Chapter 2
Introduction to Model Estimation and Selection Methods

When conducting interdisciplinary research, the employed methods may not be common knowledge in all involved fields. This chapter serves to make this work accessible to a wide audience by describing in detail the methods used for model estimation and selection, which are the fundamental concepts of model-based analyses (Mars et al. 2012). The first sections aim at making the reader accustomed to the terminology and concepts underlying the different kinds of hypotheses and models classically used in (neuro-)psychology. After showing cases where classical methods are inadequate for model selection, state of the art in Bayesian methods are presented.

This chapter is structured as follows: It starts with a simple exemplary study which is used to illustrate different types of hypotheses and models as well as statistical methods for comparing them. The presented methods are not an exhaustive list, as this would be beyond the scope of this work. Next, a detailed description of hierarchical linear models and the parametric empirical Bayes (PEB) schemes employed for all analyses in this work is given. A tutorial with an example experiment that inspired Kolossa et al. (2016) concludes this chapter.

A short note on notation: In the following, not emphasized letters refer to scalars, bold lowercase letters to (column) vectors, bold capital letters to matrices, and the superscript $[\,]^T$ denotes the transpose.

2.1 An Example Study

This section describes a fictitious study which exemplifies the different kinds of hypotheses and models in the following sections. Has learning for an exam any effect on the achieved result? In order to answer this question, a study is carried out in which a test subject takes a total of N exams, each of which is indexed with a trial number $n \in \{1, ..., N\}$. For each trial, the time spent learning and the achieved points are recorded, forming the predictors $x(n)$ and dependent variables $y(n)$, respectively. Note that the dependent variable is simply the measured data, which are the ERP sequences in the rest of this work. The examples in this chapter are the only instances

A. Kolossa, *Computational Modeling of Neural Activities for Statistical Inference*, DOI 10.1007/978-3-319-32285-8_2

where the $y(n)$ contains data other than the ERP sequence. The single-level models described in the following section represent a single subject. Multiple-level models representing more than one subject at once are presented subsequently.

2.2 Classical Single-Level Models

Classical hypothesis testing as used in psychological research relies on p-values, which will be explained shortly (Fisher 1926; Neyman and Pearson 1933). In some cases, these tests do not give the answers the researchers actually seek, but are still used lacking more sophisticated methods (Cohen 1994; Goodman 1999a). More useful methods have been proposed, with Bayes factors being the most advantageous (see, e.g., Kass and Raftery 1995; Goodman 1999b; Friston et al. 2002; Hoijtink 2012; Penny 2012). This section shortly comments on evaluation methods for informative hypotheses before going into detail about methods for trial-by-trial models which model each individual data point. It closes with showing how Bayesian evaluation schemes work and why they were used in this work.

2.2.1 The Null and Informative Hypotheses

To test the hypothesis that learning for exams has *any* effect on the achieved points (see Sect. 2.1), the subject does not prepare for one group of exams while he/she does prepare for the other group. This yields two experimental conditions $c \in \mathcal{C} = \{1, 2\}$ with N_c trials each and the total number of trials $N = \sum_{c \in \mathcal{C}} N_c$. The relation of the points achieved in the two conditions is of interest. The null hypothesis H_0 constitutes that learning does not have any effect, i.e., that there is no difference in the condition-specific mean points μ_c. It is formalized via the equality

$$\mu_1 = \mu_2, \tag{2.1}$$

with μ_1 as the mean points for tests taken without learning and μ_2 as the mean points for tests taken with learning. Classically, the null hypothesis is tested using p-values (Rutherford 2001; Weiss 2006). They represent the probability of getting the observed (or more extreme) data in the absence of any effect, i.e., if H_0 was true, but *not* the probability of H_0 being true (Biau et al. 2010). If the p-value is sufficiently small, the null hypothesis is rejected. This is sometimes confused with the likelihood that a specific effect of interest is present, which is not valid because H_0 can be false due to *any* effect (Cohen 1994). Ronald Fisher (1926) proposed an arbitrary boundary saying "We shall not often be astray if we draw a conventional line at 0.05...", which everyone followed suit. Soon $p = 0.05$ was established as *the* significance boundary to reject the null hypothesis, which is contested because

"... surely, God loves the 0.06 nearly as much as the 0.05." (Rosnow and Rosenthal 1989) and led to a bias of p-values just below 0.05 in publications (Masicampo and Lalande 2012).

This traditional null hypothesis testing is nowadays challenged as outdated "20th century thinking" (Osborne 2010). The main argument is that the null hypothesis is never true as "the probability that an effect is exactly zero is itself zero." (Friston and Penny 2003). Consequently, by increasing the amount of data, one can always reject the null hypothesis with $p \leq 0.05$ (Cohen 1994; Van de Schoot et al. 2011; Friston 2012), which has been exaggerated to the point that the collection of data is unnecessary for rejecting the null hypothesis (Royall 1997). See Wagenmakers (2007) for an overview on the critiques against p-values and Wainer (1999) for cases where they are useful.

Hoijtink et al. (2008) propose the evaluation of informative hypotheses using Bayes factors based on accuracy and complexity terms. Bayes factors can be used to select between competing informative hypotheses like

$$\mu_1 > \mu_2, \tag{2.2}$$

which states that μ_1 is larger than μ_2, or

$$\mu_1 < \mu_2, \tag{2.3}$$

which states that μ_1 is smaller than μ_2. While these methods accommodate more sophisticated hypotheses than classical approaches, they are still not quantitatively incorporating the single-trial predictors $x(n)$ and are not capable to select between *trial-by-trial* models as proposed in this work.

2.2.2 The General Linear Model

The general linear model accommodates trial-by-trial models, which incorporate the quantitative influence of learning for each single exam as well as an error term. As defined in Sect. 2.1, the amount of learning for an exam $n \in \{1, ..., N_c\}$ is quantified as predictor $x(n)$ (for ease of presentation $x(n) \in \mathbb{R}$ is assumed in the following). The achieved points of that exam $y(n)$ (for ease of presentation $y(n) \in \mathbb{R}$ is assumed in the following) are modeled via the single-level general linear model (GLM)

$$y(n) = x(n)\theta + \epsilon(n), \tag{2.4}$$

with the exam-independent unknown parameter θ, which parameterizes the influence of learning and $\epsilon(n)$ as the exam-specific error which encompasses all deviations from this model. These may be due to the student having a particularly bad or good day, the test being especially difficult or easy, or any other influences on the achieved

grade. The term $x(n)\theta$ can be interpreted as the estimate $\widehat{s}(n)$ of the clean data $s(n)$, yielding

$$y(n) = \widehat{s}(n) + \epsilon(n) = x(n)\theta + \epsilon(n). \tag{2.5}$$

The parameter θ is classically estimated by minimizing the error $\epsilon(n)$, using linear regression (Fahrmeir and Tutz 1994). While (2.5) models only one data point, it can be expressed for all exams simultaneously in matrix notation as

$$\mathbf{y} = \widehat{\mathbf{s}} + \boldsymbol{\epsilon} = \mathbf{x}\theta + \boldsymbol{\epsilon}, \tag{2.6}$$

with data vector $\mathbf{y} = [y(n=1), ..., y(n=N)]^T \in \mathbb{R}^N$, clean data estimate $\widehat{\mathbf{s}} = [\widehat{s}(n = 1), ..., \widehat{s}(n = N)]^T \in \mathbb{R}^N$, design matrix (here a vector) $\mathbf{x} = [x(n = 1), ..., x(n = N)]^T \in \mathbb{R}^N$, and error vector $\boldsymbol{\epsilon} = [\epsilon(n = 1), ..., \epsilon(n = N)]^T \in \mathbb{R}^N$. Examining (2.6) makes the exam-independent nature of the parameter θ apparent.

For more complex models, the framework of the GLM accommodates not only one single but R different predictors $x_r(n)$ with $r \in \mathcal{R} = \{1, ..., R\}$. These predictors can, e.g., be influences like social status, private lessons, gender, etc. Consequently, R model parameters θ_r represent the weights of the predictors, giving the clean data estimate $\widehat{s}(n)$ the form

$$\widehat{s}(n) = x_{r=1}(n)\theta_{r=1} + ... + x_{r=R}(n)\theta_{r=R} \tag{2.7}$$

$$= \begin{bmatrix} x_{r=1}(n) & \cdots & x_{r=R}(n) \end{bmatrix} \begin{bmatrix} \theta_{r=1} \\ \vdots \\ \theta_{r=R} \end{bmatrix}. \tag{2.8}$$

Modeling all trials at once yields

$$\begin{bmatrix} \widehat{s}(n=1) \\ \vdots \\ \widehat{s}(n=N) \end{bmatrix} = \begin{bmatrix} x_{r=1}(n=1) & \cdots & x_{r=R}(n=1) \\ \vdots & \ddots & \vdots \\ x_{r=1}(n=N) & \cdots & x_{r=R}(n=N) \end{bmatrix} \begin{bmatrix} \theta_{r=1} \\ \vdots \\ \theta_{r=R} \end{bmatrix} \tag{2.9}$$

$$\widehat{\mathbf{s}} = \mathbf{X}\boldsymbol{\theta}, \tag{2.10}$$

which again results in the model of the measured data as in (2.6)

$$\mathbf{y} = \widehat{\mathbf{s}} + \boldsymbol{\epsilon} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \tag{2.11}$$

with the design matrix $\mathbf{X} \in \mathbb{R}^{N \times R}$ and vector $\boldsymbol{\theta} \in \mathbb{R}^R$. A common type of model consists of $R = 2$ predictors with one being a constant $x_1(n) = 1$, which models an intercept (i.e., θ_1 are the points the student would get without any preparation). Consequently, $x_2(n)$ is the amount of time spent learning for test n, while θ_2 is the influence of time spent learning on the achieved points.

