original article

Bayesian analyses: where to start and what to report

Rens van de Schoot
Utrecht University & North-West University
Sarah Depaoli
University of California

Most researchers in the social and behavioral sciences will probably have heard of Bayesian statistics in which probability is defined differently compared to classical statistics (probability as the long-run frequency versus probability as the subjective experience of uncertainty). At the same time, many may be unsure of whether they should or would like to use Bayesian methods to answer their research questions (note: all types of conventional questions can also be addressed with Bayesian statistics). As an attempt to track how popular the methods are, we searched all papers published in 2013 in the field of Psychology (source: Scopus), and we identified 79 empirical papers that used Bayesian methods (see e.g. Dalley, Pollet, & Vidal, 2013; Fife, Weaver, Cool, & Stump, 2013; Ng, Ntoumanis, Thøgersen-Ntoumani, Stott, & Hindle, 2013). Although this is less than 0.5% of the total number of papers published in this particular field, the fact that ten years ago this number was only 42 indicates that Bayesian methods are slowly beginning to creep into the social and behavioral sciences.

The current paper aims to get you started working with Bayesian statistics. We provide: (1) a brief introduction to Bayesian statistics, (2) arguments as to why one might use Bayesian statistics, (3) a reading guide used to start learning more about Bayesian analyses, and, finally (4) guidelines on how to report Bayesian results. For definitions of key words used in this paper, please refer to Table 1.

Bayesian Statistics: A brief introduction

Before providing arguments why one would use Bayesian statistics, we first provide a brief introduction. Within conventional statistical techniques, the null hypothesis is always set up to assume no relation between the variables of interest. This null hypothesis makes sense when you have absolutely no idea of the relationship between the variables. However, it is often the case that researchers do have *a priori* knowledge about likely relationships between variables, which may be based on earlier research. With Bayesian methods, we use this background knowledge (encompassed in what is called a 'prior') to aid in the estimation of the model. Within Bayesian statistics, we can learn from our data and incorporate new knowledge into future investigations. We do not rely on the notion of repeating an event (or experiment) infinitely as in the conventional (i.e., frequentist) framework. Instead, we incorporate prior knowledge and personal judgment into the process to aid in the estimation of parameters.

Thus, the key difference between Bayesian statistics and conventional (e.g., maximum likelihood) statistics concerns the nature of the unknown parameters in a statistical model. The unknown model parameters are those that are freely estimated. For example, when estimating a regression model with one dependent outcome variable (Y) and two predictors (X1 and X2), see Figure 1, the unknown parameters are: one

Table 1: *A brief definition of key words and phrases*

Key Words and Phrases	*Definition*
Background Knowledge	Knowledge about population parameter values (e.g., a regression coefficient) that can be determined based on prior research, an analysis of previous data, or expert opinions.
Bayesian Statistics	A statistical tool that can be used to combine background knowledge of population parameters with current data to obtain estimates via the resulting posterior distribution.
Credibility Interval	The Bayesian version of the traditional confidence interval. Can be interpreted as the (e.g. 95%) probability that the population parameter is between the particular upper and lower bounds determined by the Bayesian credibility interval.
Confidence Interval	Frequentist (conventional) confidence intervals are based on repeated sampling theory such that, for a 95% confidence interval, 95 out of 100 replications of exactly the same experiment capture the fixed but unknown regression coefficient.
Frequentist Statistics	A class of statistics that relies on point estimation and opposes Bayesian statistics because it does not incorporate background knowledge into the estimation process (e.g., maximum likelihood estimation methods).
Hyperparameters	The specific parameters for a prior distribution. For example, if a normal distribution is selected for the prior, then the mean and variance parameters of this normal prior are called the hyperparameters. The values specified for the hyperparameters control the amount of (un)certainty incorporated into the model about a given parameter.
Likelihood Function	Represents the observed data likelihood. This weights the prior distribution in Bayesian statistics to obtain the posterior distribution form which we draw inferences.
Markov chain Monte Carlo (MCMC)	A simulation-based estimation method that is used to make simulated draws from a distribution and form a Markov chain that represents the posterior distribution.
p-value	In frequentist statistics, this is the probability of obtaining a test statistic as or more extreme than the critical value, given that the null hypothesis is true.
Parameter	A fixed but unknown feature of the model that is estimated either through frequentist or Bayesian methods.
Posterior	The distribution that is obtained once combining the prior and the likelihood in the Bayesian estimation process.
Posterior p-value	Bayesian p-value that is based on the posterior distribution obtained.
Precision	The amount of information incorporated into a prior distribution. More information is equated to having a larger degree of precision (less uncertainty) and therefore equates to smaller variability in the prior. Precision is specifically defined as the inverse of the variance.
Prior	A statistical distribution that can be used to capture the amount of (un)certainty in a population parameter. This distribution is then weighted by the sample data to obtain the posterior, which is used to make inference.

