed $N(0, \Sigma_e)$ and strongly positively correlated with

$$\Sigma_e = \begin{pmatrix} 2 & 0.92 & 0.8 \\ 0.92 & 2 & 0.88 \\ 0.8 & 0.88 & 2 \end{pmatrix}. \quad (10)$$

Table I shows quartiles of posterior probabilities associated with the true model for the two methods considered. In order to compare the univariate and SUR approaches, for each of the three outcomes, these probabilities are estimated as the number of times the relevant subvector $\theta_{(m)}$ corresponding to (7)-(9), appears in $T$ draws from $p(\theta | D)$, that is

$$p(\theta_{(m)} | D) \approx \frac{1}{T} \sum_{i=1}^{T} I(\theta_{(m)}^{(i)} = \theta_{(m)}) \quad (11)$$

where $I(\cdot)$ is the indicator function, which is one if its argument is true, $m=1,2,3$. Results corresponding to different sizes of the generated data sets are reported where, for each combination of method and size of data set, quartiles are constructed from 200 independent replications of the Markov chain. Each chain was run for 200,000 iterations and a sample of $T=10,000$ draws was retained every 10 steps after a burn-in run of 100,000 iterations. The latter was deemed sufficient from inspection of the Markov chains for the elements of $\Sigma$. For the SUR model we set $\omega=0.01I$, $S=0.01I$ and $\alpha=0.01$, whereas in the univariate case, the only change involves the prior on the single precision (inverse variance) term for the residuals, which is Gamma(0.01,0.01).

As can be seen from Table I, for each of the three phenotypes considered, the Bayesian SUR model leads to higher median probabilities being attached to the loci that are truly causative. This

<table>
<thead>
<tr>
<th>Size of data set</th>
<th>100</th>
<th>300</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q1</td>
<td>Q2</td>
<td>Q3</td>
</tr>
<tr>
<td>Univariate analyses</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y_1$</td>
<td>0.03</td>
<td>0.09</td>
<td>0.32</td>
</tr>
<tr>
<td>$y_2$</td>
<td>0.03</td>
<td>0.05</td>
<td>0.17</td>
</tr>
<tr>
<td>$y_3$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>SUR model</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y_1$</td>
<td>0.48</td>
<td>0.68</td>
<td>0.79</td>
</tr>
<tr>
<td>$y_2$</td>
<td>0.57</td>
<td>0.81</td>
<td>0.89</td>
</tr>
<tr>
<td>$y_3$</td>
<td>0.09</td>
<td>0.36</td>
<td>0.54</td>
</tr>
</tbody>
</table>

*Data are simulated from the model given by expressions (7)–(9) and (10). Values refer to 200 replications.*
advantage becomes more apparent for smaller sizes of the generated data sets. Interestingly, because of the low heterozygosity at loci 8 and 12 in the real data from which the simulated data are sampled, the univariate analysis for \( y_3 \) fails completely to select the true model for sample size \( N=100 \) and does very poorly even for \( N=300 \) whereas the SUR model leads to a higher probability of selecting the true model.

The univariate and SUR methods were also contrasted in terms of Bayes factors comparing the true model versus the most frequently visited model among the rest, using (11) to approximate (5). This gives a measure of the confidence with which the true model is chosen in each case. Results are shown in Table II. It can be seen that, for small values of \( N \), it is very difficult to successfully select the true model using a univariate analysis of the data. This is in contrast with the SUR approach, which, at least for the first two phenotypes, convincingly favours the correct model over any other model.

Finally, we also considered quartiles of the number of models visited, which showed that the SUR approach tends to favour the model generating the data more decisively than the univariate approach, thus leading to a smaller probability of accepting a model other than the true one (data not shown).

The advantage of the multivariate SUR approach in selecting the true causative loci decreases as the inter-residual correlations get smaller, as shown under our second simulation scenario. Here, the true effects are still given by expressions (7)–(9) while the matrix \( \Sigma_e \) of correlation between residuals is now

\[
\Sigma_e = \begin{pmatrix}
2 & 0.5 & 0.5 \\
0.5 & 2 & 0.45 \\
0.5 & 0.45 & 2
\end{pmatrix}
\]  \hspace{1cm} (12)

Results are shown in Tables III and IV. The two methods perform rather poorly for \( N=100 \) while the multivariate SUR model has higher probability of selecting the causative loci for \( y_3 \) compared to the univariate analysis for \( N=300 \). As in the previous simulation setting, we found that the number of models visited is smaller under the

