Stratified medicine as a statistical modelling problem: learning finite mixture models for disease subtyping with the Bayesian inference program Stan

Paul McKeigue

Usher Institute of Population Health Sciences and Informatics

Stratified medicine: the next big thing?

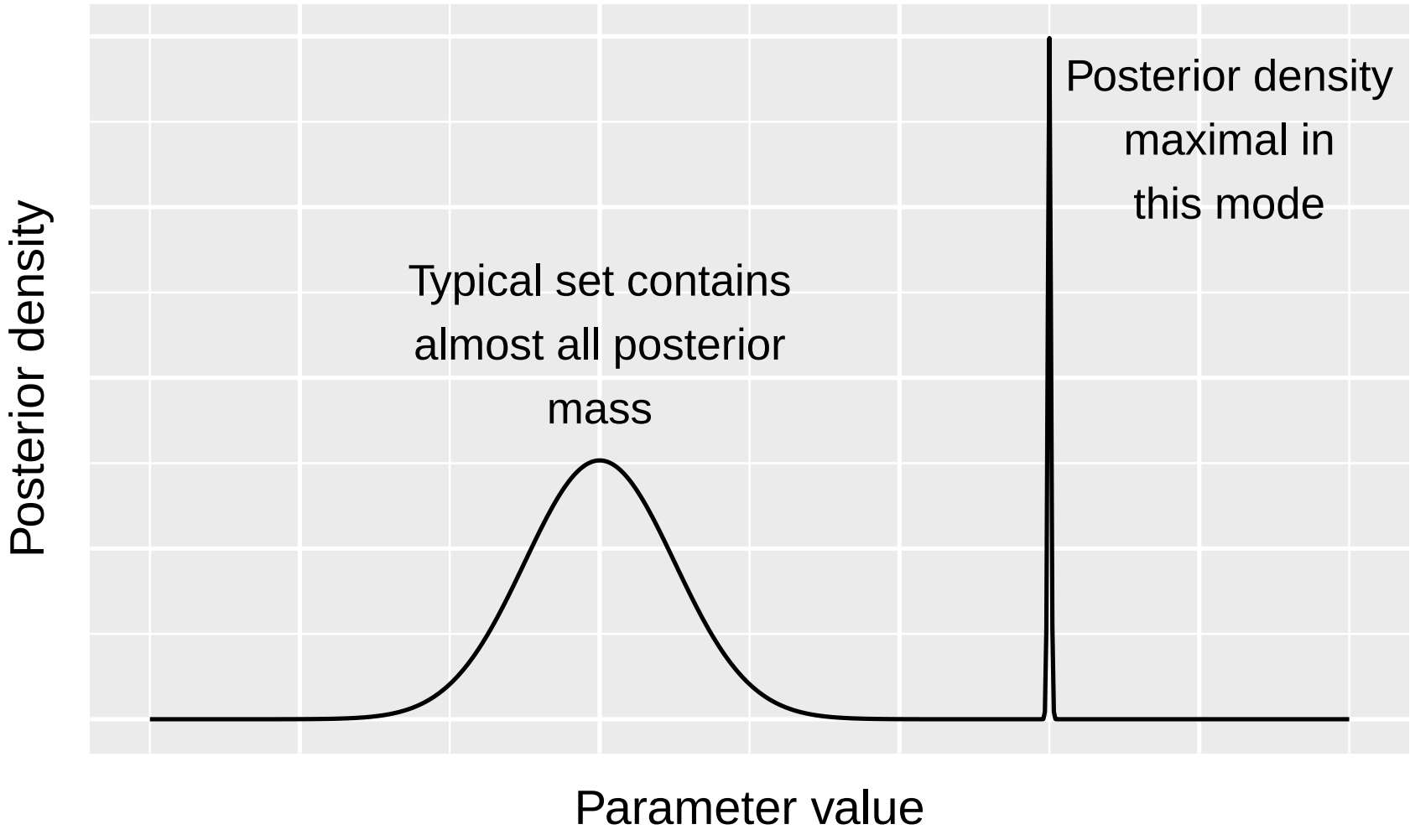
- 2015: US government announced research initiative on "**precision medicine**, an innovative approach to disease prevention and treatment that takes into account individual differences in people's genes, environments, and lifestyles".
    - Synonyms: personalized medicine (EU), stratified medicine (MRC)
- MRC (2013) workshop on stratified medicine favoured "identifying groups of patients with distinct **endotypes** (subtypes of a condition defined by a distinct functional or pathobiological mechanism)"
    - Endotypes should predict response not just to drugs in use now, but to new therapies not yet developed,
    - Soft classification: "patient may traverse more than one endotype during the course of their disease"

