Bayesian model selection with applications in social science
Wetzels, R.M.

Citation for published version (APA):
Wetzels, R. M. (2012). Bayesian model selection with applications in social science

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
9 Discussion

9.1 Discussion

In this thesis we have proposed Bayesian alternatives to frequentist null hypothesis tests. More specifically, we have outlined a Bayesian t test, a Bayesian correlation test, a Bayesian test for partial correlations, and a Bayesian one-way and two-way ANOVA. All these tests are essential tools for empirical research in psychology.

In this thesis we also compared the proposed Bayesian null hypothesis tests to their frequentist counterparts, discussed their behavior, and explained how and when these tests can be applied. In the second part of this thesis we discussed the practical benefits of the Bayesian null hypothesis tests, that is, we explain how social science can benefit from applying Bayesian methods. This dissertation points out several Bayesian solutions to problems that concern $p$ value hypothesis testing, such as the inability to gather evidence in favor of the null hypothesis, the asymmetry between the null hypothesis and the alternative hypothesis, the fallacy of the transposed conditional, and the consequences of optional stopping.

In the remainder of this discussion, we will first recap four main problems with $p$ value hypothesis testing that social science research is confronted with. Then we will discuss how the application of Bayesian can be of help, or even solve the problem.

Bayesian Methods Allow Evidence in Favor of the Null Hypothesis

Bayesian methods allow researchers to gather support in favor of the null hypothesis. This is an important feature because the current social science literature has serious problems with ad-hoc theories and models that are being discussed in the literature as being “true” while they might very well be false. This is a concern to psychological scientists, because when a certain theory is established in the literature, it is relatively difficult to overthrow. The reason is that, within the frequentist framework of null hypothesis testing, it is impossible to gather evidence in favor of the null. This makes it difficult to eliminate false results.

Fortunately, researchers can gather evidence in favor of the null hypothesis when they compute a Bayes factor that contrasts the null hypothesis with a specific alternative hypothesis. Hence, the Bayes factor allows researchers to disprove false theories more easily. This will greatly benefit the social sciences, as it enables researchers to publish the results of replication attempts for well-known experiments, even when the original finding is not replicated and evidence in favor of the null hypothesis is found instead. By encouraging replication studies, the evaluation of psychological theories and models becomes easier, something scientists can only benefit from.

Bayesian Methods Treat the Alternative and Null Hypothesis Alike

When a frequentist null hypothesis test is conducted, the alternative hypothesis is not evaluated. More specifically, the question that is not evaluated is whether the data are likely under the alternative hypothesis. In psychology, the alternative hypothesis is usually (implicitly) considered to be the theory or model that is being tested. It only
seems sensible to evaluate the plausibility of the data under both the null hypothesis and
the alternative hypothesis.

The Bayes factor does compare the two models directly. The alternative theory is
being evaluated just as the null hypothesis, resulting in a balanced, comparative measure
of evidence. Consider cases where the data are highly unlikely under the null hypothesis
but also highly unlikely under the alternative hypothesis. In such cases the \( p \) value rejects
the null hypothesis whereas the Bayes factor indicates that the data are inconclusive. We
believe that the latter approach is more sensible and it is more closely linked to the
research question at hand.

**Bayes Factors are More Easily Interpreted Than the \( p \) Value**

Many people misinterpret the \( p \) value as the probability of the null hypothesis being true.
This is not correct as the \( p \) value is a conditional probability: It is the probability of the
data, or data more extreme, given that the null hypothesis is true and given a specific
design. This is a complicated probability to interpret, as shown by a famous quote of
Jeffreys: “What the use of \( p \) implies, therefore, is that a hypothesis that may be true
may be rejected because it has not predicted observable results that have not occurred.”

Because of this confusing definition, many researchers seem to confuse this probability
with its transposed probability, the probability of the null hypothesis given the data.
These two probabilities are not the same, as the \( p \) value is calculated given that the null
hypothesis is true, therefore it can not also be the probability that the null hypothesis
is true. A clear example that \( p(D \mid L) \neq p(L \mid D) \) is that when \( D \) is the event of
someone dying, and \( L \) is the event of someone being hit by lightning. It is clear that the
\( p(D \mid L) \), the probability of someone dying when she was hit by lightning is much larger
than \( p(L \mid D) \), the probability that someone who is dead was hit by lightning (there are
of course many other ways to die, and not many people die by being hit by lightning).