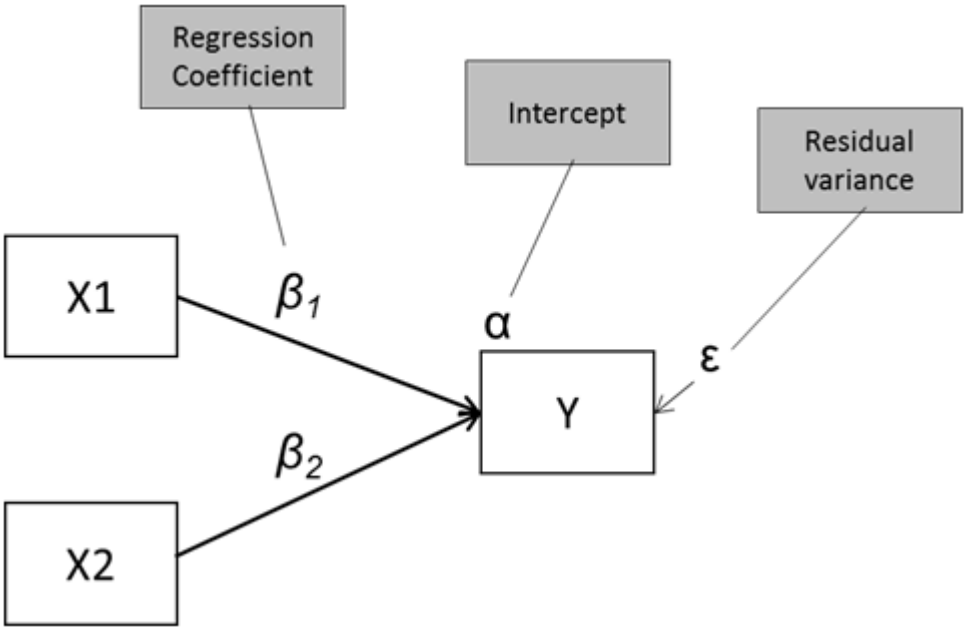

Figure 1. Regression model with the unknown parameters.

intercept (α), two regression coefficients (β_1, β_2), and the residual variance of the dependent variable (ε). With conventional statistics it is assumed that in the population there is only one true population parameter, for example, one true regression coefficient that is fixed but unknown. In the Bayesian view of probability, all unknown parameters can incorporate (un)certainty that can be defined by a probability distribution. Thus, Bayesian methods do not provide one outcome value but rather an interval ('distribution') with a probability that this interval contains the regression coefficient. That is, each parameter is believed to have a distribution that captures (un)certainty about that parameter value. This (un)certainty is captured by a distribution that is defined *before* observing the data and is called the *prior distribution* (or *prior*). Next, the *observed* evidence is expressed in terms of the *likelihood function* of the data. The data likelihood is then used to weigh the prior and this product yields the *posterior distribution*, which is a compromise of the prior distribution and the likelihood function. These three ingredients constitute the famous Bayes' theorem.

