Random-effects model selection

Matteo Lisi

January 25, 2019

Definition

In random-effects model selection [2], models are treated as random effects that could differ between subjects and have a fixed (unknown) distribution in the population. The relevant statistical quantity is the frequency with which any model prevails in the population. Note that this is different from the definition of random-effects in classical statistic where random effects models have multiple sources of variation, e.g. within- and between- subject variance. An useful way to summarize the results is by reporting the model's *exceedance probabilities*, which measures how likely it is that any given model is more frequent than all other models in the set.

An example

Let's say we have an experiment with $(1,\dots,N)$ participants. Their performance is quantitatively predicted by a set $(1,\dots,K)$ competing models. The behaviour of any subject n can be fit by the model k by finding the value(s) of the parameter(s) θ_k that maximize the marginal likelihood of the data y_n given the model, that is

$$p\left(y_n \mid k\right) = \int p\left(y_n \mid k, \theta_k\right)\, p\left(\theta_k\right) d\theta. \tag{1}$$

In a frequentist setting we ignore the prior $p\left(\theta_k\right)$ and simply find the parameter values $\hat{\theta}_k$ that maximizes the likelihood of the data, $\hat{\theta}_{nk} = \arg\max_\theta p\left(y_n | k, \theta_k\right)$. By integrating over the prior probability of parameters the marginal likelihood takes into account the complexity of the model. In the following I will adopt a simpler approach and approximate the marginal likelihood model evidence using the Akaike information criterion.

We are interested in finding which model does better at predicting behavior, however we allow for different participants to use different strategies which can be represented by different models. To achieve that we treat the model as random effects and we assume that the frequency or probability of models in the population, $(r_1,\dots,r_K)$, is described by a Dirichlet distribution with parameters $\boldsymbol{\alpha} = \alpha_1,\dots,\alpha_k$,

$$\begin{aligned} p\left(r \mid \boldsymbol{\alpha}\right) &= \mathrm{Dir}\left(r, \boldsymbol{\alpha}\right) \\ &= \frac{1}{\mathbf{B}\left(\boldsymbol{\alpha}\right)} \prod_{i=1}^{K} r_i^{\alpha_i - 1} \end{aligned} \tag{2}$$

.

Where the normalizing constant $\mathbf{B}\left(\boldsymbol{\alpha}\right)$ is the multivariate Beta function. The probabilities r generates 'switches' or indicator variables $m_n = m_1,\dots,m_N$ where $m \in \{0,1\}$ and $\sum_1^K m_{nk} = 1$.

These indicator variables prescribe the model for the subjects n, $p\left(m_{nk}=1\right)=r_k$. Given the probabilities r, the indicator variables have thus a multinomial distribution, that is

$$p\left(m_n \mid \mathbf{r}\right)=\prod_{k=1}^{K} r_k^{m_{nk}}. \tag{3}$$

The graphical model that summarizes these dependencies is shown in Fig. 1.

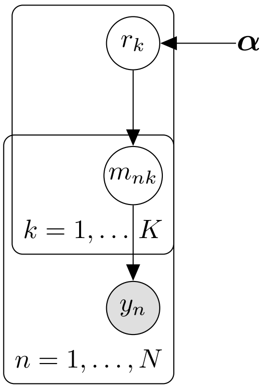

Figure 1: Random-effect generative model for multi-subject data represented as Bayesian graphical model.

Variational Bayesian approach

The goal is to estimate the parameters $\boldsymbol{\alpha}$ that define the posterior distribution of model frequencies given the data, $p\left(r|y\right)$. To do so we need an estimate of the model evidence $p\left(m_{nk}=1 \mid y_n\right)$, that is the posterior belief that the model k generated data from subject m. There are many possible approach that can be used to estimate the model evidence, either exactly or approximately. For simplicity here I will approximate the model evidence by the Akaike Information Criterion, which takes into account model complexity by applying a penalty that is a function of the number of free parameters in the model ($\mathrm{AIC}=2g-2\log\mathcal{L}$, where g is the number of free parameters and $\mathcal{L}$ its maximized likelihood). Our posterior belief needs to be normalized, that is to sum to 1, therefore we convert the Akaike criteria into Akaike weights [1]. First, we transform AIC scores into differences with respect to the AIC of the best candidate model, $\Delta_{nk}=\mathrm{AIC}_{nk}-\min\mathrm{AIC}_{\mathrm{n}}$. Next, we transform back the AIC differences back onto a likelihood scale, and normalize them by their sum to make sure the sum to 1