### Table II. Quartiles of Bayes factors comparing the true model versus the most probable among the rest of the models visited for different sizes of data sets

<table>
<thead>
<tr>
<th>Size of data set</th>
<th>100</th>
<th>300</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q1</td>
<td>Q2</td>
<td>Q3</td>
</tr>
<tr>
<td><strong>Univariate analyses</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( y_1 )</td>
<td>0.6</td>
<td>1.0</td>
<td>6.9</td>
</tr>
<tr>
<td>( y_2 )</td>
<td>1.5</td>
<td>4.0</td>
<td>8.5</td>
</tr>
<tr>
<td>( y_3 )</td>
<td>0.0</td>
<td>0.1</td>
<td>0.6</td>
</tr>
<tr>
<td><strong>SUR model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( y_1 )</td>
<td>1.9</td>
<td>7.4</td>
<td>19.0</td>
</tr>
<tr>
<td>( y_2 )</td>
<td>13.2</td>
<td>67.0</td>
<td>194.5</td>
</tr>
<tr>
<td>( y_3 )</td>
<td>1.0</td>
<td>2.8</td>
<td>8.6</td>
</tr>
</tbody>
</table>

*Data are simulated from the model given by expressions (7)–(9) and (10). Values refer to 200 replications.

### Table III. Quartiles of posterior probabilities associated with the true model for different sizes of data sets

<table>
<thead>
<tr>
<th>Size of data set</th>
<th>100</th>
<th>300</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q1</td>
<td>Q2</td>
<td>Q3</td>
</tr>
<tr>
<td><strong>Univariate analyses</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( y_1 )</td>
<td>0.17</td>
<td>0.29</td>
<td>0.48</td>
</tr>
<tr>
<td>( y_2 )</td>
<td>0.02</td>
<td>0.10</td>
<td>0.33</td>
</tr>
<tr>
<td>( y_3 )</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td><strong>SUR model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( y_1 )</td>
<td>0.25</td>
<td>0.42</td>
<td>0.65</td>
</tr>
<tr>
<td>( y_2 )</td>
<td>0.06</td>
<td>0.18</td>
<td>0.47</td>
</tr>
<tr>
<td>( y_3 )</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
</tr>
</tbody>
</table>

*Data are simulated from model given by expressions (7)–(9) and (12). Values refer to 200 replications.
SUR approach than under the univariate analyses although the differences are not as large as those under scenario 1. Finally, the two approaches yield the same results apart from simulation noise when decreasing the true inter-residuals correlations further (results not shown).

APOE GENOTYPE DATA

In this section, the Bayesian SUR approach is applied to data from 327 subjects genotyped for 12 SNPs in the APOE locus. There is a wealth of literature reporting associations between variants in the APOE region and metabolic regulation of cholesterol as well as Alzheimer’s disease that mainly focus on the three common variants (for reviews see Eichner et al. [2002] and Tanzi and Bertram [2001] whereas recent findings on the molecular mechanisms underlying the cholesterol-AD connection are reviewed in Puglielli et al. [2003]). However, some authors have recently argued that there could be important substructures within the ε2, ε3, and ε4 alleles and further investigation is needed in order to characterise the full extent of allelic heterogeneity in the APOE gene [Fullerton et al., 2000; Nickerson et al., 2000]. The objective of the study reported here was to investigate any association between the polymorphic sites and response to treatment with atorvastatin at 10 mg/day for 52 weeks in terms of changes in total cholesterol (TC), low density lipoprotein cholesterol (LDL-C), and triglycerides (TG). In the analysis, we use logarithm-transformed triglycerides values as the empirical distribution of the untransformed values was highly skewed. Data on non-genetic variables were also available, namely each patient’s age, sex, and body mass index, and were included in our analysis allowing for interactions with additive genetic effects only. For some patients, genotype data at some loci are missing. In this setting, the mechanism driving the missing data process is likely to be at random [Rubin, 1976]; that is, for any subject, the probability of not observing his or her genotype data at any locus does not depend on the true, unobserved genotype at that locus. Therefore, the missing data mechanism could be ignored without affecting the correctness of the results at more cost in efficiency. However, we make a fuller use of the available data by devising an imputation scheme that exploits the genetic nature of the data at hand. In particular, we use a nested haplotype phasing algorithm to impute the missing genotype data. Various authors have recently proposed Bayesian approaches to haplotype reconstruction from population data [Stephens et al., 2001; Niu et al., 2002; Lin et al., 2002]. A review of these methods is given in Stephens and Donnelly [2003]. Our approach uses the “naïve” Gibbs sampler described in appendix A of Stephens et al. [2001] with the adaptation for missing genotype data given in Lin et al. [2002]. This entails adding an extra step to the algorithm outlined in the previous section. Informally, given current values of β, Σ, and θ, for a generic subject with missing genotype data, if H_i are his/her current reconstructed haplotypes compatible with his/her multilocus genotype, newly sampled haplotypes H_0i are accepted with probability given by

