What to Do When Your Hessian Is Not Invertible

Alternatives to Model Respecification in Nonlinear Estimation

JEFF GILL
University of California, Davis

GARY KING
Harvard University

What should a researcher do when statistical analysis software terminates before completion with a message that the Hessian is not invertible? The standard textbook advice is to respecify the model, but this is another way of saying that the researcher should change the question being asked. Obviously, however, computer programs should not be in the business of deciding what questions are worthy of study. Although noninvertable Hessians are sometimes signals of poorly posed questions, nonsensical models, or inappropriate estimators, they also frequently occur when information about the quantities of interest exists in the data through the likelihood function. The authors explain the problem in some detail and lay out two preliminary proposals for ways of dealing with noninvertable Hessians without changing the question asked.

Keywords: *Hessian; generalize inverse; importance sampling; generalized Cholesky*

INTRODUCTION

In social science applications of nonlinear statistical models, researchers typically assume the accuracy of the asymptotic normal approximation to the likelihood function or posterior distribution. For maximum likelihood analyses, only point estimates and the variance

AUTHORS' NOTE: *Our thanks to Mike Alvarez, Neal Beck, Federico Girosi, Andrew Gelman, and Langche Zeng for helpful comments, as well as the National Science Foundation (IIS-9874747), the National Institutes of Aging (P01 AG17625-01), and the World Health Organization for research support. Software, written in Gauss and Splus/R, to implement the methods in this article, is available from our Web sites: http://psfaculty.ucdavis.edu/jgill/ and http://GKing.Harvard.Edu.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 33, No. 1, August 2004 54-87
DOI: 10.1177/0049124103262681

at the maximum are normally seen as necessary. For Bayesian posterior analysis, the maximum and variance provide a useful first approximation.

Unfortunately, although the negative of the Hessian (the matrix of second derivatives of the posterior with respect to the parameters and named for its inventor, German mathematician Ludwig Hesse) must be positive definite and hence invertible to compute the variance matrix, invertible Hessians do not exist for some combinations of data sets and models, and so statistical procedures sometimes fail for this reason before completion. Indeed, receiving a computer-generated "Hessian not invertible" message (because of singularity or nonpositive definiteness) rather than a set of statistical results is a frustrating but common occurrence in applied quantitative research. It even occurs with regularity during many Monte Carlo experiments when the investigator is drawing data from a known statistical model.

When a Hessian is not invertible, no computational trick can make it invertible, given the model and data chosen, since the desired inverse does not exist. The advice given in most textbooks for this situation is to rethink the model, respecify it, and rerun the analysis (or, in some cases, get more data). This is important and appropriate advice in some applications of linear regression since a noninvertible Hessian has a clear substantive interpretation: It can only be caused by multicollinearity or by including more explanatory variables than observations (although even this simple case can be quite complicated; see Searle 1971). As such, a noninvertible Hessian might indicate a substantive problem that a researcher would not otherwise be aware of. It is also of interest in some nonlinear models, such as logistic regression, in which the conditions of noninvertability are also well known. In nonlinear models, however, noninvertible Hessians are related to the shape of the posterior density, but how to connect the problem to the question being analyzed can often be extremely difficult.

In addition, for some applications, the textbook advice is disconcerting, or even misleading, since the same model specification may have worked in other contexts and really is the one the researcher wants estimates from. Furthermore, one may find it troubling that dropping variables from the specification substantially affects the estimates of the remaining variables and therefore the interpretation of the findings (Leamer 1973). Our point is that although a noninvertible

Hessian means the desired variance matrix does not exist, the likelihood function may still contain considerable information about the questions of interest. As such, discarding data and analyses with this valuable information, even if the information cannot be summarized as usual, is an inefficient and potentially biased procedure.[1] In situations when one is running many parallel analyses (say one for each U.S. state or population subgroup), dropping only those cases with noninvertible Hessians, as is commonly done, can easily generate selection bias in the conclusions drawn from the set of analyses. And restricting all analyses to the specification that always returns an invertible Hessian risks other biases. Similarly, Monte Carlo studies that evaluate estimators risk severe bias if conclusions are based (as usual) on only those iterations with invertible Hessians.