$$p\left(m_{nk} \mid y_n\right) \approx \frac{\exp\left(-\frac{1}{2}\Delta_{nk}\right)}{\sum\limits_{k=1}^{K}\exp\left(-\frac{1}{2}\Delta_{nk}\right)} \tag{4}$$

Given the graphical model shown in Fig. 1, the joint probability of parameters and data can be expressed as

$$
\begin{aligned}
p\left(y, r, m\right) &= p\left(y \mid m\right)\, p\left(m \mid r\right)\, p\left(r \mid \boldsymbol{\alpha}\right) \\
&= p\left(r \mid \boldsymbol{\alpha}\right) \left[\prod_{n=1}^{N} p\left(y_n \mid m_n\right)\, p\left(m_n \mid r\right)\right] \\
&= \frac{1}{\mathbf{B}\left(\boldsymbol{\alpha}\right)} \left[\prod_{k=1}^{K} r_k^{\alpha_k - 1}\right] \left[\prod_{n=1}^{N} p\left(y_n \mid m_n\right) \prod_{k=1}^{K} r_k^{m_{nk}}\right] \\
&= \frac{1}{\mathbf{B}\left(\boldsymbol{\alpha}\right)} \prod_{n=1}^{N} \left[\prod_{k=1}^{K} \left[p\left(y_n \mid m_{nk}\right)\, r_k\right]^{m_{nk}}\, r_k^{\alpha_k - 1}\right].
\end{aligned} \tag{5}
$$

And the log probability is

$$
\log p\left(y, r, m\right) = -\log \mathbf{B}\left(\boldsymbol{\alpha}\right) + \sum_{n=1}^{N} \sum_{k=1}^{K} \left[\left(\alpha_k - 1\right) \log r_k + m_{nk} \left(p\left(\log y_n \mid m_{nk}\right) + \log r_k\right)\right]. \tag{6}
$$

In order to fit this hierarchical model following the variational approach described in [2] one needs to define an approximate posterior distribution over model frequencies and assignments, $q\left(r, m\right)$, which is assumed to be adequately described by a mean-field factorisation, that is $q\left(r, m\right) = q\left(r\right) q\left(m\right)$. The two densities are proportional to the exponentiated *variational energies* $I(m), I(r)$, which are the un-normalized approximated log-posterior densities, that is

$$
q\left(r\right) \propto e^{I(r)},\ q\left(m\right) \propto e^{I(m)} \tag{7}
$$

$$
I(r) = \left\langle \log p\left(y, r, m\right) \right\rangle_{q(r)} \tag{8}
$$

$$
I(m) = \left\langle \log p\left(y, r, m\right) \right\rangle_{q(m)} \tag{9}
$$

For the approximate posterior over model assignment $q(m)$ we first compute $I(m)$ and then an appropriate normalization constant. From Eq. 6, removing all the terms that do not depend on m we have that the un-normalized approximate log-posterior (the variational energy) can be expressed as

$$
\begin{aligned}
I(m) &= \int p\left(y, r, m\right)\, q(r)\, dr \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} m_{nk} \left[p\left(\log y_n \mid m_{nk}\right) + \int q(r_k) \log r_k \, dr_k\right] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} m_{nk} \left[p\left(\log y_n \mid m_{nk}\right) + \psi(\alpha_k) - \psi\left(\alpha_S\right)\right]
\end{aligned} \tag{10}
$$

where $\alpha_S = \sum_{k=1}^{K} \alpha_k$ and ψ is the digamma function. The digamma function appears here due to a property of the Dirichlet distribution, which says that the expected value of $\log r_k$ can be computed as

$$
\mathbb{E}\left[\log r_k\right] = \int p(r_k) \log r_k \, dr_k = \psi(\alpha_k) - \psi\left(\sum_{k=1}^{K} \alpha_k\right) \tag{11}
$$

