Journal of Mathematical Psychology 55 (2011) 331–347

Contents lists available at SciVerse ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Review

A tutorial on Bayes factor estimation with the product space method

Tom Lodewyckx [a,*], Woojae Kim [b], Michael D. Lee [c], Francis Tuerlinckx [a], Peter Kuppens [a], Eric-Jan Wagenmakers [d]

[a] University of Leuven, Belgium
[b] Ohio State University, United States
[c] University of California, Irvine, United States
[d] University of Amsterdam, United States

ARTICLE INFO

Article history:
Received 2 October 2009
Received in revised form
23 May 2011
Available online 6 August 2011

Keywords:
Bayes factor
Bayesian statistics
Graphical modeling
Hierarchical modeling
Hypothesis testing
Model selection
Product space method
Transdimensional MCMC

ABSTRACT

The Bayes factor is an intuitive and principled model selection tool from Bayesian statistics. The Bayes factor quantifies the relative likelihood of the observed data under two competing models, and as such, it measures the evidence that the data provides for one model versus the other. Unfortunately, computation of the Bayes factor often requires sampling-based procedures that are not trivial to implement. In this tutorial, we explain and illustrate the use of one such procedure, known as the product space method (Carlin & Chib, 1995). This is a transdimensional Markov chain Monte Carlo method requiring the construction of a "supermodel" encompassing the models under consideration. A model index measures the proportion of times that either model is visited to account for the observed data. This proportion can then be transformed to yield a Bayes factor. We discuss the theory behind the product space method and illustrate, by means of applied examples from psychological research, how the method can be implemented in practice.

Contents

1. Introduction 332
2. Understanding and estimating Bayes factors 332
   2.1. Understanding Bayes factors 332
   2.2. Estimating Bayes factors 333
3. Theoretical background of the product space method 334
   3.1. The product space method as a mixture model 334
   3.2. The Gibbs sampler 334
   3.3. Dimension matching and reversible jump MCMC 334
4. Practical implementation of the product space method 335
   4.1. WinBUGS implementation of the transdimensional model 335
      4.1.1. The model index 335
      4.1.2. The model likelihood 336
      4.1.3. The priors and pseudopriors 336
   4.2. Updating prior model probabilities with the bisection algorithm 336
   4.3. Monitoring the sampling behavior of the model index 336
   4.4. Comparison of multiple models 337
5. Applications in psychology 337
   5.1. Application 1: Comparing multiple models of emotion dynamics 337
      5.1.1. Emotion dynamics 337
      5.1.2. Experience sampling data 338

* Corresponding author.
E-mail address: tom.lodewyckx@psy.kuleuven.be (T. Lodewyckx).

0022-2496/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmp.2011.06.001

5.1.3. Modeling emotion dynamics ... 338
5.1.4. Model selection ... 338
5.2. Application 2: Testing for subliminality in the mass at chance model ... 339
5.2.1. The assumption of subliminality ... 339
5.2.2. The experimental setup ... 339
5.2.3. The mass at chance model ... 339
5.2.4. Model selection ... 340
5.3. Application 3: Testing visual discriminability in a hierarchical model ... 341
5.3.1. The effect of enhanced discriminability ... 341
5.3.2. Picture identification task ... 342
5.3.3. Model selection ... 342
6. Discussion ... 343
Appendix A. WinBUGS code for applications ... 344
A.1. Application 1 (emotion dynamics) ... 344
A.2. Application 2 (subliminality) ... 344
A.3. Application 3 (enhanced discriminability) ... 344
Appendix B. The bisection method to optimize the prior model probabilities ... 344
Appendix C. A Markov approach to monitor the sampling behavior of the model index ... 345
References ... 346

1. Introduction

A key to progress in psychology is the ability to evaluate theoretical ideas quantitatively against empirical observations. There are many formal and quantitative ways to compare and choose between models. Frequentist hypothesis testing relies on *p*-values, confidence intervals, and other devices developed within the sampling distribution statistical approach. This approach still remains the dominant one, despite well-known and well-documented problems (see Wagenmakers, 2007, for a recent overview). More recently, research in mathematical psychology and psychometrics has followed the lead of modern statistics and other empirical sciences in adopting Bayesian methods to evaluate models (e.g., Lee, 2008; Pitt, Myung, & Zhang, 2002; Shiffrin, Lee, Kim, & Wagenmakers, 2008). The Bayesian approach has the advantage of being a conceptually simple, theoretically coherent, and generally applicable way to make inferences about models from data (see Lee & Wagenmakers, 2005).

In this paper, we focus on a well-established and well-known Bayesian model selection tool, known as the *Bayes factor* (Jeffreys, 1961; Kass & Raftery, 1995). Intuitively, Bayes factors simply measure the relative level of evidence data provide for one model over another, in the form of a likelihood ratio. Bayes factors automatically account for model complexity, rewarding simple models and penalizing complicated ones. This property is important to avoid choosing models that overfit data (Myung & Pitt, 1997; Pitt et al., 2002).

The psychological literature has a number of recent applications of the Bayes factor, including in general statistical settings (e.g., Hoijtink, 2001; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wetzels, Raaijmakers, Jakab, & Wagenmakers, 2009), and to specific psychological models (e.g., Gallistel, 2009; Kemp & Tenenbaum, 2008; Lee, 2002, 2004; Pitt et al., 2002; Steyvers, Lee, & Wagenmakers, 2009), but it could hardly be described as a widely used approach. There are a few possible reasons for the lack of application of Bayes factors. Most obviously, there is a strong temptation to stay with known methods for analyzing data while they remain acceptable practice, whatever the limitations those methods impose.

More interestingly, even among those who accept the need to use the Bayesian approach it is understood that it can be difficult to calculate Bayes factors in practice. Sometimes, easily calculated but theoretically limited approximations to the Bayes factor, such as those based on the Bayesian Information Criterion (BIC), have been used (e.g., Vickers, Lee, Dry, & Hughes, 2003). In practice, Bayesian statistical methods have mainly been limited to the estimation of model parameters, especially when models are relatively complex (e.g., Kuss, Jäkel, & Wichmann, 2005; Lee, 2006, 2008; Rouder & Lu, 2005; Rouder, Lu, Morey, Sun, & Speckman, 2008; Rouder, Lu et al., 2007), leaving Bayesian model selection as a future challenge.

The aim of this paper is to demonstrate a method for estimating Bayes factors using the computational approach developed by Carlin and Chib (1995). The method is general, in the sense that it can be applied to compare any set of two or more models, including non-nested and hierarchical models. Non-nested models are not formed from incremental developments of the same theory, but originate from very different theories. Bayesian hierarchical models recently have been popular in various research domains because of their flexibility and conceptual consistency (Lee, 2011).

We first provide a formal account of the Bayes factor, and its estimation using the method developed by Carlin and Chib (1995). Then, we focus on relevant implementational issues and formulate guidelines for proper use of the method. Finally, we demonstrate in three applications how Bayes factors are estimated in psychological research, and conclude with a discussion about the strengths, weaknesses, and niche of application for the method.

2. Understanding and estimating Bayes factors

2.1. Understanding Bayes factors

The Bayes factor compares two models by considering on average how well each can fit the observed data, where the (prior weighted) average is taken with respect to all of the possible values of the parameters. It is this averaging that accounts for differences in model complexity, because more complicated models (i.e., those that can fit many data patterns by changing their parameter values) often have lower average levels of fit than simple models.

Formally, if Model A (M_a) with parameter vector θ_a is being compared to Model B (M_b) with parameter vector θ_b using data D, the Bayes factor is defined as

$$B_{ab} = \frac{p(D \mid M_a)}{p(D \mid M_b)} = \frac{\int p(D \mid \theta_a, M_a)\, p(\theta_a \mid M_a)\ \mathrm{d}\theta_a}{\int p(D \mid \theta_b, M_b)\, p(\theta_b \mid M_b)\ \mathrm{d}\theta_b}. \quad (1)$$

Eq. (1) shows that the Bayes factor is the ratio of two marginal likelihoods, $p(D \mid M_a)$ and $p(D \mid M_b)$, representing how likely the data are under each model, and that these likelihoods are found by averaging or marginalizing the likelihood across the parameter space of each model. For the marginal likelihood to be high, a model must not only be able to fit the observed data well, but also must not predict data different from those observed.

Table 1
Interpretation scheme for values of the Bayes factor, the logarithm of the Bayes factor, and the corresponding posterior model probability, according to Raftery (1995).

Interpretation	B_{ab}	$\log(B_{ab})$	$p(M_a \mid D)$
Very strong support for M_b	<0.0067	<−5	<0.01
Strong support for M_b	0.0067 to 0.05	−5 to −3	0.01 to 0.05
Positive support for M_b	0.05 to 0.33	−3 to −1	0.05 to 0.25
Weak support for M_b	0.33 to 1	−1 to 0	0.25 to 0.50
No support for either model	1	0	0.50
Weak support for M_a	1 to 3	0 to 1	0.50 to 0.75
Positive support for M_a	3 to 20	1 to 3	0.75 to 0.95
Strong support for M_a	20 to 150	3 to 5	0.95 to 0.99
Very strong support for M_a	>150	>5	>0.99

An alternative interpretation of the Bayes factor is evident from the following equation,

$$\frac{p\,(M_a \mid D)}{p\,(M_b \mid D)} = B_{ab} \times \frac{p\,(M_a)}{p\,(M_b)}, \tag{2}$$

which reads "Posterior model odds $=$ B_{ab} $\times$ Prior model odds". This gives a second interpretation of the Bayes factor as the change in the model odds resulting from observing the data. That is, whatever the prior odds in favor of Model A, the Bayes factor B_{ab} is the multiple that describes the increase or decrease in those odds following from the new evidence provided by the data D. Since the compared models may or may not have a nested structure, the Bayes factor represents "the standard Bayesian solution to the hypothesis testing and model selection problems" (Lewis & Raftery, 1997, p. 648).

Raftery (1995) proposed a useful interpretation scheme for values of the Bayes factor, as presented in Table 1 (a similar scheme was proposed by Jeffreys, 1961). This table includes a verbal expression of the strength of evidence, and corresponding ranges for the Bayes factor B_{ab} itself, for its logarithmic rescaled version $\log B_{ab}$, and for the posterior probability $p(M_a \mid D)$ (assuming equal prior probabilities for the models). Expressing Bayes factors on the logarithmic scale has the advantages of making zero the point of indifference between the two models being compared (i.e., the point at which the Bayes factor is 1, and the data provide no more evidence for one model than the other), and making equal increments correspond to equal changes in the relative probabilities (i.e., $\log B_{ab} = +2$ is the same level of evidence in favor of Model A as $\log B_{ab} = -2$ is in favor of model B). The posterior probability is a convenient and easily interpreted value in cases where the two models being compared are the only ones of theoretical interest.

2.2. Estimating Bayes factors

For all but the simplest model comparisons, the integrations required to calculate Bayes factors are analytically intractable. Accordingly, a large number of methods have been developed to approximate Bayes factors. The earliest methods focused on analytic approximations to the required integration (see Kass & Raftery, 1995, for a review). Many of these approaches continue to be refined (e.g., Myung, Balasubramanian, & Pitt, 2000), and remain useful and applicable methods for many simple statistical and psychological models.

More recently, Bayes factor estimation has been approached within a computational (i.e., sampling-based) framework for inference, mirroring the shift in inferences about parameters from analytic to computational methods. Within the computational framework, there are at least two quite different approaches for estimating Bayes factors. The first approach is based on estimating the marginal model likelihoods for both models separately, as per Eq. (1). This approach includes methods such as prior simulation (Kass & Raftery, 1995), importance sampling (DiCiccio, Kass, Raftery, & Wasserman, 1997; Geweke, 1989), candidate estimation (Chib, 1995), and the Laplace (Tierney & Kadane, 1986) and Laplace–Metropolis (Lewis & Raftery, 1997) methods.

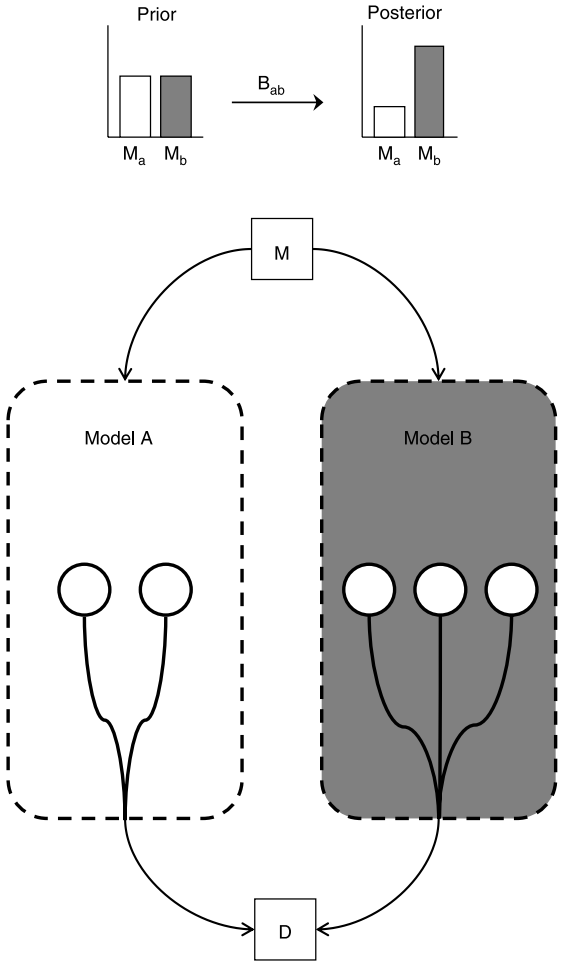

Fig. 1. Visualization of the framework of transdimensional MCMC for two models. The model index M is able to jump between Model A and Model B. Each model has a different constellation of model parameters, symbolized by the white nodes. Over MCMC iterations, the activated model and its corresponding model parameters are connected to the observed data D. The Bayes factor B_{ab} is quantified by the change from prior model odds to posterior model odds, as illustrated at the top part of the figure.

