Prior distribution

Andrew Gelman

Volume 3, pp 1634–1637

in

Encyclopedia of Environmetrics
(ISBN 0471 899976)

Edited by

Abdel H. El-Shaarawi and Walter W. Piegorsch

Prior distribution

The *prior distribution* is a key part of Bayesian inference (*see* **Bayesian methods and modeling**) and represents the information about an uncertain parameter θ that is combined with the probability distribution of new data to yield the **posterior distribution**, which in turn is used for future inferences and decisions involving θ. The existence of a prior distribution for any problem can be justified by axioms of decision theory; here we focus on how to set up a prior distribution for any given application. In general, θ can and will be a vector, but for simplicity we will focus here on prior distributions for parameters one at a time.

The key issues in setting up a prior distribution are:

- what information is going into the prior distribution;
- the properties of the resulting posterior distribution.

With well-identified parameters and large sample sizes, reasonable choices of prior distributions will have minor effects on posterior inferences. This definition of 'well identified' and 'large' sample size might seem to be circular, but in practice one can check the dependence on prior distributions by a sensitivity analysis: comparing posterior inferences under different reasonable choices of prior distribution (and, for that matter, different reasonable choices of probability models for data).

If the sample size is small, or available data provide only indirect information about the parameters of interest, the prior distribution becomes more important. In many cases, however, models can be set up hierarchically, so that clusters of parameters have shared prior distributions, which can themselves be estimated from data.

Example

We illustrate with an example from a model in pharmacokinetics, the study of the absorption, distribution, and elimination of drugs from the body. For this particular study, about 20 measurements were available on six young adult males, and a model was fit with 15 parameters per person (which we label θ_{kl} for person k and parameter l), along with two variance parameters, σ_1^2 and σ_2^2, indicating the scale of measurement/modeling error. The data (concentrations of a compound in blood and exhaled air over time) are only indirectly informative of the individual level parameters, which refer to equilibrium concentrations, volumes, and metabolic rates inside the body.

This is a nice example to use here because different principles for assigning prior distributions are relevant for different parameters in the model, as we now discuss.

Noninformative Prior Distributions

We first consider the variance parameters σ_1^2 and σ_2^2, which are actually quite well identified in the posterior distribution. For these, a noninformative uniform prior distribution works fine. (A uniform distribution on the log standard deviations was used, but enough information was available from the data that the choice of noninformative prior distribution was essentially irrelevant, and one could just as well have assigned a uniform prior distribution on the variances or the standard deviations.) The uniform prior distribution here is *improper* – that is, the function used as a 'prior probability density' has an infinite integral and is thus not, strictly speaking, a probability density at all. However, when formally combined with the data likelihood it yields an acceptable proper posterior distribution.

Highly Informative Prior Distributions

At the other extreme, fairly precise scientific information is available on some of the parameters θ_{kl} in the model. For example, parameter 8 represents the mass of the liver as a fraction of lean body mass; from previous medical studies, the liver is known to be about 3.3% of lean body mass for young adult males, with little variation. The prior distribution for $\log\theta_{k,8}$ (for persons $k = 1, \ldots, 6$) is assumed normal with mean μ_8 and standard deviation Σ_8; μ_8 was given a normal prior distribution with mean $\log(0.033)$ and standard deviation $\log(1.1)$, and Σ_8 was given an inverse χ^2 prior distribution with scale $\log(1.1)$ and two degrees of freedom. This setup sets the parameters $\theta_{k,8}$ approximately to their prior estimate, 0.033, with some variation allowed between persons.

Moderately Informative Hierarchical Prior Distributions

Finally, some of the physiological parameters θ_{kl} are not well estimated by the data – thus, they require informative prior distributions – but scientific information on them is limited. For example, in this particular study, parameter 14 represents the maximum rate of metabolism of a certain compound; the best available estimate of this parameter for healthy humans is 0.042, but this estimate is quite crude and could easily be off by a factor of 10 or 100. The maximum rate of metabolism is not expected to vary greatly between persons, but there is much uncertainty about the numerical value of the parameter. This information is encoded in a *hierarchical* prior distribution: $\log \theta_{k,14} \sim \mathrm{N}(\mu_{14}, \Sigma_{14}^2)$, with $\mu_{14} \sim \mathrm{N}\{\log(0.042), [\log(10)]^2\}$ and $\Sigma_{14} \sim \mathrm{Inv}\chi^2\{2, [\log(2)]^2\}$. Thus, the parameters $\theta_{k,14}$ for the different persons k are expected to vary by about a factor of 2, with their overall level estimated at about 0.042, with a multiplicative uncertainty of about a factor of 10. To express this subtle statement of prior uncertainty, the hierarchical prior distribution is needed.