Rather than discarding information or changing the questions of interest when the Hessian does not invert, we discuss some methods that are sometimes able to extract information in a convenient format from problematic likelihood functions or posterior distributions without respecification.[2] This has always been possible within Bayesian analysis by using algorithms that enable one to draw directly from the posterior of interest. However, the algorithms, such as those based on Monte Carlo Markov chains or higher order analytical integrals, are often more difficult to use than point estimates and asymptotic variance approximations to which social scientists have become accustomed, and so they have not been widely adopted.

Our goal in this work has been to develop an easy-to-use, off-the-shelf method for dealing with nonivertable Hessians—one that can be used by the vast majority of social scientists in real applied empirical research. We believe we have made progress, but we do not believe we are there yet since the methods we offer involve more than the usual degree of care in application, and some types of models and noninvertable Hessians will not be ameniable to the methods we introduce. As such, although the intended audience for a more mature version of the methods we introduce here is applied social scientists, we presently hope to appeal primarily to other methodologists who we hope to spur on to achieve the goal we set out. To facilitate the path for others who might also wish to get involved in this pursuit, we provide as clear as possible an introduction to basic issues involved as well.

Since the vast majority of social science is done in the frequentist or likelihood frameworks, we try to provide methods and answers reasonably close to what they would want. Thus, although our approach is Bayesian, we stick where possible to noninformative, flat priors. We therefore provide credible intervals instead of confidence intervals, as well as posterior variances instead sampling distribution variances, but the quantities we intend to compute should nevertheless be reasonably comfortable for most social scientists. In addition, all the methods we discuss are appropriate when the Hessian actually does invert and, in many cases, may be more appropriate than classical approaches in those circumstances. We begin in Section 2 by providing a summary of the posterior that can be calculated, even when the mode is uninteresting and the variance matrix nonexistent. The road map to the rest of the article concludes that motivating section.

MEANS VERSUS MODES

When a posterior distribution contains information but the variance matrix cannot be computed, all hope is not lost. In low-dimensional problems, plotting the posterior is an obvious solution that can reveal all relevant information. In a good case, this plot might reveal a narrow plateau around the maximum or collinearity between two relatively unimportant control variables (as represented by a ridge in the posterior surface). Unfortunately, most social science applications have enough parameters to make this type of visualization infeasible, and so some summary is needed (indeed, this was the purpose of maximum likelihood estimates, as opposed to the better justified likelihood theory of inference, in the first place; see King 1998).

We propose an alternative strategy. We do not follow the textbook advice by asking the user to change the *substantive question* they ask but instead ask researchers to change their *statistical summary* of the posterior so that useful information can still be elicited without changing their substantive questions, statistical specification, assumptions, data, or model. All available information from the specified model can thus be extracted and presented, at which point one may wish to stop or instead respecify the model on the basis of substantive results.

In statistical analyses, researchers collect data, specify a model, and form the posterior. They then summarize this information, essentially by posing a question about the posterior distribution. The question answered by the standard maximum likelihood (or maximum posterior) estimates is, “What is the mode of the posterior density and the variance around the mode?” In cases when the mode is on a plateau or at a boundary constraint, or the posterior’s surface has ridges or saddlepoints, the curvature will produce a noninvertible Hessian. In these cases, the Hessian also suggests that the mode itself may not be of use, even if a reasonable estimate of its variability were known. That is, when the Hessian is noninvertible, the mode may not be unique and is, in any event, not an effective summary of the full posterior distribution. In these difficult cases, we suggest that researchers pose a different but closely related question: “What is the mean of the posterior density and the variance around the mean?”

When the mode and mean are both calculable, they often give similar answers. If the likelihood is symmetric, which is guaranteed if n is sufficiently large, the two are identical, and so switching questions has no cost. Indeed, the vast majority of social science applications appeal to asymptotic normal approximations for computing the standard errors and other uncertainty estimates, and for these, the mode and the mean are equal. As such, for these analyses, our proposals involve no change of assumptions.

If the maximum is not unique or is on a ridge or at the boundary of the parameter space, then the mean and its variance can be found, but a unique mode and its variance cannot. At least in these hard cases, when the textbook suggestion of substantive respecification is not feasible or desirable, we propose switching from the mode to the mean.