The second computational approach to Bayes factor estimation is rooted in transdimensional Markov chain Monte Carlo (MCMC) methods. It involves estimating posterior model odds for chosen prior model odds, as per Eq. (2). Reversible jump MCMC (Green, 1995) is one widely used transdimensional MCMC method. A less popular method is one developed by Carlin and Chib (1995), known as the product space method. Both methods are conceptually very simple, and rely on combining the models to be compared within one hierarchical "supermodel".

Fig. 1 presents the basic framework of this approach graphically for two models: Model A and Model B. The hierarchical combination of these models is achieved using a single binary model index variable M that controls which model generates the observed data D. The prior of the model index corresponds to the prior model odds. The posterior of the model index corresponds to the posterior model odds, and can be estimated by MCMC posterior sampling methods. Combining these two odds (the first exact, the second estimated) according to Eq. (2) then gives an estimate of the Bayes factor. In the schematic demonstration in Fig. 1, for example, both models are equally likely in the prior, but Model B is about three

times more likely in the posterior. This change from prior to posterior odds corresponds to a Bayes factor B_{ab} of about 1/3.

3. Theoretical background of the product space method

After this intuitive sketch of transdimensional MCMC methodology, comprising both the product space approach (Carlin & Chib, 1995) and reversible jump MCMC (Green, 1995), we now focus on the theoretical background of the product space method (later, we make a comparison to reversible jump MCMC). A clear understanding of the method is crucial to deal with its practical aspects, which are discussed in the next section.

3.1. The product space method as a mixture model

Suppose that Model A and B are two Bayesian models under comparison. For instance, Model A is defined by a joint probability distribution of data and model parameters:

$$p(D, \theta_a \mid M_a) = p(D \mid \theta_a, M_a)p(\theta_a \mid M_a).$$

To use the product space method, we set up a mixture model in which the parameter vectors of the two models are combined in one mixture parameter vector $\theta = (\theta_a, \theta_b)$, which takes any value from the Cartesian product of the two models' parameter spaces, $\theta \in \Theta_a \times \Theta_b$. The Model A part of the mixture model is defined by the joint distribution,

$$p(D, \theta \mid M_a) = p(D \mid \theta, M_a)p(\theta \mid M_a), \tag{3}$$
$$= p(D \mid \theta_a, M_a)p(\theta_a \mid M_a)p(\theta_b \mid M_a), \tag{4}$$

provided that $p(\theta_b \mid M_a)$ is a proper distribution that integrates to 1. Writing Eq. (3) as Eq. (4) is allowed since θ_b is not relevant under M_a and independent of θ_a, and the Model B part is specified similarly. The full mixture model is now written as

$$p(D, \theta) = p(D, \theta \mid M_a)p(M_a) + p(D, \theta \mid M_b)p(M_b). \tag{5}$$

The marginal likelihood for Model A under the mixture model can now be written as follows:

$$\begin{aligned} p(D \mid M_a) &= \int p(D \mid \theta, M_a)p(\theta \mid M_a)\mathrm{d}\theta \\ &= \iint p(D \mid \theta_a, M_a)p(\theta_a \mid M_a) \\ &\quad p(\theta_b \mid M_a)\, d\theta_a \mathrm{d}\theta_b \\ &= \int p(D \mid \theta_a, M_a)p(\theta_a \mid M_a) \\ &\quad \int p(\theta_b \mid M_a)\, d\theta_b\, \mathrm{d}\theta_a \\ &= \int p(D \mid \theta_a, M_a)p(\theta_a \mid M_a)\mathrm{d}\theta_a. \end{aligned} \tag{6}$$

This means that given M_a, the model defined in Eq. (4), even with added parameters θ_b, becomes essentially Model A with respect to its marginal likelihood, and the same holds for Model B. This ensures that the ratio of the two marginal likelihoods, $p(D \mid M_a)$ and $p(D \mid M_b)$, under this mixture model is the Bayes factor we seek to obtain.

The prior distribution $p(\theta_b \mid M_a)$, or likewise $p(\theta_a \mid M_b)$, is not given by any of the two models under comparison, but needs to be specified in order to define the mixture model with parameters in a product space. For this reason, these priors may be called pseudopriors or linking densities. Given that these pseudopriors are integrated out, they have no influence on the Bayes factor and can be arbitrarily chosen by the researcher (although we point out in the next section that the choice is important for the sampling efficiency of the procedure).

3.2. The Gibbs sampler

With a model set up as above, we need to devise a way to generate samples from the joint posterior distribution for model index and all model parameters. Particularly, we are interested in samples from the marginal posterior distribution of the model index M, which will be used to estimate the Bayes factor. Carlin and Chib (1995) suggest using the Gibbs sampler. First, a Gibbs step for sampling model parameters is based on the full conditional distribution:

$$p(\theta_a \mid \theta_b, M_k, D) \propto \begin{cases} p(D \mid \theta_a, M_a)p(\theta_a \mid M_a) & \text{if } k = a \\ p(\theta_a \mid M_b) & \text{if } k = b, \end{cases} \tag{7}$$

and $p(\theta_b \mid \theta_a, M_k, D)$ is specified similarly. This means that a sample of θ_k is generated from the posterior distribution of Model k only when the model index takes the value k; otherwise, it is generated from the corresponding pseudoprior. Next, to sample the model index, we derive another conditional distribution from Eq. (4) with prior model odds factored in:

$$p(M_k \mid \theta, D) \propto \begin{cases} p(D \mid \theta_a, M_a)p(\theta_a \mid M_a)p(\theta_b \mid M_a)p(M_a) & \text{for } M_a \\ p(D \mid \theta_b, M_b)p(\theta_b \mid M_b)p(\theta_a \mid M_b)p(M_b) & \text{for } M_b. \end{cases} \tag{8}$$

Generating values from this categorical distribution is straightforward, once the (normalized) full conditional probabilities for M_a and M_b have been derived. This sampling scheme, iterating between the model parameter vector θ_a and θ_b and the model index M, will produce samples from the correct joint posterior distribution under the regularity conditions for convergence (Roberts & Smith, 1994). The posterior probability of each model is estimated by the following Monte Carlo estimator:

$$\hat{P}(M_k \mid D) = \frac{\text{Number of occurrences of } M_k}{\text{Total number of iterations}}, \tag{9}$$

which will be translated to an estimated Bayes factor by factoring out the prior model odds, as per Eq. (2).

3.3. Dimension matching and reversible jump MCMC

Any transdimensional sampling scheme for computing the Bayes factor requires that the dimensionalities of all compared models' parameter spaces are matched in some way to form a single mixture model as defined above. One valid way to do so is to attach to each model the other model's parameters in a Cartesian product, as proposed by Carlin and Chib (1995) and described above. These additional parameters are regarded as pseudoparameters that are independent of data prediction.

This is not the only way, however. Sometimes, parameters have a strong conceptual similarity (i.e., they are interpreted in the same way) and statistical similarity (i.e., they have a similar marginal posterior distribution) across models. These sorts of parameters do not have to be taken as pseudoparameters for either model. In this case, combining those with the rest of unique parameters in a Cartesian product will form a parameter space that can apply to either model (Carlin & Chib, 1995). This can improve the efficiency of the sampling process because it decreases the dimensionality of the space that needs to be sampled. In this sense, the product space method does not always employ a purely product space when some parameters are shared between the compared models. For this reason, there is no precise conceptual boundary between the product space method and reversible jump MCMC.

Nevertheless, the product space method and the reversible jump MCMC method are generally regarded as two different MCMC approaches to the problem of jumping between model spaces of different dimensionalities. When proposed initially,

Table 2
Observed field goals (y) and attempts (n) by Kobe Bryant during the NBA seasons of 1999 to 2006.

Year	y	n	y/n
1999	554	1183	0.47
2000	701	1510	0.46
2001	749	1597	0.47
2002	868	1924	0.45
2003	516	1178	0.44
2004	573	1324	0.43
2005	978	2173	0.45
2006	399	845	0.47

the key difference between the two methods was that the reversible jump MCMC method provided a general, theoretical framework in which the number of parameters of the highest-dimensional model becomes the dimension of a transdimensional model, whereas the product space method focused more on a simple, intuitive way to construct a transdimensional model whose dimensionality is simply that of the product space of all compared models. Another difference was that the reversible jump MCMC method employed the more general Metropolis–Hastings sampling algorithm, whereas the product space method relied on Gibbs sampling.

These differences, however, turned out to be not fundamental, as shown by subsequent studies. Besag (1997), Dellaportas, Forster, and Ntzoufras (2002) and Godsill (2001) showed independently that the generality of the reversible jump MCMC method with regard to transdimensional model specification can also be entertained with the product space method. Dellaportas et al. (2002) and Godsill (2001) also demonstrated that the product space method can be combined with the Metropolis–Hastings algorithm. Conversely, the reversible jump MCMC method may be used with the Gibbs sampler, as implemented by Lunn, Best, and Whittaker (2009). This means that one approach can be viewed as a special case of the other. It might be better to view these methods as two slightly different representations of the same solution to the problem of Bayesian model uncertainty.

4. Practical implementation of the product space method

Having reviewed the theoretical background of the product space method, we now focus on its implementation. We do this by providing details of the specific formulation of the Bayesian transdimensional model in WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), and explaining several fine-tuning techniques for improving the estimated Bayes factors.

4.1. WinBUGS implementation of the transdimensional model

To illustrate the implementation of the transdimensional model, we build on the Kobe Bryant example presented by Ntzoufras (2009, Section 11.4.1.). In particular, we show how this analysis is programmed in WinBUGS, a user-friendly, accessible and widely used software package for Bayesian analysis (Lunn et al., 2000). In this example, two competing models are proposed for the field goals by Kobe Bryant in the NBA. The observations consist of the observed successes $y = \{y_{1999}, \ldots, y_{2006}\}$ and the number of attempts $n = \{n_{1999}, \ldots, n_{2006}\}$ for field goals by Kobe Bryant during eight consecutive basketball seasons from 1999 to 2006. These data are listed in Table 2.

Ntzoufras (2009) calculated the Bayes factor to compare two competing Binomial models, in order to learn about the consistency of the success probabilities $\pi = \{\pi_{1999}, \ldots, \pi_{2006}\}$ in the eight basketball seasons. The null model M_1 assumes one fixed probability π^{fixed} for all seasons, whereas the alternative model M_2 assumes unique and independent success probabilities π_i^{free} for each season:

$$M_1 : y_i \sim \text{Binomial}(\pi^{\text{fixed}}, n_i) \quad \text{for } i = 1999, \ldots, 2006$$
$$M_2 : y_i \sim \text{Binomial}(\pi_i^{\text{free}}, n_i) \quad \text{for } i = 1999, \ldots, 2006.$$

The parameters of M_1 (π^{fixed}) and M_2 ($\pi_{1999}^{\text{free}}, \ldots, \pi_{2006}^{\text{free}}$) are all assigned Beta(1, 1) priors, corresponding to a uniform prior over the range [0, 1]. The Bayes factor B_{12} quantifies the relative evidence in favor of M_1 when compared to M_2 and has a closed form solution, as the marginal model likelihoods $P(M_1 \mid D)$ and $P(M_2 \mid D)$ can be calculated straight from the data. The analytic result for the log Bayes factor $\log(B_{12})$ is found to be equal to 18.79, providing very strong support for the hypothesis that success probabilities are equal over all seasons. Our product space implementation estimated this log Bayes factor to be 18.80. The details of this implementation are given by the following WinBUGS script:

```
model{

 # 1) MODEL INDEX
 #    Model index is 1 or 2.
 #    Prior probabilities based on argument prior1.
 #    Posterior probabilities obtained by averaging
 #    over postr1 and postr2.
      M ~ dcat(p[])
      p[1] <- prior1
      p[2] <- 1-prior1
      postr1 <- 2-M
      postr2 <- 1-postr1

 # 2) MODEL LIKELIHOOD
 #    For each year, successes are Binomially distributed.
 #    In M1, the success rate is fixed over years.
 #    In M2, the success rate is year-specific.
      for (i in 1:n.years){
        successes[i] ~ dbin(pi[M,i], attempts[i])
        pi[1,i]  <- pi.fixed
        pi[2,i]  <- pi.free[i]
      }

 # 3) MODEL 1 (one single rate)
 #    The fixed success rate is given a Beta prior and
      pseudoprior.
 #    Whether it is a prior or pseudoprior depends on the
      Model index.
      pi.fixed ~ dbeta(alpha.fixed[M],beta.fixed[M])
      alpha.fixed[1] <- alpha1.prior
      beta.fixed[1] <- beta1.prior
      alpha.fixed[2] <- alpha1.pseudo
      beta.fixed[2] <- beta1.pseudo

 # 4) MODEL 2 (multiple independent rates)
 #    The year-specific success rate is given a Beta prior
      and pseudoprior.
 #    Whether it is a prior or pseudoprior depends on the
      Model index.
      for (i in 1:n.years){
        pi.free[i] ~ dbeta(alpha.free[M,i],beta.free[M,i])
        alpha.free[2,i] <- alpha2.prior
        beta.free[2,i] <- beta2.prior
        alpha.free[1,i] <- alpha2.pseudo[i]
        beta.free[1,i] <- beta2.pseudo[i]
      }
}
```

Four separate sections can be distinguished in the script: The model index, the model likelihood and the prior and pseudoprior specification of respectively M_1 and M_2. To clarify the interrelations between the various components of the transdimensional model, we will refer to Fig. 2.