Learning to subtype disease: a mixture modelling problem

- Soft classification should allow each individual to be a mix of different endotypes
- Different models for prediction of outcome from covariates may be fitted to each endotype:
- Such models have been described in different contexts
  - in social science as *latent class models* (Lazarsfeld 1968)
  - in biostatistics as *finite mixture models* (Everitt 1981)
  - in machine learning as *mixtures of experts* (Jordan 1994].

Why learning mixture models from data is hard

- Model parameters may not be identified
- Likelihood surface is multimodal: maximizing the likelihood or posterior density may find an atypical mode



Bayesian learning

- Bayesian inference based on the full posterior is valid even if the posterior is multimodal
- Computing a multimodal posterior is hard
  - Markov chain Monte Carlo (MCMC) samplers tend to get stuck in one of the modes.
- To test hypotheses or compare models we need to be able to compute the **evidence** (marginal likelihood) of each model.
  - even harder than sampling the posterior.

Stan: a platform for Bayesan inference and imputation: Gelman, Lee and Guo (2015)

- Like BUGS or JAGS, Stan uses MCMC to sample the posterior distribution given the data and the model.
- Stan uses *Hamiltonian MCMC* (Duane, Kennedy, Pendleton & Roweth 1987) - to update all parameters jointly.
  - momentum as a randomized auxiliary variable
  - algorithmic differentiation to compute gradients
- As an alternative to MCMC sampling, Stan can use a faster variational Bayes algorithm to approximate the posterior.
  - this also generates a lower bound approximation to the evidence (ELBO).
- Development of Stan continues

Stan or William?

Stan Ulam (USA)

- Markov chain Monte Carlo sampling algorithm
- Method of initiating a hydrogen bomb

William Rowan Hamilton (Ireland)

- Hamiltonian dynamics, variational principle of least action
- Quaternions: efficient representation of 3D rotation

Type 1 diabetes as an exemplar of a disease with underlying endotypes

- Type 1 diabetes is now recognized to be a heterogeneous condition:
  - classic juvenile-onset Type 1 cases with rapid autoimmune destruction of islet cells
  - late-onset cases in whom loss of beta-cell function progresses slowly, some of whom have features of Type 2 diabetes including obesity
- Residual insulin secretion (measured as C-peptide) may persist years after diagnosis even in early-onset cases.

Scottish Diabetes Research Network Type 1 Bioresource

- Cohort of people clinically diagnosed as Type 1 diabetes over wide ranges of age at onset and duration.
- 5998 individuals with median duration of diabetes 20 years at enrolment (inter-quartile range 11 to 31).
- C-peptide measured at clinic visit, autoantibodies measured in half the cohort
- genotyped with Illumina chip, untyped SNPs imputed from UK10K reference panel

Calculation of genotypic risk scores from summary statistics

- Genotypic risk scores for Type 1 diabetes and Type 2 diabetes computed using the GENOSCORES platform.
  - genotype vector $\boldsymbol{g}$, genotype correlations $\boldsymbol{\Sigma}$ (estimated from reference panel), univariate regression coefficients $\boldsymbol{\alpha}$ from publicly available summary statistics
  - genotypic risk score is computed as $\boldsymbol{g}^{\intercal}\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}$
  - Coefficients approximate the weights that would be obtained by fitting a multivariate regression model to the individual-level data.
  - Score computed for each diabetes-associated region

