SOM: Bayesian analysis of genetic association across tree-structured routine healthcare data in the UK Biobank

Adrian Cortes, Calliope A. Dendrou, Allan Motyer, Luke Jostins, Damjan Vukcevic, Alexander Dilthey, Peter Donnelly, Stephen Leslie, Lars Fugger & Gil McVean

May 18, 2017

Contents

1 Model description

Consider a data set of N individuals, each of which is annotated with a series of categorical observations, which are themselves organised in a hierarchical structure reflecting increasing levels of resolution; *i.e.,* annotations are associated with nodes within a classification tree. Observations may be made at both terminal and internal nodes depending on resolution. We define an indicator, Z_{ij}, for the presence of at least one annotation $j \in T$ for individual i; where T is the set of all annotations (organised as a tree). We model the distribution of Z_{ij}, conditional on the genotype of individual i at variant s, $G_{is} \in \{0, 1, 2\}$, using a logistic model, with an intercept (β_j^0) and separate coefficients for the heterozygous (β_j^1) and homozygous (β_j^2) states:

$$Y_{ijs} = \beta_j^0 + \beta_j^1 * I(G_{is} == 1) + \beta_j^2 * I(G_{is} == 2), \tag{1}$$

$$P(Z_{ij} = 1|Y_{ijs}) = \frac{e^{Y_{ijs}}}{(1 + e^{Y_{ijs}})}. \tag{2}$$

To model the correlation structure of the genetic coefficients across categories, we allow the coefficient pair $\{\beta^1, \beta^2\}$ to evolve down the tree in a Markovian fashion. The coefficients attached to a parent node x can either be inherited by a child node y, with probability $e^{-\theta}$, or can transition to a new pair of values, with probability $1 - e^{-\theta}$. With probability $1 - \pi_1$ the new values are $\{0, 0\}$, and with probability π_1 they are drawn from a joint prior on β^1 and β^2, $f(\beta^1, \beta^2)$. The state of the ancestral node in the tree is drawn from the stationary distribution of this process; *i.e.,* $\{0, 0\}$ with probability $1 - \pi_1$ or from $f(\beta^1, \beta^2)$ with probability π_1. We use a non-local prior for $f(\beta^1, \beta^2)$, such that:

$$f(\boldsymbol{\beta}) = N_2(\mathbf{0}, \Sigma) * |\boldsymbol{\beta}|^k * e, \tag{3}$$

with

$$\Sigma = \begin{bmatrix} \sigma_1^2 & r\sigma_1\sigma_2 \\ r\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \tag{4}$$

and,

$$e = \begin{cases} 0.10, & \text{if } \beta_1 * \beta_2 < 0 \\ 0.10, & \text{if } \beta_1 > \beta_2 \\ 1, & \text{otherwise.} \end{cases} \tag{5}$$

The density of the prior is illustrated in Supplementary Figure 1. The mixture prior on the coefficients (including the point mass at 0) is referred to as $f^*(\beta^1, \beta^2)$. Unless stated otherwise we use parameter values $\pi_1 = 0.001$, $\theta = 1/3$, $\sigma_1 = 2$, $\sigma_2 = 4$, $k = 1/2$ and $r = 0.5$, throughout. The unknown intercept term β_j^0 is chosen, for each value of $\{\beta_j^1, \beta_j^2\}$, to maximise the likelihood. That is,

$$L(\beta_j^1, \beta_j^2 | \boldsymbol{Z_j}) = max_{\beta_j^0} L(\beta_j^0, \beta_j^1, \beta_j^2 | \boldsymbol{Z_j}), \tag{6}$$

where $\boldsymbol{Z_j}$ is $\{Z_{1j}, Z_{2j}, \ldots, Z_{Nj}\}$.

The joint distribution of different annotations across individuals has substantial non-independence. For example, the same individual might be recorded as having different subtypes of a disorder on separate visits to a hospital, the recording of a specific disease subtype will mean that other subtypes are less likely to be recorded for the same individual and a disease may have multiple diagnostic features. However, rather than attempt to capture such structure, we make the approximation that annotations are independent conditional on an individual's genotype (and evaluate the impact of this approximation). Hence, the likelihood for a given vector of $\{\beta^1, \beta^2\}$ values across annotations, $\boldsymbol{\beta}$, is given by the product over all nodes in the tree T:

$$L(\boldsymbol{\beta} | \boldsymbol{Z}) = \prod_{j \epsilon T} L(\boldsymbol{\beta_j} | \boldsymbol{Z_j}), \tag{7}$$

where $\boldsymbol{\beta_j} = \{\beta_j^1, \beta_j^2\}$. The prior density for $\boldsymbol{\beta}$ can be calculated by considering the state of the ancestral node, A, and all transitions between parent and child nodes:

$$P(\boldsymbol{\beta}) = p(\boldsymbol{\beta}_A) \prod_{p,c} q(\boldsymbol{\beta}_p, \boldsymbol{\beta}_c), \tag{8}$$

where $q(\boldsymbol{\beta}_p, \boldsymbol{\beta}_c)$ is the transition probability between the coefficients of the parent and child nodes. Because of the structure of the model, it is possible to sum the likelihood over all possible values of $\boldsymbol{\beta}$ using dynamic programming. To achieve this, for each node j we calculate an integrated likelihood

$$L_j = \int P_j(D|\boldsymbol{\beta}) f^*(\boldsymbol{\beta}) d\boldsymbol{\beta}, \tag{9}$$

where $P_j(D|\boldsymbol{\beta})$ is given by the likelihood function in Equation 6 when j is a terminal node, or by

$$P_j(D|\boldsymbol{\beta}) = \prod_{i \epsilon \gamma(j)} [e^{-\theta} P_i(D|\boldsymbol{\beta}) + (1 - e^{-\theta}) L_i], \tag{10}$$

when j is an intermediate node. Here, $e^{-\theta}$ is the stay transition probability in $\boldsymbol{\beta}$ and $(1 - e^{-\theta})$ is the switch transition probability in $\boldsymbol{\beta}$, which results in uncorrelated genetic coefficients between nodes. Note that in practice we evaluate the functions over a grid of values for $\boldsymbol{\beta}$.

The full likelihood (i.e. by summing over all possible coefficients) is given by summing the values at the ancestral node A:

$$L_{full} = L_A. \tag{11}$$

The likelihood under the model of no genetic association across all nodes in the tree, $L_\varnothing$, is calculated by summing the likelihood over all nodes with $\boldsymbol{\beta} = \mathbf{0}$, and the prior on this:

$$L_j(\boldsymbol{\beta} = \mathbf{0}) = \prod_{i \epsilon \gamma(j)} p_{00} L_i(\boldsymbol{\beta} = \mathbf{0}), \tag{12}$$

where $p_{00} = e^{-\theta} + (1 - e^{-\theta})(1 - \pi_1)$. For terminal nodes, $L_j(\boldsymbol{\beta} = \mathbf{0})$ is calculated directly from the likelihood function by evaluating Equation 6 at $\boldsymbol{\beta} = \mathbf{0}$. It follows that the null likelihood $L_\varnothing$ is given by

$$L_\varnothing = (1 - \pi_1) L_A(\boldsymbol{\beta} = \mathbf{0}). \tag{13}$$

2 Model fitting and Bayes factor calculation

There are two objectives to the analysis. First, to calculate the evidence for association between a genetic variant and any of the annotations, thus identifying variants that have association to at least one annotation. Second, for variants with some association, to identify those annotations with non-zero coefficients.

Our first objective can be met by calculating a Bayes factor that compares the likelihood integrated over all possible values of $\boldsymbol{\beta}$ in which at least one node is active, L^+, to the likelihood under which all nodes are inactive. By noting that there is only one way in which all nodes can be inactive and that it is easy to calculate both the prior, $\pi_\varnothing$, and likelihood, $L_\varnothing$, for this state, we can obtain the Bayes factor as follows. First, note the we can rewrite the full likelihood function L_{full} in Equation 11 as:

$$L_{full} = \pi_\varnothing L_\varnothing + \sum_{p \epsilon \varnothing'} \pi_p L_p, \tag{14}$$

which sums over the path where all nodes are inactive and all possible path with at least one active node ($p \in \varnothing'$). Then, we can solve for the likelihood L^+:

$$L^+ = \frac{L_{full} - \pi_\varnothing L_\varnothing}{(1 - \pi_\varnothing)}. \tag{15}$$

The desired Bayes factor is then calculated by taking the ratio of the two likelihoods:

$$\mathrm{BF}_{\mathrm{tree}} = \frac{L^+}{L_\varnothing} = \frac{L_{full} - \pi_\varnothing L_\varnothing}{(1 - \pi_\varnothing) L_\varnothing}. \tag{16}$$

Using the same framework, it is also possible to compute Bayes factors for the cases where there is no correlation in state between parent and child nodes (*i.e.,* $\theta \to \infty$), and where all states are active and either share a single set of coefficients (*i.e.,* $\pi_1 \to 1$, $\theta \to 0$) or are independent (*i.e.,* $\pi_1 \to 1$, $\theta \to \infty$). In theory it would be possible either to estimate π_1 and θ or to integrate over a hyper-prior.

For those variants where there is evidence for association within the annotation tree, it is possible to identify active nodes and estimate coefficients of association for each node by using the forward and backward algorithms, also

known as the inside and outside algorithms when applied to tree-like Markov models. The forward (inside) algorithm has been described above, though for completeness and consistency of notation, it is repeated below.