4.1.1. The model index

The model index M has a categorical distribution over the domain $\{M_1, M_2\}$ with prior model probabilities determined by the argument `prior1`. The function of the model index within the transdimensional model is to connect model elements, as visualized at three locations in Fig. 2. The MCMC average of `postr1` is an estimate of $P(M_1 \mid D)$ and, after factoring out the prior model probabilities, this gives an estimate of the Bayes factor.

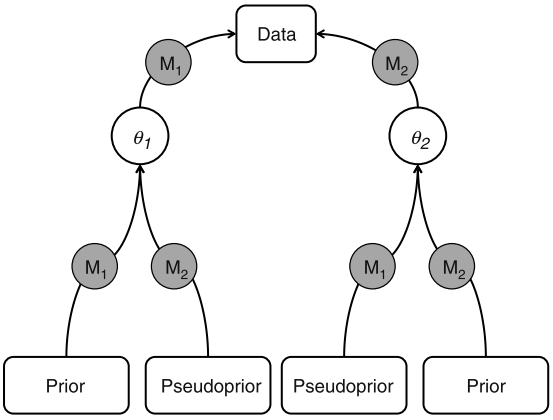

Fig. 2. Visualization of the general structure of priors and pseudopriors within a transdimensional model for comparing two models. The model index M activates one of two models, M_1 or M_2, at each MCMC iteration. Model activation determines how data, parameters and (pseudo)priors are connected to each other through a selection mechanism that occurs at two levels. First, the parameter vector of the activated model is given a prior distribution, while the parameter vector of the non-activated model is given a pseudoprior (bottom of the figure). Second, only the parameter vector of the activated model is connected to the observations (top of the figure). This way, only the "connected" parameters are assigned a prior distribution, while the "disconnected" parameters are assigned a pseudoprior distribution.

4.1.2. The model likelihood

In this part of the script, the common structure of both models is represented. In the Kobe Bryant example, the common model structure for each observation y_i is a Binomial distribution with success probability π_i : $y_i \sim \text{Binomial}(\pi_i, n_i)$. The further specification of π_i is defined in the parameter vectors of the models. The parameter vector θ_1 contains the overall success probability of M_1, whereas θ_2 contains the unique success probabilities that are assumed under M_2:

$$\theta_1 = \{\pi^{\text{fixed}}\}$$
$$\theta_2 = \{\pi^{\text{free}}_{1999}, \pi^{\text{free}}_{2000}, \pi^{\text{free}}_{2001}, \pi^{\text{free}}_{2002}, \pi^{\text{free}}_{2003}, \pi^{\text{free}}_{2004}, \pi^{\text{free}}_{2005}, \pi^{\text{free}}_{2006}\}.$$

The parameter space of the transdimensional model now consists of the model index and the parameter vectors: $\{M, \theta_1, \theta_2\}$. The behavior of the model index induces model activation: The value of M determines which parameter vector is connected to the likelihood, and thus which model is "active". This is illustrated in the upper part of Fig. 2.

4.1.3. The priors and pseudopriors

In the last two sections of the script, M is used to decide for each parameter vector whether it should be assigned a prior or pseudoprior distribution. For example, if M_1 is activated, the corresponding parameter vector θ_1 is connected to the model likelihood. This parameter vector is assigned a prior distribution such that the parameter vector can be updated based on prior and observed information. However, if M_1 is not activated, it cannot update the parameters properly as it is disconnected from the model likelihood. Therefore, it is assigned a pseudoprior distribution such that sampling continues. A similar reasoning can be formulated for the distribution of θ_2.

This intuition is illustrated in the bottom part of Fig. 2. The parameters of the pseudoprior distributions are estimated by running the models in separate runs and using the MCMC samples to estimate distributions.[1] The script for the transdimensional model can be used for this action by setting the prior model probability for the model that one wants to estimate equal to 1, since this is equivalent to estimating the model without the transdimensional framework. The goal of specifying the pseudopriors is to find good approximations of the true posterior distribution. This can be done, for example, by comparing the histogram of MCMC values to the proposed pseudopriors.

It is important that pseudopriors are chosen from a known family of probability distributions. WinBUGS automatically derives full conditional distributions, such as the one for the model index (see Eq. (8)) that clearly depend on the pseudoprior distribution. An alternative technique, which seems to be logical at first sight, would be to include additional, independent runs of each model's posterior simulation of parameters within the same WinBUGS script. One could then regard samples from these runs as if they were from pseudopriors, and supply them to the main, transdimensional routine simultaneously. However, this approach does not work because the main purpose of using pseudopriors is not to generate samples when the corresponding model is inactive, but they are used for the conditional probabilities of model indexes to be computed, as shown in Eq. (8). When provided with such a script, WinBUGS considers those pseudoprior samples as constant values, which eventually comes down to not using pseudopriors at all.

4.2. Updating prior model probabilities with the bisection algorithm

With the transdimensional model formulated in WinBUGS, one can obtain a posterior-simulated sample of the model index M, and thus estimate posterior model probabilities for given prior model probabilities. For some analyses, however, the available data may provide strong evidence in favor of one of the models. In practice, this will mean the less favored model is (almost) never activated. Increasing the number of iterations is one possible way of tackling this problem, but is not always feasible. For example, in the Kobe Bryant analysis, B_{12} is about equal to $\exp(18.79) \approx 144$ million. This implies that, under assumption of equal prior model probabilities, about 144 million Gibbs iterations are needed to have at least one M_2 activation.

An efficient solution to this problem is to choose prior model probabilities that make the number of posterior model activations for both models approximately equal. For example, if the data favor M_1 over M_2, we should increase $P(M_2)$ such that their posterior probabilities are more or less equal. This is conveniently done using an automatic search algorithm. We have successfully used the bisection algorithm, which was originally designed to find the root of a continuous function within a region between a positive and negative function value (Conte & De Boor, 1980). We use the algorithm to find a difference in posterior model probabilities which is close to zero. The bisection algorithm and its application to update prior probabilities is explained in Appendix B.

4.3. Monitoring the sampling behavior of the model index

It is not just the choice of the prior model probabilities that determines the quality of the Bayes factor estimates in the transdimensional model: Autocorrelation in the chains can still lead to inaccurate estimates after having obtained equal posterior model activation. Consider the three following situations. Fig. 3(a) shows the trace plot of the model index M under assumption of

[1] In the Kobe Bryant example, the prior distributions as well as the (estimated) pseudopriors for the success probabilities are Beta distributions. Choosing the same functional form for prior and pseudoprior simplifies the WinBUGS code, as it involves choosing only between different parameters of the same type of distributions, instead of choosing between different types of distributions. Assuming agreement in distributional type when specifying pseudopriors for different models may not always be desirable.

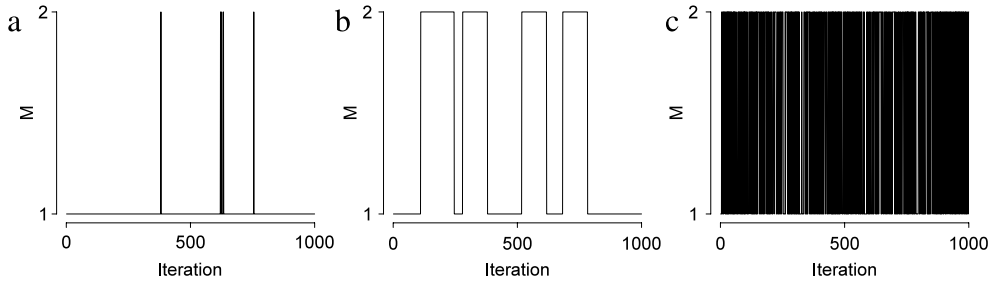

Fig. 3. Trace plots of the model index M representing three typical situations, with (a) asymmetric model activation, (b) equal model activation with few model switches, and (c) optimal sampling behavior with equal model activation and frequent model switches.

equal prior model probabilities. Clearly, M_1 is preferred strongly over M_2. The bisection algorithm is used to detect optimal prior model probabilities. Fig. 3(b) shows the trace plot of M under the assumption of optimal prior model probabilities after applying the bisection algorithm. It can be seen that posterior model activation is more or less equal, but there are only a few model switches. This situation also leads to a low quality of the Bayes factor estimates. The optimal situation is visualized in the trace plot in Fig. 3(c), where both models are activated equally often and model switches occur frequently.

Frequent model switching can be facilitated by considering two key aspects of the problem. One is concerned with the efficiency of posterior simulation of parameters within each model. Good mixing or low autocorrelation within each model is a prerequisite for a successful transdimensional simulation. Many useful techniques, suggested so far, for improving standard MCMC chains can be utilized for this purpose (e.g., Gelman, Carlin, Stern, & Rubin, 2004). The second approach deals directly with the transdimensional scheme. This may include changing the prior model probabilities, reparameterizing models for (more) parameters to be shared between models, and improving pseudoprior estimation. Once adequately efficient mixing within each model is confirmed, problems in a transdimensional scheme can be diagnosed by monitoring model switching behavior within a framework we call a Markov approach. More details can be found in Appendix C.

4.4. Comparison of multiple models

The presented WinBUGS implementation compares two statistical models with each other. The script can be easily extended to the comparison of multiple models by allowing more than one (integer) value for the model index M. For each model, a prior model probability and necessary pseudopriors are formulated.

The bisection method, as explained above and in Appendix B, is not generalizable to the situation of comparing more than two models. Manual calibration of the prior model probabilities might be very intensive or even impossible when one of the models is supported strongly. One might change the multiple-model comparison into several comparisons of two models, where the bisection method can still be applied for each of the comparisons separately. Even better would be to develop a general bisection method for more than two models, but this requires more sophisticated implementation.

As for the Markov approach explained above and in Appendix C, the two-dimensional visualization is not generalizable. However, we can still obtain the $M \times M$ transition matrix for the M models under comparison and use it to make decisions to improve model transitions.

5. Applications in psychology

In this section, we discuss three applications of the product space method, handling research questions in psychology. Each application focuses on a particular issue related to the product space method. In the first application, we generalize the method to comparison of more than two models. In the second application, we illustrate how the bisection method calibrates prior model probabilities. In the third application, we illustrate how the Markov approach is applied to monitor the sampling behavior of the model index.

The results are reported in terms of log Bayes factors (and posterior model probabilities). The Savage–Dickey density ratio is used as an alternative Bayes factor estimation method to validate our findings. The Savage–Dickey method is a straightforward Bayes factor estimation technique for null hypothesis testing on a particular parameter. The Bayes factor B_{01} that compares the null model M_0, with $\alpha = c$, to the full model M_1, with α given some prior distribution $p(\alpha)$ that includes c, can be estimated with the ratio of the prior density $P(\alpha = c \mid M_1)$ and posterior density $P(\alpha = c \mid M_1, D)$. More information on the Savage–Dickey density ratio can be found in Wagenmakers, Lodewyckx, Kuriyal, and Grasman (2010) and Wetzels et al. (2009).

All analyses have been performed in R 2.11.1 (R Development Core Team, 2010) and WinBUGS 1.4.3 (Lunn et al., 2000). Appendix A contains the WinBUGS scripts of the transdimensional models that are discussed in the applications. A file containing all R and WinBUGS scripts can be downloaded at http://ppw.kuleuven.be/okp/people/Tom_Lodewyckx/.

5.1. Application 1: Comparing multiple models of emotion dynamics

5.1.1. Emotion dynamics

People's feelings and emotions show continuous changes and fluctuations across time, reflecting the ups and downs of daily life. Studying the dynamics of emotions offers a unique window on how people emotionally respond to events and regulate their emotions, and provides crucial information about their psychological well being or maladjustment. Here we focus on two processes underlying emotion dynamics.