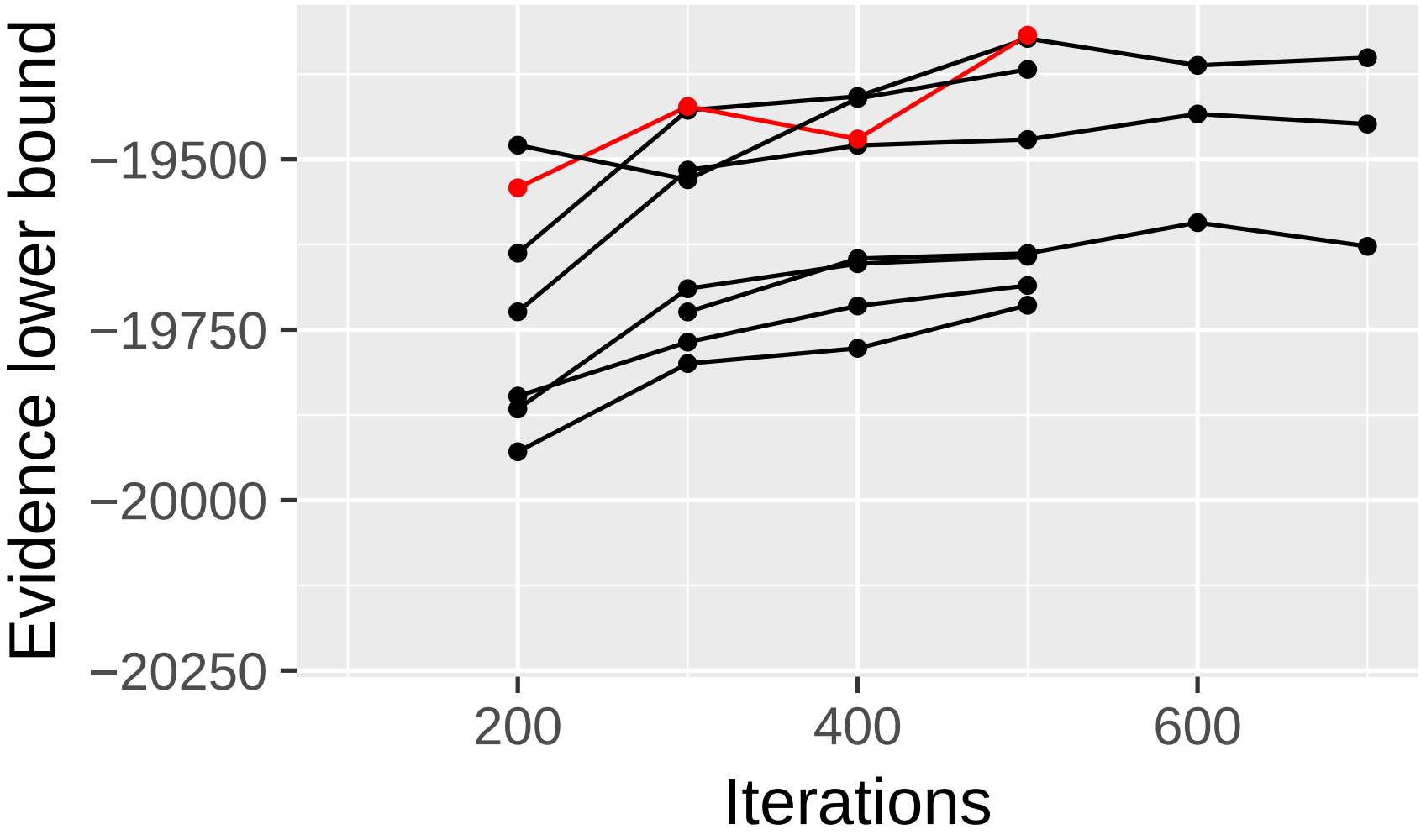
Statistical model

- logistic regression of each individual's mixture component on covariates $\boldsymbol{Z}$: age at onset, genotypic scores for Type 1 and Type 2 diabetes.
  - $\text{logit}\,(\boldsymbol{\lambda}) = \boldsymbol{Z}^{\intercal}\boldsymbol{\gamma}$
- linear regressions of J outcome variables on covariates $\boldsymbol{X}$ given kth mixture component:
  - $\langle \boldsymbol{y}_j \mid k \rangle = \boldsymbol{X}^{\intercal}\beta_{jk}$
- y_{ij}: jth outcome variable in ith individual distributed as mixture of component-specific distributions with mixture weights $\lambda_i, (1 - \lambda_i)$