More specifically, in the specific case of null hypothesis testing we can show that
these two conditional probabilities are not the same by Lindley’s paradox. In his article,
D. V. Lindley (1957) gives an example that shows that if \( H \) is a simple hypothesis, and
\( y \) the result of an experiment, the following two phenomena can occur at the same time:

1. a significance test for \( H \) reveals that \( y \) is significant at the 5% level
2. the posterior probability of \( H \), given \( y \), is, for quite small prior probabilities of \( H \),
as high as 95%.

It might be that the \( p \) value is so often misinterpreted as the probability of the null
hypothesis given the observed data, because this is often the probability that researchers
want to calculate. Therefore, computing the Bayes factor might be a convenient solution
to this misinterpretation. The Bayes factor has a clear interpretation as the change
from prior to posterior odds, brought about by the data. Assuming that this is what
researchers are interested in, applying Bayesian methods would not only yield a measure
that is easier to interpret, but it would also yield a measure that gives an answer to the
question that is being asked.

**Bayes Factors are Not Vulnerable to Optional Stopping**

The optional stopping problem is usually defined as the fact that the \( p \) value has an exact
interpretation as the probability of the observed data or data more extreme, given that
the null hypothesis is true and given a pre-defined design and sample size. Hence, after data collection starts one is not allowed to stop before the experiment was supposed to be finished and interpret the $p$ value. In the same vain, one is not allowed to continue testing after the predefined sample size is reached.

In sum, researchers are not allowed to monitor the $p$ value as the data come in and stop when it falls below .05 (the usual critical value below which the data is considered significant). At the same time, researchers are not allowed to continue testing when the planned sample size is reached. When one computes a $p$ value, both early stopping and continued testing are considered to be statistical cheating.

However, in some situations it can even be unethical not to practice optional stopping. Let’s assume that a researcher conducts an experiment to investigate whether a new medicine to treat a disease has a positive effect. She constructs a control group with participants receiving placebo treatment and an experimental group with participants receiving the new medicine. Each participant receives one pill each week for a total of 20 weeks. Now, what if the new medicine is so successful that after 10 weeks it is obvious that it cures the disease much more effectively than the old treatment? Then, according to the statistical rules, the researcher is not allowed to stop experimenting and make the new medicine available to the patients in the control group. However, it is arguably not ethical to prevent the control group from using the medicine, as the patients in the control group may experience needless discomfort (or even death) before the experiment is finished.

Bayesian model selection is not vulnerable to the optional stopping problem. Edwards, Lindman, and Savage (1963, p. 193) note that “the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience”. Hence, the researcher can monitor the Bayes factor when the data come in. If the data are convincing enough, she can stop data collection and vice versa, if the data are not convincing enough, she can keep on adding data until she finds that her point has been proven.

9.2 Future Directions

In this dissertation, we have shown that researchers in social science can benefit greatly from adding Bayesian methods to their statistical toolbox. We have pointed out various pitfalls of frequentist $p$ value null hypothesis testing and we have shown how computing Bayes factors can be used to circumvent these pitfalls.

However, Bayesian methods have only recently become popular. Hence, there is still a lot of room for development of Bayesian methods for mainstream scientific purposes. In the remaining part of the discussion, we list a few remaining issues for future research.

How to Make a Choice Between the Various Default Prior Distributions?

This thesis deals with the application of default priors for Bayesian model selection. Default priors are the preferred choice for standard testing situations, because we feel that a statistical test should be as objective as possible. Moreover, a statistical test can yield a reference point for the behavior of the Bayes factor when other prior distributions are used. Hence, a personal (i.e., subjective) prior distribution would be difficult to use
for standard testing situations, and the same holds for a prior distribution that is based on the data.

However, in recent years, different default priors have been proposed, each having slightly different properties and differing asymptotic behavior. This induces a new question that is slightly paradoxical, is the choice between various objective priors a subjective choice? We see at least two ways to study this problem, one is to investigate the behavior of the various default prior distribution which answers the more pragmatic question whether the choice for a particular prior makes a practical difference (see next subsection). Another way to study this problem is to investigate whether it would be possible to create one overarching prior distribution that encompasses all the default options.

What is the Behavior of the Various Default Priors for Different Linear Models?

Much statistical research is focused on linear models. As discussed earlier, there are different default prior distributions that can be used. We already indicated that it may be hard to choose one of these various “default” choices (e.g., the original $g$ prior with $g = n$, $g = \#\text{parameters}$, $g = \ldots$; the Jeffreys-Zellner-Siow prior; one of the Liang et al. scale mixture priors). It is interesting to investigate whether in practice, these priors result in substantially differing conclusions. We have conducted an extensive simulation study comparing the most common default prior distributions, for linear models, generalized linear models, and generalized linear mixed models (results not reported in this thesis). The interim conclusion is that when a reasonable sample size is used, there is not much difference between the default priors. If this result holds more generally, maybe an Occam’s razor for the specification of priors should be proposed: If various priors yield the same results, the least complex prior should be chosen.