First, Suls, Green, and Hillis (1998) introduced *affective inertia* as a concept that describes how strong one's affective state carries over from one moment to the next. Kuppens, Allen, and Sheeber (2010) elaborated on this concept and found that emotional inertia, quantified as the first order autoregression effect of the emotional process, was higher for depressed individuals than for non-depressed individuals. This suggests that the fluctuations in people's emotions and moods is characterized by an autoregressive component. Second, apart of autocorrelation, emotion dynamics are also thought to be subjected to *circadian rhythms*. Various studies indicate the existence of circadian rhythms for emotions and their relevance in the explanation for psychological problems (e.g., Boivin, 2006; Kahneman, Krueger, Schwartz, & Stone, 2004; Peeters, Berkhof, Delespaul, Rottenberg, & Nicolson, 2006). The goal of this application is to study the relative role of these two processes in emotion dynamics using a time series of positive affect. To this end, we will estimate a model that involves an autocorrelation effect, a model that involves a circadian effect, and a model that involves both.

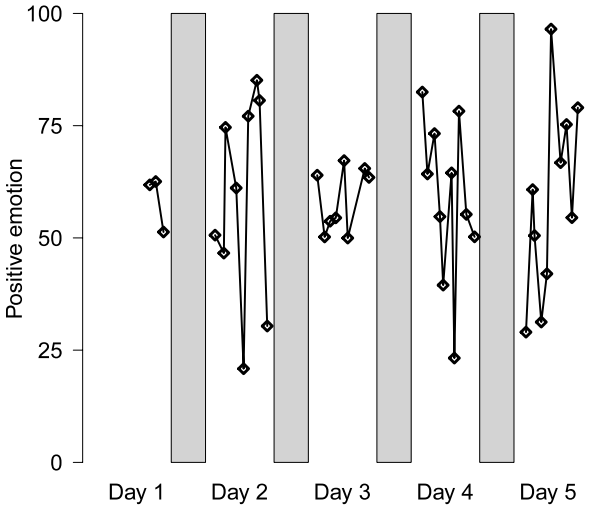

Fig. 4. Measurements of positive emotion during five consecutive days for one participant. The gray rectangles correspond to the nights (from 12 to 8 am).

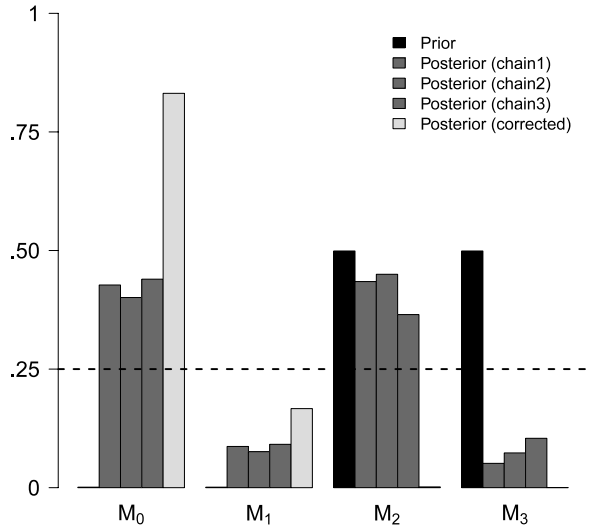

Fig. 5. Optimal prior probabilities, observed posterior probabilities and corrected posterior probabilities for the four emotion models, obtained with the product space method.

5.1.2. Experience sampling data

The observations were obtained in an experience sampling study (Kuppens et al., 2010), in which participants' emotions were assessed for ten times a day over a period of about two weeks during their daily life (for an introduction in experience sampling methods, see Bolger, Davis, & Rafaeli, 2003). On semi-random occasions within a day, the participant was alerted by a palmtop computer and asked to answer a number of questions about their current affective state.

We focus on a particular subset of observations, involving the time evolution of positive emotion for one of the participants during the first five days of the study, as visualized in Fig. 4. Positive emotion is an average of four diary items (relaxed, satisfied, happy, cheerful) and reflects the intensity of positive emotions on a 0 (no intensity) to 100 (high intensity) scale.[2] As can be seen in the figure, mere visual inspection of the data does not allow to guess whether an autoregressive or circadian process might be the underlying mechanism.

5.1.3. Modeling emotion dynamics

We formulate four candidate models for the observed time series described above, which we denote as y_t, with t being an index for discrete time (i.e., $t = 1, 2, \ldots$, ignoring the fact that the measurements were unequally spaced in time).

$$M_0 : y_t \sim \text{Normal}(\mu, \sigma^2)$$
$$M_1 : y_t \sim \text{Normal}(\mu + \phi_{I(r_t>1)}[y_{t-1} - \mu], \sigma^2)$$
$$M_2 : y_t \sim \text{Normal}(\mu + \alpha\text{time}_t + \beta\text{time}_t^2, \sigma^2)$$
$$M_3 : y_t \sim \text{Normal}(\mu + \phi_{I(r_t>1)}[y_{t-1} - \mu] + \alpha\text{time}_t + \beta\text{time}_t^2, \sigma^2).$$

The *null model* M_0 assumes that positive emotions fluctuate around some average level μ with error variance σ^2. In the *autoregressive model* M_1, the fixed effects part of the model is extended with an autoregression coefficient $\phi_{I(r_t>1)}$, modeling the relation between the current value y_t and the previous y_{t-1} (conditional on μ). The index function $I(\cdot)$ in the subscript of ϕ acts as a selection mechanism: The estimate for the autoregression coefficient ϕ only depends on observations that satisfy the specified condition within $I(\cdot)$, or $\phi = 0$ when the condition is not satisfied. Since r_t represents the within-day rank of the observation ($r = 1, 2, 3, \ldots$ for the first, second, third,…observations within a day), $\phi_{I(r_t>1)}$ is interpreted as the autoregression coefficient for all observations except for those observations preceded by a night. The *circadian model* M_2 assumes a parabolic day pattern, in line with findings from various studies that have found an inverted U-shaped day rhythm for positive emotion (e.g., Boivin, 2006; Peeters et al., 2006). This was modeled with a second degree polynomial, with α the linear coefficient and β the quadratic coefficient. In this model, time is represented with variable time_t, the time of the day expressed in hours, including minutes and seconds rescaled to the decimal hour scale. Finally, in the *combined model* M_3, the autoregressive and the circadian models are aggregated into a model containing all critical parameters ϕ, α and β. The prior distributions for the parameters are

$$\sigma \sim \text{Uniform}(0, 100)$$
$$\mu \sim \text{Normal}(0, 100^2)$$
$$\phi \sim \text{Normal}(0, 1^2)$$
$$\alpha, \beta \sim \text{Normal}(0, 10^2).$$

5.1.4. Model selection

The product space method was implemented to estimate posterior model probabilities and log Bayes factors for the four candidate models in the light of the observed emotion data.[3] Fig. 5 visualizes various aspects of the analysis for each of the models. The left bars in black represent the chosen *prior model probabilities*. The bisection method was not applicable since more than two models are being compared, and hence the prior model probabilities were updated manually (which took about ten iterations). The obtained prior for the model index is strongly asymmetric as almost all the prior mass is divided over M_2 and M_3. The three middle bars in dark gray show the *estimated posterior model probabilities* for the three Markov chains, using the optimal prior model probabilities. We find that posterior probabilities are estimated consistently, with small differences reflecting the

[2] To eliminate unwanted effects of day level of positive emotion, for each day, the day average was changed to the same overall five-day average by adding or subtracting a constant to all observations within that day.

[3] Three chains of 501 000 iterations were obtained. The final sample size was 10 000, after removing a burn in of 1000 iterations and thinning each chain with a factor 50. The log Bayes factor estimates were validated with the Savage–Dickey method. WinBUGS code for the transdimensional model can be found in Appendix A.1.

probabilistic and autodependent nature of the Gibbs sampler. Although equal posterior model activation is not obtained in the strict sense (indicated with the dashed line), activation is sufficient for all models to obtain stable estimates. To facilitate the interpretation of these prior and posterior probabilities, the right bars in light gray indicate the *corrected posterior model probabilities*: These are the posterior probabilities we would have obtained in case we had chosen a uniform prior for the model index.[4]

To explain the fluctuations of this participant's positive emotions during the observed five days, the null model seems to be the dominant model with $P(M_0 \mid y) = 0.8330$, whereas the autoregressive model seems to be a less supported option with $P(M_1 \mid y) = 0.1649$. The two models that contain the quadratic trend seem to be poor candidates for explaining the data with $P(M_2 \mid y) = 0.0017$ and $P(M_3 \mid y) = 0.0004$.

By calculating the corresponding log Bayes factors, we quantify the relative evidence between the models. For instance, there is positive support in favor of the null model when compared to the autoregressive model ($\log B_{10} = -1.62$), and very strong support in favor of the null model when comparing it to the circadian model and the combined model (respectively $\log B_{20} = -6.18$ and $\log B_{30} = -7.66$). Also, the autoregressive model is given strong and very strong support when comparing it to the models that contain the circadian pattern (respectively $\log B_{21} = -4.56$ and $\log B_{31} = -6.04$). When considering the circadian and the combined model, there is positive support in favor of the circadian model ($\log B_{32} = -1.47$).

This example shows clearly how strong inferences based on model selection may depend on the initial model choice. Imagine the situation where only M_2 and M_3 would have been considered. In that case, we would conclude that the circadian model is positively supported above the combined model ($\log B_{32} = -1.47$), leaving the impression that the circadian model is a good model. However, when considering all four models, the circadian model merely has a posterior probability of 0.0017.

Posterior inference for model parameters is possible with the MCMC output of the transdimensional output, but should be performed with caution. One should always consider the posterior distribution conditional on the value of the model index, also when a parameter is shared between models. In certain cases, however, unconditional posterior distributions for shared parameters may be of interest since one can incorporate model uncertainty into the inference and resulting interpretation of those parameters.

5.2. Application 2: Testing for subliminality in the mass at chance model

5.2.1. The assumption of subliminality

Priming studies have investigated the effect of consciously undetectable stimuli on human behavior. This is known as the subliminal priming effect (Lepore & Brown, 1997; Massar & Buunk, 2010; Mikulincer, Hirschberger, Nachmias, & Gillath, 2001). Although most studies concern visual priming, researchers have also experimented in the auditory domain (Kouider & Dupoux, 2005), and even explored the neurological basis of subliminal priming (Dehaene et al., 2001, 1998). However, these studies have one common fundamental assumption, which is that it is impossible to process the presented stimuli on a conscious level. To test the validity of this assumption experimentally, participants are presented a stimulus repeatedly and asked to indicate whether or not they perceived it. Rouder, Morey, Speckman, and Pratte (2007) have criticized the analysis of these performances and illustrate various problematic situations. Some procedures formulate an arbitrary cut-off value for the detection performance, whereas other analyses lack power or ignore individual differences by aggregating the observations over individuals. The implications are crucial: If stimuli are assumed to be undetectable while they are actually weakly detectable, inferences about subliminal priming effects are not valid.

Table 3
Observations and model selection results for the prime identification task, with the number of successes K_i, the number of attempts N_i, the proportion of successes K_i/N_i, the estimated log Bayes factors with the product space method $\log \hat{B}_i^{ps}$, and the Savage–Dickey method $\log \hat{B}_i^{sd}$ for individuals $i = 1, \ldots, 27$. Negative values for the log Bayes factors indicate support for the subliminal hypothesis, positive values indicate support for the supraliminal hypothesis.

i	K_i	N_i	K_i/N_i	$\log \hat{B}_i^{ps}$	$\log \hat{B}_i^{sd}$
1	150	284	0.53	−1.60	−1.66
2	142	288	0.49	−2.82	−2.79
3	154	287	0.54	−1.28	−1.27
4	155	288	0.54	−1.15	−1.16
5	136	288	0.47	−3.21	−3.19
6	138	288	0.48	−3.12	−3.10
7	211	288	0.73	30.39	28.61
8	140	288	0.49	−2.93	−2.96
9	148	285	0.52	−2.03	−2.01
10	159	287	0.55	−0.31	−0.27
11	164	288	0.57	0.85	0.87
12	150	288	0.52	−1.89	−1.95
13	158	288	0.55	−0.64	−0.60
14	138	288	0.48	−3.12	−3.10
15	148	288	0.51	−2.18	−2.19
16	146	288	0.51	−2.41	−2.39
17	163	288	0.57	0.64	0.56
18	145	288	0.50	−2.51	−2.52
19	180	288	0.62	7.18	6.96
20	155	288	0.54	−1.15	−1.16
21	148	287	0.52	−2.14	−2.12
22	147	287	0.51	−2.24	−2.24
23	134	288	0.47	−3.33	−3.33
24	134	286	0.47	−3.26	−3.26
25	167	288	0.58	1.76	1.72
26	149	288	0.52	−2.05	−2.07
27	147	288	0.51	−2.25	−2.30

5.2.2. The experimental setup

We discuss observations that were collected in an experiment conducted by Rouder, Morey et al. (2007). Visual stimulus material consisted of the set of numbers {2, 3, 4, 6, 7, 8}. In each trial, one of these numbers was presented on the computer screen as a 22 ms prime stimulus, followed by a 66 ms mask "#####" and another number from the same set as a 200 ms target stimulus. The participant had to indicate whether the 22 ms prime stimulus in the current trial was higher or lower than 5. The dependent measure was the accuracy of the answer, such that the experiment resulted in K_i successes out of N_i trials. All 27 participants were presented 288 trials. Table 3 lists the observed individual successes K_i and attempts N_i, and the corresponding proportion of successes K_i/N_i.[5] Most individuals perform around chance level ($K_i/N_i \approx 0.50$), suggesting that subliminality is plausible.