Using the mean and its variance seems obviously useful when the mode or its variance does not exist, but in many cases when the two approaches differ and both exist, the mean would be preferred to the mode. For an extreme case, suppose the posterior for a parameter θ is truncated normal with mean 0.5, standard deviation 10, and truncation on the [0, 1] interval (cf. Gelman, Carlin, Stern, and Rubin 1995:114, problem 4.8). In this case, the posterior, estimated from a sample of data, will be a small segment of the normal curve. Except when the unit interval captures the mode of the normal posterior (very unlikely given the size of the variance), the mode will almost always be a corner solution (0 or 1). The mean posterior, in contrast, will be some number within (0,1). In this case, it seems clear that 0 or 1 does not make good single-number summaries of the posterior, whereas the mean is likely to be much better.

In contrast, when the mean is not a good summary, the mode is usually not satisfactory either. For example, the mean will not be very helpful when the likelihood provides little information at all, in which case the result will effectively return the prior. The mean will also not be a very useful summary for a bimodal posterior since the point estimate would fall between the two humps in an area of low density. The mode would not be much better in this situation, although it does at least reasonably characterize one part of the density.[3]

In general, when a point estimate makes sense, the mode is easier to compute, but the mean is more likely to be a useful summary of the full posterior. We believe that if the mean were as easy to compute as the mode, few would choose the mode. Thus, we hope to reduce the computational advantage of the mode over the mean by proposing some procedures for computing the mean and its variance.

When the inverse of the negative Hessian exists, we compute the mean and its variance by importance resampling. That is, we take random draws from the exact posterior in two steps. We begin by drawing a large number of random numbers from a normal distribution, with mean set at the vector of maximum posterior estimates and variance set at the estimated variance matrix. Then we use a probabilistic rejection algorithm to keep only those draws that are close enough to the correct posterior. These draws can then be used directly to study some quantity of interest, or they can be used to compute the mean and its variance.

When the inverse of the negative Hessian does not exist, we suggest two separate procedures to choose from. One is to create a *pseudo-variance matrix* and use it, in place of the inverse, in our importance resampling scheme. In brief, applying a generalized inverse (when necessary, to avoid singularity) and generalized Cholesky decomposition (when necessary, to guarantee positive definiteness) together often produces a pseudo-variance matrix for the mode that is a reasonable summary of the curvature of the posterior distribution. (The generalized inverse is a commonly used technique in statistical analysis, but the generalized Cholesky has not before been used for statistical purposes, to our knowledge.) Surprisingly, the resulting matrix is not usually ill conditioned. In addition, although this is a "pseudo" rather than an "approximate" variance matrix (since the matrix that would be approximated does not exist), the calculations change the resulting variance matrix as little as possible to achieve positive definiteness. We then take random draws from the exact posterior using importance resampling as before but using two diagnostics to correct problems with this procedure.[4]

A second proposal introduces a way to draw random numbers directly from a singular (truncated) multivariate normal density and to use these numbers in an importance sampling stage to draw from the exact posterior.

We now first describe in substantive terms what is "wrong" with a Hessian that is noninvertible (Section 3), describe how we create a pseudo-variance matrix (in Section 4, with algorithmic details and numerical examples in Appendix 10.1), and outline the concept of importance resampling to compute the mean and variance (in Section 5). We give our preliminary alternative procedure in Section 7, as well as an empirical example (Section 6) and other possible approaches (in Sections 7 and 8). Section 9 concludes.

WHAT IS A NONINVERTIBLE HESSIAN?

In this section, we describe the Hessian and problems with it in intuitive statistical terms. Given a joint probability density $f(y|\boldsymbol{\theta})$, for an $n \times 1$ observed data vector y and unknown $p \times 1$ parameter vector $\boldsymbol{\theta}$, denote the $n \times p$ matrix of first derivatives with respect to $\boldsymbol{\theta}$ as $g(\boldsymbol{\theta}|y) = \partial \ln[f(y|\boldsymbol{\theta})]/\partial\boldsymbol{\theta}$ and the $p \times p$ matrix of second derivatives as $\mathbf{H} = \mathbf{H}(\boldsymbol{\theta}|y) = \partial^2 \ln[f(y|\boldsymbol{\theta})]/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$. Then the Hessian matrix is $\mathbf{H}$, normally considered to be an estimate of $E[g(\boldsymbol{\theta}|y)g(\boldsymbol{\theta}|y)'] = E[\mathbf{H}(\boldsymbol{\theta}|y)]$. The maximum likelihood or maximum posterior estimate, which we denote $\hat{\boldsymbol{\theta}}$, is obtained by setting $g(\boldsymbol{\theta}|y)$ equal to zero and solving analytically or numerically. When $-\mathbf{H}$ is positive definite in the neighborhood of $\hat{\boldsymbol{\theta}}$, the theory is well known, and no problems arise in application.