The three ingredients underlying Bayesian statistics are summarized in Figure 2 for one of the regression coefficients (β) pulled from Figure 1. The first ingredient of Bayesian statistics is knowledge about this parameter before observing the data, as is captured in the prior distribution. Often this knowledge stems from systematic reviews, meta-analyses or previous studies on similar data (see O'Hagan et al., 2006). In Figure 2 five different priors are displayed for β. The variance, or precision (inverse of the variance), of the prior distribution reflects one's level of (un)certainty about the value of the parameter of interest: the smaller the prior variance, the more certain one is about the parameter value. There are three main classes of priors that differ in the amount of certainty they carry about the population parameter. These different priors are called: (1) non-informative priors, (2) informative priors, and (3) weakly-informative priors. Non-informative priors are used to reflect a great

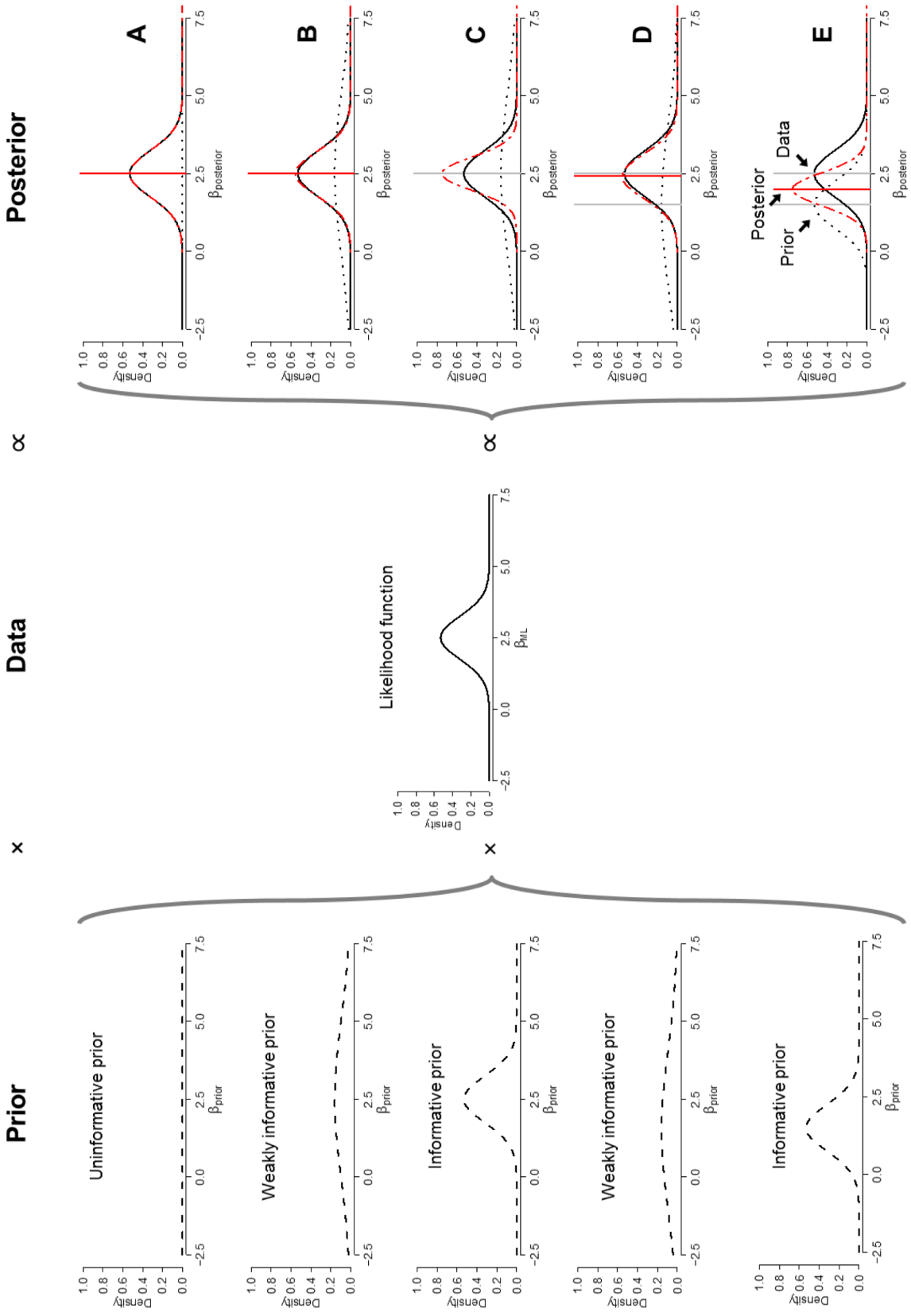

Figure 2. The three ingredients of Bayesian statistics for the regression coefficient.

deal of uncertainty in what the population parameter looks like. Weakly-informative priors incorporate some information into the model and reflect more certainty about the population parameter compared to a non-informative prior. This prior contains some useful information, but it does not typically have much influence on the final parameter estimate. Finally, the prior that contains the most amount of certainty about the population parameter is an informative prior. Informative priors contain strict numerical information that is crucial to the estimation of the model and can have a large impact on final estimates. These three levels of informativeness are created by modifying the parameters of the prior, called hyperparameters. Specifically, the hyperparameters for these priors (e.g., the prior mean and prior variance) are fixed to express specific information and levels of (un)certainty about the model parameters being estimated.