5.2.3. The mass at chance model

The Mass At Chance (MAC) model, introduced by Rouder, Morey et al. (2007), offers a clever Bayesian approach for testing the validity of the subliminality assumption for observed

[4] In theory, the ratio of posterior to prior model odds (the Bayes factor) does not depend on prior model probabilities. Therefore, chosen prior and estimated posterior model probabilities are easily transformed into corrected posterior model probabilities.

[5] For some of the participants, the data were incomplete such that $N_i < 288$.

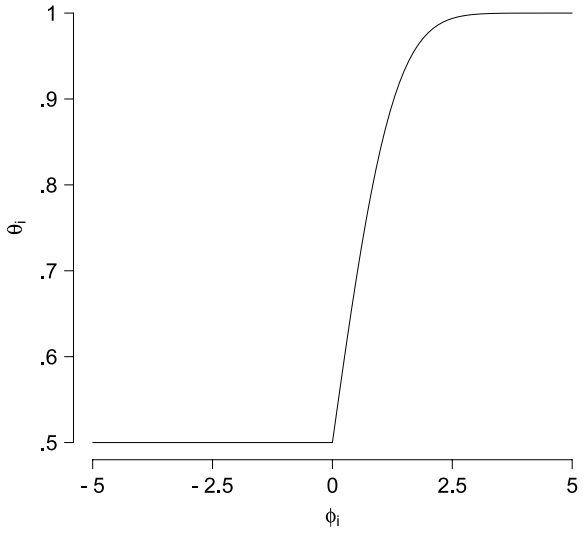

Fig. 6. The MAC transformation function of the mass at chance model.

success counts. The model assumes that a Binomial rate parameter θ_i underlies the generation of failures and successes, so that $K_i \sim (\theta_i, N_i)$. That Binomial rate is determined by an individual latent detection ability ϕ_i. The MAC transformation function, visualized in Fig. 6, quantifies the relation between θ_i and ϕ_i and makes an important difference between positive and negative ϕ_i values. A participant with a *negative ability* is unable to detect the prime stimulus consciously and his performance will be at chance level ($\theta_i = 0.5$).[6] On the other hand, a participant with a *positive ability* is able to detect the prime stimulus consciously ($0.5 < \theta_i \leq 1$), and, the more positive ϕ_i, the better the performance. The cumulative standard normal density function serves as a continuously increasing transformation function that maps $\Re^+ \mapsto [0.5, 1[$. We can now say that $\phi_i = \Phi^{-1}(\theta_i)$ is the probit transformation of the rate θ_i, with $\Phi^{-1}(\cdot)$ denoting the inverse cumulative standard normal density function.

Fig. 6 shows that only positive detection abilities ϕ_i can lead to performance above chance level. It also explains "mass at chance" since, after transformation, the mass over the negative domain of ϕ_i is squeezed together on the value $\theta_i = 0.5$. Whereas the distribution of ϕ_i is fully continuous, the distribution of θ_i is a mix of discrete (for $\theta_i = 0.5$) and continuous (for $0.5 < \theta_i \leq 1$) components. Therefore, an appropriate prior distribution for the latent ability ϕ_i is the standard normal distribution, $\phi_i \sim N(0, 1)$. The corresponding prior distribution on the rate scale is a (normalized) combination of a point mass probability $P(\theta_i = 0.50) = 0.50$ and a uniform distribution over the range of $0.50 < \theta_i \leq 1$ (see Rouder, Morey et al., 2007).

The MAC model is visualized in Fig. 7, using the notation provided by graphical modeling. Graphical models are a standard language for representing probabilistic models, widely used in statistics and machine learning (e.g., Gilks, Thomas, & Spiegelhalter, 1994; Jordan, 2004; Koller, Friedman, Getoor, & Taskar, 2007), and recently gained popularity in psychological modeling (e.g., Kemp, Shafto, Berke, & Tenenbaum, 2007; Lee, 2008; Shiffrin et al., 2008). The graphical model presented in Fig. 7 uses the same notation as Lee (2008). Nodes in the graph correspond to variables, and the graphical structure is used to indicate dependencies between the variables, with child nodes depending on parent nodes. Continuous variables are represented with circular nodes and discrete variables with square nodes. Observed variables (usually data) are shaded and unobserved variables (usually model parameters) are not shaded. Deterministic variables (variables that are simply functions of other nodes, and included for conceptual clarity) are shown as double-bordered nodes.

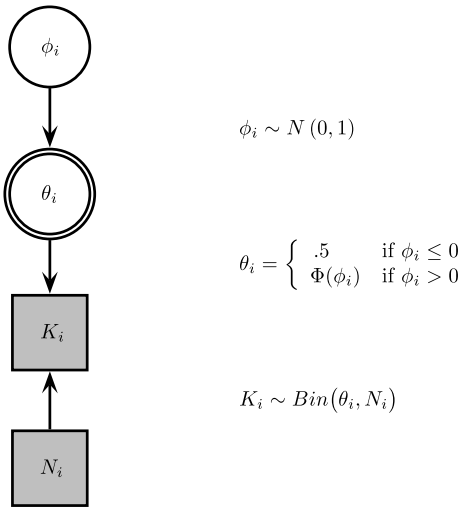

Fig. 7. Graphical model for the mass at chance model.

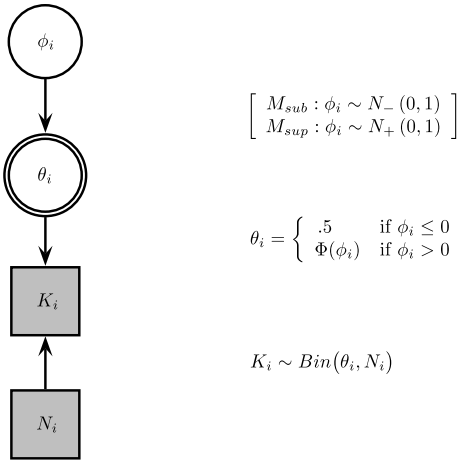

Fig. 8. Graphical model for the model comparison in the mass at chance model, representing the subliminal model, M_{sub}, and the supraliminal model, M_{sup}.

5.2.4. Model selection

Rouder, Morey et al. (2007) estimated posterior distributions for the latent abilities for each of the 27 subjects using the MAC model. It was concluded that perception was subliminal when 95% of the posterior mass for ϕ_i was located below zero. Using this criterion, they selected three out of the 27 subjects as subliminal perceivers, and found marginal evidence for another two subjects. For the remaining 22 subjects, they concluded that "Although many of these participants may be truly at chance, we do not have sufficient evidence from the data to conclude this".

Another way of testing for subliminality in the MAC model is by estimating a Bayes factor for each subject that compares the models of subliminal ($M_{\text{sub}} : \phi_i < 0$) and supraliminal ($M_{\text{sup}} : \phi_i > 0$) perception. Both competing models are formally described in Fig. 8. The notation is very similar to the one in Fig. 7, with the difference that, in this figure, two models are presented in one graphical model. This notation is practical for presenting models with the same basic structure of parameters, but differences in

[6] Performance below chance level is unrealistic, since it would mean that one knows the correct response, but gives the incorrect response on purpose.

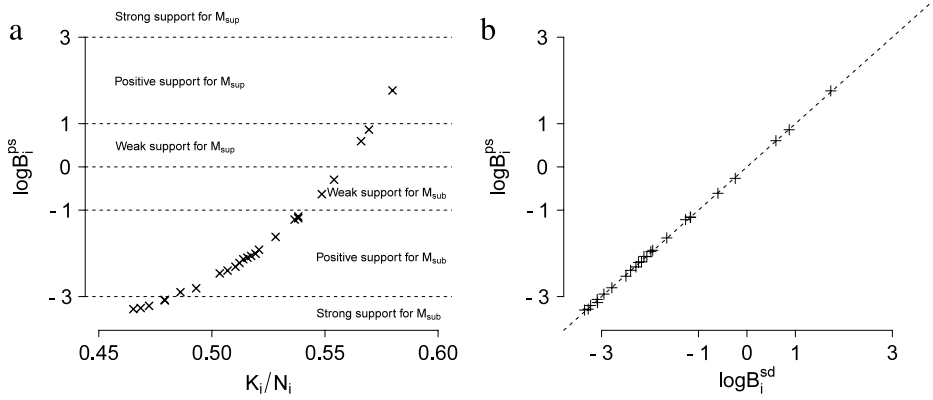

Fig. 9. Visualization of model selection results. (a) The log Bayes factor obtained with the product space method $\log \hat{B}_i^{\text{ps}}$ is compared to proportion of correct answers K_i/N_i. (b) The log Bayes factor obtained with the product space method $\log \hat{B}_i^{\text{ps}}$ is compared to the log Bayes factor obtained with the Savage–Dickey method $\log \hat{B}_i^{\text{sd}}$. Note that the figures do not include subject 7, since the corresponding log Bayes factor estimate is an outlier.

prior assumptions about parameters. The order restrictions are quantified by restricting the standard normal prior for ϕ_i to the negative value domain ($M_{\text{sub}} : \phi_i \sim N_-(0, 1)$) or the positive value domain ($M_{\text{sup}} : \phi_i \sim N_+(0, 1)$).

We estimated the log Bayes factors in favor of the supraliminal model using the product space method, denoted $\log \hat{B}_i^{\text{ps}}$.[7] Fig. 9(a) shows the estimated log Bayes factors, obtained with the product space method, as a function of the proportion of correct trials K_i/N_i. As expected, the evidence in favor of the supraliminal model increases with the proportion of correct responses. We might take $\log \hat{B}_i^{\text{ps}} < -3$, interpreted as "at least strong evidence in favor of M_{sub}", as a criterion to select subjects for subliminal priming tasks. This leads us to the selection of five subjects. As already suggested by Rouder, Morey et al. (2007), it might be plausible that other subjects are at the subliminal level as well, but that there is not enough evidence to make such an inference. Observing the curve that is revealed by the individual points in Fig. 9(a), we might formulate a cut-off value for the proportion correct, such as $K_i/N_i < 0.48$, or fit a function that models the relation between proportion correct and log Bayes factor (at least, under the assumption of a fixed sample size N_i).

In Fig. 9(b), the estimates obtained with the product space method are compared to those obtained with the Savage–Dickey density ratio. The estimates are as good as equal, which suggests that log Bayes factors are estimated correctly with both methods.

To illustrate how the bisection method operates, Fig. 10 shows the iterative history of prior model probabilities for each individual. An initial prior model probability is chosen at 0.5. If the corresponding difference in posterior probabilities $\delta = \pi_0^{\text{postr}} - \pi_1^{\text{postr}}$ is positive, M_0 is dominant so its prior model probability should be decreased (otherwise, if δ is negative, π_0^{prior} should be increased). This step is repeated until δ is within a reasonable region of tolerance $[-0.10, 0.10]$. Each of the lines represent the updating history for one of the individuals. It shows that even in extreme situations, the bisection algorithm works: For one of the individuals, 44 bisection iterations were necessary to find an optimal prior model probability, resulting in a log Bayes factor of about 30. Without the automatic prior calibration, it would be impossible to perform model selection for such extreme data.

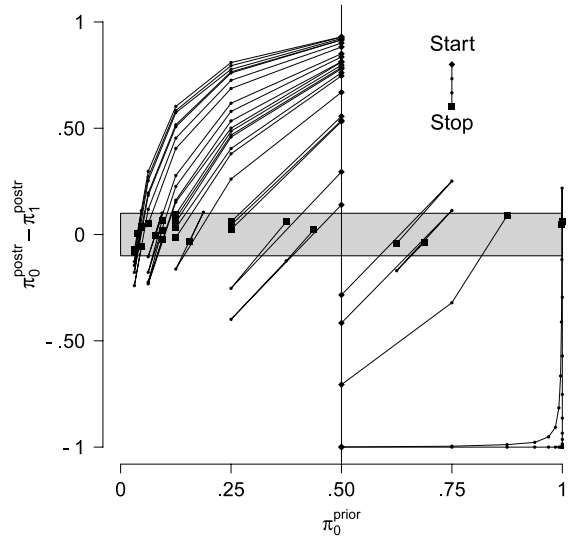

Fig. 10. Visualization of the prior calibration process with the product space method. Each connected line represents the subsequent values for the prior model probability $P(M_0)$ and the posterior difference $P(M_0 \mid y) - P(M_1 \mid y)$ for one of the 27 individuals, as obtained with the bisection method. The full vertical line connects all the starting points at $P(M_0) = 0.5$. The gray area represent the acceptance region $[-0.10, 0.10]$ for the difference in posterior model activation.

5.3. Application 3: Testing visual discriminability in a hierarchical model

5.3.1. The effect of enhanced discriminability

It is assumed that prior exposure to a stimulus – whether in the real world, or priming in an experimental context – leads to better processing of that stimulus in the future. This has been investigated in various implicit memory tasks, such as the picture identification paradigm (Reinitz & Alexander, 1996). In this paradigm, studying a target stimulus in a preceding phase increases the accuracy of identifying that stimulus when it is briefly presented as a prime stimulus in a forced-choice task against a foil with very similar characteristics. This effect is referred to as enhanced discriminability.

