Improving estimation in genetic models using prior information

Espen Moen Eilertsen

Domain for Mental and Physical Health, Norwegian Institute of Public Health, Oslo, Norway

E-mail: EspenMoen.Eilertsen@fhi.no

Abstract

Statistical models used to investigate research questions in behavioral genetics often require large amounts of data. This paper introduces some key concepts of Bayesian analysis and illustrates how these methods can aid model estimation when the data does not provide enough information to reliably answer research questions. The use of informative prior distributions is discussed as a method of incorporating information from other sources than the data at hand. The procedure is illustrated with an ACE model decomposition of the variance of antisocial personality disorder. The data originates from the Norwegian Twin Registry, and includes adult twins assessed with the Structured Interview for DSM Personality (SIDP-IV). Inclusion of prior information lead to a shift with respect to conclusions about the presence of shared environmental effects compared to a traditional analysis. Small and medium sized studies should consider use of prior information to aid estimation of population parameters.

Introduction

In behavioral genetics, it is often of interest to investigate the relative genetic and environmental contribution to phenotypic variance. These analyses have been valuable in understanding the etiological basis of a variety of behavioral traits such as cognitive abilities, psychopathology and personality (Plomin et al., 2013).

Twin designs are one of the major methods used for disentangling the variance components attributable to genetics and environments (Plomin et al., 2013). Because it is known that monozygotic (MZ) twins share all their genes whereas dizygotic (DZ) twins on average share half of their genes, genetic contributions to a phenotype can be investigated by assuming equal environmental influences in MZ twins and in DZ twins (Plomin et al., 2013). Decades of research has accumulated substantial evidence regarding the influence of genetics and environment in a variety of traits. A recent meta-analysis (Polderman et al., 2015) summarized a majority of the traits that have been studied using the classical twin design. In total they investigated close to 18000 traits based on approximately 15 million twin pairs. This makes clear that in many situations, researchers already possess information about the contribution of genetics and environment in the phenotypes under study.

Bayesian methods allow such prior information to be incorporated in the analysis and thereby combine historical information with new data. In small samples and/or complex models with many parameters, there is often not enough data to provide estimates of satisfactory precision. When the available data is not a satisfactory information source, incorporation of prior information can aid precision in model estimation.

Bayesian approaches to genetic models have been described by others (Eaves & Erkanli, 2003; Eaves et al., 2005; van den Berg et al., 2006a; van den Berg et al., 2006b). None of these papers explicitly utilized prior information. The aim of the current paper is to give a practical example on how prior information can be incorporated in analyses using Bayesian methods, and discuss some methodological benefits and limitations. The method will be described in relation to the popular ACE model, but is generalizable to more complex models. As an illustration, the ACE model is fit to a sample of MZ and DZ male twins, assessed for antisocial personality disorder (ASPD). First, a Bayesian model that does not incorporate prior information is fitted and contrasted with a maximum likelihood (ML) analysis in order to demonstrate their similarity. Second, a model that incorporates prior information based on a meta-analysis is fitted.

ACE model

The ACE model splits phenotypic variance into three components:

$$\sigma_P^2 = \sigma_A^2 + \sigma_C^2 + \sigma_E^2.$$

σ_A^2 is the additive effects of genes. σ_C^2 represents environmental factors that create differences between pairs of twins, known as the “shared environment”. Shared environmental factors could be social class, parental norms, or the local school. σ_E^2 represents environmental factors that create differences within pairs of twins (in addition to everything else that is not captured by σ_A^2 and σ_C^2, such as random measurement error), known as the “unique environment”. Such factors could be everything from birth weight or education not shared by twins, to occupation and spouse. What separate σ_E^2 from σ_C^2 is not a property of the factors, but how the twins respond to them. If twins respond similar – it is a shared factor, and if twins respond dissimilar – it is a unique factor. Often these variance components are presented on a standardized scale by dividing each component by the estimate of total variation, making them interpretable as proportion of total variation. These are referred to by the capital letters, A, C and E.