In the forward (inside) algorithm we are iterating up from the terminal nodes towards the root of the tree calculating the joint likelihood of the subtree each node subtends. To initialise, let j be a terminal node, so $F_j(\boldsymbol{\beta})$ is the probability of the observed data at node j for a given value of $\boldsymbol{\beta}$,

$$F_j(\boldsymbol{\beta}) = P_j(D|\boldsymbol{\beta}). \tag{17}$$

We can then integrate over the values of $\boldsymbol{\beta}$ to calculate the integrated likelihood at node j as in Equation 9,

$$L_j = \int F_j(\boldsymbol{\beta}) f^*(\boldsymbol{\beta}) d\boldsymbol{\beta}. \tag{18}$$

For intermediate nodes we calculate F_j recursively up the tree. First, let j be an intermediate node and $\gamma(j)$ the set of child nodes of j. For each $i \in \gamma(j)$ we define,

$$G_i(\boldsymbol{\beta}) = e^{-\theta} F_i(\boldsymbol{\beta}) + (1 - e^{-\theta}) L_i, \tag{19}$$

It follows that for an internal node

$$F_j(\boldsymbol{\beta}) = \prod_{i \in \gamma(j)} G_i(\boldsymbol{\beta}). \tag{20}$$

We can then calculate the integrated likelihood at node j using Equation 18 as for the terminal nodes and continue the algorithm up the tree until the ancestral node, A, is reached.

In the backward (outside) algorithm we calculate the probability density of $\boldsymbol{\beta}$ starting from the root of the tree and moving recursively down the tree. The quantity we are aiming to calculate is the likelihood for the data not subtended by the node of interest.

To initialise, let A be the ancestral node, so $B_A(\boldsymbol{\beta})$ is given by the prior on $\boldsymbol{\beta}$,

$$B_A(\boldsymbol{\beta}) = f^*(\boldsymbol{\beta}). \tag{21}$$

We then iterate down the tree. Let i and j be such that $j \in \gamma(i)$ (that is i is the parent of j), then

$$B_j(\boldsymbol{\beta}) = \int B_i(\boldsymbol{\beta}') q(\boldsymbol{\beta}, \boldsymbol{\beta}') \frac{F_i(\boldsymbol{\beta}')}{G_j(\boldsymbol{\beta})} d\boldsymbol{\beta}', \tag{22}$$

where $q(\boldsymbol{\beta}, \boldsymbol{\beta}')$ is the (transition) probability of state $\boldsymbol{\beta}$ in the daughter node given state $\boldsymbol{\beta}'$ in the parent node. Note that because of the structure of the model there are only two types of transition, which enables efficient calculation. The posterior density for $\boldsymbol{\beta}$ in node j can then be calculated from

$$\pi_j(\boldsymbol{\beta}|D) = \frac{F_j(\boldsymbol{\beta}) \times B_j(\boldsymbol{\beta})}{L_{full}}, \tag{23}$$

and from this distribution we can integrate to estimate the probability of $\boldsymbol{\beta} \neq \mathbf{0}$ and the 95% credible sets for $\boldsymbol{\beta}$.

3 Conditional analysis

To account for linkage disequilibrium in the MHC and to identify independent associations with HLA alleles we performed conditional analysis. For each of the datasets (SR and HES) we first analysed each imputed HLA and identified the allele with the strongest evidence of association, as measured by the $\mathrm{BF}_{\mathrm{tree}}$ statistic. We then continue to analyse the remaining HLA alleles in an iterative approach, where at each iteration we controlled for previous identified HLA alleles, through conditional analysis. To account for these covariates in the analysis we use an approximation to the likelihood function. Let Δ_{ij} quantify the aggregated risk effects due to covariates in individual i in annotation j,

$$\Delta_{ij} = \sum_k [\hat{\beta}^1_{jk} \times I(G_{ik} == 1) + \hat{\beta}^2_{jk} \times I(G_{ik} == 2)], \tag{24}$$

where the genetic coefficients $\{\hat{\beta}^1_{jk}, \hat{\beta}^2_{jk}\}$ are the MAP estimates inferred for the HLA allele with the largest $\mathrm{BF}_{\mathrm{tree}}$ in round k for annotation j, and $G_{ik} \in \{0, 1, 2\}$ are the genotypes for individual i in the HLA allele identified in round k.