From this, we have that the un-normalized posterior belief that model k generated data from subject n is

$$
u_{nk} = \exp\left[p\left(\log y_n \mid m_{nk}\right) + \psi(\alpha_k) - \psi\left(\alpha_S\right)\right] \tag{12}
$$

and the normalized belief is

$$g_{nk} = \frac{u_{nk}}{\sum_{k=1}^{K} u_{nk}} \tag{13}$$

We need also to compute the approximate posterior density $q(r)$, and we begin as above by computing the un-normalized, approximate log-posterior or variational energy

$$\begin{aligned} I(r) &= \int p\left(y, r, m\right)\, q(m)\, dm && (14) \\ &= \sum_{k=1}^{K} \left[\log r_k \left(\alpha_{0k} - 1\right) + \sum_{n=1}^{N} g_{nk} \log r_k \right] && (15) \end{aligned}$$

The logarithm of a Dirichlet density is $\log \mathrm{Dir}(r, \boldsymbol{\alpha}) = \sum_{k=1}^{K} \log r_k \left(\alpha_{0k} - 1\right) + \ldots$, therefore the parameters of the approximate posterior are

$$\boldsymbol{\alpha} = \boldsymbol{\alpha}_0 + \sum_{n=1}^{N} g_{nk} \tag{16}$$

Iterative algorithm

The algorithm [2] proceeds by estimating iteratively the posterior belief that a given model generated the data from a certain subject, by integrating out the prior probabilities of the models (the r_k predicted by the Dirichlet distribution that describes the frequency of models in the population) in log-space as described above. Next the parameters of the approximate Dirichlet posterior are updated, which gives new priors to integrate out from the model evidence, and so on until convergence.Convergence is assessed by keeping track of how much the vector $\boldsymbol{\alpha}$ change from one iteration to the next, i.e. is common to consider that the procedure has converged when $\|\boldsymbol{\alpha}_{t-1} \cdot \boldsymbol{\alpha}_t\| < 10^{-4}$ (where $\cdot$ is the dot product).

Exceedance probabilities

After having found the optimised values of $\boldsymbol{\alpha}$, one popular way to report the results and rank the models is by their exceedance probability, which is defined as the (second order) probability that participants were more likely to choose a certain model to generate behavior rather than any other alternative model, that is

$$\forall j \in \{1, \ldots, K, j \neq k\}\,, \;\; \varphi_k = p\left(r_k > r_j \mid y, \boldsymbol{\alpha}\right). \tag{17}$$

In the case of $K > 2$ models, the exceedance probabilities φ_k are computed by generating random samples from univariate Gamma densities and then normalizing. Specifically, each multivariate Dirichlet sample is composed of K independent random samples $(x_1, \ldots, x_K)$ distributed according to the density $\mathrm{Gamma}\left(\alpha_i, 1\right) = \frac{x_i^{\alpha_i - 1} e^{-x_i}}{\Gamma(\alpha_i)}$, and then set normalize them by taking $z_i = \frac{x_i}{\sum_{i=1}^{K} x_i}$. The exceedance probability φ_k for each model k is then computed as

$$\varphi_k = \frac{\sum \mathbf{1}_{z_k > z_j, \forall j \in \{1,\ldots,K, j \neq k\}}}{\text{n. of samples}} \tag{18}$$

where $\mathbf{1}_{\ldots}$ is the indicator function ($\mathbf{1}_{x>0} = 1$ if $x > 0$ and 0 otherwise), summed over the total number of multivariate samples drawn.

References

[1] Kenneth P. Burnham and David R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* Springer New York, New York, US, 2nd editio edition, 2002.

[2] Klaas Enno Stephan, Will D. Penny, Jean Daunizeau, Rosalyn J. Moran, and Karl J. Friston. Bayesian model selection for group studies. *NeuroImage*, 46(4):1004–1017, 2009.