The ACE model is often estimated as a structural equation model (Neale & Cardon, 1992), but it can also be formulated as a mixed effects model (Rabe-Hesketh et al., 2008). Rabe-Hesketh et al. (2008) clearly outlined ways of specifying the mixed effects model, and it is this approach that will be followed here. Some repetition of their specification is however necessary in order to demonstrate how prior information is incorporated in the Bayesian analysis, but see their paper for a thorough description. The mixed effects ACE model is specified as

$$y_{ij} = u + a_{ij}z_{1ij} + a_j z_{2ij} + c_j + e_{ij}.$$

Here y_{ij} is the phenotype score for twin i in family j. u is the overall mean. a_{ij} is a random coefficient that varies across twins nested in families. a_j and c_j are random coefficients that vary across families, and e_{ij} is the residual term.

The additive genetic variance is estimated from the two random effects a_{ij} and a_j which are constrained to have equal variance,

$$Var(a_{ij}) = Var(a_j) = \sigma_A^2.$$

a_{ij} represents unique genetic effects among twin pairs whereas a_j represents common genetic effects among twin pairs. MZ twin pairs share all genetic effects, whereas DZ twin pairs share half of the genetic effects while the other half is unique. The different genetic covariance in MZ twins and DZ twins is induced from the covariates z_{1ij} and z_{2ij},

$$z_{1ij} = \begin{cases} 0 \; if \; MZ \\ \sqrt{.5} \; if \; DZ \end{cases},$$

$$z_{2ij} = \begin{cases} 1 \; if \; MZ \\ \sqrt{.5} \; if \; DZ \end{cases}.$$

c_j is a random intercept at the family level and estimates shared environmental variance. The random effects are assumed normally distributed and independent with mean zero and variance to be estimated, representing the σ_A^2, σ_C^2 and σ_E^2 component.

$$a_{ij} \sim N(0, \sigma_A^2),$$
$$a_j \sim N(0, \sigma_A^2),$$
$$c_j \sim N(0, \sigma_C^2),$$
$$e_{ij} \sim N(0, \sigma_E^2).$$

In sum, the model has four parameters: u, σ_A^2, σ_E^2 and σ_E^2.

BAYESIAN ANALYSES

Bayesian analysis typically starts with specifying information about model parameters that is derived from other information sources than the data itself. This information is represented in probability distributions known as prior distributions. Prior distributions are then combined with the likelihood of the data given the parameters to obtain the posterior distribution. The posterior distribution is thus a compromise between data and prior information (Gelman, 2003), and forms the basis for statistical inference. In many applications, prior distributions are specified as noninformative in order to let the data guide estimation. This is accomplished by assigning equal probabilities over all possible parameter values. In such cases, inferences from posterior distributions are often close to those obtained from maximum likelihood based techniques. However, if information already exists about model parameters, for instance based on previous studies or meta-analysis, prior distributions can be specified to convey this information. Depending on the strength of prior certainty, the resulting posterior distribution is then to a larger extent dominated by the prior distributions.

In practice, Markov Chain Monte Carlo (MCMC) techniques are often used to approximate the posterior distribution. These techniques allow samples from the posterior distribution to be simulated even in complicated statistical models. These samples then form the basis for inferences about quantities of interest. For instance, the mean of the posterior samples can be used as a point estimate, the standard deviation as an estimate of uncertainty, etc. (for an introduction see Kruschke, 2010; Schoot & Depaoli, 2014). MCMC sampling also makes it convenient to estimate statistics that are functions of parameters. For instance, the narrow-sense heritability is a function of the three variance parameters in the ACE model:

$$h^2 = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_C^2 + \sigma_E^2}.$$

By calculating this quantity under each simulation, the posterior distribution of the narrow-sense heritability could also be approximated.