The second ingredient is the information in the data itself. It is the observed evidence expressed in terms of the likelihood function of the data (β_{ML}). Thirdly, both *prior* and *data* are combined via Bayes' theorem. The *posterior distribution* reflects one's updated knowledge, balancing background knowledge (the prior) with observed data (the likelihood). With a non or weakly informative prior, the posterior estimate may not be influenced by the choice of the prior much at all, see Figure 2A, 2B and 2C. With informative (or subjective) priors, the posterior results will have a smaller variance, see Figure 2C. If the prior disagrees with the information in the data, the posterior will be a compromise between the two, see Figure 2E, and then one has truly learned something new about the data or the theory.

Why would one use Bayesian Statistics?

There are four main reasons as to why one might choose to use Bayesian statistics: (1) complex models can sometimes not be estimated using conventional methods, (2) one might prefer the definition of probability, (3) background knowledge can be incorporated into the analyses, and (4) the method does not depend on large samples.

First, some complex models simply cannot be estimated using conventional statistics. In these cases of rather complex models, numerical integration is often required to compute estimates based on maximum likelihood estimation, and this method is intractable due to the high dimensional integration needed to estimate the maximum likelihood. For example, conventional estimation is not available for many multilevel latent variable models, including those with random effect factor loadings, random slopes when observed variables are categorical, and three-level latent variable models that have categorical variables. As a result, alternative estimation tools are needed. Bayesian estimation can also handle some commonly encountered problems in orthodox statistics. For example, obtaining impossible parameters estimates, aiding in model identification (Kim, Suh, Kim, Albanese, & Langer, 2013), producing more accurate parameter estimates (Depaoli, 2013), and aiding in situations where only small sample sizes are available (Zhang, Hamagami, Wang, Grimm, & Nesselroade, 2007).

Second, many scholars prefer Bayesian statistics because of the different definition of probability. Consider for example the interpretation of confidence intervals (CIs). The frequentist CI is based on the assumption of a very large number of repeated samples from the

population. For any given data set, a regression coefficient can be computed. The correct frequentist interpretation for a 95% CI is that 95 out of 100 replications of exactly the same experiment capture the fixed but unknown regression coefficient. Often this 95% CI is misinterpreted as meaning there is a 95% probability that the regression coefficient resides between the upper and lower limit, which is actually the Bayesian interpretation. Thus, Bayesian confidence intervals may be more intuitively appealing.

Third, as described above, with Bayesian statistics one can incorporate (un)certainty about a parameter and update this knowledge. Let background knowledge be the current state of affairs about a specific theoretical model, which can be operationalized by means of a statistical model, see for example Figure 1. Everything that is already known about the parameters in the model based on, for example, previous publications, can be used to specify informative priors, see Figure 2. When the priors are updated with current data, something can be learned, especially if the priors (i.e., current state of affairs) disagree with the current data. Let us explain this conflict between the prior and the current data using a simplified example where two groups were generated (M_1=0, M_2=0.45, SD=2; n=100) using an exact data set. Obviously, when no prior knowledge is specified (using non-informative prior distributions), there is no difference between the population difference ($\Delta M_{population}$ = 0.45) and the estimated difference obtained with the Bayesian analysis ($\Delta M_{posterior}$ = 0.45). Next, we specified informative priors that were inaccurate to the population; that is, for M_1 we specified a prior mean of .50 and for M_2 we specified a prior mean of .05. We varied the precision of the prior distribution to obtain weakly informative (low precision) and highly informative priors (high precision). The relation between the precision and the prior-data conflict (i.e., the difference between $\Delta M_{population}$ and $\Delta M_{posterior}$) is shown in Figure 3. In conclusion, the higher the precision, the more influence the prior specification has on the posterior results. If there is a large prior-data conflict, apparently the current state of affairs about the statistical model does not match with the current data. This is what a Bayesian would call: fun! Because