In contrast to classical or informative hypotheses, competing models $m \in \mathcal{M}$, with m as model index and $\mathcal{M}$ as model space, are not specified by different inequality constraints of means μ_c, but by different predictors in the design matrices $\mathbf{X}$ (e.g., $m =$ time spent learning versus $m =$ time spent playing video games). Note that if the design matrix of one model consists of multiple predictors x_r, the estimated parameters θ_r allow for inference regarding the influence of the specific predictors (Friston et al. 2002). Methods for parameter estimation and model selection are described in Sect. 2.4.

2.3 Hierarchical Multiple-Level Models

A single-level GLM (2.11) can be extended by additional levels which allow for the parameters of a lower level themselves to be modeled by a higher level. In these multiple-level GLMs, the first-level models the data $\mathbf{y}$ as a linear combination of predictors $\mathbf{X}^{(1)}$ weighted by parameters $\boldsymbol{\theta}^{(1)}$, and an additive error $\boldsymbol{\epsilon}^{(1)}$, which is exactly (2.11) with an additional superscript $(\,)^{(1)}$ indicating the first level (Kiebel and Holmes 2003). The second level sets priority on the first-level parameters by modeling them as consisting of a design matrix $\mathbf{X}^{(2)}$, parameters $\boldsymbol{\theta}^{(2)}$, and errors $\boldsymbol{\epsilon}^{(2)}$ (Friston and Penny 2003). The second level parameters can again be modeled by a third level, which consists of a design matrix $\mathbf{X}^{(3)}$, parameters $\boldsymbol{\theta}^{(3)}$, and errors $\boldsymbol{\epsilon}^{(3)}$ (Friston et al. 2002). This section elaborates on multiple-level GLMs within the scope to which they are applied in this work. Readers interested in a more generalized introduction to linear hierarchical models are referred to Fahrmeir and Tutz (1994).

2.3.1 The First Level

The first level of a multiple-level GLM is the same as the single-level GLM (2.11)

$$\begin{bmatrix} y(n=1) \\ \vdots \\ y(n=N) \end{bmatrix} = \begin{bmatrix} x^{(1)}_{r=1}(n=1) & \cdots & x^{(1)}_{r=R}(n=1) \\ \vdots & \ddots & \vdots \\ x^{(1)}_{r=1}(n=N) & \cdots & x^{(1)}_{r=R}(n=N) \end{bmatrix} \begin{bmatrix} \theta^{(1)}_{r=1} \\ \vdots \\ \theta^{(1)}_{r=R} \end{bmatrix} + \begin{bmatrix} \epsilon^{(1)}(n=1) \\ \vdots \\ \epsilon^{(1)}(n=N) \end{bmatrix} \tag{2.12}$$

or, in matrix notation,

$$\mathbf{y} = \mathbf{X}^{(1)}\boldsymbol{\theta}^{(1)} + \boldsymbol{\epsilon}^{(1)}, \tag{2.13}$$

with data vector $\mathbf{y} \in \mathbb{R}^N$, design matrix $\mathbf{X}^{(1)} \in \mathbb{R}^{N \times R}$, parameter vector $\boldsymbol{\theta}^{(1)} \in \mathbb{R}^R$, and error vector $\boldsymbol{\epsilon}^{(1)} \in \mathbb{R}^N$.

2.3.2 The Second Level

The second level models the first-level parameters $\boldsymbol{\theta}^{(1)}$ following

$$\boldsymbol{\theta}^{(1)} = \mathbf{X}^{(2)}\boldsymbol{\theta}^{(2)} + \boldsymbol{\epsilon}^{(2)}, \tag{2.14}$$

with design matrix $\mathbf{X}^{(2)} \in \mathbb{R}^{R\times R}$, parameters $\boldsymbol{\theta}^{(2)} \in \mathbb{R}^{R}$, and error $\boldsymbol{\epsilon}^{(2)} \in \mathbb{R}^{R}$ (Penny et al. 2003). If the second level design matrix $\mathbf{X}^{(2)}$ is chosen to be all zeros $\mathbf{X}^{(2)} = \mathbf{0} \in \mathbb{R}^{R\times R}$, an unconstrained prior is set on the first-level parameters $\boldsymbol{\theta}^{(1)}$ (Friston et al. 2007)

$$\boldsymbol{\theta}^{(1)} = \boldsymbol{\epsilon}^{(2)}, \tag{2.15}$$

which allows for single-level Bayesian inference (Ostwald et al. 2012). This two-level GLM

$$\begin{aligned} \mathbf{y} &= \mathbf{X}^{(1)}\boldsymbol{\theta}^{(1)} + \boldsymbol{\epsilon}^{(1)} \\ \boldsymbol{\theta}^{(1)} &= \boldsymbol{\epsilon}^{(2)}. \end{aligned} \tag{2.16}$$

is used for Bayesian model estimation throughout this work.

2.3.3 The Third Level

For studies with L different subjects $\ell \in \{1, ..., L\}$, the second level can be used to facilitate a mixed effects analysis (Friston et al. 2002). In this case, each subject is modeled with an individual set of first-level parameters $\boldsymbol{\theta}_{\ell}^{(1)}$ which are shrunk towards subject-independent second level group parameters $\boldsymbol{\theta}^{(2)}$. These parameters are consequently modeled by a third level

$$\boldsymbol{\theta}^{(2)} = \mathbf{X}^{(3)}\boldsymbol{\theta}^{(3)} + \boldsymbol{\epsilon}^{(3)}, \tag{2.17}$$

with design matrix $\mathbf{X}^{(3)} \in \mathbb{R}^{R\times R}$, parameters $\boldsymbol{\theta}^{(3)} \in \mathbb{R}^{R}$, and error $\boldsymbol{\epsilon}^{(3)} \in \mathbb{R}^{R}$. Such three-level GLMs are used in the studies by Mars et al. (2008) and Kolossa et al. (2013), where the data from all L subjects is modeled simultaneously. These studies use an unconstrained prior on the second level parameters $\boldsymbol{\theta}^{(2)}$ by employing an all-zero third-level design matrix $\mathbf{X}^{(3)} = \mathbf{0} \in \mathbb{R}^{R\times R}$. The detailed composition of the vectors and matrices is as follows: The subject-specific data vectors $\mathbf{y}_{\ell} \in \mathbb{R}^{N_{\ell}}$ with N_{ℓ} as the number of trials of a subject ℓ, design matrices $\mathbf{X}_{\ell} \in \mathbb{R}^{N_{\ell}\times R}$, first-level parameters $\boldsymbol{\theta}_{\ell}^{(1)} \in \mathbb{R}^{R}$, and error vectors $\boldsymbol{\epsilon}_{\ell}^{(1)} \in \mathbb{R}^{N_{\ell}}$ are augmented to yield the first level of the GLM (Friston et al. 2002)

$$\begin{bmatrix} \mathbf{y}_{\ell=1} \\ \vdots \\ \mathbf{y}_{\ell=L} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{(1)}_{\ell=1} & \mathbf{0}_{N_{\ell=1}\times R} & \cdots & \mathbf{0}_{N_{\ell=1}\times R} \\ \mathbf{0}_{N_{\ell=2}\times R} & \mathbf{X}^{(1)}_{\ell=2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0}_{N_{\ell=L-1}\times R} \\ \mathbf{0}_{N_{\ell=L}\times R} & \cdots & \mathbf{0}_{N_{\ell=L}\times R} & \mathbf{X}^{(1)}_{\ell=L} \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}^{(1)}_{\ell=1} \\ \vdots \\ \boldsymbol{\theta}^{(1)}_{\ell=L} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}^{(1)}_{\ell=1} \\ \vdots \\ \boldsymbol{\epsilon}^{(1)}_{\ell=L} \end{bmatrix}, \tag{2.18}$$

with all-zero matrices $\mathbf{0}_{N_\ell \times R} \in \mathbb{R}^{N_\ell \times R}$ specifying inter-subject independence of the parameters $\boldsymbol{\theta}^{(1)}_\ell$. Note how this model allows for unbalanced data sets, i.e., variable N_ℓ over subjects. The first level (2.18) is written in condensed form as

$$\mathbf{y} = \mathbf{X}^{(1)}\boldsymbol{\theta}^{(1)} + \boldsymbol{\epsilon}^{(1)}, \tag{2.19}$$

with data vector $\mathbf{y} \in \mathbb{R}^N$, where $N = \sum_{\ell=1}^{L} N_\ell$ is the total number of trials over subjects, design matrix $\mathbf{X}^{(1)} \in \mathbb{R}^{N \times LR}$, parameter vector $\boldsymbol{\theta}^{(1)} \in \mathbb{R}^{LR}$, and error vector $\boldsymbol{\epsilon}^{(1)} \in \mathbb{R}^N$. The second level models the subject-individual first-level parameters $\boldsymbol{\theta}^{(1)}_\ell$ as samples from subject-independent group parameters $\boldsymbol{\theta}^{(2)} \in \mathbb{R}^R$. This requires a second level design matrix $\mathbf{X}^{(2)} \in \mathbb{R}^{LR \times R}$ that consists of L stacks of identity matrices $\mathbf{I}_R \in \mathbb{R}^{R \times R}$, giving the second level the form

$$\begin{bmatrix} \boldsymbol{\theta}^{(1)}_{\ell=1} \\ \vdots \\ \boldsymbol{\theta}^{(1)}_{\ell=L} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_R \\ \vdots \\ \mathbf{I}_R \end{bmatrix} \boldsymbol{\theta}^{(2)} + \begin{bmatrix} \boldsymbol{\epsilon}^{(2)}_{\ell=1} \\ \vdots \\ \boldsymbol{\epsilon}^{(2)}_{\ell=L} \end{bmatrix}, \tag{2.20}$$

which is summarized as the standard second level (2.14)

$$\boldsymbol{\theta}^{(1)} = \mathbf{X}^{(2)}\boldsymbol{\theta}^{(2)} + \boldsymbol{\epsilon}^{(2)}. \tag{2.21}$$

