The Tension Between Generality and Accuracy

YU XIE
University of Michigan

Statistics is full of trade-offs. In sampling, efficient samples are often expensive to collect. In estimation, robust estimators tend to be inefficient. In data analysis, parsimonious models may not fit observed data. These trade-offs create tensions because we would ideally prefer efficiency and low cost in sampling, robustness and efficiency in estimation, and parsimony and good fit in data analysis.

David L. Weakliem (1999 [this issue]) has done us a service by pointing out two major limitations of the Bayesian information criterion (BIC) developed by Raftery (1986, 1995), which closely resembles the Schwartz (1978) criterion. First, Weakliem shows that Bayes factors depend on priors pertaining to the distribution of unknown parameters, whereas the BIC does not. Second, the BIC takes into consideration only the total sample size but not its distributions across tables, categories, or cells.

Weakliem makes a good case that the BIC has certain limitations. However, this point is neither surprising nor new. *Any* index, test statistic, or criterion has limitations. The key question is whether the limitations are so severe as to render the BIC useless in applied research. My answer to this question is a definite no. In the following, I will elaborate my answer.

AUTHOR'S NOTE: *I wish to thank Mark Becker, Susan Murphy, and Adrian Raftery for helpful suggestions.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 27 No. 3, February 1999 428-435

THE VALUE OF GENERALITY: SIMPLICITY

What Weakliem sees as drawbacks of the BIC are precisely where its virtues lie. The BIC requires neither the input of a prior on the part of the researcher nor adjustment according to frequency distributions. The simplicity with which the BIC can be readily applied to social science research is of high value in practice, for it allows researchers with different theoretical and statistical orientations to communicate research results without the danger of "twisting" statistics to fit one's own preconceptions. This point can be best seen from Weakliem's first criticism, summarized as follows.

Let x denote observed data and B_{01} denote the Bayes factor, which is the ratio in the "likelihood" of observing the data between models M_0 and M_1. By definition,

$$B_{01} = p(x|M_0)/p(x|M_1). \tag{1}$$

Models M_0 and M_1 each imply an unknown parameter vector, denoted as θ_0 and θ_1, respectively. Because θ_0 and θ_1 can have different values in their parameter space, it is necessary to integrate the likelihood function across all possible values, leading to

$$p(x|M_0) = \int p(x|\theta_0, M_0)p(\theta_0|M_0)d\theta_0 \tag{2a}$$

$$p(x|M_1) = \int p(x|\theta_1, M_1)p(\theta_1|M_1)d\theta_1, \tag{2b}$$

where the integration is typically a multiple integral for the entire parameter space.

Weakliem's first criticism is based on an analysis of a 2×2 cross-classified table on anomia and gender from the 1993-94 General Social Survey. Using a log linear model, he focuses on the interaction parameter contained in M_1 that measures the log odds ratio (θ) measuring the association between anomia and gender, with the independence model as M_0. For convenience, he ignores the parameter space for the two marginal parameters. Thus, $p(x|M_0)$ reduces to a single number, the likelihood under M_0. He then assumes a normal distribution with mean of zero and standard deviation of σ as the prior for θ to calculate $p(x|M_1)$. His key result, presented in his Figure 1, is that B_{01}

increases, nearly linearly after a threshold, with σ. This is understandable, since a large σ means that small likelihoods away from ($\theta = 0$) for model M_1 are weighted more heavily, reducing the integrated likelihood $p(x|M_1)$, and thus favoring the null hypothesis.

From this exercise, Weakliem drew the conclusion that the Bayes factor is sensitive to the choice of priors for θ. Different researchers may choose different priors and arrive at different Bayes factors. This is well known and indeed acknowledged by Raftery (1995:129). To applied researchers, however, this kind of flexibility is often a vice rather than a virtue, since it potentially leaves too much room for subjectivity and arbitrariness. In theory, a careful researcher could consider a range of plausible priors and evaluate their plausibility objectively. In practice, researchers would be too tempted to go along with the prior that would yield results in support of their own theories. In Weakliem's example, a researcher may choose a particular value for σ to support his or her theory that the association between gender and anomia exists or does not exist. That is, the "prior" can predetermine the conclusion.

In contrast, the BIC can be interpreted as assuming a particular prior; a multivariate normal distribution with maximum likelihood (ML) estimates as mean and the inverse of the Fisher information matrix as variance (Raftery 1995:132). Note that this prior is determined by data, not by the researcher, and has a nice interpretation that it contains "the same amount of information as would, on average, a single observation" (p. 132). Hence, the BIC provides a unifying framework for all researchers, those with strong priors, those with weak priors, and those with no priors. It makes specification of the prior automatic.

What happens if the applied researcher is unwilling to accept the prior that is implied by the BIC? Evaluation of the Bayes factor becomes very difficult (for a recent review, see Kass and Raftery 1995). Not only is Weakliem's General Social Survey example extremely simple, but Weakliem also fixes the two marginal parameters at their ML estimates. For most applied researchers studying real problems in practice, both (2a) and (2b) are too complicated to calculate. It is clear from Weakliem's article that he does not recommend giving up model selection entirely; nor does he recommend giving up

the Bayes factor as a criterion for model selection. However, it is unclear how applied researchers will benefit from Weakliem's first criticism of the BIC in practice.

