Bayesian modelling of whole-brain cell count data

Neuroscientists can now collect cell-count data across the whole animal brain providing them with powerful approaches to measuring recent neuronal activity, gene expression and anatomical connectivity. However, these methods are expensive and time-intensive; the resulting datasets tend to be under-sampled with fewer animals than the number of brain regions. Consequently, these data are a challenge for traditional frequentist approaches that rely on large sample sizes for statistical inference.

Cell count data typically come with an explicit structure of experimental groups, with data subdivided first by experimental condition, then by animal subject and finally by brain region. We present a novel example application on real experimental data that demonstrates how hierarchical Bayesian methods are better suited to these data through their ability to capture this structure, providing a clean, parameterised description of the experiment.

We discover that common choices for prior distributions over correlation matrices are over-regularising. This means that despite principled efforts to capture correlations, the properties of the chosen prior distribution suppress them in the posterior. To fix this, we investigate methods for targeting correlations in the prior. Initial results show better recovery of parameters against ground truth in simulated data tasks, and an improvement in the accuracy of imputation for held-out data points.

Our Bayesian approach is highly flexible and can be easily applied to other datasets using different expression markers or experimental groupings. Brain-atlas connectivity matrices and projections studies, for example, can provide excellent prior information to facilitate this targeted approach to correlation modelling, and indeed for other approaches we are investigating more generally.

Model Structure

Bayesian modelling combines prior knowledge with data to update a parameterised, generative model of the observations [1, 2]. Our model uses a negative binomial distribution with a log link function for the mean:

$$y_i \sim \text{NegativeBinomial}(\lambda_i, \phi) \tag{1}$$
$$\log \lambda_i = E_i + \theta_{r[i],g[i]} + \gamma_{a[i],r[i]} \tag{2}$$

The priors are organised in a hierarchy:

$$\boldsymbol{\gamma}_{ar} \sim \text{MvT}(\nu, 0, \Sigma_g) \tag{3}$$
$$\Sigma_g = \text{diag}(\boldsymbol{\tau}_g) \cdot \Omega \cdot \text{diag}(\boldsymbol{\tau}_g) \tag{4}$$
$$\Omega \sim \text{LKJ-T}(1, \boldsymbol{\rho}) \tag{5}$$

where r, g, and a represent brain regions, experimental groups and animals respectively. Parameters are indexed by observation number i, so, for example, $r[i]$ corresponds to the brain region associated with data point y_i.

The model assumes that brain regions have some underlying log-mean expression level θ_{rg}. These are often the main parameters of interest and are used to form comparisons across or within experimental groups. For each animal a vector of correlated random effects $\boldsymbol{\gamma}_{ra}$ is estimated, allowing the model flexibility to capture the unique responses of each animal to the stimulus. These are partially pooled within each region by τ_r to estimate a measure of the variability about the baseline level while also providing soft regularisation.

Results

We found that the LKJ distribution, the usual prior for correlation matrices in hierarchical models, tends to overregularise posterior correlations by pulling them towards zero. Figure 1 gives a visual guide to our new targeted prior and highlights some of its benefits for a simulated data study.

We applied the presented Bayesian model to experimental data (Figure 2A) that examined the role of the *nucleus reuniens* (NRe) in the recognition memory circuit: lesion of the NRe impair some memory processes [4]. The data consisted of c-Fos cell counts for 23 brain regions from 39 male rats. Rats were required to discriminate novel or familiar objects subject to a sham or lesion of the NRe giving a two by two experimental design. We sampled the posterior using Hamiltonian Monte-Carlo in `Stan` [5]. Our model results are compared with a frequentist analysis in Figure 2B.

References

[1] A. Gelman et al. *Bayesian Workflow*. 2020. arXiv: 2011.01808 [stat.ME].

[2] Rens van de Schoot et al. In: *Nature Reviews Methods Primers* 1.1 (2021), pp. 1–26.

[3] B. M. S. Exley. "The role of the nucleus reuniens of the thalamus in the recognition memory network". 2019.

[4] G. Barker and E. Warburton. In: *The Journal of Neuroscience* 38 (Feb. 2018), pp. 1802–17.

[5] Stan Development Team. *RStan: the R interface to Stan*. R package version 2.21.2. 2020.