There exist (at least) two competing theories that can account for this facilitation effect. A first theory assumes that prior exposure to a stimulus increases its *encoding efficiency*, as is

[7] Three chains of 110 000 iterations were obtained. The final sample size was 100 000, after removing a burn in of 10 000 iterations (without thinning). The log Bayes factor estimates were validated with the Savage–Dickey method and denoted as $\log \hat{B}_i^{\text{sd}}$. The Savage–Dickey method could be used for this non-nested model selection problem by comparing both models to the same null model $\phi_i = 0$ and using the transitivity property. WinBUGS code for the transdimensional model can be found in Appendix A.2.

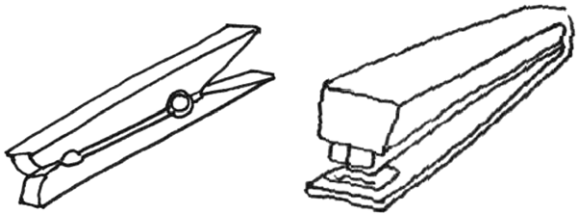

Fig. 11. Example of a stimulus pair of visually similar objects (Zeelenberg et al., 2002).

discussed in the perceptual representation system by Schacter (1992). A second line of research interprets the facilitation effect as a mere *bias* toward the exposed stimulus, as argued in Ratcliff and McKoon (1995, 1996). Interestingly, these frameworks make contradicting predictions when both target stimulus and foil stimulus are previously studied. From the encoding efficiency perspective, this situation would lead to enhanced discriminability as the encoding of the target stimulus has become more efficient. However, the bias perspective predicts no effect of enhanced discriminability, since exposure to the foil stimulus eliminates the bias effect toward the target stimulus. Zeelenberg, Wagenmakers, and Raaijmakers (2002) investigated this prediction (and others) in a series of three experiments, using both auditory and visual stimulus modalities. We focus on experiment three, using the picture identification task.

5.3.2. Picture identification task

Zeelenberg et al. (2002) conducted an experiment with 74 subjects, using 42 pairs of visually similar pictures, such as the clothes peg and stapler shown in Fig. 11. In the study block, subjects were familiarized with the pictures from 21 picture pairs, with each stimulus being presented three times for 2 s. This within-subjects manipulation assigned half of the picture pairs to the "Study Both" (*SB*) condition and the other half to the "Study Neither" (*SN*) condition. In each of the 42 trials in the test block, one of the picture pairs was used as stimulus material. One of the pictures was used as a target stimulus and briefly presented for 40 ms. Subjects were presented with both pictures from the picture pair and had to identify which one was used as a prime in a two-alternative forced-choice task. For each subject i, this resulted in counts of correct identifications K_i^{SB} and K_i^{SN}, with corresponding trial counts $N_i^{SB} = N_i^{SN} = 21$. Fig. 12 shows the relation between proportions for all 74 subjects. Enhanced discriminability is expressed as a higher proportion of correct identifications in the *SB* condition when compared to the *SN* condition.

5.3.3. Model selection

Zeelenberg et al. (2002) found a significant effect of the within subject manipulation, using a paired t-test. The proportion of correct trials was higher in the Study Both condition, 74.7%, than in the Study Neither condition, 71.5%, with $t(73) = 2.19$, $p < 0.05$. This result was taken to support increased encoding efficiency.

We present an alternative strategy using Bayesian hierarchical modeling, in which differences in proportions between the experimental conditions are treated as random effects and the hypothesis test is applied on the level of the hierarchical distribution. In Fig. 13, the graphical model is presented. For both experimental conditions, we assume that the counts of correct identifications are Binomially distributed with success rates θ_i^{SB} and θ_i^{SN}. As in the analysis of the data by Rouder, Morey et al. (2007), we work with probit transformations $\phi_i^{SB} = \Phi^{-1}(\theta_{SB})$ and $\phi_i^{SN} = \Phi^{-1}(\theta_{SN})$. The crucial part of the analysis then concerns the difference between the transformed success rates for the two

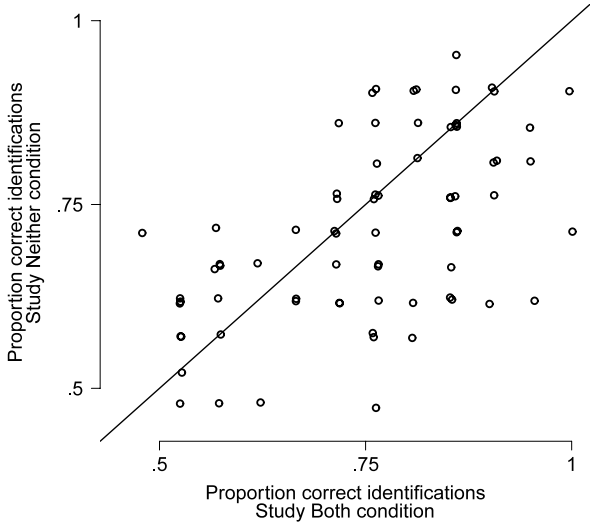

Fig. 12. Proportions of correct identifications of the 74 subjects in the Study Both and Study Neither conditions. Jitter has been added to distinguish participants with exactly the same proportions.

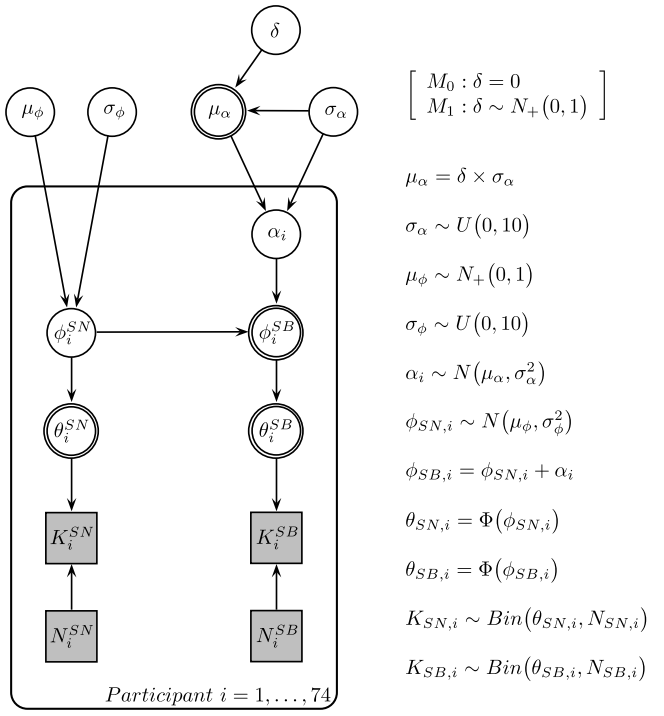

Fig. 13. Graphical model for the hierarchical model for the Zeelenberg et al. (2002) data.

conditions, formalized as the difference $\alpha_i = \phi_i^{SB} - \phi_i^{SN}$. Positive values of α_i indicate an effect of enhanced discriminability for individual i.

With the trial count in each condition for each subject being as small as 21 but the total number of subjects being as large as 74, this model is an ideal candidate for a hierarchical extension. By introducing a hierarchical structure, it becomes possible to take evidence from other subjects' responses and make more accurate inferences about effects at individual level. The plate in the graphical model in Fig. 13 is a common way for hierarchical models to visualize that the model part within the plate is repeated for all subjects. Hierarchical distributions are formulated for $\phi_i^{SN} \sim$

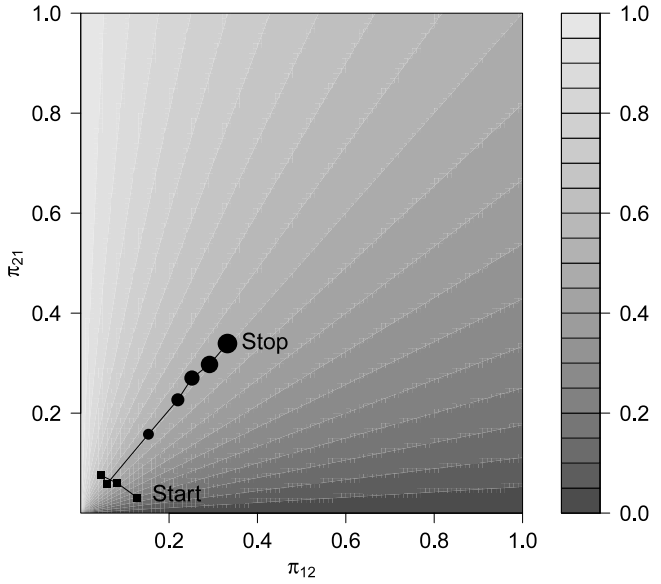

Fig. 14. Visualization of the Markov approach to monitor the sampling behavior of the model index.

$N\left(\mu_\phi, \sigma_\phi^2\right)$ and for $\alpha_i \sim N\left(\mu_\alpha, \sigma_\alpha^2\right)$. By running a posterior simulation with this model and inspecting the distributions of α_i, we would be able to draw inferences about the enhanced discriminability of each subject.

While the ability to make accurate inferences at the individual level is very useful, particularly in a situation where half of the subjects exhibit positive effects and the other half negative effects, a test of a single hypothesis at the group level can be informative as well. In this case, we are interested in $\delta = \mu_\alpha/\sigma_\alpha$, which is the standardized group effect of α_i. We compare the "no effect model" M_0, with $\delta = 0$, to the "effect model" M_1, with $\delta > 0$. The corresponding prior distribution for δ under M_1 is a standard normal prior that has been restricted to the positive real domain.

We applied the product space method to estimate the log Bayes factor $\log \hat{B}_{10}^{\text{ps}}$, comparing M_1 to M_0, and obtained an estimate equal to 1.45.[8] With this result, we find positive evidence in favor of an enhanced discriminability group effect. This is consistent with the conclusion from Zeelenberg et al. (2002), although the evidence is less strong than the p-value may suggest (Wetzels et al., 2011).

The Markov approach to monitor the sampling behavior of the model index was applied and visualized in Fig. 14. The transition probabilities between the models are crucial for the quality of the log Bayes factor estimate and are optimized in a two step approach. In a first step, the bisection method calibrates the prior model probabilities to approximate equal posterior model activation, using a minimal sample size of 10 000 iterations, after removing a burn in of 1000 samples. Four iterations are needed for this step, represented with squares in the visualization. Although posterior model activation is about equal, transition probabilities are rather low ($\pi_{12}, \pi_{21} \approx 0.06$). In the second step, we change MCMC settings to increase these transition probabilities, while using the calibrated set of prior model probabilities. We simultaneously increase the sample size (50 000, 100 000, 150 000, 200 000, 250 000) and the thinning factor (5, 10, 15, 20, 25) such that the thinned sample size is always 10 000. The circles with increasing diameter in the visualization represent these iterations with increasing thinning factor, and it clearly shows that model transitions are increased (the final transition probabilities are $\pi_{12}, \pi_{21} \approx 0.34$). We should remark that increasing the sample size and thinning is just one of the possibilities to increase model transitions (more details are provided in Appendix C).

6. Discussion

In Bayesian statistics, the Bayes factor is one of the most important and widely used methods for the quantitative evaluation of hypotheses and models. Bayes factors have an important role to play in the psychological sciences, which regularly seeks to test statistical hypotheses and substantive psychological models. We have explained, demonstrated and validated a general computational method by Carlin and Chib (1995) for estimating Bayes factors. This method can be applied to any statistical hypothesis test or model comparison, including comparison of multiple models, non-nested models and hierarchical models.

An attractive feature of the method is its conceptual simplicity. Like all transdimensional MCMC methods, the basic approach is to estimate the posterior distribution of a model index that controls which model generates predictions about the observations. This index directly corresponds to our intuitions about model selection: We start from a prior belief about the model probabilities and use the observations to update our belief into posterior model probabilities. The direction and strength of this update from prior to posterior model probabilities is quantified by the Bayes factor.

It is the case, however, that the product space method requires some sophistication with regard to various implementational issues. The WinBUGS implementation is based on a conceptual understanding of the method. In addition, the quality of the Bayes factor estimate depends on the choice of the prior model probabilities and the sampling behavior of the model index. In this paper, we tried to give some general guidelines and specific examples to help with these implementational issues.

Overall, we believe that the product space method occupies a useful niche between alternative approaches, based on a trade-off between ease of implementation and generality of application. Two alternative approaches for model selection that were discussed in this paper are the Savage–Dickey density ratio and reversible jump MCMC. The Savage–Dickey method is relatively easy to implement, but only applicable to a restricted class of nested comparisons.[9] In addition, testing a null hypothesis for multiple parameters simultaneously can bring about computational issues of multidimensional density estimation. On the other hand, reversible jump MCMC (Green, 1995) is actually more similar to the product space method than it appears from the model selection literature. However, seeking maximum sampling efficiency, its implementation usually requires complex analytic derivation of a mapping function and a Jacobian matrix. Achieving the same level of efficiency with the product space approach amounts to finding suitable reparameterizations of compared models and performing corresponding transformations of their posterior distributions, which is not a routine procedure. Very often in the psychological sciences, it suffices to compare only a few alternative formal models against available data, and the highest algorithm efficiency is not a critical factor. In these circumstances, we believe the product space method provides a relatively powerful and easily implemented approach for quantifying the evidence the data provide for and against the competing models in a general setting.