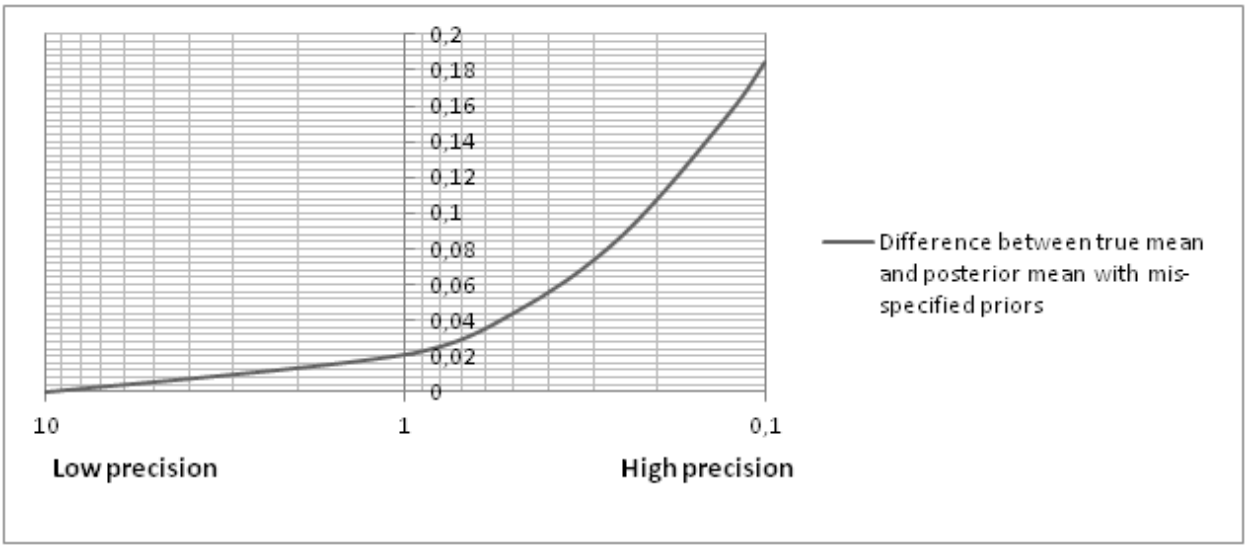

Figure 3. The relation between precision and bias.