The third level (2.17) sets an unconstrained prior on the parameters of the second level via an all-zero third-level design matrix $\mathbf{X}^{(3)} = \mathbf{0} \in \mathbb{R}^{R \times R}$, which yields

$$\boldsymbol{\theta}^{(2)} = \boldsymbol{\epsilon}^{(3)}, \tag{2.22}$$

and, in summary, the complete three-level GLM

$$\begin{aligned} \mathbf{y} &= \mathbf{X}^{(1)}\boldsymbol{\theta}^{(1)} + \boldsymbol{\epsilon}^{(1)} \\ \boldsymbol{\theta}^{(1)} &= \mathbf{X}^{(2)}\boldsymbol{\theta}^{(2)} + \boldsymbol{\epsilon}^{(2)} \\ \boldsymbol{\theta}^{(2)} &= \boldsymbol{\epsilon}^{(3)}. \end{aligned} \tag{2.23}$$

2.4 Model Estimation and Selection

After having specified the linear models in the previous sections, final results are obtained in two steps: (1) Estimation of the unknown parameters $\boldsymbol{\theta}$, and (2) calculation of the model likelihoods for model selection. This section starts with motivating the use of Bayesian model estimation methods, which is followed by a detailed description of the estimation schemes for the two-level GLM used in this work (see Sect. 2.3.2). It is concluded by instructions for Bayesian model selection. Readers interested in generalized Bayesian estimation schemes for GLMs of any order are referred to Friston et al. (2002).

A well-known method for single-level model estimation and selection is based on the mean squared error (MSE) between the clean data estimates $\widehat{\mathbf{s}}$ (2.5) and the measured data $\mathbf{y}$ (Kleinbaum et al. 2013)

$$\mathrm{MSE}\,(\widehat{\mathbf{s}}, \mathbf{y}) = \frac{1}{N}\sum_{n=1}^{N}(\widehat{s}(n) - y(n))^2. \tag{2.24}$$

(1) For each model $m \in \mathcal{M}$, the parameters $\boldsymbol{\theta}_m$ are optimized by finding the parameters which yield the smallest MSE

$$\boldsymbol{\theta}_{m,\mathrm{opt}} = \arg\min_{\boldsymbol{\theta}_m} \left\{\mathrm{MSE}\,(\widehat{\mathbf{s}}_m, \mathbf{y})\right\}. \tag{2.25}$$

(2) Now the models employ $\boldsymbol{\theta}_{m,\mathrm{opt}}$ and the model

$$m_{\mathrm{opt}} = \arg\min_{m} \left\{\mathrm{MSE}(\widehat{\mathbf{s}}_{m,\mathrm{opt}}, \mathbf{y})\right\}, \tag{2.26}$$

with the smallest MSE is selected as the best model m_{opt}, as it offers the best fit (or accuracy) of the measured data. A major shortcoming of this comparison scheme is its blindness to the complexity of the models, i.e., the number R of model parameters θ_r, which are used to calculate $\widehat{s}(n)$.

Bayesian evaluation schemes take the complexity of the models into account by employing a penalty factor for complexity, which is often referred to as *Occam's razor* (MacKay 1992). The reason is to choose the least complex model that offers a good explanation of the data (Myung and Pitt 1997). The most vivid factor influencing model complexity is the number R of model parameters. In the context of general linear models, the design matrix of the least complex model consists of only one constant predictor per trial (i.e., $R = 1$), yielding an all-one vector $\mathbf{x}^{(1)} = [1 \ldots 1]^T \in \mathbb{R}^N$, which is equivalent to the classical null hypothesis. Consequently, the most complex model fits the data perfectly and has as much predictors as trials, i.e., $R = N$, and an identity design matrix $\mathbf{X}^{(1)} = \mathbf{I}_N \in \mathbb{R}^{N\times N}$. Obviously, this model is overfitted and offers no theory behind the data-generating process (MacKay 1992; Pitt and Myung 2002). In this work, these shortcomings are addressed by parametric empirical Bayesian (PEB) methods used for parameter estimation and model selection (Friston et al. 2002).

In this framework, competing models are selected based on their log-likelihoods, which are derived by taking a complexity-accuracy trade-off into account (Friston et al. 2007). It is important to note that complexity is not solely based on the number of model parameters, but on other factors like parameter independence as well (see Stephan et al. 2009 for details). This Bayesian framework is proven to be a useful tool for model selection in many fields (Hoeting et al. 1999; Pitt and Myung 2002; Penny et al. 2010) and has been able to significantly advance neuroimaging research (Woolrich 2012).

2.4.1 Collapsing and Augmenting the Hierarchical Model

The first step for parameter estimation and evidence calculation via PEB are alterations to the model structure of the two-level GLM (see Sect. 2.3.2). Specifically, the model is first collapsed to a non-hierarchical form and subsequently augmented, so that all parameters appear in the error vector and can be estimated at once using expectation maximization (EM) (Friston et al. 2002). For the two-level GLM (2.13), (2.14)

$$\mathbf{y} = \mathbf{X}^{(1)}\boldsymbol{\theta}^{(1)} + \boldsymbol{\epsilon}^{(1)} \tag{2.27}$$

$$\boldsymbol{\theta}^{(1)} = \mathbf{X}^{(2)}\boldsymbol{\theta}^{(2)} + \boldsymbol{\epsilon}^{(2)} \tag{2.28}$$

the errors on both levels are assumed to be normally distributed

$$\boldsymbol{\epsilon}^{(1)} \sim \mathcal{N}(0, \Sigma_{\epsilon}^{(1)}) \tag{2.29}$$

$$\boldsymbol{\epsilon}^{(2)} \sim \mathcal{N}(0, \Sigma_{\epsilon}^{(2)}), \tag{2.30}$$

with zero mean and isotropic error covariance matrices

$$\Sigma_{\epsilon}^{(1)} = \lambda^{(1)}\mathbf{I}_N \tag{2.31}$$

$$\Sigma_{\epsilon}^{(2)} = \lambda^{(2)}\mathbf{I}_R. \tag{2.32}$$

The hyper-parameters $\lambda^{(1)}$ and $\lambda^{(2)}$ control the variances at level (1) and (2), respectively (Mars et al. 2008). They are called hyper-parameters because they parameterize the covariance of the errors $\boldsymbol{\epsilon}^{(1)}$ and $\boldsymbol{\epsilon}^{(2)}$ (Friston and Penny 2003; Penny 2012). Note that this section covers only models where there are no hyper-priors on the hyper-parameters (see Friston et al. 2007 for details). Identity matrices $\mathbf{I}_N \in \mathbb{R}^{N\times N}$ and $\mathbf{I}_R \in \mathbb{R}^{R\times R}$ place independence assumptions over trials and parameters, respectively (Ostwald et al. 2012). While there can be multiple covariance components on any level, only one covariance component per level is assumed in this work. Substitution of (2.28) in (2.27) yields the non-hierarchical form

$$\mathbf{y} = \mathbf{X}^{(1)}\boldsymbol{\epsilon}^{(2)} + \mathbf{X}^{(1)}\mathbf{X}^{(2)}\boldsymbol{\theta}^{(2)} + \boldsymbol{\epsilon}^{(1)} \tag{2.33}$$

$$= \left[\mathbf{X}^{(1)}\ \mathbf{X}^{(1)}\mathbf{X}^{(2)}\right]\begin{bmatrix}\boldsymbol{\epsilon}^{(2)}\\ \boldsymbol{\theta}^{(2)}\end{bmatrix} + \boldsymbol{\epsilon}^{(1)}, \tag{2.34}$$

which is augmented so that the parameters appear in the error vector

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{0}_{R\times 1}\\ \mathbf{0}_{R\times 1}\end{bmatrix} = \begin{bmatrix}\mathbf{X}^{(1)} & \mathbf{X}^{(1)}\mathbf{X}^{(2)}\\ -\mathbf{I}_R & \mathbf{0}_{R\times R}\\ \mathbf{0}_{R\times R} & -\mathbf{I}_R\end{bmatrix}\begin{bmatrix}\boldsymbol{\epsilon}^{(2)}\\ \boldsymbol{\theta}^{(2)}\end{bmatrix} + \begin{bmatrix}\boldsymbol{\epsilon}^{(1)}\\ \boldsymbol{\epsilon}^{(2)}\\ \boldsymbol{\theta}^{(2)}\end{bmatrix}. \tag{2.35}$$

The augmented model can be expressed condensedly as

$$\widetilde{\mathbf{y}} = \widetilde{\mathbf{X}}\,\widetilde{\boldsymbol{\theta}} + \widetilde{\boldsymbol{\epsilon}}. \tag{2.36}$$

This reformulation of the hierarchical model is computationally efficient and allows an instructive form of the EM algorithm (Friston et al. 2002; Friston et al. 2007). The error covariance matrix of the augmented form is assembled following

$$\Sigma_{\widetilde{\epsilon}} = \sum_{i=1}^{2}\lambda_i\mathbf{Q}_i + \Sigma_{\widetilde{\theta}}, \tag{2.37}$$

with $\lambda_i = \lambda^{(i)}$, as each level is modeled with a single covariance component. The matrices $\mathbf{Q}_1 \in \mathbb{R}^{(N+2R)\times(N+2R)}$ and $\mathbf{Q}_2 \in \mathbb{R}^{(N+2R)\times(N+2R)}$ are the augmented forms of the identity matrices in (2.31) and (2.32), more precisely

$$\mathbf{Q}_1 = \begin{bmatrix}\mathbf{I}_N & \mathbf{0}_{N\times R} & \mathbf{0}_{N\times R}\\ \mathbf{0}_{R\times N} & \mathbf{0}_{R\times R} & \mathbf{0}_{R\times R}\\ \mathbf{0}_{R\times N} & \mathbf{0}_{R\times R} & \mathbf{0}_{R\times R}\end{bmatrix} \tag{2.38}$$

and

$$\mathbf{Q}_2 = \begin{bmatrix}\mathbf{0}_{N\times N} & \mathbf{0}_{N\times R} & \mathbf{0}_{N\times R}\\ \mathbf{0}_{R\times R} & \mathbf{I}_R & \mathbf{0}_{R\times R}\\ \mathbf{0}_{R\times N} & \mathbf{0}_{R\times R} & \mathbf{0}_{R\times R}\end{bmatrix}. \tag{2.39}$$