The problem described as "a noninvertible Hessian" can be decomposed into two distinct parts. The first problem is *singularity*, which means that $(-\mathbf{H})^{-1}$ does not exist. The second is *nonpositive definiteness*, which means that $(-\mathbf{H})^{-1}$ may exist, but its contents do not make sense as a variance matrix. (A matrix that is positive definite is nonsingular, but nonsingularity does not imply positive definiteness.) Statistical software normally describes both problems as "noninvertibility" since their inversion algorithms take computational advantage of the fact that the negative of the Hessian must be positive definite if the result is to be a variance matrix. This means that these programs do not bother to invert nonsingular matrices (or even to check whether they are nonsingular) unless it is established first that they are also positive definite.

We first describe these two problems in single-parameter situations, in which the intuition is clearest but our approach does not add much of value (because the full posterior can easily be visualized). We then move to more typical multiple-parameter problems, which are more complicated but we can help more. In one dimension, the Hessian is a single number measuring the degree to which the posterior curves downward on either side of the maximum. When all is well, $\mathbf{H} < 0$, which indicates that the mode is indeed at the top of the hill. The variance is then the reciprocal of the negative of this degree of curvature, $-1/\mathbf{H}$, which of course is a positive number as a variance must be.

The first problem, singularity, occurs in the one-dimensional case when the posterior is flat near the mode—so that the posterior forms a plateau at best or a flat line over $(-\infty, \infty)$ at worst. Thus, the curvature is zero at the mode, and the variance does not exist since $1/0$ is not defined. Intuitively, this is as it should be since a flat likelihood indicates the absence of information, in which case any point estimate is associated with an (essentially) infinite variance (to be more precise, $1/\mathbf{H} \rightarrow \infty$ as $\mathbf{H} \rightarrow 0$).

The second problem occurs when the "mode" identified by the maximization algorithm is at the bottom of a valley instead of at the top of a hill ($g(\boldsymbol{\theta}|y)$ is zero in both cases), in which case the curvature will be positive. (This is unlikely in one dimension, except for seriously defective maximization algorithms, but the corresponding problem in high-dimensional cases of "saddlepoints," in which the top of the hill for some parameters may be the bottom for others, is more common.)

The difficulty here is that $-1/\mathbf{H}$ exists but is negative (or, in other words, is not positive definite), which obviously makes no sense as a variance.

A multidimensional variance matrix is composed of variances, which are the diagonal elements and must be positive, and correlations, which are off-diagonal elements divided by the square root of the corresponding diagonal elements. Correlations must fall within the $[-1, 1]$ interval. Although invertibility is an either/or question, it may be that information about the variance or covariances exists for some of the parameters but not for others.

In the multidimensional case, singularity occurs whenever the elements of $\mathbf{H}$ that would map to elements on the diagonal of the variance matrix, $(-\mathbf{H})^{-1}$, combine in such a way that the calculation cannot be completed (because they would involve divisions by zero). Intuitively, singularity indicates that the variances to be calculated would be (essentially) infinite. When $(-\mathbf{H})^{-1}$ exists, it is a valid variance matrix only if the result is positive definite. Nonpositive definiteness occurs in simple cases either because the variance is negative or the correlations are exactly -1 or 1.[5]