\[
\min \left\{ 1, \frac{p(y|X(H_i), \beta, \Sigma, \theta)}{p(y|X(H_0), \beta, \Sigma, \theta)} \right\}.
\]

In the previous expression, we use X(H_i) to highlight the fact that the reconstructed haplotypes uniquely determine the subject’s missing genotypes or the entries of the covariate matrix X. Thus, although our model is a model for geno-

|TABLE IV. Quartiles of Bayes factors comparing the true model versus the most probable among the rest of the models visited for different sizes of data sets^a |
|---|---|---|---|---|---|---|---|---|
|Size of data set | 100 | 300 | 500 |
| | Q1 | Q2 | Q3 | Q1 | Q2 | Q3 | Q1 | Q2 | Q3 |
| Univariate analyses |  |  |  |  |  |  |  |  |  |
| y1 | 1.0 | 1.75 | 8.5 | 2.2 | 7.1 | 19.4 | 2.1 | 8.0 | 21.4 |
| y2 | 3.2 | 4.0 | 23.0 | 19.0 | 47.5 | 109.6 | 27.5 | 98.4 | 104.3 |
| y3 | 0.0 | 0.2 | 1.5 | 1.0 | 3.7 | 13.0 | 2.0 | 4.4 | 6.0 |
| SUR model |  |  |  |  |  |  |  |  |  |
| y1 | 2.1 | 5.1 | 12.4 | 4.0 | 8.9 | 23.5 | 3.1 | 10.4 | 22.2 |
| y2 | 4.0 | 6.2 | 25.3 | 9.2 | 64.4 | 159.7 | 130.9 | 156.6 | 232.3 |
| y3 | 0.0 | 0.5 | 2.5 | 2.6 | 9.8 | 19 | 5.1 | 5.6 | 8.8 |

^aData are simulated from (7)–(9) and (12). Values refer to 200 replications.
typic effects, we use the nested haplotype phasing as a convenient way of imputing the missing data. A detailed description of this additional step is given in the Appendix.

The results presented refer to a sample of 20,000 models obtained after a burn-in run of 20,000 iterations with a total chain length of 500,000 iterations. Figure 1 shows, for each of the three outcomes considered, the probability that the genotype effects at each locus are different from zero conditional on the observed data, $\Pr(\beta_i \neq 0|D)$, estimated as the proportion of sampled models containing the corresponding terms. For each of the 12 loci considered, the additive and dominance effects are plotted side-by-side. Also shown are analogous probabilities for the non-genetic predictors age, BMI and sex. Similar plots for two-locus interaction terms did not indicate any important effects and are not presented. None of the polymorphisms considered has any important effect on all three outcomes simultaneously, apart from locus L650, where the probability of effect is small, and L4870 is one of the polymorphisms involved in the APOE major alleles. Among the non-genetic effects considered, sex is an important predictor of changes in LDL-C.

Figure 2 shows posterior densities of regression coefficients for those predictors that appear to be important in Figure 1. These were obtained from the subset of sampled models that contained the corresponding terms (shown as proportions of the total by the heavily shaded parts of the horizontal bars above each graph). Males and subjects homozygote for the mutant allele at locus L4870 show a smaller decrease in TC and LDL-C over the 52-week treatment period compared to females and heterozygote or homozygote wild type. Also, we report in Table V those models with posterior frequencies greater than or equal to 5%: for changes in triglycerides and TC, the null model containing just the intercept term is the first and second most frequently visited one, respectively.

Table VI shows, for each of the three phenotypes considered, the Bayes factors for the composite hypothesis of any genetic effect being different from zero versus the null hypothesis of no genetic effects. Overall, conditional on the observed data, there is no evidence of any genetic effects on the outcomes considered. We also report Bayes factors testing the hypothesis of a non-zero effect at locus L4870 (which is one of the loci involved in the APOE major alleles), which again show no real evidence for important effects in this data set. With regard to these findings, it should be noted that, despite the consistent association of the APOE locus with LDL-C concentrations in the general population, results from lipid pharmacogenetic studies are less clear-cut [Ordovas, 2004].