[8] Three chains of 251 000 iterations were obtained. The final sample size was 10 000, after removing a burn in of 1000 iterations and thinning each chain with a factor of 25. The log Bayes factor estimate was validated with the Savage–Dickey method and was equal to 1.43. WinBUGS code for the transdimensional model can be found in Appendix A.3.

[9] Non-nested models that can be connected with a common nested model, like those in the second application, are an exception.

Appendix A. WinBUGS code for applications

A.1. Application 1 (emotion dynamics)

```
model{

# MODEL INDEX
  M ~ dcat(p[])
  for(m in 1:4){p[m] <- prior[m]}

# MODEL LIKELIHOOD
  for(t in 2:n){y[t] ~ dnorm(mean[t],tau)
                mean[t] <- mu[M]
                         + st[t]*phi[M]*(y[t-1]-mu[M])
                         + a[M]*time[t]
                         + b[M]*pow(time[t],2)}
  tau <- pow(sd,-2)
  sd ~ dunif(0,100)

# MODEL 1: null model
  mu[1] <- mu1[M];   phi[1] <- 0; a[1] <- 0; b[1] <- 0
  mu1[1] <- mu1.pr
  mu1[2] <- mu1.ps
  mu1[3] <- mu1.ps
  mu1[4] <- mu1.ps

# MODEL 2: autoregressive model
  mu[2]  <- mu2[M]; phi[2]  <- phi2[M]; a[2] <- 0; b[2] <- 0
  mu2[1] <- mu2.ps; phi2[1] <- phi2.ps;
  mu2[2] <- mu2.pr; phi2[2] <- phi2.pr;
  mu2[3] <- mu2.ps; phi2[3] <- phi2.ps;
  mu2[4] <- mu2.ps; phi2[4] <- phi2.ps;

# MODEL 3: circadian model
  mu[3]  <- mu3[M]; phi[3] <- 0; a[3]  <- a3[M]; b[3]  <- b3[M]
  mu3[1] <- mu3.ps;              a3[1] <- a3.ps; b3[1] <- b3.ps
  mu3[2] <- mu3.ps;              a3[2] <- a3.ps; b3[2] <- b3.ps
  mu3[3] <- mu3.pr;              a3[3] <- a3.pr; b3[3] <- b3.pr
  mu3[4] <- mu3.ps;              a3[4] <- a3.ps; b3[4] <- b3.ps

# MODEL 4: combined model
  mu[4]  <- mu4[M]; phi[4]  <- phi4[M]; a[4]  <- a4[M]; b[4]  <- b4[M]
  mu4[1] <- mu4.ps; phi4[1] <- phi4.ps; a4[1] <- a4.ps; b4[1] <- b4.ps
  mu4[2] <- mu4.ps; phi4[2] <- phi4.ps; a4[2] <- a4.ps; b4[2] <- b4.ps
  mu4[3] <- mu4.ps; phi4[3] <- phi4.ps; a4[3] <- a4.ps; b4[3] <- b4.ps
  mu4[4] <- mu4.pr; phi4[4] <- phi4.pr; a4[4] <- a4.pr; b4[4] <- b4.pr

# PRIORS AND PSEUDOPRIORS
  mu1.pr ~ dnorm(0,.0001); mu1.ps ~ dnorm(mu1.psm,mu1.pst)
  mu2.pr ~ dnorm(0,.0001); mu2.ps ~ dnorm(mu2.psm,mu2.pst)
  mu3.pr ~ dnorm(0,.0001); mu3.ps ~ dnorm(mu3.psm,mu3.pst)
  mu4.pr ~ dnorm(0,.0001); mu4.ps ~ dnorm(mu4.psm,mu4.pst)
  phi2.pr ~ dnorm(0,1);  phi2.ps ~ dnorm(phi2.psm,phi2.pst)
  phi4.pr ~ dnorm(0,1);  phi4.ps ~ dnorm(phi4.psm,phi4.pst)
  a3.pr ~ dnorm(0,.01); a3.ps ~ dnorm(a3.psm,a3.pst)
  a4.pr ~ dnorm(0,.01); a4.ps ~ dnorm(a4.psm,a4.pst)
  b3.pr ~ dnorm(0,.01); b3.ps ~ dnorm(b3.psm,b3.pst)
  b4.pr ~ dnorm(0,.01); b4.ps ~ dnorm(b4.psm,b4.pst)
}
```

A.2. Application 2 (subliminality)

```
model{

# MODEL INDEX
  M ~ dcat(p[])
  p[1] <- prior1
  p[2] <- prior2
  postr1 <- 2-M
  postr2 <- M-1

# MODEL LIKELIHOOD
  K ~ dbin(theta,N)
  theta <- Q*phi(phi[M]) + (1-Q)*.5
  Q <- step(phi[M])

# MODEL 1: subliminal model
  phi[1] <- phi.sub[M]
  phi.sub[1] <- phisub.prior
  phi.sub[2] <- phisub.pseudo
  phisub.prior  ~ dnorm(0,1)I(,0)
  phisub.pseudo ~ dnorm(phisub.psm,phisub.pst)I(,0)

# MODEL 2: supraliminal model
  phi[2] <- phi.supra[M]
  phi.supra[2] <- phisupra.prior
  phi.supra[1] <- phisupra.pseudo
  phisupra.prior  ~ dnorm(0,1)I(0,)
  phisupra.pseudo ~ dnorm(phisupra.psm,phisupra.pst)I(0,)

}
```

A.3. Application 3 (enhanced discriminability)

```
model{

# MODEL INDEX
  M ~ dcat(p[])
  p[1] <- prior1
  p[2] <- prior2
  postr1 <- 2-M
  postr2 <- M-1

# MODEL LIKELIHOOD
  for(subj in 1:nsubj){
    K1[subj] ~ dbin(theta1[subj],N)
    K2[subj] ~ dbin(theta2[subj],N)
    theta1[subj] <- phi(phi1[subj])
    theta2[subj] <- phi(phi2[subj])
    phi1[subj] <- phi2[subj] + alpha[subj]
    phi2[subj] ~ dnorm(phi.mu,phi.tau)
    alpha[subj] ~ dnorm(alpha.mu,alpha.tau)
  }
  phi.mu ~ dnorm(0,1)I(0,)
  phi.tau <- pow(phi.std,-2)
  phi.std ~ dunif(0,10)
  alpha.mu <- delta[M] * alpha.std
  alpha.tau <- pow(alpha.std,-2)
  alpha.std ~ dunif(0,10)

# MODEL 1: null model
  delta[1] <- delta.null
  delta.null <- 0

# MODEL 2: full model
  delta[2] <- delta.full[M]
  delta.full[1] <- deltafull.pseudo
  delta.full[2] <- deltafull.prior
  deltafull.pseudo ~ dnorm(deltafull.psm,deltafull.pst)I(0,)
  deltafull.prior ~ dnorm(0,1)I(0,)

}
```

Appendix B. The bisection method to optimize the prior model probabilities

The choice of prior model probabilities in the transdimensional model is important for the quality of the Bayes factor estimate. Prior model probabilities π_1^{prior} and π_2^{prior} should be chosen such that approximate equal posterior model activation is obtained: That is, $\hat{\pi}_1^{\text{postr}} \approx \hat{\pi}_2^{\text{postr}}$. It is convenient to formalize this goal as wanting to specify prior model probabilities for which the difference in posterior model probabilities, $\hat{\delta} = \hat{\pi}_2^{\text{postr}} - \hat{\pi}_1^{\text{postr}}$, is approximately 0.

The bisection method (Conte & De Boor, 1980) is used to find the root (a function value of 0) of a continuous function within a specified interval of values for the function argument. Because of continuity, the function values of the interval bounds should have opposite signs to guarantee that the root is inside the interval. Translating the prior specification problem to the bisection method, the function we want to find the root for is f_{ps}. This function has π_1^{prior} as a function argument, it applies the product space method using the set of prior model probabilities $\{\pi_1^{\text{prior}}, \pi_2^{\text{prior}}\}$, and gives as output the difference in posterior model probabilities $\hat{\delta} = \hat{\pi}_2^{\text{postr}} - \hat{\pi}_1^{\text{postr}}$. The value of $\hat{\delta}$ has a range of -1 (when M_1 is exclusively activated) to 1 (M_2 is exclusively activated), with the root being the desired position of equal model activation.

Under normal circumstances, the bisection method is able to find prior model probabilities when applying the bisection method

to this function f_{ps}. By systematically scanning the function values over the region of possible values [0, 1] for the function argument π_1^{prior}, the algorithm finally stops when the function value is close enough to the root. One can distinguish three actions in the algorithm:

1. *Initialization*. Set the initial search interval for π_1^{prior} equal to $I = [I_{lower}, I_{upper}] = [0, 1]$. The corresponding set of function values for these lower and upper boundaries are $[-1, 1]$, reflecting full dominance of respectively M_1 and M_2.
2. *Bisection*: Estimate the function value for the midpoint of the interval $I_{mid} = (I_{lower} + I_{upper})/2$. Based on the sign of the function value, shrink the interval I to one of the bisections of the original I. If $f_{PS}(I_{mid})$ is negative, set $I_{lower} = I_{mid}$. If $f_{PS}(I_{mid})$ is positive, set $I_{upper} = I_{mid}$. This way, the function values of the borders of the new interval always have opposite signs (and thus contain the root).
3. *Evaluation*: The algorithm repeats the bisection step until $|f_{PS}(I_{mid})| < \epsilon$, with ϵ set to some arbitrary, small, positive precision value. The value of ϵ defines the preferred degree of equal model activation. For instance, setting ϵ equal to 0.10 makes the algorithm stop once estimated posterior model probabilities are within the region of [0.45, 0.55], with a maximum absolute difference of 0.10. Once that condition is obtained, the optimal prior model probability is approximated by I_{mid}.

We should be aware of the fact that f_{ps} is a stochastic function: Repeated runs of the function, while keeping the function argument π_1^{prior} constant, will return different results. This kind of variability can be reduced by changing MCMC settings, such as collecting more MCMC samples, or using a thinning factor. This is worth doing, in our experience, since variability can form a fundamental problem for the method. In particular, if the estimated difference in posterior model probabilities does not have the same sign as the true difference in posterior model probabilities, then the chosen bisection interval does not contain the root of f_{ps}. Monitoring the sampling behavior of the model index is crucial to obtain good estimates of the posterior model probabilities (see Appendix C).

The bisection method can deal relatively well with situations of strong asymmetry in evidence, when one of the models is preferred much more than the other. This is illustrated by the application of the bisection method in the Kobe Bryant analysis. Here, the extreme value of the best prior probability $\pi_1^{prior} = 0.000000007451$ is obtained only after 27 bisection iterations. A maximum can be specified for the number of bisections since, at some point, the computational precision boundaries of a computer are reached.

Appendix C. A Markov approach to monitor the sampling behavior of the model index

Well chosen prior model probabilities are necessary to obtain equal posterior model activation within the product space method. However, equal posterior model activation does not automatically imply good sampling behavior of the model index. As illustrated in Fig. 3(b), equal posterior model activation can be obtained with only a few model switches. For a categorical parameter, the lack of model switches in its Markov chain is comparable to a high level of autocorrelation for the Markov chain of a continuous parameter. To improve model switching behavior, various practical actions can be taken, such as reparameterization of the model, changing prior distributions, using a thinning factor, and so on. In this appendix, we discuss an approach using Markov transition matrices to monitor the sampling behavior of the model index.

The reason why we name it a Markov approach is not because the posterior samples of the model index are actually a Markov chain of a fixed order, but rather because we focus on the first order dependency in the series of model index samples to learn about their switching behavior. While it is true that the Gibbs sampler for the full transdimensional model generates a Markov chain of order 1, the model index alone, looked at marginally, does not. One sufficient condition for it to be a Markov chain of order 1 is that the within-model transition of parameters is performed by an independent sampler.[10] Of course, this cannot be true for MCMC simulations. However, it can be said that the Markov approach presented here will be a good approximation to model switching behavior when the MCMC sampling of parameters within each model exhibits good mixing with a reasonably low degree of autocorrelation throughout the chain.

For the Markov chain of the model index M, the 2×2 *transition matrix* π^{trans} is defined. This matrix contains the transition probabilities (π_{12}^{trans} and π_{21}^{trans}) on the off-diagonal elements and the non-transition probabilities ($\pi_{11}^{trans} = 1 - \pi_{12}^{trans}$ and $\pi_{22}^{trans} = 1 - \pi_{21}^{trans}$) on the diagonal elements:

$$\pi^{trans} = \begin{bmatrix} \pi_{11}^{trans} & \pi_{12}^{trans} \\ \pi_{21}^{trans} & \pi_{22}^{trans} \end{bmatrix}. \tag{C.1}$$

These probabilities describe the level of persistency of model activation, once a particular model has been activated. For example, $\pi_{11}^{trans} = 0.99$ and $\pi_{12}^{trans} = 0.01$ indicates that, once M_1 has been activated, there is a strong tendency that M_1 will stay activated over several MCMC iterations. The optimal situation would be that the probabilities of activating M_1 or M_2 at the next MCMC iteration are equal, and that these probabilities are independent of the currently activated model. This corresponds to a transition matrix with all values equal to 0.5.