CREATING A PSEUDO-VARIANCE MATRIX

Below, we use a generalized inverse procedure to address singularity in the $-\mathbf{H}$ matrix. The classic inverse $\mathbf{A}^{-1}$ of $\mathbf{A}$ can be defined as meeting five intuitive conditions: (1) $\mathbf{HA}^{-1}\mathbf{A} = \mathbf{H}$, (2) $\mathbf{A}^{-1}\mathbf{AA}^{-1} = \mathbf{A}^{-1}$, (3) $(\mathbf{AA}^{-1})' = \mathbf{A}^{-1}\mathbf{A}$, (4) $(\mathbf{A}^{-1}\mathbf{A}) = \mathbf{AA}^{-1}$, and (5) $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ (with 1-4 implied by 5). In contrast, the Moore-Penrose generalized inverse matrix $\mathbf{A}^{-}$ of $\mathbf{A}$ meets only the first four conditions (Moore 1920; Penrose 1955). We also apply a generalized Cholesky (detailed in Appendix 10.1) to address cases in which the inverse or generalized inverse of $(-\mathbf{H})$ is not positive definite. The generalized inverse is primarily changing the parts of $-\mathbf{H}$ that get mapped to the variances (so they are not infinities), and the generalized Cholesky adjusts what would get mapped to the correlations (by slightly increasing variances in their denominator) to keep them within $(-1, 1)$. More specifically, our pseudo-variance matrix is calculated as $\mathbf{V}'\mathbf{V}$, where $\mathbf{V} = \text{GCHOL}(\mathbf{H}^{-})$, $\text{GCHOL}(\cdot)$ is the generalized Cholesky, and

$\mathbf{H}^-$ is the generalized inverse of the Hessian matrix. The result is a pseudo-variance matrix that is in most cases well conditioned (i.e., not nearly singular).[6] If the Hessian is invertible, the pseudo-variance matrix is the usual inverse of the negative Hessian.

IMPORTANCE RESAMPLING

"Sampling importance resampling" (SIR), or simply "importance resampling," is a simulation technique used to draw random numbers directly from an exact (finite sample) posterior distribution.[7] The chief requirement for a successful implementation of importance resampling is simulations from a distribution that is a reasonable approximation to the exact posterior. If this requirement is not met, the procedure can take too long to be practical or can miss features of the posterior. The approximating distribution is required but need not be normalized.

In our case (and indeed in most cases), we use the multivariate normal distribution as our approximation in general or the multivariate t distribution when the sample size is small. Using the normal or t should be relatively uncontroversial since our proposal is addressed to applications for which the asymptotic normal approximation was assumed appropriate from the start, and for most applications, it probably would have worked except for the failed variance matrix calculation. This first approximation thus retains as many of the assumptions of the original model as possible. Other distributions can easily be used if that seems necessary.

For either the normal or t distributions, we set the mean at $\hat{\boldsymbol{\theta}}$, the vector of maximum likelihood or maximum posterior estimates (i.e., the vector of point estimates reported by the computer program that failed when it got to the variance calculation). For the normal, we set the variance equal to our pseudo-variance matrix. For the t, the pseudo-variance is adjusted by the degrees of freedom to yield the scatter matrix.

The idea of importance resampling is to draw a large number of simulations from the approximation distribution, decide how close each is to the target posterior distribution, and keep those close with higher probability than those farther away. To be more precise, denote $\tilde{\boldsymbol{\theta}}$ as one random draw of $\boldsymbol{\theta}$ from the approximating normal distribution and use it to compute the importance ratio, the ratio of the posterior $P(\cdot)$ to the normal approximation, both evaluated at $\tilde{\boldsymbol{\theta}} : P(\tilde{\boldsymbol{\theta}}|y)/N(\tilde{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}}, \mathbf{V}'\mathbf{V})$. We then keep $\tilde{\boldsymbol{\theta}}$ as a random draw from the posterior with probability proportional to this ratio. The procedure is repeated until a sufficiently large number of simulations have been accepted.

The simulations can be displayed with a histogram to give the full marginal distribution of a quantity interest (see Tanner 1996; King, Tomz, and Wittenberg 2000) or parameter of the model. By taking the sample average and sample standard deviation of the simulations, they can also be used to compute the mean and standard error or full-variance matrix of the parameters, if these kinds of more parsimonious summaries are desired. A computed variance matrix of the means will normally be positive definite, so long as more simulations are drawn than there are elements of the mean vector and variance matrix (and normally, one would want at least 10 times that number). It is possible, of course, that the resulting variance matrix will be singular, even when based on many simulations, if the likelihood or posterior contains exact dependencies among the parameters. But in this case, singularity in the variance matrix (as opposed to the Hessian) poses no problem since all that will happen is that some of the correlations will be exactly 1 or -1, which can be very informative substantively; standard errors, for example, will still be available.