To model the distribution of Z_{ij} we modified the logistic model in Equation 1 and 2 to account for the aggregate risk effect due to HLA alleles identified in previous rounds:

$$Y^c_{ijs} = \beta^0 + \beta^1_j * I(G_{is} == 1) + \beta^2_j * I(G_{is} == 2) + \Delta_{ij}, \tag{25}$$

and,

$$P(Z_{ij} = 1|Y^c_{ijs}) = \frac{e^{Y^c_{ijs}}}{1 + e^{Y^c_{ijs}}}. \tag{26}$$

The conditional likelihood function is then given by the binomial distribution,

$$L^c_j(\boldsymbol{\beta}|\boldsymbol{Z_j}) = \prod_{i=1}^{N} p_{ij}^{c\,Z_{ij}} (1 - p^c_{ij})^{1-Z_{ij}}, \tag{27}$$

where we let $p^c_{ij} = P(Z_{ij} = 1|Y^c_{ijs})$.

To compute the above conditional likelihood we use an approximation by taking the 2nd order Taylor expansion around $\Delta = 0$. After evaluation of the first and second derivatives of $log(L^c_j(\boldsymbol{\beta}|\boldsymbol{Z_j}))$ at $\Delta = 0$ and simplifying terms we obtain:

$$log(L_j^c(\boldsymbol{\beta}|\boldsymbol{Z_j})) \approx log(L_j(\boldsymbol{\beta}|\boldsymbol{Z_j})) + \sum_{i=1}^{N}[\Delta_{ij}(Z_{ij} - p_{ij}) - \frac{\Delta_{ij}^2}{2}p_{ij}(1-p_{ij})], \quad (28)$$

where $L_j(\boldsymbol{\beta}|\boldsymbol{Z_{ij}})$ and p_{ij} are given by the equivalent functions when we don't account for covariates. We note that while the approximation works well for early rounds, its accuracy is likely to decrease after multiple rounds of conditioning. Extensions that enable re-estimation at later steps will be explored in subsequent work.

4 Predicting the expected magnitude of genetic dilution due to the winners curse

The magnitude of the estimated effect of the GRS on any given diagnostic term is a measure of how consistent the phenotypic diagnosis criterion is between the GWAS used to derive the GRS and the group of individuals identified with the diagnostic term in the UK Biobank. A decrease from a value of 1 represents a dilution of the GRS, and the extent of this dilution is related to several factors, including: misclassification, misdiagnosis, miscoding, disease heterogeneity, and an expected dilution from the winner's curse.

However, because the effect sizes are typically estimated in those papers where the effect was first discovered, they are subject to the winner's curse bias [1], which would lead to apparent dilution even in a cohort with identical phenotyping. For each of the IMDs for which GRSs were constructed, we performed simulations to estimate the amount of expected dilution due to the winner's curse. For each study we simulated 50,000 case-control datasets with sample sizes matching those reported in the paper from which the effect sizes were estimated. Allele frequencies at risk loci and effect sizes were sampled with replacement from the empirical distribution of genome-wide significant SNPs (from the same paper). Genotypes were sampled from a multinomial distribution and phenotypes were simulated with an additive genetic risk. Simulated datasets were analysed with logistic regression and, for any given replicate, we repeated the simulation if the genotype to phenotype association was not genome-wide significant (P-value < 5x10-8). The expected dilution was then calculated as the average over replicates of the sum of the estimated genetic effects over the sum of the true genetic effects.

For each of the studies analysed we estimated the expected dilution to be no more than 15% (Supplementary Table 11).

5 Extent of genetic dilution due to misclassification

To assess how misclassification between related traits can affect dilution and associated diagnostic terms, we performed a series of simulations where we mis-

classified individuals in the UK Biobank with a diagnosis of type 1 diabetes (T1D) to a diagnosis of type 2 diabetes (T2D) - or the reverse - and we performed TreeWAS analysis on these permuted datasets. This was performed in both the SR and the HES datasets, and the T1D and T2D genetic risk scores were analysed against each dataset.

When we simulated a misclassification from T2D to T1D we observed that the evidence of association of the T1D GRS with the T1D term was not affected (Supplementary Figure 6c,d) and remained highly significant ($PP = 1$) for all simulated misclassification rates, but there was increased dilution of the estimated genetic effect with increasing misclassification rates (Supplementary Figure 6e,f). Therefore, misclassification is one of the factors that can affect the extent of dilution observed for the genetic effect of a GRS on its respective diagnostic term. When misclassification was performed in the reverse direction, no significant increase in the dilution was observed for the T2D GRS on the T2D diagnostic term.

In our simulation analysis we did not observe an association between the T1D GRS and T2D diagnostic terms (Supplementary Table 7): through the simulations we estimated that we would require at least a 10% misclassification rate of T1D onto T2D to observe an association between the T1D GRS and the T2D diagnostic term in the HES dataset (Supplementary Figure 7d).

References

[1] John PA Ioannidis. Why most discovered true associations are inflate. *Epidemiology*, 19(5):640–648, 2008.