Although Bayesian estimation routines are implemented in some general purpose statistical software packages, the techniques are often only available for standard models. For more uncommon models, such as those often found in behavioral genetics, MCMC estimation can be programmed in most common computer languages with some programming effort. Perhaps most useful are programs such as BUGS (Spiegelhalter et al., 2003) and Stan (Stan Development Team, 2014a), which are fully devoted to Bayesian estimation and offers much more flexibility in model specifications than general purpose software. The examples in the current paper were estimated using Stan (see http://folk.uio.no/espenmei/bayesianACE.html for example code). Stan can run through the commandline terminal, but also interfaces to general statistical software packages that are suited for pre- and post-processing of data, such as R (Stan Development Team, 2014b).

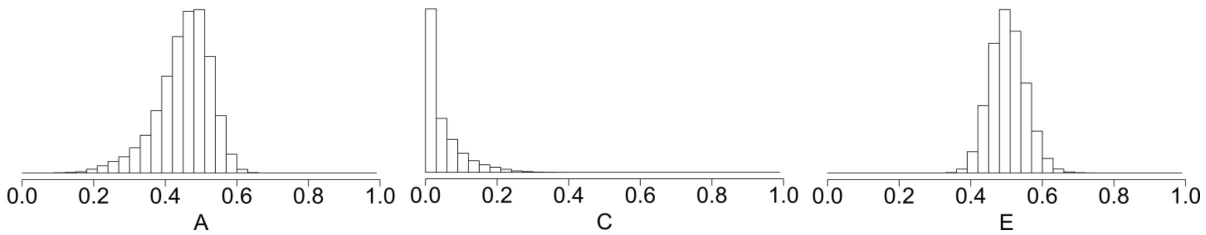

Figure 1. Histograms of posterior samples with uninformative priors.

EXAMPLE DATA ANALYSIS

To illustrate the procedure, the ACE model is fit to a dataset containing information on ASPD from the Norwegian Institute of Public Health Twin Panel (see Torgersen et al., 2012 for a description of the material). In order to later illustrate the value of including prior information when there are limited amounts of available data, only a sub-sample consisting of the male MZ and DZ twins is considered in the analysis. The sample consists of responses from 445 MZ twins (220 pairs) and 236 DZ twins (116 pairs) born between 1967 and 1979.

The outcome measure is an aggregate of seven 4-level items from the Structured Interview for DSM-IV Personality (SIDP-IV), which is a comprehensive semi-structured diagnostic interview for DSM-IV Axis II diagnoses. Because the purpose of the analysis is illustration, complicating considerations such as the information level in the dependent variable are not considered. The outcome is here treated as continuous.

UNINFORMATIVE PRIOR DISTRIBUTIONS

To demonstrate that a Bayesian analysis with non-informative prior distributions coincides with a ML analysis, the data is first analyzed using uninformative prior distributions, and parameter estimates contrasted with those obtained from a ML analysis. The latter analysis was carried out using the Gllamm program in STATA which allows constraining the variances of a_{ij} and a_j to equal (Rabe-Hesketh et al., 2004).

For the mean (μ), a normal distribution centered at zero, with a large standard deviation (relative to the scale of data) was used. There exists a comprehensive literature on which distributions are suited for uninformative prior distributions on variance parameters. Here, a uniform distribution with a lower bound at zero and no upper bound on the standard deviation (square root of the variance components) was used, which has been recommended as a general approach (Gelman, 2006). These priors express no particular knowledge about the standard deviations, other than that they are greater than zero.

$$u \sim Normal(0{,}10),$$
$$\sigma_A \sim Uniform(0, \infty),$$
$$\sigma_C \sim Uniform(0, \infty),$$
$$\sigma_E \sim Uniform(0, \infty).$$

As seen from table 1, the Bayesian and ML estimates closely resemble each other, as expected when using noninformative priors. The main difference is that the Bayesian analysis does not estimate σ_C^2 to be exactly zero, but slightly higher. Conversely σ_A^2 is estimated slightly higher in the ML analysis. These minor discrepancies are likely a result of σ_C^2 being very close to zero, which can be problematic in linear mixed models (Chung et al., 2012). The uniform prior can also result in overestimation when the variance is small (Gelman, 2006). Additionally, point estimates based on the posterior samples will depend on which summary statistic is being used as long as the distribution is not symmetrical. In this case, the median would lead to a considerably lower estimate of σ_C^2. However, it is perhaps more informative to visually inspect the posterior samples. Figure 1 display histograms of the standardized variance components. Although C estimates as high as .2 are not completely unlikely, most of the mass is located very close to zero. Consequently, there is not much evidence suggesting that a C effect is present in the investigated population. In this situation, the c_j term could alternatively be dropped from the model in favor of retaining a more parsimonious AE model. This approach is often followed in this type of analysis, but will not be further investigated here (see Vehtari, Gelman and Gabry (2015) for a discussion of model comparison in a Bayesian context).