THE COST OF GENERALITY: APPROXIMATION

Making the BIC a generally applicable criterion for assessing goodness of fit is not cost free. As is well known and emphasized by Weakliem, the BIC is based on an approximation. One message that Weakliem tries to convey is that the approximation is so crude that the BIC should be avoided or at best modified in practice. Is this claim justified?

To evaluate Weakliem's claim, let us revisit his equation (2) (essentially equation (15) of Raftery 1995), which approximates minus twice the logged integrated likelihood (equation (2)) with (omitting subscript 0 or 1)

$$\begin{aligned} -2\log p(x|M) &= \underset{[1]}{-2\log p(x|\theta^*)} - \underset{[2]}{2\log p(\theta^*|M)} \\ &\quad \underset{[3]}{-k\log(2\pi)} + \underset{[4]}{k\log(n)} + \underset{[5]}{\log(|\mathbf{i}|)} - \underset{[6]}{O(n^{-1/2})}, \end{aligned} \tag{3}$$

where θ^* denotes the ML estimates, k is the number of parameters, n is the sample size, and $\mathbf{i}$ is the expected Fisher information matrix. The numbers in the line below the equation identify the order of the terms. The BIC is based on omitting terms 2, 3, 5, and 6 and changing (3) into

$$-2\log(x|\theta^*) + k\log(n). \tag{4}$$

Weakliem does not have any problem with omitting term 6, since it is of order $n^{-1/2}$ and will go to zero as sample size increases. However, he is unhappy with omitting terms 2, 3, and 5, since they will not diminish as sample size increases. His examples show that, in particular, omitting terms 2 and 5 can be problematic: Different researchers may want more realistic priors for term 2, and frequency distributions can affect term 5 drastically. He demonstrates the relevance of omitting terms 2 and 5 with empirical examples and accordingly draws two criticisms of the BIC.

Although Weakliem's empirical examples are thought provoking, it is not clear that his conclusions hold true asymptotically. He overlooks the fact that equation (3) is dominated by terms 1 and 4 in the sense that they go to infinity, whereas terms 2, 3, and 5 do not, as sample size increases. In this sense, his observation that the value of equation (3) (my (4)) will converge to the true value in addition to a constant belies the fact that it does not converge to anything. Will the differences due to omitting terms 2 and 5 diminish as sample size increases? Weakliem claims that they will not, but offers no proof. From equation (3), it appears that the *relative* importance of these terms will diminish as sample size increases. Further investigation into this matter is needed.

It should be emphasized that the approximation of equation (4) is not used by itself, but rather is used to calculate the Bayes factor of equation (1). The amount of error should be smaller as a result. To see this, let us combine equations (1) through (3):

$$\begin{aligned} -2\log(B_{01}) &= -2\log[p(x|M_0)] + 2\log[p(x|M_1)] \\ &= 2\log[p(x|\theta_1^*)/p(x|\theta_0^*)] + 2\log[p(\theta_1^*)/p(\theta_0^*)] + (k_1 - k_0)\log(2\pi) \\ &\quad -(k_1 - k_0)\log(n) - \log(|\mathbf{i}_1|/|\mathbf{i}_0|) + O(n^{-1/2}), \end{aligned} \tag{5}$$

where k_0 and k_1 denote the number of parameters, respectively, for model M_0 and M_1, and $\mathbf{i}_0$ and $\mathbf{i}_1$ similarly denote the expected Fisher information matrix for M_0 and M_1. Thus, the approximation for the BIC requires that the relative importance of $\log[p(\theta_1^*)/p(\theta_0^*)]$, $(k_1 - k_0)$, and $\log(|\mathbf{i}_1|/|\mathbf{i}_0|)$ diminishes. In typical situations, these quantities measuring differences between two models are smaller than their absolute values in a single model. The difference in degrees of freedom, $(k_1 - k_0)$, for example, is usually a much smaller number than either k_1 or k_0.

Weakliem's second major criticism of the BIC stems from his observation that a large sample with a skewed distribution of frequencies may contribute little information pertaining to hypothesis testing, whereas the BIC only adjusts for the total sample size. He interprets his criticism in light of the fact that term 5 ($\log(|\mathbf{i}|)$) in equation (3) is omitted for the BIC's approximation. To correct for this omission, Weakliem proposes a modified BIC measure, MBIC, that effectively adjusts sample size downward due to an uneven distribution of frequencies.

Although Weakliem's second criticism has some validity, two points can be raised in defense of the BIC. First, it is not true that frequency distribution does not enter the BIC, for it does through the ratio in likelihood, $p(x|\theta_1^*) / p(x|\theta_0^*)$, in equation (5). The less information in the data, the closer to one the ratio in likelihood. Second, it is not the level of $|\mathbf{i}|$ that matters; rather, what matters is the ratio in $|\mathbf{i}|$ between two models, as shown in equation (5). Because the effect of frequency distribution on $|\mathbf{i}_0|$ is likely to be similar to that on $|\mathbf{i}_1|$, it seems to me that Weakliem's adjustment for MBIC penalizes uneven distributions too severely.