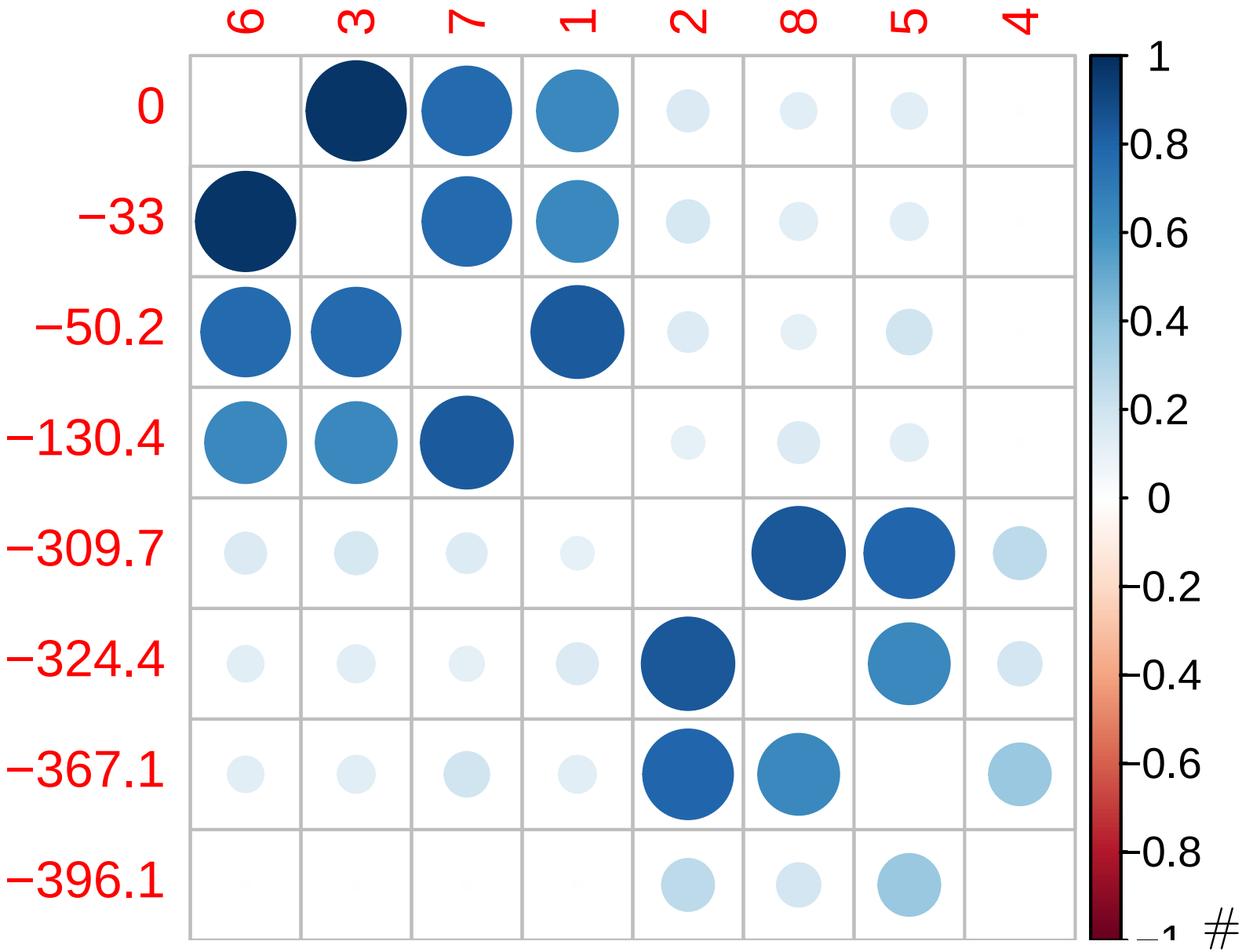
Learning the model

- Effect of age at onset on mixture probability constrained to be negative: ensures identifiability
- Posterior is multimodal: different runs of variational Bayes converge to different modes with evidence lower bound differing by > 50 natural log units
- Posterior means from best variational Bayes run used as initial values for MCMC sampler: 4 chains, convergence diagnostics show adequate mixing