![Figure 1](image-url)
Finally, the posterior densities of the inter-residual correlations are shown in Figure 3. Whereas the correlations between regression residuals for TC and triglycerides and for triglycerides and LDL-C appear to be modest, there is strong positive correlation between the residuals of the TC and LDL-C regressions.

**DISCUSSION**

In this article, we have described the use of Bayesian Seemingly Unrelated Regressions to model multivariate quantitative traits in genetic association studies. In simulation studies, the

---

**TABLE V. SUR models with posterior probability greater than or equal to 5%**

<table>
<thead>
<tr>
<th>Trait</th>
<th>Model</th>
<th>Posterior model probability (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total cholesterol</td>
<td>L4870a</td>
<td>23.3</td>
</tr>
<tr>
<td></td>
<td>Intercept only</td>
<td>20.3</td>
</tr>
<tr>
<td></td>
<td>L650a</td>
<td>6.4</td>
</tr>
<tr>
<td>Triglycerides</td>
<td>Intercept only</td>
<td>36.2</td>
</tr>
<tr>
<td></td>
<td>L650a</td>
<td>8.6</td>
</tr>
<tr>
<td></td>
<td>L650d</td>
<td>5.0</td>
</tr>
<tr>
<td>LDL cholesterol</td>
<td>L4870a</td>
<td>Gender</td>
</tr>
<tr>
<td></td>
<td>Intercept only</td>
<td>Gender</td>
</tr>
<tr>
<td></td>
<td>L650a</td>
<td>Gender</td>
</tr>
<tr>
<td></td>
<td>L650a</td>
<td>Gender:L650a</td>
</tr>
<tr>
<td></td>
<td>L650d</td>
<td>Gender:L650a</td>
</tr>
<tr>
<td></td>
<td>Intercept only</td>
<td>Gender:L650a</td>
</tr>
</tbody>
</table>

*For each of the three phenotypes, marginal probabilities are shown. Values were estimated from a posterior sample of 10,000 models.

**TABLE VI. Bayes factors using the SUR model**

<table>
<thead>
<tr>
<th>Trait</th>
<th>Any genetic effect vs. no genetic effects</th>
<th>Additive effect at locus L4870</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total cholesterol</td>
<td>&lt;0.01</td>
<td>0.86</td>
</tr>
<tr>
<td>Triglycerides</td>
<td>&lt;0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>LDL</td>
<td>&lt;0.01</td>
<td>1.10</td>
</tr>
</tbody>
</table>

---
The proposed approach leads to increased probability of selecting the true models compared to Bayesian univariate methods, whenever the residuals from the univariate models are highly correlated. The latter circumstance is expected to hold with many real datasets, as it is quite common in association studies to collect data on two or more related quantitative phenotypes. There could, therefore, be advantages in adopting a multivariate approach in such cases whereas, as expected, the univariate and multivariate SUR models give similar results when inter-residual correlations are low.

The SUR model offers a flexible way of parameterising genotypic effects. In particular, the method is suited to cases where one is not prepared to assume that polymorphic loci will have an effect on all the traits considered simultaneously, this being the underlying assumption in the more common multivariate regression model. At the same time, the SUR model encompasses the multivariate regression model as a particular case as discussed in Methods. The Bayesian formulation then has several advantages. Namely, it is straightforward in this framework to accommodate missing genotype data by iteratively imputing them using, for instance, the nested haplotype phasing step described in the Appendix. It is easy to envisage wider applications of this approach as, for example, in haplotype-based association studies with unphased genotype data. In addition, complex hypotheses can be tested using Bayes factors estimated from the sampled models. Finally, if prior information is available this can be easily incorporated. We may, for instance, modify the probability of inclusion for certain model terms or exclude certain interaction terms if this is corroborated by prior knowledge. The Bayesian SUR model presented here has been implemented as an R package [R Development Core Team, 2004], which is available on request from the first author; the model definition requires no extra effort on the part of the user than the univariate counterparts.

