

cvLME

A multi-language library to perform
cross-validated Bayesian model selection

Joram Soch, BCCN Berlin

Toolbox URL: <https://github.com/JoramSoch/cvLME>
Support Contact: Joram Soch <joram.soch@bccn-berlin.de>
Current Version: cvLME V0.1.1
Related Paper: Soch J, Allefeld C (2018). MACS – a new SPM toolbox for model assessment, comparison and selection. *Journal of Neuroscience Methods*, vol. 306, pp. 19-31; DOI: 10.1016/j.jneumeth.2018.05.017.

Contents

1	Model spaces and model selection	1
1.1	Log model evidence	1
1.2	Log Bayes factor	1
1.3	Posterior probabilities	1
1.4	Log family evidence	2
1.5	Implementation	3

1 Model spaces and model selection

1.1 Log model evidence

A model space is defined as a set of models. In the context of these tools, a model space is always initialized with a set of *log model evidences* (LME)

$$\text{LME}(m) = \log p(y|m) = \log \int p(y|\theta, m) p(\theta|m) d\theta \quad (1.1)$$

or *cross-validated log model evidences* (cvLMEs)

$$\text{cvLME}(m) = \sum_{i=1}^S \log \int p(y_i|\theta, m) p(\theta| \cup_{j \neq i} y_j, m) d\theta \quad (1.2)$$

where S is the number of data subsets.

1.2 Log Bayes factor

The *Bayes factor* (BF) is defined as the ratio of two model evidences,

$$\text{BF}_{12} = \frac{p(y|m_1)}{p(y|m_2)}, \quad (1.3)$$

such that the *log Bayes factor* (LBF) is the difference of two log model evidences,

$$\text{LBF}_{12} = \log \text{BF}_{12} = \log \frac{p(y|m_1)}{p(y|m_2)} = \text{LME}(m_1) - \text{LME}(m_2). \quad (1.4)$$

1.3 Posterior probabilities

Given more than two models, one can also calculate *posterior model probabilities* (PPs) by simply applying Bayes' theorem to the model evidences

$$p(m_i|y) = \frac{p(y|m_i) p(m_i)}{\sum_{j=1}^M p(y|m_j) p(m_j)} \quad (1.5)$$

or, equivalently, to the exponentiated log model evidences (LME)

$$p(m_i|y) = \frac{\exp[\text{LME}(m_i)] p(m_i)}{\sum_{j=1}^M \exp[\text{LME}(m_j)] p(m_j)} \quad (1.6)$$

where $p(m_i)$ are prior model probabilities and M is the number of models.

Note that posterior probabilities do not on depend on absolute LME values, but only on relative LME difference. For this reason, the mean LME over models is subtracted from all LMEs before PPs are calculated.

1.4 Log family evidence

The *family evidence* (FE) is obtained by marginalizing over “model” within “family”, i.e. as the marginal probability over the model evidences from all models within one family

$$p(y|f) = \sum_{m \in f} p(y|m) p(m|f) \quad (1.7)$$

and the *log family evidence* (LFE) is the natural logarithm of this quantity

$$\text{LFE}(f) = \log p(y|f) = \log \sum_{m \in f} p(y|m) p(m|f) \quad (1.8)$$

where $p(m|f)$ is a (most likely uniform) within-family prior distribution.

Note that, with a uniform within-family prior, the family evidence is the average of model evidences, but the log family evidence is not the average of the log model evidences! In particular, the problem is that we usually cannot access model evidences $p(y|m)$ directly, but only deal with log model evidences $\log p(y|m)$. LMEs are used to avoid computational problems with very small model evidences that could not be stored in standard computers, e.g. $p(y|m) = 10^{-100} \Rightarrow \log p(y|m) \approx -230$. However, just exponentiating LMEs does not work, because they often fall below a specific underflow threshold $-u$, e.g. $u = 745$, so that all model evidences would be 0.

The solution is to select the maximum LME within a family

$$L^*(f) = \max_{m \in f} [\text{LME}(m)] \quad (1.9)$$

and define differences between LMEs and maximum LME as

$$L'(m) = \text{LME}(m) - L^*(f) . \quad (1.10)$$

Then, the log family evidence can be written as

$$\text{LFE}(f) = \log p(y|f) = \log \left[\frac{1}{M_f} \sum_{i=1}^{M_f} \exp [\text{LME}(m_i)] \right] \quad (1.11)$$

which can be further developed in the following way:

$$\begin{aligned} \text{LFE}(f) &= \log \left[\frac{1}{M_f} \sum_{i=1}^{M_f} \exp [L'(m_i) + L^*(f)] \right] \\ &= \log \left[\frac{1}{M_f} \exp L^*(f) \sum_{i=1}^{M_f} \exp L'(m_i) \right] \\ &= L^*(f) + \log \sum_{i=1}^{M_f} \exp L'(m_i) - \log M_f . \end{aligned} \quad (1.12)$$

1.5 Implementation

In **MATLAB**, (log) Bayes factors, posterior model probabilities and log family evidences are implemented via the functions `MS_LBF`, `MS_PP` and `MS_LFE` which have to be called with an $M \times N$ matrix `LME` as input.

In **Python**, a model space object has to be initiated via `ms = cvBMS.MS(LME)` and (log) Bayes factors, posterior model probabilities and log family evidences are calculated via `ms.LBF`, `ms.BF`, `ms.PP`, and `ms.LFE`.