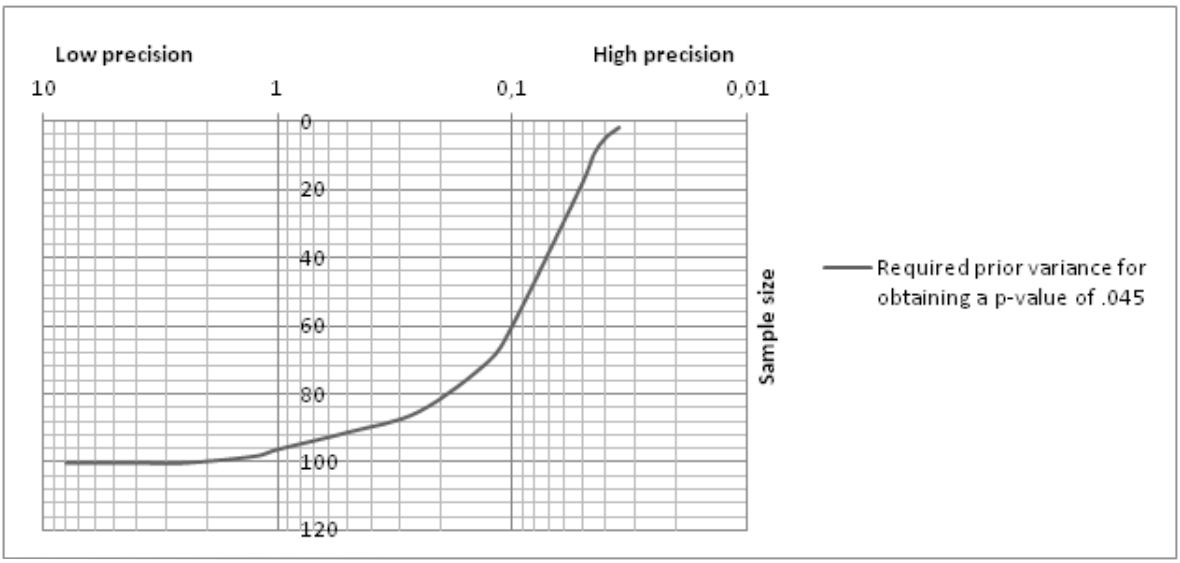

Figure 4. The relation between precision and the possible gain in sample size.

now, finally, something new has been discovered and one should discuss in the paper how it could be that there is a prior-data conflict. Is it the theory that needs to be adjusted? Or, was the data not a random sample from the population? Or does the theory not hold for the specific population used for the current study? All of these questions are related to updating knowledge.

Fourth, Bayesian statistics is not based on large samples (i.e., the central limit theorem) and hence large samples are not required to make the math work. Many papers have shown the benefits of Bayesian statistics in the context of small data set (e.g., Zhang et al., 2007).To illustrate the decrease in required sample size we performed a small simulation study. Multiple exact data sets with two groups, see above, were generated with the goal to obtain for every data set the same *p*-value for a *t*-test. With n = 100 the *t*-test produced a just significant effect of p = .045. Also, when using objective Bayesian statistics with an infinitive low prior precision (non-informative prior) the Bayesian *p*-value was .045. Next, we specified weakly and highly informative priors with a prior mean equal to the population values (data based prior), but we varied the precision. The relation between the precision and the required sample size to obtain the same significant effect of p = .045 is shown in Figure 4 showing that the higher the precision, the smaller the sample size needed to obtain the same effect. In conclusion, the more precision a researcher is willing to specify before seeing the data, the smaller the sample size needed to obtain the same effect compared to an analysis without specifying any prior knowledge.

1 When using exact data sets the data characteristics are exactly the same as the population statistics. For example, if the population mean is specified as being zero with a standard deviation of 2, the data set generated from this population also has exactly a mean of zero and a SD of 2. The software BIEMS (Mulder, Hoijtink, & de Leeuw, 2012) was used for generating such an exact data. The t-tests for mean differences were performed in the software Mplus.

Where to start?

Of course, the introduction offered in the current paper is not enough to start working with Bayesian statistics, therefore we provide a step-by-step reading guide as well as resources for statistical programs that can implement

Bayesian methods. For a gentle introduction to Bayesian estimation, we recommend the following: Kaplan and Depaoli (2013); Kruschke (2011); and van de Schoot et al. (2013). For a more advanced treatment of the topic, readers can be referred to a variety of sources, which include Gelman, Carlin, Stern, and Rubin (2004).

There are many different software programs that can be used to implement Bayesian method in a variety of contexts and we list the major programs here. Various packages in the R programming environment (e.g., Albert, 2009) implement Bayesian estimation, with the number of Bayesian packages steadily increasing. Likewise, AMOS (Arbuckle, 2006), BUGS (Ntzoufras, 2009), and M*plus* (Muthén, 2010) can be used for estimating Bayesian latent variable models, which can also include multilevel or mixture extensions. BIEMS (Bayesian inequality and equality constrained model selection; Mulder, Hoijtink, & de Leeuw, 2012) is a Bayesian program for multivariate statistics and Bayesian hypothesis testing. Standard statistical models estimated through the SAS software program can now be used for Bayesian methods. Finally, SPSS incorporates Bayesian methods for imputing missing data.

What to include in an empirical Bayesian paper?