The parameter covariance matrix $\Sigma_{\widetilde{\theta}} \in \mathbb{R}^{(N+2R)\times(N+2R)}$ is of the form

$$\Sigma_{\widetilde{\theta}} = \begin{bmatrix}\mathbf{0}_{N\times N} & \mathbf{0}_{N\times R} & \mathbf{0}_{N\times R}\\ \mathbf{0}_{R\times N} & \mathbf{0}_{R\times R} & \mathbf{0}_{R\times R}\\ \mathbf{0}_{R\times R} & \mathbf{0}_{R\times R} & \Sigma_\theta^{(2)}\end{bmatrix}, \tag{2.40}$$

with $\Sigma_\theta^{(2)} = e^{16}\mathbf{I}_R \in \mathbb{R}^{R\times R}$ specifying an unconstrained prior on the second-level parameters, with Euler's number $e = \sum_{\kappa=0}^{\infty}\frac{1}{\kappa!}$.

2.4.2 Model Parameter Optimization and Likelihood Calculation

Based on the augmented model (2.36), PEB methods commence to compute the conditional *posterior probability* densities of the *parameters* using *empirical* data, coining the term *parametric empirical Bayes* (Friston et al. 2002). In this framework, the parameters $\widetilde{\boldsymbol{\theta}}$ are modeled to be normally distributed random variables $\mathrm{p}(\widetilde{\boldsymbol{\theta}}|\mathbf{y}) = \mathcal{N}(\widetilde{\boldsymbol{\theta}};\, \boldsymbol{\mu}_{\widetilde{\theta}|y},\, \Sigma_{\widetilde{\theta}|y})$ with conditional means $\boldsymbol{\mu}_{\widetilde{\theta}|y}$ and covariance matrix $\Sigma_{\widetilde{\theta}|y}$. This approach is fundamentally different from classical parameter optimization, where the parameters are assumed to be fixed values.

The parameter densities $\mathrm{p}(\widetilde{\boldsymbol{\theta}}|\mathbf{y})$ are estimated via expectation maximization (EM). The EM algorithm estimates the parameter densities by maximizing the free energy $F_{\widetilde{\theta}}$, which gives a lower bound approximation of the log-likelihood of the data conditioned on the hyper-parameters (Friston et al. 2002)

$$F_{\widetilde{\theta}} \leq \log \mathrm{p}(\mathbf{y}|\boldsymbol{\lambda}) = \log \int \mathrm{p}(\widetilde{\boldsymbol{\theta}}, \mathbf{y}|\boldsymbol{\lambda}) d\widetilde{\boldsymbol{\theta}}. \tag{2.41}$$

See Friston et al. (2002, 2007) for a detailed and thorough derivation. For the augmented general linear model (2.36), the free energy is calculated according to (Friston et al. 2007)

$$F_{\widetilde{\theta}} = -\frac{N}{2}\log(2\pi) - \frac{1}{2}(\mathbf{G}\widetilde{\mathbf{y}})^T \Sigma_{\widetilde{\epsilon}}(\mathbf{G}\widetilde{\mathbf{y}}) + \frac{1}{2}\log|\Sigma_{\widetilde{\epsilon}}^{-1}| + \frac{1}{2}\log|\Sigma_{\widetilde{\theta}|y}|, \tag{2.42}$$

with $|\cdot|$ denoting the determinant of a matrix and $\mathbf{G} = \Sigma_{\widetilde{\epsilon}}^{-1} - \Sigma_{\widetilde{\epsilon}}^{-1}\widetilde{\mathbf{X}}\Sigma_{\widetilde{\theta}|y}\widetilde{\mathbf{X}}^T\Sigma_{\widetilde{\epsilon}}^{-1}$. The composition of all the terms in (2.42) will be described shortly during the summary of the EM algorithm. Generally, EM is an iterative algorithm for maximum likelihood estimation of data conditional on unobserved parameters (Neal and Hinton 1998). Readers completely unfamiliar with EM algorithms are referred to Fahrmeir and Tutz (1994) for a general introduction to EM in the context of linear models. In the PEB framework, the EM algorithm alternates between maximizing the free energy with regard to the parameter distribution $\mathrm{p}(\widetilde{\boldsymbol{\theta}}|\mathbf{y}) = \mathcal{N}(\widetilde{\boldsymbol{\theta}};\, \boldsymbol{\mu}_{\widetilde{\theta}|y},\, \Sigma_{\widetilde{\theta}|y})$ in the E-step and with regard to the hyper-parameters $\boldsymbol{\lambda} = [\lambda_1\ \lambda_2]^T$ in the M-step. Under Gaussian assumptions, the E-step is simply the calculation of the conditional mean and covariance of the model parameters while keeping the hyper-parameters fixed, following (Friston 2002)

$$\Sigma_{\widetilde{\epsilon}} = \Sigma_{\widetilde{\theta}} + \sum_{i=1}^{2} \lambda_i \mathbf{Q}_i \tag{2.43}$$

$$\Sigma_{\widetilde{\theta}|y} = \left(\widetilde{\mathbf{X}}^T \Sigma_{\widetilde{\epsilon}}^{-1} \widetilde{\mathbf{X}}\right)^{-1} \tag{2.44}$$

$$\boldsymbol{\mu}_{\widetilde{\theta}|y} = \Sigma_{\widetilde{\theta}|y} \widetilde{\mathbf{X}}^T \Sigma_{\widetilde{\epsilon}}^{-1} \widetilde{\mathbf{y}}. \tag{2.45}$$

The M-step serves to estimate the error covariances $\Sigma_{\tilde{\epsilon}}$ which rest upon the hyper-parameters $\boldsymbol{\lambda}$, as can be seen in (2.43). They are obtained by maximizing the free energy $F_{\tilde{\theta}}$ while keeping the conditional mean and covariance of the parameters fixed (Friston et al. 2002)

$$\mathbf{G} = \Sigma_{\tilde{\epsilon}}^{-1} - \Sigma_{\tilde{\epsilon}}^{-1} \widetilde{\mathbf{X}} \Sigma_{\tilde{\theta}|y} \widetilde{\mathbf{X}}^T \Sigma_{\tilde{\epsilon}}^{-1} \tag{2.46}$$

$$\boldsymbol{h} = \begin{bmatrix} \frac{\partial F_{\tilde{\theta}}}{\partial \lambda_1} \\ \frac{\partial F_{\tilde{\theta}}}{\partial \lambda_2} \end{bmatrix} = \begin{bmatrix} -\frac{1}{2}\mathrm{tr}\{\mathbf{G}\mathbf{Q}_1\} + \frac{1}{2}\widetilde{\mathbf{y}}^T \mathbf{G}^T \mathbf{Q}_1 \mathbf{G}\widetilde{\mathbf{y}} \\ -\frac{1}{2}\mathrm{tr}\{\mathbf{G}\mathbf{Q}_2\} + \frac{1}{2}\widetilde{\mathbf{y}}^T \mathbf{G}^T \mathbf{Q}_2 \mathbf{G}\widetilde{\mathbf{y}} \end{bmatrix} \tag{2.47}$$

$$\mathbf{H} = \begin{bmatrix} \langle -\frac{\partial^2 F_{\tilde{\theta}}}{\partial \lambda_1 \partial \lambda_1} \rangle & \langle -\frac{\partial^2 F_{\tilde{\theta}}}{\partial \lambda_1 \partial \lambda_2} \rangle \\ \langle -\frac{\partial^2 F_{\tilde{\theta}}}{\partial \lambda_2 \partial \lambda_1} \rangle & \langle -\frac{\partial^2 F_{\tilde{\theta}}}{\partial \lambda_2 \partial \lambda_2} \rangle \end{bmatrix} = \begin{bmatrix} \frac{1}{2}\mathrm{tr}\{\mathbf{G}\mathbf{Q}_1\mathbf{G}\mathbf{Q}_1\} & \frac{1}{2}\mathrm{tr}\{\mathbf{G}\mathbf{Q}_1\mathbf{G}\mathbf{Q}_2\} \\ \frac{1}{2}\mathrm{tr}\{\mathbf{G}\mathbf{Q}_2\mathbf{G}\mathbf{Q}_1\} & \frac{1}{2}\mathrm{tr}\{\mathbf{G}\mathbf{Q}_2\mathbf{G}\mathbf{Q}_2\} \end{bmatrix} \tag{2.48}$$

$$\boldsymbol{\lambda} = \boldsymbol{\lambda} + \mathbf{H}^{-1}\boldsymbol{h} = \boldsymbol{\lambda} + \Delta\boldsymbol{\lambda}. \tag{2.49}$$

The operator $\mathrm{tr}\{\cdot\}$ denotes the trace of a matrix, $\langle\cdot\rangle$ is the expectation operator, $\mathbf{h}$ is a gradient vector, and $\mathbf{H}$ is referred to as Fisher's information matrix (Friston et al. 2002). Note that (2.46) and (2.47) use the first and expected second partial derivatives of the free energy with regard to the hyper-parameters, which are known as *Fisher scoring* (Friston et al. 2002). Steps (2.43)–(2.48) are repeated until convergence, which can be a specific value of $\Delta\boldsymbol{\lambda}$ or some fixed number of iterations. Readers interested in the formal derivation of (2.43)–(2.48) and more generalized applications, such as multiple covariance constraints on any level or higher order models, are referred to Friston et al. (2002).