Table 1. Summary statistics from ML and Bayesian analysis with uninformative priors. Estimate is posterior mean for the Bayesian method. SD is posterior standard deviation. SE is standard error. Regression coefficients are the "fixed-effects" in the model. Variance components are the variances of the "random-effects" on the scale of the data and the standardized variance components.

	Bayesian		ML	
	Estimate	SD	Estimate	SE
Regression coefficients				
u	.61	.07	.61	.07
Variance components				
σ_A^2	.99	.18	1.08	.14
σ_C^2	.09	.11	.00	.00
σ_E^2	1.09	.10	1.07	.10
A	.46		.50	
C	.04		.00	
E	.50		.50	

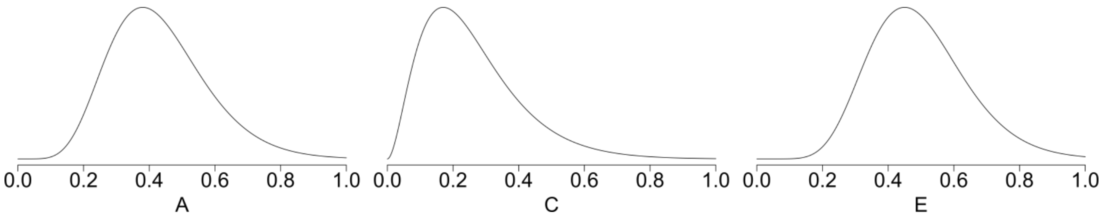

Figure 2. Illustration of prior distributions.

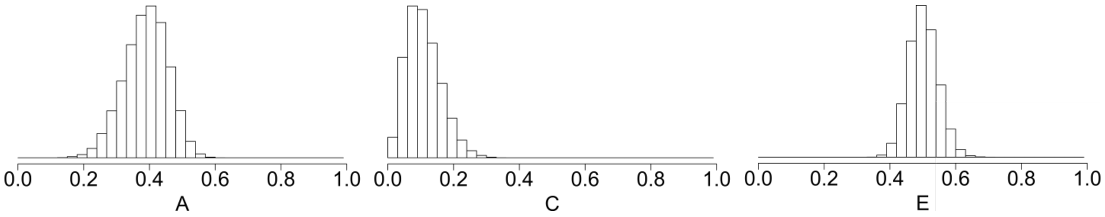

Figure 3. Histograms of posterior samples with informative priors.

From this analysis, additive genetic effects seem to account for roughly half of the variation in ASPD for males, whereas there is not much evidence suggesting the presence of shared environmental effects. Either, this is because shared environments do not contribute to variation in the population investigated, or, the data is too noisy to accurately estimate a variance component attributable to shared environmental effects. Assuming the latter, it would be meaningful to add information to allow this entity to be more accurately estimated. This can be done by forming informative prior distributions that favors certain parameter values.

INFORMATIVE PRIOR DISTRIBUTIONS

As discussed above, informative prior distributions carry information about model parameters that is obtained from sources other than the current data. In order to form informative prior distributions for the variance parameters, results from Rhee and Waldman's (2002) meta-analysis on the genetic and environmental influences on antisocial behavior were used. In total, 51 studies were analyzed using varying methods for examination of heritability. They report genetic and environmental variance estimates as a function of different covariates. For this application, their estimates of A, C and E for males were considered the most appropriate background information. Their aggregate estimates across studies were: A = .38, C = .17 and E = .45.