How to Interpret the Bayes Factor Scale?

How to interpret the Bayes factor scale? Jeffreys proposed a scale for the interpretation of the Bayes factor, a scale that is used throughout this thesis. However, many prior distributions can be considered a valid choice and this choice influences the Bayes factor. Hence, if the choice for a prior is somewhat ad-hoc, then the resulting Bayes factor scale is also somewhat ad-hoc.

One potential solution to this problem is to interpret the Bayes factor scale in terms of statistical power. For example, Cohen’s $d$ is interpreted as follows; an effect size $d$ below 0.3 is considered to be a “small” effect size, a $d$ of 0.5 is a “medium” effect size and a $d$ of 0.8 or higher is a “large” effect size. A researcher could combine the scale of Cohen and Jeffreys. One could calculate the sample sizes needed to obtain a certain Bayes factor, assuming a certain effect size. For example, if one expects a small effect size, $d = 0.3$ and one wishes to obtain a Bayes factor of 3, the expected sample size could be $n = 40$. However, if the researchers wish to obtain a higher Bayes factor of, say 30, the expected sample size could be $n = 200$. This information could be used to calibrate the Bayes factor and give it a scale that is interpretable across different prior distributions.

How to Choose a Prior Distribution that is Based on an Experimental Question or Design?

There are many different choices to make when it comes to choosing a prior distribution. In some situations, a default prior distribution is the preferred choice, but there might be
9.2. Future Directions

situations in which one would be better off choosing a subjective or an empirical prior. This depends on the scientific question that is being asked. We can imagine situations where implementing as much prior information in the model as possible makes sense. For example, when one is comparing two models that have a clear psychological interpretation. Then, when comparing these two models (or theories), the prior distributions are a substantive part of the psychological theory and hence should be chosen in line with this theory.

Vanpaemel (2010) discusses an example where he formalized three hypotheses that concern a decision maker who has to choose between two alternatives. The first hypothesis states total indifference between the alternatives. Hence, the probability of choosing alternative 1 over alternative 2 is equal to 0.5, $\theta = 0.5$. Hypothesis two is that the decision maker is biased towards one alternative over the other. The probability of choosing alternative 1 over alternative 2 could then be anything between zero and one, $\theta \sim \text{dbeta}(1, 1)$. If one assumes that there are correct and incorrect alternatives, one could formulate a the third hypothesis; the decision maker performs better than chance $\theta > 0.5$. Notice that the model equations of these three models are assumed to be the same for all hypotheses. The only difference, the difference that is psychologically relevant, is implemented in the model through the prior distributions on the parameter $\theta$.

It would be convenient if the psychological community would decide when it is appropriate to use one of the various prior choices that are available, defining a choice protocol and distinguishing various experimental settings and combine these settings with a heuristic for choosing a prior distribution.

**What are the Pros and Cons of Calculating the Bayes Factor Versus Parameter Estimation?**

This thesis focuses on the calculation of the Bayes factor, comparing different models or hypotheses. However, the calculation of the Bayes factor is often not easy, and the influence of the prior on the Bayes factor is considerable. Hence, Bayes factors should be used and interpreted with care. In comparison, the influence of the prior distribution on parameter estimation is far less. To avoid complex discussions on the merits of Bayes factors, one could also revert to Bayesian parameter estimation. Note however, these two approaches are by no means mutually exclusive. A study giving a roadmap on how to combine these two approaches in psychological research would be valuable to researchers interested in applying Bayesian methods.

**How to Handle the Prior Probability of a Model or Hypothesis?**

It is difficult to interpret or specify the probability of a model or a hypothesis. For example, what does it mean that a model has a probability $p$ of being true? Moreover, it is difficult to define how multiple comparisons should be taken into account. If one is comparing many different models at the same time, it might be important to take the prior probability of a specific model into account, based on the number of models, or maybe even based on the number of parameters in the model (Scott & Berger, 2010). There are ideas on how to do this for well-defined models but for psychological process models this is much more complicated.

Furthermore, what is the prior probability of the null model? For example, considering the $t$ test, there is no definite agreement about whether the null hypothesis is ever exactly true. If it is true, is the null equally probable as the alternative hypothesis that there is
a difference between the means? Again, for psychologically relevant models the situation becomes even more complex. For the calculation of the Bayes factor these questions are not directly relevant, as the prior model probability is not taken into account. However, the questions themselves remain interesting and relevant to psychological researchers that are using Bayes for their inference.