After convergence of the EM algorithm, the parameter densities are estimated and the free energy $F_{\tilde{\theta}}$ is adjusted to yield the variational free energy F used for model selection (Friston and Penny 2003; Friston et al. 2007)

$$F = \underbrace{-\frac{N}{2}\log(2\pi) - \frac{1}{2}(\mathbf{G}\widetilde{\mathbf{y}})^T \Sigma_{\tilde{\epsilon}}(\mathbf{G}\widetilde{\mathbf{y}}) + \frac{1}{2}\log|\Sigma_{\tilde{\epsilon}}^{-1}|}_{\text{accuracy term}} + \underbrace{\frac{1}{2}\log|\Sigma_{\tilde{\theta}|y}| + \frac{1}{2}\log|-\mathbf{H}^{-1}|}_{\text{complexity term}}. \tag{2.50}$$

The variational free energy is a lower bound approximation of the usually not directly computable log-likelihood $\log(p(\mathbf{y}|m))$, i.e., the logarithm of the probability of the data $\mathbf{y}$ given the model $m \in \mathcal{M} = \{1, ..., M\}$ (Penny et al. 2010). Many approximations to the log-likelihood have been proposed, with the variational free energy F being superior and commonly used in neuroimaging (Beal 2003; Beal and Ghahramani 2003; Friston et al. 2007; Penny 2012).

The only values of further interest besides the variational free energy are the conditional means of the parameters which serve as their point estimates. They can be used for model fitting or comparison of effect sizes, with the latter being only sound if the data and regressors have been normalized (Hoijtink 2012). Using the

vector of the conditional means of the augmented model $\boldsymbol{\mu}_{\tilde{\theta}|y} = \left[\boldsymbol{\mu}_{\epsilon|y}^{(2)}\ \boldsymbol{\mu}_{\theta|y}^{(2)}\right]^T$ from (2.45), the conditional means of the first-level parameters are calculated according to

$$\boldsymbol{\mu}_{\theta|y}^{(1)} = \mathbf{X}^{(2)}\boldsymbol{\mu}_{\theta|y}^{(2)} + \boldsymbol{\mu}_{\epsilon|y}^{(2)}. \tag{2.51}$$

The described methods for parameter estimation and variational free energy calculation are implemented in the `spm_PEB.m` function of the freely available Statistical Parametric Mapping (SPM8) software (Dempster et al. 1981; Friston et al. 2002, 2007), which was used for model estimation in this work.

2.4.3 Model Selection Using Bayes Factors and Posterior Model Probabilities

After parameter estimation and calculation of the variational free energy F (2.50) in the previous chapter, the best model m can be selected. If the choice is solely among two models $\mathcal{M} = \{1, 2\}$, the Bayes factor (BF) is a suitable measure (Kass and Raftery 1995). It is the ratio of the model likelihoods

$$\mathrm{BF}_{1\to 2} = \frac{\mathrm{p}(\mathbf{y}|m=1)}{\mathrm{p}(\mathbf{y}|m=2)}, \tag{2.52}$$

whose natural logarithm, log(BF), equals the difference of variational free energy (Penny et al. 2004)

$$\log\left(\mathrm{BF}_{1\to 2}\right) = \log\left(\frac{\mathrm{p}(\mathbf{y}|m=1)}{\mathrm{p}(\mathbf{y}|m=2)}\right) = F_{m=1} - F_{m=2}. \tag{2.53}$$

Positive values reflect evidence in favor of model 1 over 2. The interpretation of log-Bayes factors is often unintuitive, and their bilateral nature makes the description of selection procedures with $M > 2$ models unnecessarily complex. The interpretation becomes most vivid and independent of the number of evaluated models by computing posterior model probabilities (PMP) $\mathrm{P}(m|\mathbf{y})$, which are calculated based on the model likelihoods following Bayes' rule (Penny et al. 2010)

$$\mathrm{P}(m|\mathbf{y}) = \frac{\mathrm{p}(\mathbf{y}|m)\mathrm{P}(m)}{\sum\limits_{\mu\in\mathcal{M}} \mathrm{p}(\mathbf{y}|\mu)\mathrm{P}(\mu)}, \quad \forall m \in \mathcal{M}, \tag{2.54}$$

with $\mathrm{P}(m)$ as the prior probability of model m. Assuming equal prior probabilities of $\mathrm{P}(m) = \frac{1}{M}, \forall m \in \mathcal{M}$, (2.54) simplifies to

Table 2.1 Bayes factors BF, log-Bayes factors log(BF), posterior model probabilities $P(m|\mathbf{y})$, and how to interpret them (Kass and Raftery 1995; Penny et al. 2004)

$BF_{1\to 2}$	$\log(BF_{1\to 2})$	$P(m|\mathbf{y})$	Significance
1–3	0–1.1	0.50–0.75	Weak
3–20	1.1–3	0.75–0.95	Positive
20–150	3–5	0.95–0.99	Strong
>150	>5	>0.99	Very strong

The Bayes factor and log-Bayes factor compare model $m = 1$ with model $m = 2$. The significance boundaries as indicated apply only for the comparison of two models. As the posterior model probability $P(m|\mathbf{y})$ allows for simultaneous selection among any number of models, the size of the model space has to be taken into account for interpretation

$$P(m|\mathbf{y}) = \frac{p(\mathbf{y}|m)}{\sum_{\mu\in\mathcal{M}} p(\mathbf{y}|\mu)} = \frac{e^{F_m}}{\sum_{\mu\in\mathcal{M}} e^{F_\mu}}. \tag{2.55}$$

The posterior model probability is normalized to the model space $\mathcal{M} = \{1, ..., M\}$ with $\sum_{m\in\mathcal{M}} P(m|\mathbf{y}) = 1$ and is interpreted as the probability that model m is the best model, given the observed data *and* given all evaluated models. Table 2.1 gives an overview for the interpretation of Bayes factors BF, log-Bayes factors log(BF), and posterior model probabilities $P(m|\mathbf{y})$. Note that the indicated significance boundaries apply for the comparison of two models only. To calculate $P(m|\mathbf{y})$ for a model space $\mathcal{M}$ with $M > 2$ is feasible, but the interpretation of the achieved posterior model probabilities has to account for the size of the model space. For $M = 3$ models, $P(m|\mathbf{y}) \approx 0.33$ states equal posterior model probabilities.

2.4.4 Group Studies

Studies typically contain data from more than one subject. While Mars et al. (2008) and Kolossa et al. (2013) model all subjects at once using third-order GLMs as described in Sect. 2.3.3, there is an emerging tendency in neuroimaging to use subject-specific model estimation with subsequent group-level selection (see, e.g., Garrido et al. 2009; Ostwald et al. 2012; Lieder et al. 2013; Kolossa et al. 2015).

This strict separation of single-subject estimation and group-level selection enables the specification of assumptions about the model distribution over subjects. Specifically, fixed-effects (Stephan et al. 2007) and random-effects analyses (Stephan et al. 2009) can be applied. Penny et al. (2010) give an overview on both approaches: Fixed-effects analyses assume that all subjects use the same model, and should be applied for models of basic functions which are not expected to differ across subjects. Random-effects analyses allow for individual subjects to use different models, and should be used for cognitive tasks which can be solved with different learned strategies. An important drawback of both approaches is the so-called brittleness:

For fixed-effects analyses, the results can become ambiguous if different subjects use different models, or if the model space consists of a large number of models, while for random-effects analyses the addition of just one model to the model space may alter the mutual ranking of all other models (Penny et al. 2010).

Probabilistic inference is assumed to be a basic function and not a consciously available learned scheme. Throughout this work the model space remains small with $M < 20$. Based on these considerations, fixed-effects analyses for group studies will now be introduced and applied in this work.

Given the subject-specific model likelihoods $\mathrm{p}(\mathbf{y}_\ell|m)$, the group-level likelihood $\mathrm{p}(\mathbf{y}|m)$ is obtained following (Stephan et al. 2007)

$$\mathrm{p}(\mathbf{y}|m) = \prod_{\ell=1}^{L} \mathrm{p}(\mathbf{y}_\ell|m), \tag{2.56}$$

which can be equivalently expressed in log-likelihood and variational free energy as

$$\log\left(\mathrm{p}(\mathbf{y}|m)\right) = \sum_{\ell=1}^{L} \log\left(\mathrm{p}(\mathbf{y}_\ell|m)\right) = F_m = \sum_{\ell=1}^{L} F_{m,\ell}. \tag{2.57}$$

Based on the group log-likelihoods, the group log-Bayes factor $\log(\mathrm{GBF})$ (2.53) between two models can be calculated according to

$$\log\left(\mathrm{GBF}_{1\to 2}\right) = \log\left(\frac{\mathrm{p}(\mathbf{y}|m=1)}{\mathrm{p}(\mathbf{y}|m=2)}\right) = F_{m=1} - F_{m=2}, \tag{2.58}$$

while the posterior model probability $\mathrm{P}(m|\mathbf{y})$ (2.55) for any number of models follows

$$\mathrm{P}(m|\mathbf{y}) = \frac{\mathrm{p}(\mathbf{y}|m)}{\sum_{\mu\in\mathcal{M}} \mathrm{p}(\mathbf{y}|\mu)} = \frac{e^{F_m}}{\sum_{\mu\in\mathcal{M}} e^{F_\mu}}. \tag{2.59}$$

Table 2.1 can be used for the interpretation of group (log-)Bayes factors and posterior model probabilities.

2.5 A Transfer Example Experiment—Setup

This section introduces an exemplary experiment which illustrates parameter estimation and model selection as employed in this work. The experiment transfers the methods to the field of speech communication in order to show how widely applicable they are, but the origins of event-related potentials will not be left completely out of sight. This example provided the initial idea for (Kolossa et al. 2016).