Gamma distributions were used to convey this information. Gamma distributions have been proposed as appropriate prior distributions for variance parameters when prior information is available (Chung et al., 2012). A difficulty arises in parameterizing these distributions to reflect prior knowledge. In this case, they were parameterized with mode equal to the aggregate estimates from the meta-analysis. Ideally, information on between-study variance of the A, C and E components would be available and incorporated as prior uncertainty in the point estimates. However, the author has not been able to locate such information. Therefore, it was reasoned that neither of the variance components was likely to differ more on average than ± .15 on a standardized scale (Analysis was also carried out setting the standard deviations to .10 and .20. This did not lead to any substantial change in parameter estimates.). Based on this decision, gamma distributions with mode equal to the meta-analytic aggregate results and standard deviations equal to .15 were used as prior distributions for the variance components (see figure 2). Because the data analyzed is not on a standard scale, the distributions were scaled according to the total variance in the outcome variable.

From table 2, it can be seen that when prior information is included in the model, the resulting posterior estimates are moved towards the prior distributions, and the difference between these estimates and those obtained from the ML approach increases. σ_E^2 remains almost the same, but σ_A^2 is estimated lower and σ_C^2 higher. The histogram (figure 3) of the C component shows that the mass of the posterior has clearly moved away from zero, peaking around .1. From these results, it seems more appropriate to retain C in the model and thus accept the full ACE model as an adequate description of the data. In many ways this also seems like a less drastic decision, because it seems unreasonable to conclude that there exist no common environmental effects in the population.

DISCUSSION

The current paper has discussed how Bayesian analysis allows prior information to be included in estimation of parameters of a heritability model, and illustrated this with a real data example. The initial conclusion from a conventional analysis of no effect of shared environment, was altered to a conclusion that shared environment contribute to ASPD. Because the purpose of the analysis has been to illustrate the procedure, the analysis was restricted to the ACE model which is likely to be familiar to most readers. However, it is when estimating models of higher complexity the method are likely to be most advantageous. By including information in the prior distributions, parameters that would otherwise be hard to identify might be possible to estimate. Consequently, this allows researchers to investigate hypothesis of higher complexity.

Advantages of Bayesian analysis have been highlighted, specifically the possibility to include prior information in the analysis as well as the relatively straight forward steps of obtaining posterior estimates of functions of parameters, such as standardized variance components. There are however also issues specific to this type of analysis that should be considered.

One of the most controversial issues in Bayesian methods relates to the subjective nature of building prior distributions. Although the prior distributions used here were partially based on results from a large meta-analysis, the spread of the prior distributions was based on personal experience with heritability research. Other researchers might have different preferences, yielding different posterior estimates. This could in turn lead to different conclusions. However, there are always subjective decisions that have to be made when building a statistical model. It is the analysts that decide which covariates to include in the model, how to model multivariate scales, etc. These decisions, as well as chosen priors, always need to be justified.

Although the great flexibility in model specification is one of the advantages of MCMC based Bayesian analysis, it can also work as a disadvantage. More flexibility also means more chances of committing errors, and most software offers little protection against misspecification of models. It is therefore suggested to always validate models by comparing results against those obtained from standard packages, or to simulate datasets where the true parameters are known.

The goal of the current paper has been to illustrate Bayesian estimation and the use of informative priors in heritability research. Because of the large amount of previous research, these methods seems particularly useful in the field of behavioral genetics when considering complicated research questions that typically require large amounts of data.

ACKNOWLEDGMENTS

I would like to thank Line C. Gjerde for help obtaining and preparing the dataset, as well as feedback on the manuscript. Data on zygosity on the twins used in this study was obtained from the Norwegian Twin Registry, the Norwegian Institute of Public Health. I would like to thank the twins for their participation. This work was supported by the Norwegian Research Council – Medicine, Health Sciences and Biology (FRIMEDBIO, No. 231105/F20), the Norwegian Foundation for Health and Rehabilitation, and the European Commission under the program "Quality of Life and Management of the Living Resources" of the 5th Framework Program (No. QLG2-CT-2002-01254).

