Bayesian inference on high-dimensional Seemingly Unrelated Regression, applied to metabolomics data

M. Banterle, A. Lewin

Increasingly, epidemiologists are collecting multiple high-dimensional molecular data sets on large cohorts of people. The interest is in finding associations between these data sets and with genetic variants. In order to do this effectively these multi-variate data sets should be modelled jointly, taking into account correlations in the data. Sparse solutions are usually required, and performing variable selection in this setting is critical.

We present a Bayesian Seemingly Unrelated Regressions (SUR) model for associating metabolomics outcomes with genetic variants, allowing for both sparse variable selection and correlation between the outcomes.

$$\underset{n\times 1}{y_k} = \underset{n\times d_k}{X_{\gamma_k}}\ \underset{d_k\times 1}{\beta_{\gamma_k}} + \underset{n\times 1}{\epsilon_k} \qquad \text{for } k = 1, \cdots, q \quad \text{and} \quad Cov[\epsilon_k \epsilon_l] = R_{kl} \neq 0$$

A binary matrix Γ encodes variable selection between all possible pairs of outcomes and predictors.

$$\gamma_{jk} = \begin{cases} 1 & \text{outcome } k \text{ associated with predictor } j \\ 0 & \text{else} \end{cases}$$

A simple sparsity inducing prior is used, for example $\gamma_{jk} \sim Bern(\omega_{jk})$, $\omega_{jk} \sim Beta$, but more structured priors can also be employed.

This model can be fit using a Gibbs sampler, but this quickly becomes computationally unfeasible as the dimensions of the problem grow. Previously people have made use of two alternate simplifying assumptions, either assuming independence between the outcomes (Bottolo et al. 2011, Lewin et al. 2015) or selecting predictors jointly for all the outcomes (Bhadra and Mallik 2013, Bottolo et al. 2013). In both this simplified cases conjugate priors can be used on the regression coefficients and variance/covariances.

In order to overcome some of the computational difficulty with the general SUR model, Zellner and Ando (2010) proposed a reparametrisation of the model in which the likelihood factorises completely into a product of conditional distributions, and used a Direct Monte Carlo (DMC) approach to estimate the posterior. This would improves computational time, however their method requires re-sampling of the regression coefficients in order to obtain the correct posterior distribution.

We extend their work by allowing for a more general prior distribution, and we show that it is possible to build a Gibbs-DMC sampler without the need for re-sampling, improving considerably on the computational aspect of the method.

Our current work is on approximate inference procedures such as Expectation Propagation (Minka 2001, Andersen et al. 2015, Hernndez-Lobato et al. 2015) applied to our SUR model, in order to tackle the inherent difficulties of stochastic search algorithms in very high-dimensional model spaces.

The proposed methods are applied to simulated data to illustrate the computational gains and further demonstrated on a metabolomics (highly structured, with strong correlations) v. genetic variants data set from the North Finnish Birth Cohort.