One diagnostic often used to detect a failure of importance resampling is when too many candidate values of $\tilde{\boldsymbol{\theta}}$ are rejected, in which case the procedure will take an unreasonably long time; to be useful, a better approximation would be needed. That is, a long runtime indicates a problem, but letting it run longer is a reasonable solution from a statistical perspective, although not necessarily from a practical one. The danger of relying on only this procedure occurs when the approximation distribution entirely misses a range of values of $\boldsymbol{\theta}$ that have a posterior density systematically different from the rest. Since the normal has support over $(-\infty, \infty)$, the potential for this problem to occur vanishes as the number of simulations grows. Thus, one check would be to compute a very large number of simulations (since the coverage will be greater) with an artificially large variance matrix, such as the pseudo-variance matrix multiplied by a positive factor, which we label F. Of course, it is impossible to cover the continuum of values that $\boldsymbol{\theta}$ can take, and the procedure can therefore miss features such as pinholes in the surface, very sharp ridges, or other eccentricities.

Moreover, this procedure cannot be completely relied on in our case since we know that the likelihood surface is nonstandard to some degree. The normal approximation, after all, requires an invertible Hessian. The key to extracting at least some information from the Hessian via our pseudo-variance matrix is determining whether the problems are localized or instead affect all the parameters. If they are localized, or the problem can be reparameterized so they are localized, then some parameters effectively have infinite standard errors or pairs of parameters have perfect correlations. Our suggestion is to perform two diagnostics to detect these problems and to alter the reported standard errors or covariances accordingly.

For small numbers of parameters, using cross-sectional likelihood plots of the posterior can be helpful, and trying to isolate the noninvertibility problem in a distinct set of parameters can be very valuable in trying to understand the problem. This, of course, is not always possible.

To make the normal or t approximation work better, it is advisable to reparameterize so that the parameters are unbounded and approximately symmetric. (This strategy will also normally make the maximization routine work better.) For example, instead of estimating $\sigma^2 > 0$ as a variance parameter directly, it would be better to estimate γ, where $\sigma^2 = e^\gamma$, since γ can take on any real number. It is also helpful to rescale the problem so that the estimated parameters are approximately of the same size.

EMPIRICAL EXAMPLE

We now discuss an illustration of the methods discussed, using public policy data on poverty and its potential causes, measured by the state at the county level (Federal Information Processing Standard [FIPS]). This example highlights a common and disturbing problem in empirical model building. Suppose a researcher seeks to apply a statistical model specification to multiple data sets for the purpose of comparison. Examples might come from comparing models across 50 U.S. states, 18 Organization for Economic Cooperation and Development (OECD) countries, 12 Commonwealth of Independent States (CIS) countries, or even the same unit across some time series. If the Hessian fails to invert in even a small proportion of the cases, then generally the researcher is forced to respecify the model for nonsubstantive, technical reasons. Either the researcher respecifies only the problem cases, in which case differences among the results are contaminated by investigator-induced omitted variable bias, or all equations are respecified in an effort to get comparable results, in which case the statistical analyses are not of the substantive question posed. Neither approach is satisfactory from a substantive perspective.

Our example data are 1989 county-level economic and demographic data for all 2,276 nonmetropolitan U.S. counties ("ERS Typology") hierarchically organized by state such that each state is a separate unit of analysis with counties as cases. The government (U.S. Bureau of the Census, U.S. Department of Agriculture [USDA], state agencies) collects these data to provide policy-oriented information about conditions leading to high levels of rural poverty. The dichotomous outcome variable indicates whether 20 percent or more of the county's residents live in poverty. Our specification includes the following explanatory variables: `Govt`, a dichotomous factor indicating whether government activities contributed a weighted annual average of 25 percent or more labor and proprietor income over the three previous years; `Service`, a dichotomous factor indicating whether service activities contributed a weighted annual average of 50 percent or more labor and proprietor income over the three previous years; `Federal`, a dichotomous factor indicating whether federally owned lands make up 30 percent or more of a county's land area; `Transfer`, a dichotomous factor indicating whether income from transfer payments (federal, state, and local) contributed a weighted annual average of 25 percent or more of total personal income over the past three years; `Population`, the log of the county population total for 1989; `Black`, the proportion of black residents in the county; and `Latino`, the proportion of Latino residents in the county.