COMPLIANCE WITH ETHICAL STANDARDS

Approval was received from the Norwegian Data Inspectorate and the regional Ethical Committee, and written informed consent was obtained from all participants after they were given a complete description of the study. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. The author declares no conflict of interest.

Table 2. Summary statistics from Bayesian analysis with informative priors. Estimate is posterior mean and SD is posterior standard deviation. Regression coefficients are the "fixed-effects" in the model. Variance components are the variances of the "random-effects" on the scale of the data and the standardized variance components.

	Estimate	SD
Regression coefficients		
u	.61	.07
Variance components		
σ_A^2	.86	.16
σ_C^2	.23	.12
σ_E^2	1.09	.09
A	.39	
C	.11	
E	.50	

REFERENCES

Chung Y, Rabe-Hesketh S, Gelman A, Liu J, Dorie V (2012). Avoiding boundary estimates in linear mixed models through weakly informative priors. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 284. http://biostats.bepress.com/ucbbiostat/paper284.

Eaves L, Erkanli A (2003). Markov Chain Monte Carlo approaches to analysis of genetic and environmental components of human developmental change and G× E interaction. *Behav Genet* **33**(3):279-299.

Eaves L, Erkanli A, Silberg J, Angold A, Maes HH, Foley D (2005). Application of Bayesian inference using Gibbs sampling to item-response theory modeling of multi-symptom genetic data. *Behav Genet* **35**(6):765-780.

Gelman A (2005). Analysis of variance – why it is more important than ever. *Ann Statist* **33**(1):1-53.

Gelman A (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayes Anal* **1**(3):515-534.

Kruschke J (2010). Doing Bayesian data analysis: A tutorial introduction with R. Academic Press.

Neale M, Cardon L (1992). Methodology for genetic studies of twins and families. Springer Science & Business Media.

Plomin R, DeFries JC, Knopik VS, Neiderheiser J (2013). Behavioral genetics. Palgrave Macmillan.

Polderman TJ, Benyamin B, de Leeuw CA, Sullivan PF, van Bochoven A, Visscher PM, Posthuma D (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat Genet* **47**: 702-709.

Rabe-Hesketh S, Skrondal A, Pickles A (2004). Generalized multilevel structural equation modeling. *Psychometrika* **69**(2):167-190.

Rabe-Hesketh S, Skrondal A, Gjessing HK (2008). Biometrical modeling of twin and family data using standard mixed model software. *Biometrics* **64**(1):280-288.

Rhee SH, Waldman ID (2002). Genetic and environmental influences on antisocial behavior: a meta-analysis of twin and adoption studies. *Psychol Bull* **128**(3):490.

Spiegelhalter D, Thomas A, Best N, Lunn D (2003). WinBUGS user manual.

Stan Development Team (2014a). Stan: A C++ Library for Probability and Sampling, Version 2.5.0. http://mc-stan.org.

Stan Development Team (2014b). RStan: the R interface to Stan, Version 2.5. http://mc-stan.org/rstan.htm.

Torgersen S, Myers J, Reichborn-Kjennerud T, Røysamb E, Kubarych TS, Kendler KS (2012). The heritability of Cluster B personality disorders assessed both by personal interview and questionnaire. *J Pers Disord* **26**(6):848-866.

van den Berg SM, Beem L, Boomsma DI (2006a). Fitting genetic models using Markov Chain Monte Carlo algorithms with BUGS. *Twin Res Hum Genet* **9**(3):334-342.

van den Berg SM, Setiawan A, Bartels M, Polderman TJ, van der Vaart AW, Boomsma DI (2006b). Individual differences in puberty onset in girls: Bayesian estimation of heritabilities and genetic correlations. *Behav Genet* **36**(2):261-270.

van de Schoot R, Depaoli S (2014). Bayesian analyses: Where to start and what to report. *Eur Health Psychol* **16**(2):75-84.

Vehtari A, Gelman A, Gabry J (2015). Efficient implementation of leave-one-out cross-validation and WAIC for evaluating fitted Bayesian models. arXiv preprint arXiv:1507.04544.