A difficulty with our approach, which is common to all model-based multilocus analysis methods, is how to deal with very many loci. In such cases, a model search strategy based on uniform priors might not be effective in traversing the space of possible models as this can be of high dimension, especially if between-locus interaction terms are to be included. A possible solution would be to modify the ratio of move proposals at each iteration of the Markov chain, by giving more weight to interaction terms between loci whose main effects appear to be important in the current sample of models. A drawback of this strategy is the risk of missing important interaction effects between loci that, individually, have no important main effects.

The choice of the model space prior distribution requires careful consideration. We have used a uniform prior that incorporates relations between
 predictors by imposing certain conditional dependences. An alternative would be to consider a uniform prior on the complexity or the number of terms entering each model with models of equal complexity having the same a priori probability. A related problem is how to deal with potentially high correlation between predictors as a result of SNPs being in tight LD. This may have undesirable consequences on model search as the Markov chain will tend to mix slowly over clusters of similar models biasing posterior model probabilities away from any “good” models, if uniform model priors are used. This will not, however, affect model comparisons based on Bayes factors [Chipman et al., 2001].

The application of the Bayesian SUR model to the SNP genotype data in the APOE region showed no evidence of genetic effects on any of the three traits considered (Table VI). However, there appears to be a non-zero additive effect on changes in TC and LDL-C at a locus involved in one of the APOE major alleles (L4870) although the corresponding Bayes factor was not large (Table VI). One of the two-way interaction terms between loci and between additive effects at each locus and non-genetic variables had a large posterior probability of being greater than zero. There was, however, a significant gender effect on changes in LDL-C over the treatment period with females showing a sharper decrease.

ACKNOWLEDGMENTS

The authors thank Daniel Chasman for helpful comments on the manuscript. C.J.V. was supported by a postdoctoral research grant from the Wellcome Trust (GR 068213).

REFERENCES


APPENDIX

Let \( y_1, y_2, \ldots, y_M \) denote measurements on \( M \) continuous phenotypes taken on \( n \) subjects, with \( y_m \) a \( n \times 1 \) vector, \( m = 1, \ldots, M \). Indicate with \( G = (G_1, \ldots, G_m) \) the set of multilocus, possibly incomplete genotypes with \( G_i \in \{ q_i q_i', q_i Q_i', Q_i q_i', Q_i Q_i' \} \), \( i = 1, \ldots, n \), \( l = 1, \ldots, L \). \( H = (H_1, \ldots, H_n) \) is the corresponding set of (unknown) haplotype pairs where \( H_i = (h_i, h_i') \). Notice that these are treated as latent variables and we sample from their posterior distribution for the purposes of imputing unknown genotype data. We assume independent priors for \( \beta, \Sigma^{-1}, \theta \) and \( H \) and partition the joint prior as

\[
p(\beta, \Sigma^{-1}, \theta, H) = p(\theta) p(\Sigma) p(\beta) p(H)
\]

\[
= N(0, \Omega^{-1}) Wi(\alpha, S) p(\theta) Di(\gamma).
\]

In the expression above, we have assumed a uniform distribution for \( \theta \), \( Wi(\alpha, S) \) is a Wishart distribution with parameters \( \alpha \), a scalar, and \( S \) a \( M \times M \) positive-definite matrix and \( Di(\gamma) \) is a Dirichlet prior on the \( 2^d = D \) possible haplotype assignments with \( d \) the number of segregating sites in the dataset. The choice of model space prior \( p(\theta) \) is discussed in Methods. We set \( \gamma_1 = \gamma_2 = \cdots = \gamma_p = 1 \), which corresponds to a uniform prior density on the unit simplex of dimension \( D - 1 \) [Denison et al., 2002]; that is, we assume that all haplotype assignments are, a priori, equally probable. This choice of prior for haplotype assignments leads to the simplest version of

algorithm 2 in the appendix of Stephens et al. [2001], which has been shown to give phasing accuracy that is similar to that of the EM algorithm. A hybrid algorithm is then used to obtain samples from the target posterior density \( p(\beta, \Sigma^{-1}, \theta, H | y) \). In particular, starting from a random imputation of missing genotype data using site-specific frequencies, a random guess of \( H \) and some initial values \( \{ \beta^{(0)}, \Sigma^{(0)}, \theta^{(0)} \} \), one iterates between the following steps:

- Draw new values of \( \beta \) from \( N(\beta^*, A) \) with \( A = (\Omega + X' \Sigma^{-1} X)^{-1} \) and \( \beta^* = A X' \Sigma^{-1} y \);
- Draw a new value of \( \Sigma^{-1} \) from \( Wi(\alpha + 1/n, S + R'R) \) where \( R \) is the \( n \times M \) matrix of residuals given by

\[
R = \begin{pmatrix}
\epsilon_{11} & \epsilon_{12} & \cdots & \epsilon_{1M} \\
\epsilon_{21} & \epsilon_{22} & \cdots & \epsilon_{2M} \\
\vdots & \vdots & \ddots & \vdots \\
\epsilon_{n1} & \epsilon_{n2} & \cdots & \epsilon_{nM}
\end{pmatrix}
\]

and \( \epsilon_{im} \) is the error defined as \( \epsilon_{im} = y_{im} - x_{im}' \beta^{(t)} \);
- Draw a new value of the parameter vector \( \theta \), using a reversible Metropolis step. Specifically, with equal probability, a move is proposed to modify the current model by adding an explanatory variable, deleting an explanatory variable or replacing a term that is currently included in the model with a new one. All these moves have to satisfy the restriction that the resulting model nests or is nested in the current one for the reasons mentioned in Methods. Any move from \( \theta \) to \( \theta' \) is then accepted with probability given by

\[
\zeta = \min \left\{ \frac{1}{p(y | \Sigma, \theta, \theta')} \frac{p(\theta')}{p(\theta)} q(\theta' | \theta) \right\}
\]

\[
= \min \left\{ \frac{1}{p(y | \Sigma, \theta') \frac{p(\theta')}{p(\theta)} Z} \right\}
\]

\[
= \min \left\{ \frac{\Omega^{1/2} | \lambda |^{1/2} \exp(-\frac{1}{2} \mathbf{d}')}{\Omega^{1/2} | \lambda |^{1/2} \exp(-\frac{1}{2} \mathbf{d})} Z \right\}
\]

where in the expressions above, \( q(\cdot | \cdot) \) is a proposal distribution that is asymmetric because of the restrictions on the model space and

\[
a = \text{tr}(S \Sigma_{\zeta}^{-1}) + y' \Sigma^{-1} y - \beta^* \Lambda^{-1} \beta^*.
\]

- Update the haplotype reconstruction of each individual in a random order, using a different order at each iteration of the Markov chain. In particular, for individual \( i \), remove his or her current haplotype pair from \( H \) and, from \( H_{-i} \),

form a list \( h=(h_1,\ldots,h_R) \) of distinct haplotypes with corresponding counts \( c=(c_1,\ldots,c_R) \) of the number of times they appear in \( H_i \). Calculate the vector \( p=(p_1,\ldots,p_R) \) where, for \( r=1,\ldots, R, \) \( p_r=0 \) if no haplotypes can be found in \( h \) that are compatible with the observed genotype data for subject \( i \) \( (g_i^{obs}) \), \( p_r = \sum_{s=1}^S (c_r + 1)(c_s + 1) - 1 \) if \( h_r \) is compatible with \( g_i^{obs} \) and there are \( S \) possible complement haplotypes in \( h \) and \( p_r=c_r \), otherwise. Then, if \( k \) is the number of sites at which the individual is heterozygous or has missing genotype data, with probability \( 2^k/(\sum_r c_r + 2^k) \) reconstruct his or her haplotypes completely at random (by randomly choosing the phase at each observed heterozygous site and randomly imputing missing genotypes using locus specific frequencies). Otherwise, update the haplotype reconstruction \( h_{r1} \) to \( h_r \) with probability \( c_r/\sum_r c_r \). This automatically updates the complement haplotype \( h_{r2} \) at observed heterozygous sites. For loci with missing data, if there are \( U \) possible complements of the newly proposed haplotype \( h_{r1} \) in the list \( h \), then choose \( h_{r2} \) with probability \( c_u/\sum_u c_u \). If no complement can be found in \( h \), then unknown sites for both \( h_{r1} \) and \( h_{r2} \) are reconstructed randomly. Finally accept the proposed haplotype pair \( H_i \) with probability

\[
Q = \min \left\{ 1, \frac{p(y|X(H_i), \beta, \Sigma, \theta)}{p(y|X(H_i), \beta, \Sigma, \theta)} \right\}.
\]

where we highlight the fact that the newly proposed haplotypes reconstruction may change the entries of the design matrix \( X \) for subject \( i \) at loci with missing data. Also, the acceptance ratio above only applies to subjects with missing genotype data, as in all other cases we have \( Q=1. \)