01.06.17

The problem of non-equal number of regressors between sessions

Problem 1

The cross-validated log model evidence (cvLME) – which is part of cross-validated Bayesian model selection (cvBMS) – assumes that each GLM regressor – and thus, each model parameter – is present in each CV fold / fMRI sessions. This is violated by some GLMs, because they include modelling of e.g. missed responses or error trials which may not occur in all sessions.

Solution 1

A solution to this is to include "empty" regressors consisting only of zeros into the design matrix of the respective sessions. This does not prohibit SPM model estimation, because SPM uses the pseudo-inverse for model inversion, so that cvLME model assessment can also be performed. Practically, the cvLME is implemented by (vertically) concatenating the design matrices (and data) from all training sessions and estimating the informative posterior which serves as the prior for the test session.

Problem 2

Critically, the (concatenated) training design matrix is used for updating the precision matrix Λ of the regression coefficients β:

$$p(\beta|\tau) = \mathcal{N}(\beta; \mu_0, (\tau\Lambda_0)^{-1})$$

$$p(\beta|\tau, y) = \mathcal{N}(\beta; \mu_n, (\tau\Lambda_n)^{-1})$$

In the training data, we have

$$\mu_0^{(n)} = 0_p \qquad \mu_n^{(n)} = \Lambda_n^{(n)-1}(X^T P y + \Lambda_0^{(n)} \mu_0^{(n)})$$

$$\Lambda_0^{(n)} = 0_{pp} \qquad \Lambda_n^{(n)} = X^T P X + \Lambda_0^{(n)}$$

Now, if the training design matrices only contain zero regressors for a specific model parameter, this will cause two problems.

$\Lambda_n^{(1)}$ will have a zero entry on its diagonal so that it is not invertible and its determinant is zero: First, this means that $\mu_n^{(1)}$ is not defined, because Λ_n^{-1} does not exist. As $\mu_n^{(1)} = \mu_0^{(2)}$, $\mu_n^{(2)}$ cannot be calculated. Second, the out-of-sample log model evidence (oosLME) for the test data contains the term $\sum \log|\Lambda_0|$. As $\Lambda_n^{(1)} = \Lambda_0^{(2)}$, the oosLME is $-\infty$.

Remark 2a

Note that the problem does not exist when there is a non-empty regressor in at least one training design matrix in each CV fold. As always one regressor is in the test set, the problem does not occur with the extra regressor present in 0 sessions (no need for empty regressors) or at least 2 sessions (always at least one regressor in the training set). In other words, when the extra regressor is only in 1 session, the problem will occur in only one CV fold – the one in which the extra regressor is in the test set and training set only has zero regressors.

Solution 2

There are several possible solutions to this problem, but the simplest one is probably to change $\Lambda_0^{(1)}$ from

$$\Lambda_0^{(1)} = 0_{pp}$$

to

$$\Lambda_0^{(1)} = \lambda_0 \cdot I_p$$

where

$$1 >> \lambda_0 \approx 0$$

is a very small number, so that

- results do not change (very much) in non-problematic cases;
- $\Lambda_n^{(1)}$ is invertible, $\mu_n^{(1)}$ is defined and $|\Lambda_0^{(2)}| \neq 0$ when the problem described above occurs.

I've analysed a single-subject single-session data set (Henson et al., 2002) with λ_0 ranging from 10^{-3} to 10^{-10} and compared results to the default case where $\lambda_0 = 0$. I've observed that changes in the cvLME are restricted to the fourth decimal place (when $\lambda_0 = 10^{-10}$). If such a change is conceived as a log Bayes factor (LBF) and recalculated into a posterior probability (PP), the result is 0.5. With multi-session data (more than ca. 350 scans), the change would probably be even smaller.

In the MACS Toolbox V1.0 (see GitHub), I've therefore implemented $\lambda_0 = \exp(-23) = 1.03 \cdot 10^{-10} \approx 10^{-10}$. Unlike the cvBMS and cvBMA toolkits, MACS therefore automatically solves the problem – given that empty regressors are specified which can be easily done by setting one onset to much longer than the duration of the session.

Remark 26