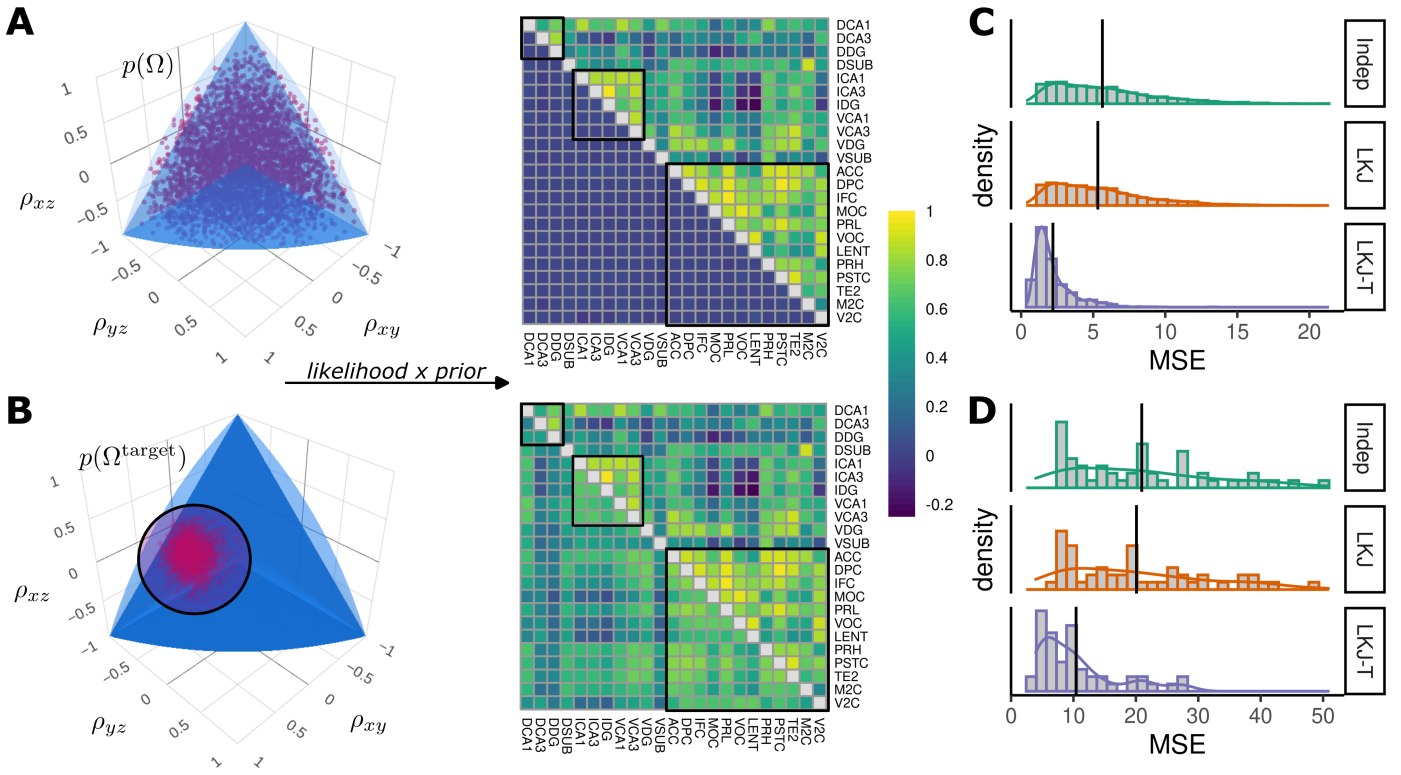

Figure 1: **Prior distributions for correlation matrices.** **A**: Draws from an LKJ(2.0) distribution for a 3 by 3 correlation matrix. Samples are plotted within the convex region defined by the positive semi-definite constraint (left). This prior is overly regularising (right): when fitting a two-level hierarchical normal model to fictive data generated using true correlation structure (upper triangular) the recovered posterior median correlation matrix is very close to the identity matrix (lower triangular). **B**: Analogous to **A** except the prior is targeted towards the true correlation structure by the LKJ-T$(1, \boldsymbol{\rho})$ distribution. **C**: To quantify the benefits of better correlation recovery the mean squared error (MSE) of animal random effects compared to true generating parameters were calculated over fifty fictive datasets. This error is markedly reduced when correlations are targeted. Vertical lines give distribution means, and 'Indep' specifies a simple model where brain regions are assumed independent ($\Omega = I$). **D**: For the same fitting process we evaluate the performance of the model on imputing a held-out data point. Distributions show the MSE of the imputed value from ground truth. The distribution mean is reduced by approximately a half when our correlation prior is used.

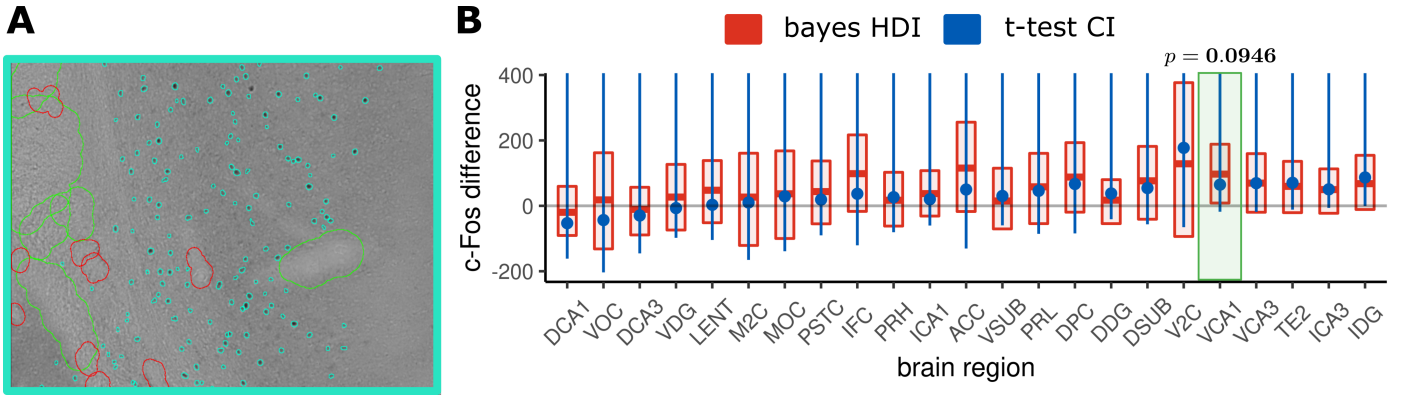

Figure 2: **Data and modelling results.** **A**: c-Fos positive nuclei (teal), with tears and voids shown in red and green [3]. **B**: The 95% confidence intervals resulting from uncorrected one-sided two-sample t-tests, and posterior distributions over the same quantity described with 90% highest density intervals. Brain regions on the x-axis are ordered in terms of *decreasing* p-value. No significant results were detected by the frequentist analysis to suggest a higher expression of c-Fos by Lesion-Familiar animals compared to Sham-Familiar ones. However, the Bayesian model gives evidence of this difference in VCA1. For this analysis the target correlation matrix was derived from prior knowledge about network structure and constructed based on conditional independence assumptions.