What Would Happen if Noninformative Prior Distributions Were Used for All the Parameters in this Example?

In our parameterization, noninformative prior distributions on the parameters θ_{kl} correspond to setting $\Sigma_k = \infty$ for each parameter k, thus allowing each person's parameters to be estimated from that person's own data. If noninformative prior distributions were assigned to all the individual parameters, then the model would fit the data very closely but with scientifically unreasonable parameters – for example, a person with a liver weighing 10 k. This sort of difficulty is what motivates a researcher to specify a prior distribution using external information.

Theory

Most of the theoretical work on prior distributions has been on two topics: first, determining the conditions that must be satisfied by the prior and data distributions so that the posterior distribution is well behaved; and second, setting up rules for noninformative prior distributions that satisfy various invariance principles. These strands of research are related, in that prior distributions set up based on invariance rules alone will make sense only if they lead to reasonable posterior distributions.

Perhaps the most well known theoretical result is that, for variance parameters σ in a linear regression model, the uniform prior distribution for $\log \sigma$ is acceptable when applied to the lowest level variance component but *not* acceptable for higher level variance components. For these, a uniform prior distribution on σ is acceptable, in the sense that the posterior distribution will be proper if sample sizes are moderate or large. By 'not acceptable', we refer to prior distributions that yield improper posterior distributions, even when the data really supply enough information to estimate the parameter accurately. These theoretical results do not give *recommended* models, but rather are useful in *ruling out* certain natural seeming models with poor statistical properties.

Conversely, one might seek to avoid theoretical considerations entirely and simply pick a 'subjective' prior distribution that best represents one's scientific knowledge about the set of uncertain parameters in the problem. In practice, however, subjective knowledge is hard to specify precisely, and so it is important to study the sensitivity of posterior inferences. In many problems, the key issue in setting up the prior distribution is the specification of the model into parameters that can be clustered hierarchically.

Literature

Recent theoretical and applied overviews of Bayesian statistics, including many examples and uses of prior distributions (mostly noninformative), appear in [3], [4] and [7]. See [2] for a review of decision theoretic foundations, and [3] and [8] for full discussions on the theoretical principles for distributions and many references on the topic. The hierarchical prior distribution for the pharmacokinetic example is discussed in [6] and, briefly, in section 18.1 of [7]. Setting up noninformative prior distributions for multivariate models is an important topic of current research; see [1] and [5].

Acknowledgments

Work supported in part by US National Science Foundation Grant SBR-9708424 and Young Investigator Award DMS-9796129.

References

[1] Barnard, J., McCulloch, R. & Meng, X.L. (1997). Modeling covariance matrices in terms of standard deviations and correlations, with applications to shrinkage, Technical Report, Department of Statistics, University of Chicago.

[2] Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd Edition, Springer-Verlag, New York.

[3] Bernardo, J.M. & Smith, A.F.M. (1994). *Bayesian Theory*, Wiley, New York.

[4] Carlin, B.P. & Louis, T.A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman & Hall, London.

[5] Daniels, M.J. & Kass, R.E. (1999). Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models, *Journal of the American Statistical Association* **94**, 1254–1263.

[6] Gelman, A., Bois, F.Y. & Jiang, J. (1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions, *Journal of the American Statistical Association* **91**, 1400–1412.

[7] Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (1995). *Bayesian Data Analysis*, Chapman & Hall, London.

[8] Kass, R.E. & Wasserman, L.A. (1996). The selection of prior distributions by formal rules, *Journal of the American Statistical Association* **91**, 1343–1370.

(*See also* **Modeling, environmental**; **Subjective probability**)

ANDREW GELMAN