There are several key components that must be included in the write-up of an empirical paper implementing Bayesian estimation methods. The statistical program used for analysis is an important detail to include since different methods (called *sampling methods*) are implemented in different Bayesian programs and these methods may lead to slightly different results. A discussion of the priors needs to be in place. The researcher should thoroughly detail and justify all prior distributions that were implemented in the model, even if default priors were used from a software program. It is important to always provide these details so that results can be replicated, a full understanding of the impact of the prior can be obtained, and future researchers can draw from (and potentially update) the priors implemented. A discussion of chain convergence must be included. Each model parameter estimated should be monitored to ensure that convergence was established for the posterior. A variety of statistical tools can be used to help monitor and evaluate chain convergence (see, Sinharay, 2004), and visual inspection of convergence plots can also aid in detecting non-convergence. Finally, researchers might also find it beneficial to run a sensitivity analysis using different forms and levels of informativeness for the priors implemented. Although we do not recommend using this as a means for updating the prior on the same data set (i.e., the original prior should still be used in the final write-up), the sensitivity analysis can help provide insight into the impact of the prior and this impact can be discussed further in the paper.

Conclusion

In our experience, we have found Bayesian methods to be incredibly useful for solving estimation problems, handling smaller sample sizes with greater accuracy, and incorporating prior judgment or knowledge into the estimation process. It is our aim that this paper will serve as a starting point for those interested in implementing Bayesian methods.

Author's note

The first author was supported by a grant from the Netherlands organization for scientific

research: NWO-VENI-451-11-008.

References

Albert, J. (2009). *Bayesian computation with R*. New York: Springer.

Arbuckle, J. L. (2006). *Amos* (Version 7.0) [Computer Program]. Chicago: SPSS.

Dalley, S. E., Pollet, T. V., & Vidal, J. (2013). Body size and body esteem in women: The mediating role of possible self expectancy. *Body Image, 10*(3), 411-414. doi:10.1016/j.bodyim.2013.03.002

Depaoli, S. (2013). Mixture class recovery in GMM under varying degrees of class separation: Frequentist versus Bayesian estimation. *Psychological Methods, 18*(2), 186-219. doi:10.1037/a0031609

Fife, B. L., Weaver, M. T., Cook, W. L., & Stump, T. T. (2013). Partner interdependence and coping with life-threatening illness: The impact on dyadic adjustment. *Journal of Family Psychology, 27*(5), 702-711. doi:10.1037/a0033871

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

Kaplan, D. & Depaoli, S. (2013). Bayesian statistical methods. In T. D. Little (ed.), *Oxford handbook of quantitative methods* (pp 407- 437). Oxford: Oxford University Press.

Kim, S. Y., Suh, Y., Kim, J. S., Albanese, M., & Langer M. M. (2013). Single and multiple ability estimation in the SEM framework: a non-informative Bayesian estimation approach. *Multivariate and Behavioral Research, 48*(4), 563-591. doi:10.1080/00273171.2013.802647

Muthén, B. (2010). *Bayesian analysis in Mplus: A brief introduction*. Technical Report. Version 3.

Mulder, J., Hoijtink, H., & de Leeuw, C. (2012). BIEMS: A Fortran 90 program for calculating Bayes factors for inequality and equality constrained models. *Journal of Statistical Software, 46*(2), 1-39.

Ng, J. Y. Y., Ntoumanis, N., Thøgersen-Ntoumani, C., Stott, K., & Hindle, L. (2013). Predicting psychological needs and well-being of individuals engaging in weight management: The role of important others. *Applied Psychology: Health and Well-being, 5*(3), 291-310. doi:10.1111/aphw.12011

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

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J.,... Rakow, T. (2006). *Uncertain judgements: Eliciting experts' probabilities*. West Sussex: Wiley.

Sinharay, S. (2004). Experiences with Markov Chain Monte Carlo Convergence Assessment in two psychometric examples. *Journal of Educational and Behavioral Statistics, 29*(4), 461-488. doi:10.3102/10769986029004461

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J. & van Aken, M. A. G. (2013). A gentle introduction to Bayesian Analysis: Applications to research in child development. *Child Development*. doi:10.1111/cdev.12169

Zhang, Z., Hamagami, F., Wang, L., Grimm, K. J., & Nesselroade, J. R. (2007). Bayesian analysis of longitudinal data using growth curve models. *International Journal of Behavioral Development, 31*(4), 374-383. doi:10.1177/0165025407077764

Rens van de Schoot
is Assistant Professor at Utrecht University, the Netherlands and extra-ordinary Professor at the Optentia research programme, North West University, South-Africa
a.g.j.vandeschoot@uu.nl

Sarah Depaoli
is Assistant Professor at University of California, Merced, USA
sdepaoli@ucmerced.edu