The *stationary distribution* π^{stat} is a two-dimensional vector, reflecting the expected posterior model activation, and is derived from the transition matrix π^{trans}.[11] The elements π_1^{stat} and π_2^{stat} represent the probabilities of respectively M_1 and M_2 being activated.

$$\pi^{stat} = \begin{bmatrix} \pi_1^{stat} \\ \pi_2^{stat} \end{bmatrix}. \tag{C.2}$$

Fig. C.15 visualizes the relation between the transition matrix and the stationary distribution. The x and y axes represent the transition probabilities π_{12}^{trans} and π_{21}^{trans} over their full range from 0 to 1. Since $\pi_{11}^{trans} = 1 - \pi_{12}^{trans}$ and $\pi_{22}^{trans} = 1 - \pi_{21}^{trans}$, all possible values for the transition matrix are represented within this two-dimensional grid. Each point within this grid represents a unique transition matrix, for which the stationary distribution can be derived. The contour surface within this grid represents the value of π_1^{stat} as a function of the transition probabilities, representing the full stationary distribution (since $\pi_2^{stat} = 1 - \pi_1^{stat}$).

Although Fig. C.15 shows the link between all possible transition matrices and their corresponding stationary distributions, this does not mean that all of these situations are plausible within an MCMC context. We discuss the three trace plots for the model index as depicted in Fig. 3, as they each represent typical situations

[10] A proof of this proposition is available upon request. The intuition is as follows. Suppose that dependency present in a Markov chain for a transdimensional model can be divided into dependency due to the within-model transition of parameters and dependency due to the transition of the model index. Consider that the Markov model presented in the paper only describes the transition of the model index. It makes sense that the model becomes an accurate description when the within-model dependency is taken out of the equation, which is done by assuming an independent sampler within each model.

[11] The derivation is based on the equality $\pi^{stat'}(\mathbf{I} - \pi^{trans} + \mathbf{U}) = \mathbf{1}$, with $\mathbf{I}$ a 2×2 identity matrix, $\mathbf{U}$ a 2×2 matrix of ones and $\mathbf{1}$ a two-dimensional vector of ones.

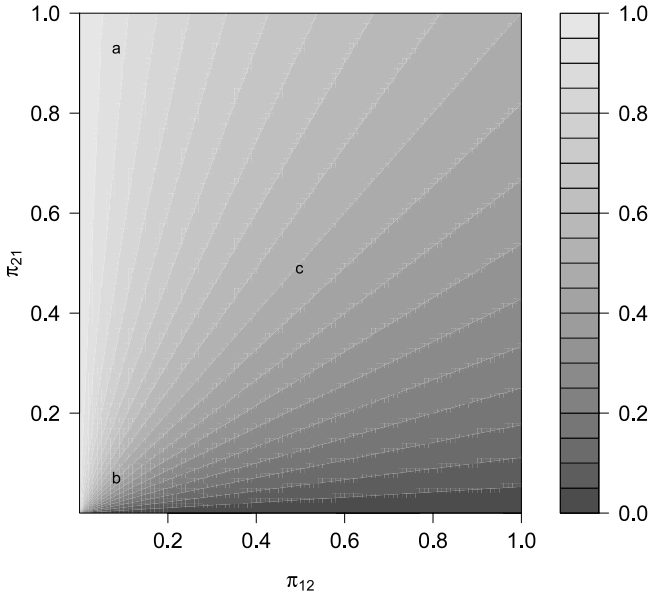

Fig. C.15. Contour plot of the stationary probability of Model 1, π_1^{stat}, as a function of the transition probabilities π_{12}^{trans} and π_{21}^{trans}. The three prototypical situations that have been illustrated in Fig. 3(a), (b) and (c) are located within this grid with the corresponding symbols a, b and c.

for the model index in transdimensional MCMC. The corresponding letters (a, b, c) in the subfigures of Fig. 3 are also located in the grid of Fig. C.15.

The situation of *strong preference for one of the models* is illustrated in Fig. 3(a). Typically, these cases are situated within the grid in the upper-left quadrant (dominance of M_1) and the lower-right quadrant (dominance of M_2). This problem can be solved by changing prior model probabilities. However, even when posterior model activation has been obtained when using an optimal prior distribution for the model index, there can still be a *lack of model switching*, as illustrated in Fig. 3(b). Fig. C.15 reveals that equal posterior model activation is obtained whenever transition probabilities are equal. However, transition probabilities close to zero lead to poor estimates of the posterior model probabilities, since there are almost no model switches. Various actions can be taken to increase the number of model switches, such as reparameterizing the model so that parameters may be shared between models, and improving the estimation of pseudopriors. In case some parameters are shared by the compared models, it is important to check whether their posterior distributions have enough overlap. The goal is to get as close to the *optimal situation* of equal posterior model activation as possible, as illustrated in Fig. 3(c). In Fig. C.15, that situation is located in the center of the grid. We also note that the upper-right quadrant is not a plausible value region within an MCMC context, since transition probabilities higher than 0.5 can be interpreted as negative autocorrelations for Markov chains for continuous parameters.

References

Besag, J. (1997). Comment on "Bayesian analysis of mixtures with an unknown number of components". *Journal of the Royal Statistical Society, Series B*, *59*, 774.

Boivin, D. B. (2006). Influence of sleep–wake and circadian rhythm disturbances in psychiatric disorders. *Journal of Psychiatry and Neuroscience*, *25*, 446–458.

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: capturing life as it is lived. *Annual Review of Psychology*, *54*, 579–616.

Carlin, B. P., & Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, *57*, 473–484.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, *90*, 1313–1321.

Conte, S. D., & De Boor, C. W. (1980). *Elementary numerical analysis: an algorithmic approach* (3rd ed.) McGraw-Hill.

Dehaene, S., Naccache, L., Cohen, L., Bihan, D. L., Mangin, J. F., Poline, J. B., & Rivière, D. (2001). Cerebral mechanisms of word masking and unconscious repetition priming. *Neuroscience*, *4*, 752–758.

Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., van de Moortele, P. F., & Le Bihan, D. (1998). Imaging unconscious semantic priming. *Nature*, *395*, 597–600.

Dellaportas, P., Forster, J. J., & Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing*, *12*, 27–36.

DiCiccio, T. J., Kass, R. E., Raftery, A. E., & Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association*, *92*, 903–915.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*, 439–453.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.) Boca Raton, FL: Chapman & Hall, CRC.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, *57*, 1317–1339.

Gilks, W. R., Thomas, A., & Spiegelhalter, D. J. (1994). A language and program for complex Bayesian modelling. *The Statistician*, *43*, 169–177.

Godsill, S. J. (2001). On the relationship between Markov chain Monte Carlo methods for model uncertainty. *Journal of Computational and Graphical Statistics*, *10*, 230–248.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*, 711–732.

Hoijtink, H. (2001). Confirmatory latent class analysis: model selection using Bayes factors and (pseudo) likelihood ratio statistics. *Multivariate Behavioral Research*, *36*, 563–588.

Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Oxford University Press.

Jordan, M. I. (2004). Graphical models. *Statistical Science*, *19*, 140–155.

Kahneman, D., Krueger, A. B., Schkade, D. A., Schwartz, N., & Stone, A. A. (2004). A survey method for characterizing daily life experience: the day reconstruction method. *Science*, *306*, 1776–1780.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 377–395.

Kemp, C., Shafto, P., Berke, A., & Tenenbaum, J. B. (2007). Combining causal and similarity-based reasoning. In *Advances in neural information processing systems*: *Vol. 19*. Cambridge, MA: MIT Press.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*, 10687–10692.

Koller, D., Friedman, N., Getoor, L., & Taskar, B. (2007). Graphical models in a nutshell. In L. Getoor, & B. Taskar (Eds.), *Introduction to statistical relational learning*. Cambridge, MA: MIT Press.

Kouider, S., & Dupoux, E. (2005). Subliminal speech priming. *Psychological Science*, *16*, 617–625.

Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. *Psychological Science*, *21*, 984–991.

Kuss, M., Jäkel, F., & Wichmann, F. A. (2005). Bayesian inference for psychometric functions. *Journal of Vision*, *5*, 478–492.

Lee, M. D. (2002). Generating additive clustering models with limited stochastic complexity. *Journal of Classification*, *19*, 69–85.

Lee, M. D. (2004). A Bayesian data analysis of retention functions. *Journal of Mathematical Psychology*, *48*, 310–321.

Lee, M. D. (2006). A hierarchical Bayesian model of human decision-making on an optimal stopping problem. *Cognitive Science*, *30*, 555–580.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*, 1–15.

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*, 1–7.

Lee, M. D., & Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: comment on Trafimow (2003). *Psychological Review*, *112*, 662–668.

Lepore, L., & Brown, R. (1997). Category and stereotype activation: is prejudice inevitable? *Journal of Personality and Social Psychology*, *72*, 275–287.

Lewis, S. M., & Raftery, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace–Metropolis estimator. *Journal of the American Statistical Association*, *92*, 648–655.

Lunn, D. J., Best, N., & Whittaker, J. C. (2009). Generic reversible jump MCMC using graphical models. *Statistics and Computing*, *19*, 395–408.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—a Bayesian modelling framework: concepts, structure and extensibility. *Statistics and Computing*, *10*, 325–337.

Massar, K., & Buunk, A. P. (2010). Judging a book by its cover: jealousy after subliminal priming with attractive and unattractive faces. *Personality and Individual Differences*, *49*, 634–638.

Mikulincer, M., Hirschberger, G., Nachmias, O., & Gillath, O. (2001). The affective component of the secure base schema: affective priming with representations of attachment security. *Journal of Personality and Social Psychology*, *81*, 305–321.

Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: differential geometry and model selection. *Proceedings of the National Academy of Sciences*, *97*, 11170–11175.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: a Bayesian approach. *Psychonomic Bulletin & Review*, *109*, 472–491.

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*. Hoboken, NJ: John Wiley & Sons Inc.

Peeters, F., Berkhof, J., Delespaul, P., Rottenberg, J., & Nicolson, N. A. (2006). Diurnal mood variation in major depressive disorder. *Emotion*, *6*, 383–391.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472–491.
Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 111–196). Cambridge: Blackwells.
Ratcliff, R., & McKoon, G. (1995). Bias in the priming of object decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 754–767.
Ratcliff, R., & McKoon, G. (1996). Bias effects in implicit memory tasks. *Journal of Experimental Psychology: General*, *125*, 403–421.
R Development Core Team. (2010). R: a language and environment for statistical computing. Organization R foundation for statistical computing. Vienna, Austria.
Reinitz, M. T., & Alexander, R. (1996). Mechanisms of facilitation in primed perceptual identification. *Memory & Cognition*, *24*, 129–135.
Roberts, G. O., & Smith, A. F. M. (1994). Simple conditions for the convergence of the Gibbs sampler and Metropolis–Hastings algorithms. *Stochastic Processes and their Applications*, *2*, 207–216.
Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*, 573–604.
Rouder, J. N., Lu, J., Morey, R. D., Sun, D., & Speckman, P. L. (2008). A hierarchical process dissociation model. *Journal of Experimental Psychology: General*, *137*, 370–389.
Rouder, J. N., Lu, J., Speckman, P. L., Sun, D., Morey, R. D., & Naveh-Benjamin, M. (2007). Signal detection models with random participant and item effects. *Psychometrika*, *72*, 621–642.
Rouder, J. N., Morey, R. D., Speckman, P. L., & Pratte, M. S. (2007). Detecting chance: a solution to the null sensitivity problem in subliminal priming. *Psychonomic Bulletin & Review*, *14*, 597–605.
Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t*-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.
Schacter, D. (1992). Priming and multiple memory systems: perceptual mechanisms of implicit memory. *Journal of Cognitive Neuroscience*, *4*, 244–256.
Shiffrin, R. M., Lee, M. D., Kim, W.-J., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*, 1248–1284.
Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*, 168–179.
Suls, J., Green, P., & Hillis, S. (1998). Emotional reactivity to everyday problems, affective inertia, and neuroticism. *Personality and Social Psychology Bulletin*, *24*, 127.
Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, *81*, 82–86.
Vickers, D., Lee, M. D., Dry, M., & Hughes, P. (2003). The roles of the convex hull and number of intersections upon performancee on visually presented traveling salesperson problems. *Memory & Cognition*, *31*, 1094–1104.
Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*, 779–804.
Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. P. P. P. (2010). Bayesian hypothesis testing for psychologists: a tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*, 158–189.
Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: an empirical comparison using 855 *t* tests. *Perspectives on Psychological Science*, *6*, 291–298.
Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: a flexible WinBUGS implementation of a default Bayesian *t*-test. *Psychological Bulletin & Review*, *16*, 752–760.
Zeelenberg, R., Wagenmakers, E.-J., & Raaijmakers, J. G. W. (2002). Priming in implicit memory tasks: prior study causes enhanced discrimability, not only bias. *Journal of Experimental Psychology: General*, *131*, 38–47.