Any proposed speech communication system has to be evaluated with regard to the degree to which it distorts the clean speech signal and how these distortions are perceived by humans. Mean opinion scores of listening quality subjective (MOS-LQS) are obtained in formal listening tests, where test subjects are presented with speech samples, the quality of which they rate from 1 (very bad) to 5 (excellent) (ITU 2006a, 2006b). These tests are important, as human perception is the absolute reference concerning speech signal quality, but they are also expensive and time consuming. ITU-T Recommendation Q.862 (ITU 2007) defines the perceptual evaluation of speech quality (PESQ) for automatized and fast speech quality assessment. These mean opinion scores of listening quality objective (MOS-LQO) range from 1 (very bad) to 4.5 (excellent).

The comparison of MOS-LQS with MOS-LQO values is not simple and straightforward. The sequences of MOS-LQS and MOS-LQO values can be interpreted as signals, and while at first glance MOS-LQS seems like a clean signal which captures the listening quality ratings of the subjects, it is in fact corrupted by noise. This noise consists of inter-subject as well as intra-subject contributions (Holmes and Friston 1998), the same as for neuroimaging data in general (Friston et al. 2002) and a sequence of ERP amplitudes in particular (Mars et al. 2008). Inter-subject noise captures subject-individual differences in speech quality perception, which are, for example, different score offsets or individual sensitivity to speech distortions. Intra-subject noise subsumes systematic score deviations which are, e.g., a dependency on the order the files were listened to, and random deviations which occur due to non-deterministic behavior of humans: the same person might rate the same file differently on multiple occasions (Carron and Bailey 1969). Listening tests are designed to reduce this noise by presenting the same file multiple times with different adjacent files and then average the results, but the noise is not eliminated completely. If an alternative model of MOS-LQO is challenging PESQ as the best approximation for the MOS-LQS values, noise has to be taken into account during the model selection phase. Parametric empirical Bayes (see Sect. 2.4) offers a useful tool for evaluating models of speech quality.

In this example experiment, PESQ is assumed to be the ground truth model for speech quality perception. Two kinds of noise are added with different signal-to-noise ratios (SNR) to simulate inter- and intra-subject variability yielding synthetic noisy MOS-LQS values. The MOS-LQO values obtained from the ground truth model, a model which is similar to but distinct from the ground truth model, the null model and the encompassing model are used as competing models, which enter model estimation and selection using parametric empirical Bayes (see Sect. 2.4), log-Bayes factors, and posterior model probabilities (see Sects. 2.4.3 and 2.4.4). It will be shown how confidently the methods used in this work identify the correct model in dependence on the SNR, and how the results depend on the number of trials N, as these are the marginal conditions for model selection (Penny 2012).

The rest of this section is structured as follows: First, an overview on the computation of the signal-to-noise ratios is given, followed by a description of the data-generating framework and test conditions. Next, the four competing models entering

model selection are introduced. The selection is done for a single subject and a group of 16 subjects to show the effect of both intra- and inter-subject variability. A summary of the evaluation results concludes this chapter.

2.5.1 Signal-to-Noise Ratio Simulation

The signal-to-noise ratio simulation for this tutorial is based on the same signal model (1.6) as used in Sect. 1.2 but with a time-variant clean signal $s(n)$. Modeling $s(n)$ to be time-variant over trials is realistic for ERP sequences (Squires et al. 1976; Mars et al. 2008; Kolossa et al. 2013, 2015), which makes the results of these simulations well applicable to SNRs obtained from real ERP sequences with the methods proposed in Sect. 1.2. To recapitulate, the measured signal $y(n)$ is modeled to be composed of a clean signal $s(n)$ and the additive noise $\epsilon(n)$:

$$y(n) = s(n) + \epsilon(n), \tag{2.60}$$

with $n \in \{1, ..., N\}$. Note that in contrast to (1.6) no stimulus-specific index is necessary. This signal model is similar to the one-level linear model (2.11), but with a known clean signal $s(n)$ instead of the estimate $\widehat{s}(n)$ and the term "noise" instead of "error". As in (1.7) the signal-to-noise power ratio (SNR) is defined as

$$\text{SNR} = \frac{P_s}{P_\epsilon}, \tag{2.61}$$

which is in decibel

$$\text{SNR}\,[\text{dB}] = 10 \log_{10} \left(\frac{P_s}{P_\epsilon} \right), \tag{2.62}$$

with the signal power

$$P_s = \frac{1}{N} \sum_{n=1}^{N} s^2(n) \tag{2.63}$$

and the noise power

$$P_\epsilon = \sigma_\epsilon^2. \tag{2.64}$$

Inserting (2.64) in (2.62) and solving for σ_ϵ^2 yields

$$\sigma_\epsilon^2 = \frac{P_s}{10^{\frac{\text{SNR}\,[\text{dB}]}{10}}}, \tag{2.65}$$

which can be used to calculate the necessary variance σ_ϵ^2 of the error for a desired SNR [dB], given a signal $s(n)$.

2.5.2 Synthetic Data and Experimental Conditions

This section describes the framework for the generation of the synthetic data and test conditions. A note on the version of PESQ which is employed in this section: When calculating MOS-LQO values PESQ integrates speech signal disturbances over frequency and time into two factors which capture symmetric and asymmetric disturbances (Rix et al. 2001). In order to give here a example that is easily transferable to ERP data, only the symmetric disturbance is included in the PESQ model. Furthermore, the disturbances are not derived by using clean and degraded speech signals but are simply sampled, which has no influence on the methods and results but enables a clear and an easy-to-follow presentation. In order to transfer this example to model estimation and selection for event-related potentials, the ERP amplitudes take the place of the MOS-LQO values, while the observer models and link functions yield the predictors.

The employed simplified version of PESQ follows (Rix et al. 2001)

$$\mathrm{PESQ}(n) = 4.5 - d_{\mathrm{sym}}(n) \cdot 0.1, \tag{2.66}$$

with $d_{\mathrm{sym}}(n) \in \{0, ..., 35\}$ as the symmetric disturbance of speech sample n. The specific formula (2.66) can be generalized to a linear model with two parameters

$$s(n) = \theta_1 - d_{\mathrm{sym}}(n)\theta_2. \tag{2.67}$$

These parameters are used to introduce inter-subject variability, with $\theta_1 \sim \mathcal{N}(4.5, \sigma^2_{\theta_1})$ and $\theta_2 \sim \mathcal{N}(0.1, \sigma^2_{\theta_2})$ as random Gaussian variables with means 4.5 and 0.1, and variances $\sigma^2_{\theta_1}$ and $\sigma^2_{\theta_2}$, respectively. While this is no noise in the classical sense, the same definitions as in Sect. 2.5.1 are used to make the magnitude of randomness vivid. Adding zero-mean Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2_\epsilon)$ to (2.67) models the intra-subject variability

$$y(n) = s(n) + \epsilon(n) = \theta_1 - d_{\mathrm{sym}}(n)\theta_2 + \epsilon(n). \tag{2.68}$$

The integer values for dynamic distortions $d_{\mathrm{sym}}(n)$ are sampled from a uniform distribution in the interval [0, 35], and (2.63) along with (2.65) are used to calculate the variance σ^2_ϵ necessary to create noise conditions of SNR [dB] $\in \{8, 6, 4, 2, 0, -2\}$. While these SNR values are not expected for intra-subject variability in speech quality perception, they are realistic for EEG data (see Sects. 3.6.1 and 4.7.1 for subject-specific $\widehat{\mathrm{SNR}}$ values obtained in the studies in this work). The number of trials varies with $N \in \{50, 100, 150, 200, 250, 300, 350, 400, 450, 500\}$, yielding 60 scenarios with different combinations of SNR and trial numbers. For the simulation with multiple subjects, the scenarios are identical for each subject. After subject-individual model estimation, fixed-effects analyses are applied in order to get the group-level results (see Sect. 2.4.4). The estimation for all scenarios is repeated a thousand times with new error and stimulus sampling. (Group) log-Bayes factors and posterior model probabilities are calculated for each repetition and subsequently

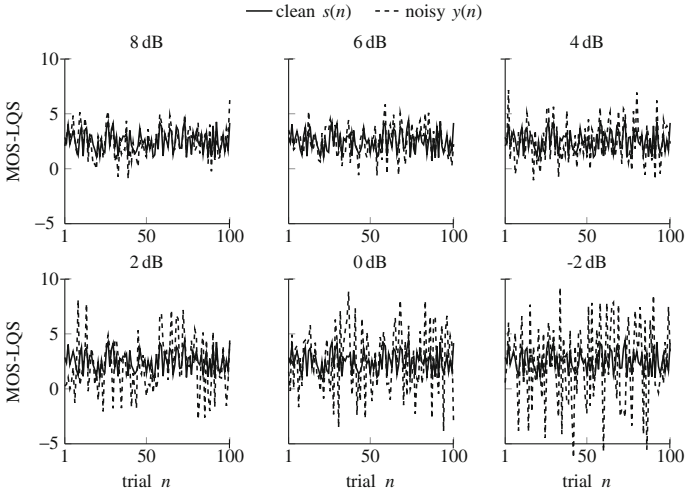

Fig. 2.1 Clean (——) and noisy (- - -) synthetic MOS-LQS values over trials $n \in \{1, ..., N = 100\}$ for decreasing SNR [dB] $\in \{8, 6, 4, 2, 0, -2\}$

averaged over repetitions (Penny 2012). Figure 2.1 shows clean (——) and noisy (- - -) synthetic MOS-LQS values with parameters $\theta_1 = 4.5$ and $\theta_2 = 0.1$ in the scenarios with $N = 100$ trials for a single repetition and all SNR conditions. The clean signal becomes gradually more disturbed and is unrecognizable for the lowest SNRs.