Our key substantive question is whether the fraction of black predicts poverty levels, even after controlling for governmental efforts and the other control variables. Since the government supposedly has a lot to do with poverty levels, it is important to know whether it is succeeding in a racially fair manner or whether there is more poverty in counties with larger fractions of African Americans, after controlling for the other measured factors. (Whether the effect, if it exists, is due to more blacks being in poverty or more whites and blacks in heavily black counties being in poverty would be interesting to know but is not material for our substantive purposes.)

We analyze these data with a logistic regression model, and so $P(Y_i = 1|X) = [1 + \exp(-X_i\beta)]^{-1}$, where X_i is a vector of all our explanatory variables. Using this specification, 43 of the states produce invertible Hessians and therefore readily available results. Rather than alter our theory and search for a new specification driven by numerical and computational considerations, we apply our approach to the remaining 7 state models. From this 43/7 dichotomy, we choose a matched pair of similar states for discussion: one case with a (barely) invertible Hessian with the model specification (Texas) and the other a noninvertible (Florida) case. These states both have large rural areas, similar demographics, and similar levels of government involvement in the local county economies, and we would like to know whether fraction of black predicts poverty in the same way in each.

The logit model for Texas counties ($n = 196$) produces the results in the first pair of columns in Table 1. The coefficient on fraction black is very large, with a relatively small standard error, clearly supporting the racial bias hypothesis. The second pair of columns reestimates the model without the `Federal` variable (the source of the noninvertibility in the Florida example, which we describe below), and the results for fraction black (and the other variables) are largely unchanged. In contrast to the *modes* and their standard deviations in the first two sets of results, the final pair of columns gives the *means* and their standard deviations by implementing our importance resampling (but obviously without the need for a pseudo-variance matrix). In Texas, the means are very close to the modes, and the standard errors in the two cases are very close as well, and so the importance resampling in this (invertible) case did not generate any important differences.

We also give the Hessian matrix from this problem here, which reveals that the `Federal` variable is a potentially problematic

TABLE 1: Logit Regression Model: Nonsingular Hessian, Texas

	Standard Results		*Minus Federal*		*Importance Resampling*	
Parameter	*Estimate*	*Standard Error*	*Estimate*	*Standard Error*	*Estimate*	*Standard Error*
Black	15.91	3.70	16.04	3.69	15.99	3.83
Latino	8.66	1.48	8.73	1.48	8.46	1.64
Govt	1.16	0.78	1.16	0.78	1.18	0.74
Service	0.17	0.62	0.20	0.63	0.19	0.56
Federal	−5.78	16.20	—	—	−3.41	17.19
Transfer	1.29	0.71	1.17	0.69	1.25	0.63
Population	−0.39	0.22	−0.39	0.22	−0.38	0.21
Intercept	−0.47	1.83	−0.46	1.85	−0.51	1.68

component of the model. Note the zeros and very small values in the fourth row and column.

$$\mathbf{H} = \begin{bmatrix} 0.13907100 & 0.00971597 & 0.01565632 & 0.00000000 & 0.01165964 & 1.27113747 & 0.01021141 & 0.03364064 \\ 0.00971597 & 0.00971643 & 0.00000000 & 0.00000000 & 0.00022741 & 0.09510282 & 0.00128841 & 0.00211645 \\ 0.01565632 & 0.00000000 & 0.01594209 & 0.00000000 & 0.00305369 & 0.14976776 & 0.00170421 & 0.00246767 \\ 0.00000000 & 0.00000000 & 0.00000000 & 0.00000003 & 0.00000000 & 0.00000044 & -0.00000001 & 0.00000000 \\ 0.01165964 & 0.00022741 & 0.00305369 & 0.00000000 & 0.01166205 & 0.10681518 & 0.00136332 & 0.00152559 \\ 1.27113747 & 0.09510282 & 0.14976776 & 0.00000044 & 0.10681518 & 11.77556446 & 0.09904505 & 0.30399224 \\ 0.01021141 & 0.00128841 & 0.00170421 & -0.00000001 & 0.00136332 & 0.09904505 & 0.00161142 & 0.00131032 \\ 0.03364064 & 0.00211645 & 0.00246767 & 0.00000000 & 0.00152559 & 0.30399224 & 0.00131032 & 0.01222711 \end{bmatrix}.$$

To see this near singularity, Figure 1 provides a matrix of bivariate cross-sectional contour plots for each pair of coefficients from the Texas data, for contours at 0.05, 0.15, . . . , 0.95 where the 0.05 contour line bounds approximately 0.95 of the data, holding constant all other parameters at their maxima. (Recall that these easy-to-compute cross-sectional likelihood plots are distinct from the more desirable but harder-to-compute marginal distributions; parameters not shown are held constant in the former but integrated out in the latter.) In these Texas data, the likelihood is concave at the global maximum, although the curvature for `Federal` is only slightly greater than zero. This produces a near-ridge in the contours for each variable paired with `Federal`, and although it cannot be seen in the figure, the ridge is gently sloping around maximum value in each cross-sectional likelihood plot.