In brief, I agree with Weakliem that the BIC is based on a crude approximation but recognize the necessity for its general use. Furthermore, it appears that Weakliem has exaggerated the inaccuracies of the BIC as a result of the approximation. We need more work on the subject before knowing for sure the full consequences of the approximation cost of the BIC.

THE PRACTICE OF GENERALITY: TRIANGULATION

I wonder if Weakliem's real aim is to urge researchers not to rely exclusively on the BIC as the sole criterion for model selection. If it is, he has succeeded, even before his critique is published. Applied researchers have for a long time used a variety of criteria in assessing models' goodness of fit. They include, among other tools, the likelihood ratio chi-square statistic (L^2), the Pearson chi-square statistic (X^2), L^2 to degrees of freedom, and the index of dissimilarity (Δ). I do not know of a single researcher who blindly applies the BIC when selecting models.

For example, in my own earlier work (Xie and Pimentel 1992), I used the following five criteria to compare our revised model for age patterns of fertility to the traditional Coale-Trussell model: L^2, X^2, Δ, the BIC, and a prediction exercise. We did not rely on the BIC exclusively for model selection. Instead, we wanted to see if the BIC would give a different conclusion than those according to the other criteria. It did not. When all the criteria yielded the same conclusion that our revised model was superior to the original model, it gave us more

confidence in the conclusion. Triangulation of this kind is often a necessary part of doing empirical research.

Weakliem suggests in his critique that researchers delete irrelevant cells in calculating n for the BIC. This sounds like good advice. In fact, I have independently come to the same realization in my recent work (Lin and Xie 1998). In log-rate models presented in Xie and Pimentel (1992) and Xie (1994), I used actual events rather than exposure in calculating n.

In reanalyzing the 16-nation data on intergenerational mobility, Weakliem has revealed some interesting features of the data. He then asserts that there is asymmetry in the data but that investigators who used the BIC did not notice it. This is at best an inaccurate characterization of the literature. Grusky and Hauser (1984) did not notice it when the BIC was not even around. My own reanalysis of the same data (Xie 1992) clearly considered asymmetry, although I did use the BIC to test models that constrain the variation in two-way association parameters across layers.

CONCLUSION

Weakliem's critique of the BIC only highlights the difficulty of model selection in applied social science research. On one hand, we desire accuracy and do not favor a criterion that is contaminated by large approximation errors. On the other hand, we would like to have criteria that are generally applicable, easy to implement, and free from researchers' subjectivity and arbitrariness. Striking a balance between the two is not easy. The merits and the drawbacks of the BIC are just two sides of the same coin.

Seen in this light, the limitations of the BIC that are the focus of Weakliem's article should not surprise us, nor should they persuade us to give up on the BIC. Like many other methods in statistics, the BIC is a double-edged sword but a useful sword nonetheless. I encourage researchers to continue its use as one of many valuable tools for model selection, while recognizing its potential drawbacks and their consequences.

REFERENCES

Grusky, David B. and Robert M. Hauser. 1984. "Comparative Social Mobility Revisited: Models of Convergence and Divergence in 16 Countries." *American Sociological Review* 49:19-38.

Kass, Robert E. and Adrian E. Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90:773-95.

Lin, Ge and Yu Xie. 1998. "Some Additional Considerations of Loglinear Modeling of Interstate Migration: A Comment on Herting, Grusky, and Rompaey." *American Sociological Review* 63:900-07.

Raftery, Adrian E. 1986. "Choosing Models for Cross-Classifications." *American Sociological Review* 51:145-46.

———. 1995. "Bayesian Model Selection in Social Research." Pp. 111-63 in *Sociological Methodology*, edited by Peter Marsden. Washington, DC: American Sociological Association.

Schwartz, G. 1978. "Estimating the Dimension of a Model." *Annals of Statistics* 6:461-64.

Weakliem, David L. 1999. "A Critique of the Bayesian Information Criterion for Model Selection." *Sociological Methods & Research* 27:359-97.

Xie, Yu. 1992. "The Log-Multiplicative Layer Effect Model for Comparing Mobility Tables." *American Sociological Review* 57:380-95.

———. 1994. "Log-Multiplicative Models for Discrete-Time, Discrete-Covariate Event History Data." Pp. 301-40 in *Sociological Methodology*, edited by Peter Marsden. Washington, DC: American Sociological Association.

Xie, Yu and Ellen Efron Pimentel. 1992. "Age Patterns of Marital Fertility: Revising the Coale-Trussell Method." *Journal of the American Statistical Association* 87:977-84.

Yu Xie is John Stephenson Perrin Professor of Sociology at the University of Michigan. He is also affiliated with the Population Studies Center and the Survey Research Center of the Institute for Social Research. His main areas of interest are social stratification, demography, statistical methods, and sociology of science. He is currently completing a book in collaboration with Kimberlee Shauman that studies the career processes and outcomes of women in science.