2.5.2.1 A Single Subject

For a single subject the parameters are set to the distribution means ($\theta_1 = 4.5$ and $\theta_2 = 0.1$), and the noisy signal is calculated according to (2.68). The results are presented later on in Sect. 2.6.1 as log-Bayes factors (log(BF), (2.53)) and posterior model probabilities (P($m|\mathbf{y}$), (2.55)), both based on the variational free energy F_m (2.50).

2.5.2.2 Multiple Subjects

For $L = 16$ subjects inter-subject variability is introduced by sampling the parameters $\theta_{\ell,1}$ and $\theta_{\ell,2}$, $\ell \in \mathscr{L} = \{1, 2, ..., L = 16\}$, from their respective distributions $\theta_1 \sim \mathcal{N}(4.5, \sigma^2_{\theta_1})$ and $\theta_2 \sim \mathcal{N}(0.1, \sigma^2_{\theta_2})$. Standard deviations are set to $\sigma_{\theta_1} = 0.45$ and $\sigma_{\theta_2} = 0.01$, i.e., the inter-subject parameter variability corresponds to an SNR of 20 dB (2.62). The variational free energy $F_{m,\ell}$ is calculated for each subject within one scenario and then summed over the subjects to obtain F_m (2.57), which is used to calculate group log-Bayes factors (log(GBF), (2.58)) and posterior model probabilities (P($m|\mathbf{y}$), (2.59)) for model selection in Sect. 2.6.2.

2.5.3 The Model Space

This section introduces the four models which make up the model space $\mathcal{M} = \{\text{TRU, SQD, NUL, ENP}\}$ and specifies the respective design matrices which are input to the PEB estimation framework (see Sect. 2.4). The ground truth (TRU) model is the data-generating model, while the squared distortion (SQD) model assumes a quadratic instead of a linear dependence on the symmetric distortion $d_{\text{sym}}(n)$. The null (NUL) model is the least complex hypothesis, claiming that all variability in the data is due to noise and will be used as reference model for the log-Bayes factors. The encompassing (ENP) model fits the data perfectly, but is the most complex conceivable model.

2.5.3.1 The Ground Truth Model (TRU)

The ground truth (TRU) model correctly assumes the estimated signal $\widehat{s}(n)$ to depend on an offset θ_1 minus the dynamic distortion $d_{\text{sym}}(n)$ multiplied with θ_2:

$$\widehat{s}(n) = \theta_1 - d_{\text{sym}}(n)\theta_2, \tag{2.69}$$

which is a two-parameter linear model (2.10). Consequently, the first-level design matrix is of the form

$$\mathbf{X}^{(1)} = \begin{bmatrix} 1 & -d_{\text{sym}}(n=1) \\ \vdots & \vdots \\ 1 & -d_{\text{sym}}(n=N) \end{bmatrix} \in \mathbb{R}^{N\times 2}. \tag{2.70}$$

2.5.3.2 The Squared Distortion Model (SQD)

The squared distortion (SQD) model states that $\widehat{s}(n)$ depends on an offset θ_1 minus the square of the dynamic distortion

$$\widehat{s}(n) = \theta_1 - d^2_{\text{sym}}(n)\theta_2, \tag{2.71}$$

yielding the first-level design matrix

$$\mathbf{X}^{(1)} = \begin{bmatrix} 1 & -d^2_{\text{sym}}(n=1) \\ \vdots & \vdots \\ 1 & -d^2_{\text{sym}}(n=N) \end{bmatrix} \in \mathbb{R}^{N\times 2}. \tag{2.72}$$

2.5.3.3 The Encompassing Model (ENP)

The encompassing (ENP) model assumes that all trials are independent of each other without any additive noise

$$y(n) = \widehat{s}(n) = \theta_n. \tag{2.73}$$

The corresponding first-level design matrix is an identity matrix

$$\mathbf{X}^{(1)} = \mathbf{I}_N = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{bmatrix} \in \mathbb{R}^{N \times N}, \tag{2.74}$$

which models each trial with an independent predictor. Note that the data can be modeled perfectly, but at the cost of highest complexity.

2.5.3.4 The Null Model (NUL)

The null (NUL) model represents the least complex hypothesis, claiming that all variation in the data is due to noise. It models $\widehat{s}(n)$ as constant over trials

$$\widehat{s}(n) = \theta_1. \tag{2.75}$$

The first-level design matrix is an all-one column vector

$$\mathbf{X}^{(1)} = \mathbf{x}^{(1)} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^N. \tag{2.76}$$

The NUL model is used as common reference model for the (group) log-Bayes factors, which is inspired by classical null hypothesis testing (see Sect. 2.2.1).

2.6 A Transfer Example Experiment—Results

This section shows the results of Bayesian model estimation and selection depending on the number of trials N and the signal-to-noise ratio. It starts with Fig. 2.2 which shows the clean (—) and fitted (- - -) MOS-LQS values for the TRU model for the scenarios depicted in Fig. 2.1. The means of the optimized parameter densities $\mu_{\theta_1|y} = \theta_1$ and $\mu_{\theta_2|y} = \theta_2$ are written in each panel and used as point estimates for the model fitting (see Sect. 2.4.2). The difference between the clean

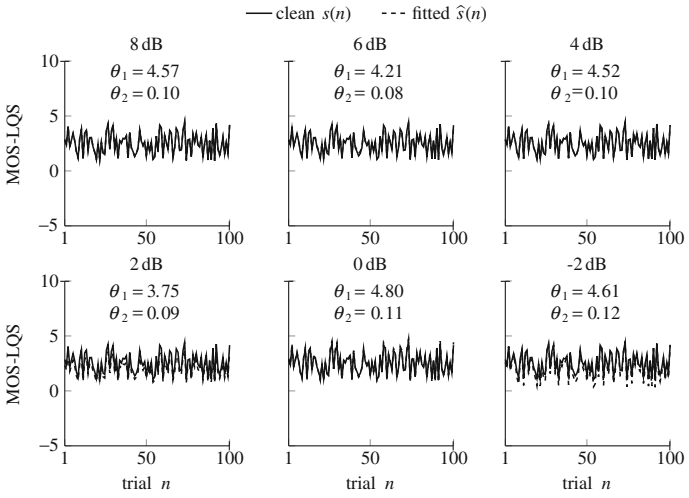

Fig. 2.2 Clean (——) and fitted (- - -) MOS-LQS values for trials $n \in \{1, ..., N = 100\}$ and SNR [dB] $\in \{8, 6, 4, 2, 0, -2\}$. The fitted MOS-LQS values are based on the TRU model and the point estimates of the parameters θ_1 and θ_2, which are shown in the plots above the respective curves

and fitted MOS-LQS values is graphically indeterminable for SNRs higher than 4 dB. The fit between clean and fitted MOS-LQS values is still good for lower SNRs, but errors become visible.

The rest of this section is split between Bayesian model selection based on a single subject and a group of $L = 16$ subjects. The respective selection results start with Figs. 2.3 and 2.5 depicting the (group) log-Bayes factors ((2.53) and (2.58)) followed by Figs. 2.4 and 2.6 showing posterior model probabilities ((2.55), (2.59)).

2.6.1 A Single Subject

Figure 2.3 shows the log-Bayes factors log(BF) of the TRU (—×—), SQD (- ○ -), and ENP (···□···) models versus the NUL model over an increasing number of trials N in noise conditions SNR [dB] $\in \{8, 6, 4, 2, 0, -2\}$. Throughout all conditions, an increasing number of trials N is accompanied by nearly linearly growing log-Bayes factors for the TRU and SQD model, while for the ENP model the log-Bayes factor is rapidly decreasing. The log-Bayes factors of the SQD model remain constantly below those the TRU model, meaning that the TRU model is favored over the SQD model for any number of trials and all SNR conditions. For an SNR of 8 dB, the TRU model is superior to all other models regardless of the number of trials. At 6 and 4 dB SNR and $N = 50$ trials, all log-Bayes factors are negative, e.g., the NUL model has the greatest variational free energy, while the TRU model is superior for $N = 100$ and more trials. The number of trials necessary for a positive log-Bayes factor for

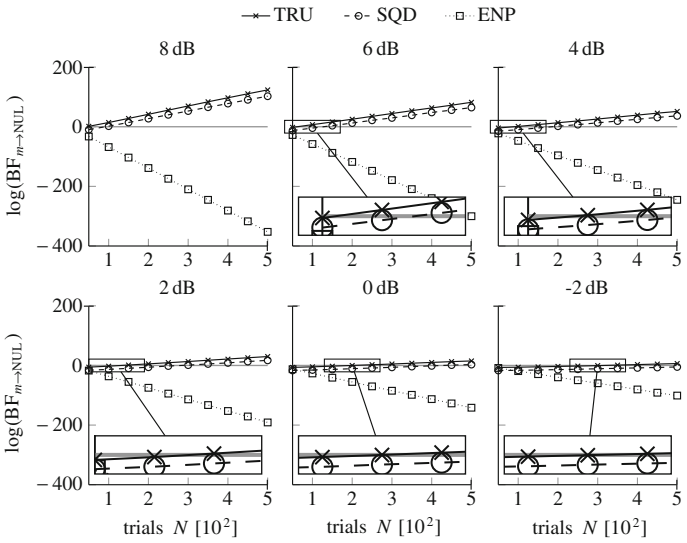

Fig. 2.3 Log-Bayes factors $\log(\mathrm{BF}_{m\to\mathrm{NUL}})$ (2.53) with $m \in \{\mathrm{TRU},\ \mathrm{SQD},\ \mathrm{ENP}\}$ in dependence on the number of trials N and $\mathrm{SNR}\,[\mathrm{dB}] \in \{8, 6, 4, 2, 0, -2\}$ for a single subject