Comparison of multiple variational Bayes runs

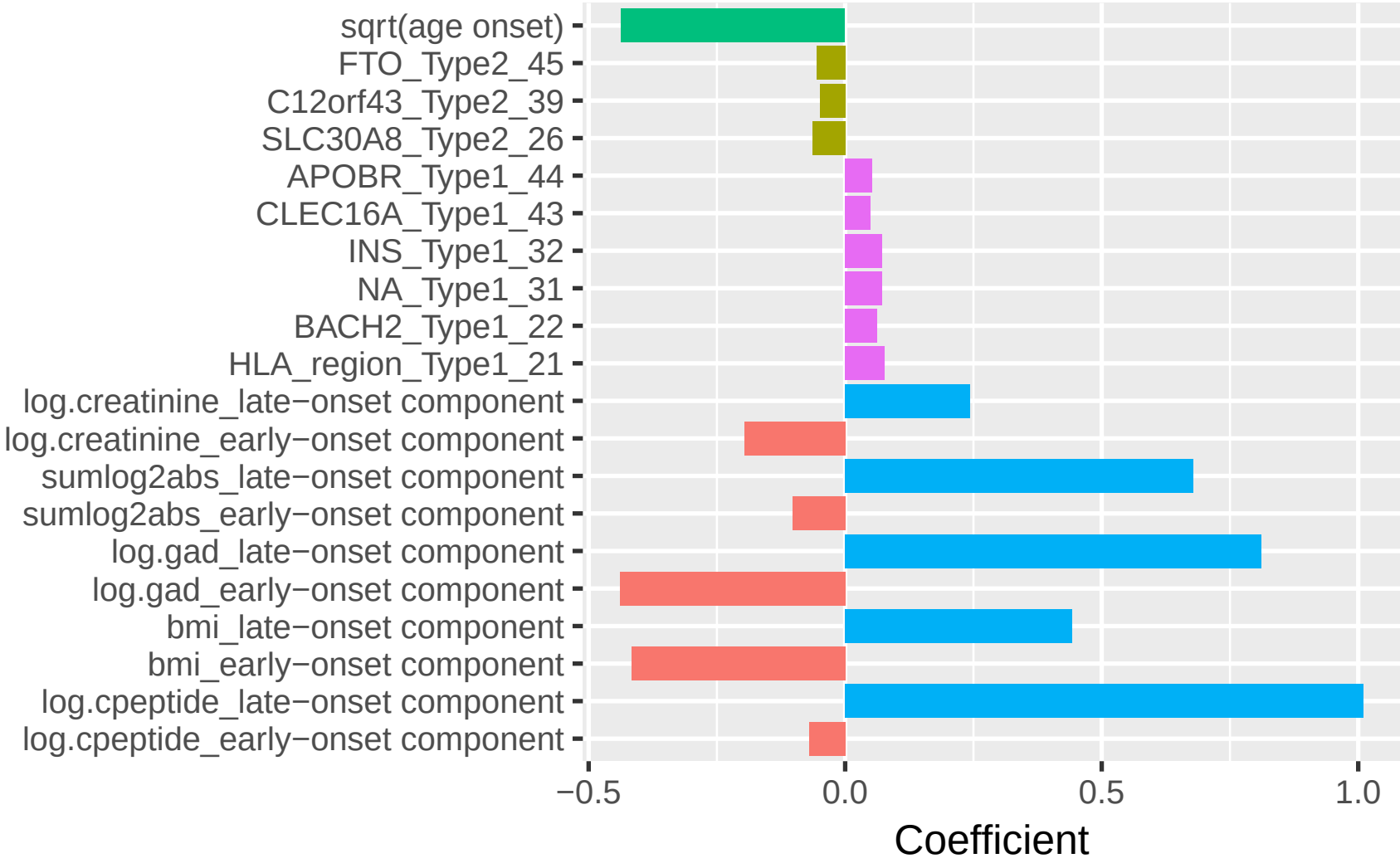


Heat map of correlations between posterior mean parameter vectors in different runs



#

Is the model biologically interpretable?



Conclusions

- With current version of Stan it is possible to learn a model of diabetes as a mix of two endotypes from which multiple clinical features secretion can be predicted :
  - classic early onset Type 1, late-onset with Type 2 like features
- Neither variational Bayes nor MCMC can be relied on to explore the multimodal posterior adequately in a single run
  - can use multiple runs of variational bayes algorithm to select best mode for MCMC sampling
- New features in the Stan development pipeline may improve learning of multimodal posteriors:-
  - annealing with adiabatic cooling (heat bath)
  - Riemannian Monte Carlo
  - nested sampling (sample from prior truncated by likelihood)