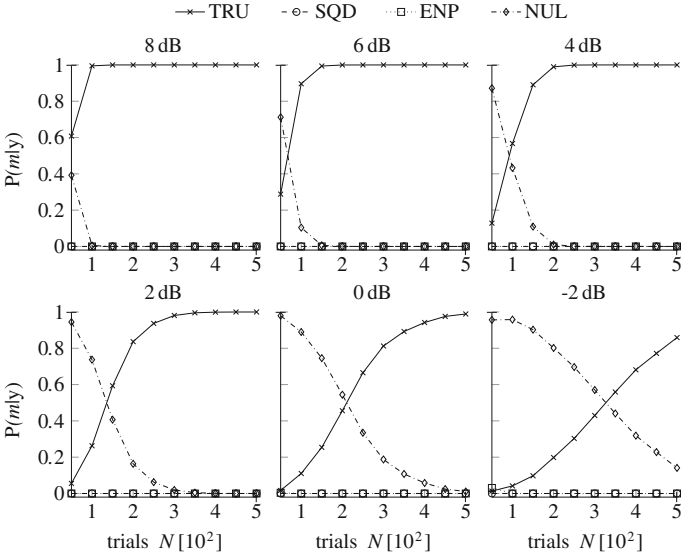

Fig. 2.4 Posterior model probabilities $\mathrm{P}(m|\mathbf{y})$ (2.54) with $m \in \{\mathrm{TRU},\ \mathrm{SQD},\ \mathrm{ENP},\ \mathrm{NUL}\}$ in dependence on the number of trials N and $\mathrm{SNR}\,[\mathrm{dB}] \in \{8, 6, 4, 2, 0, -2\}$ for a single subject

the TRU model versus the NUL model increases with decreasing SNR. For 2 dB, 0 dB, and −2 dB, the necessary number of trials is $N = 150$, $N = 250$, and $N = 350$, respectively. The area at which the NUL model ceases to be supported is magnified in the lower parts of the panels.

Posterior model probabilities $P(m|\mathbf{y})$ (2.55) offer a more intuitive way of model selection. Figure 2.4 shows posterior model probabilities for the TRU (–⋆–), SQD (-⊖-), ENP (···□···), and NUL (-◈-) models in dependence on the number of trials N in noise conditions SNR [dB] $\in \{8, 6, 4, 2, 0, -2\}$. Naturally, they mirror the results obtained from Fig. 2.3, as they are also based on the variational free energy, but the characteristic areas where the model with the highest posterior model probability changes are much clearer. In contrast to the steady increase in log-Bayes factors, posterior model probabilities change more rapidly around the critical trial number necessary for correctly identifying the TRU model as the best model. For SNRs of 0 and −2 dB, the statistical significance even at high trial numbers of $N = 250$ and $N = 350$ is still very low with posterior model probabilities of $P(\text{TRU}|\mathbf{y}) \approx 0.65$ and $P(\text{TRU}|\mathbf{y}) \approx 0.55$, respectively. Thus, inference based on a single participant in very noisy conditions is not feasible or requires a lot of trials.

2.6.2 *Multiple Subjects*

Figure 2.5 shows the group log-Bayes factors (log(GBF), i.e., log-Bayes factors summed over subjects (2.58)) of the TRU (–⋆–), SQD (-⊖-) and ENP (···□···) models versus the NUL model over an increasing number of trials N in noise conditions SNR [dB] $\in \{8, 6, 4, 2, 0, -2\}$. The results are qualitatively very similar to those obtained for a single subject in Fig. 2.3, i.e., linear evolution of group log-Bayes factors for an increasing number of trials N in all noise conditions, but the absolute values are on a larger scale. For conditions with an SNR of 6 dB or lower and only a few trials, the log(GBF) favors the NUL model. The number of trials necessary for correctly identifying the TRU model as the best model and rejecting the NUL are close to those for a single subject, i.e., $N = 100$, $N = 100$, $N = 150$, $N = 200$, and $N = 300$ for 6 dB, 4 dB, 2 dB, 0 dB, and −2 dB, respectively.

Figure 2.6 shows posterior model probabilities (2.59) for the TRU (–⋆–), SQD (-⊖-), ENP (···□···), and NUL (-◈-) models in dependence on the number of trials N for noise conditions SNR [dB] $\in \{8, 6, 4, 2, 0, -2\}$. As expected, the results mirror those depicted in Fig. 2.5, while meaningful differences are revealed in comparison to Fig. 2.4: Up to 2 dB SNR the critical trial numbers are identical to those of a single subject, but for a signal-to-noise ratio of 0 and −2 dB, $N = 200$ and $N = 300$ trials suffice for correctly identifying the TRU model as best model. In each case these are 50 trials less than for a single participant, for whom the posterior model probabilities of about 0.5 did not permit to draw meaningful conclusions at these points. Over all data points the statistical significance is greatly increased in comparison to Fig. 2.4 allowing for much more reliable conclusions (see Table 2.1 for details).

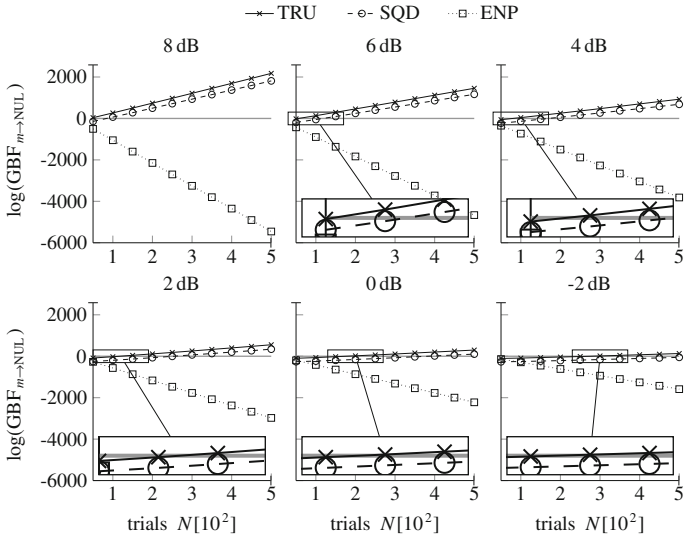

Fig. 2.5 Group log-Bayes factors $\log(\mathrm{GBF}_{m\to\mathrm{NUL}})$ (2.58) with $m \in \{\mathrm{TRU},\ \mathrm{SQD},\ \mathrm{ENP}\}$ in dependence on the number of trials N and for noise conditions SNR [dB] $\in \{8, 6, 4, 2, 0, -2\}$ for a group of $L = 16$ subjects

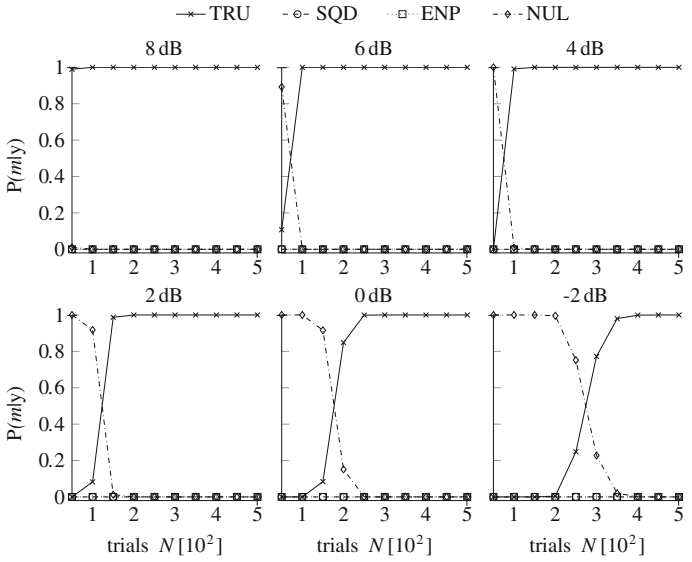

Fig. 2.6 Posterior model probabilities $\mathrm{P}(m|\mathbf{y})$ (2.59) with $m \in \{\mathrm{TRU},\ \mathrm{SQD},\ \mathrm{ENP},\ \mathrm{NUL}\}$ in dependence on the number of trials N and for noise conditions SNR [dB] $\in \{8, 6, 4, 2, 0, -2\}$ for a group of $L = 16$ subjects

2.7 Evaluation Summary

Log-Bayes factors and posterior model probabilities derived in the parametric empirical Bayes (PEB) framework are proven to be very useful for model estimation and selection. The use of this method is limited by the quality and quantity of the data, i.e., the signal-to-noise ratio (SNR) and the number of trials N. In view of the variability of SNRs across subjects, $N \geq 200$ trials per subject should be sufficient for reliably identifying the best model (see Sects. 3.6.1 and 4.7.1 for subject-specific $\widehat{\text{SNR}}$ values obtained in this work).

A major shortcoming of log-Bayes factors is that only two models are directly compared with each other simultaneously, which makes it complicated to get a clear view of the results for the complete model space. Additionally, (group) log-Bayes factors can become extremely large (see Fig. 2.5 with group log-Bayes factors in the magnitude of 10^3), which leads to a tendency to judge small (group) log-Bayes factors as not appropriately significant (see Table 2.1 with log-Bayes factors greater than five denoting very strong significance). Posterior model probabilities do not suffer from these shortcomings. All models composing the model space are simultaneously compared to each other, and the probabilities are normalized to the model space. The interpretation of statistical significance is intuitive and reminiscent of classical approaches, making Bayesian model selection also accessible to non-experts. Taking multiple subjects into account decreases the required number of trials per subject for correctly selecting the TRU model under low signal-to-noise ratios and increases the statistical power of the results. Group studies are therefore mandatory in order to report meaningful results.

http://www.springer.com/978-3-319-32284-1

Computational Modeling of Neural Activities for Statistical Inference

Kolossa, A.

2016, XXIV, 127 p. 42 illus., 20 illus. in color., Hardcover
ISBN: 978-3-319-32284-1