The point estimates and standard errors correctly pick up the unreliability of the `Federal` coefficient value by giving it a very large standard error, but, as is generally the case, the contours reveal more information. In particular, the plot indicates that distribution

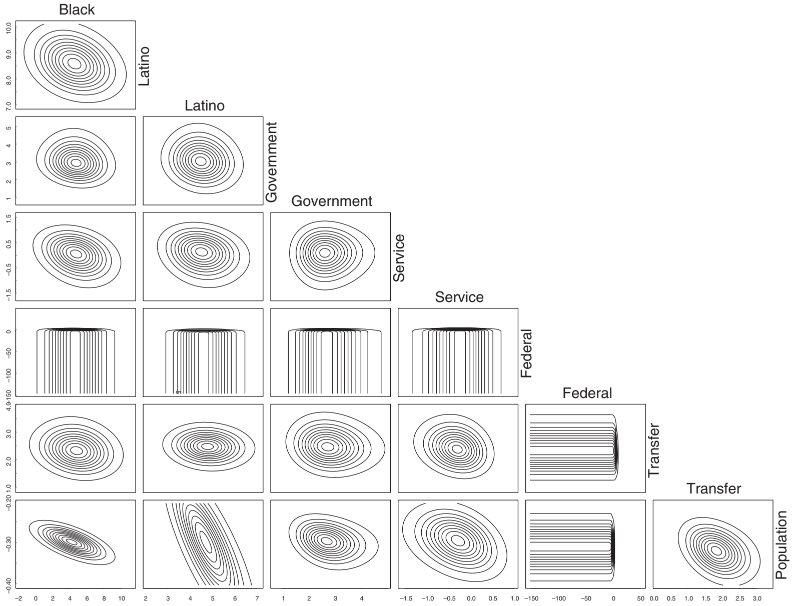

Figure 1: Contourplot Matrix, Logit Model: Texas

of the coefficient on `Federal` is quite asymmetric and indeed very informative in the manner by which the probability density drops as we come away from the near-ridge. The modes and their standard errors, in the first pair of columns in Table 1, cannot reveal this additional information. In contrast, the resampling results can easily reveal the richer information. For example, to compute the entries in the last two columns of Table 1, we first took many random draws of the parameters from their exact posterior distribution. If, instead of summarizing this information with their means and standard deviations, as in the table, we presented univariate or bivariate histograms of the draws, we would reveal all the information in Figure 1. In fact, the histograms would give the exact marginal distributions of interest (the full posterior, with other parameters integrated out) rather than merely the contours as shown in the figures, and so the potential information revealed, even in this case when the Hessian is invertible, could be substantial. We do not present the histograms in this example because they happen to be similar to the contours in this particular data set.[8]

TABLE 2: Logit Regression Model: Singular Hessian, Florida

	Standard Results		*Minus Federal*		*Importance Sampling*	
Parameter	*Estimate*	*Standard Error*	*Estimate*	*Standard Error*	*Estimate*	*Standard Error*
Black	5.86	???	5.58	5.34	5.56	2.66
Latino	4.08	???	3.21	8.10	3.97	2.73
Govt	−1.53	???	−1.59	1.24	−1.49	1.04
Service	−2.93	???	−2.56	1.69	−2.99	1.34
Federal	−21.35	???			−20.19	∞
Transfer	2.98	???	2.33	1.29	2.98	1.23
Population	−1.43	???	−0.82	0.72	−1.38	0.47
Intercept	12.27	???	6.45	6.73	11.85	4.11

We ran the same specification for Florida (33 counties), providing the maximum likelihood parameter estimates in Table 2 and the following Hessian, which is noninvertible (and represented in the table with question marks for the standard errors):

$$\mathbf{H} = \begin{bmatrix} 0.13680004 & 0.04629599 & 0.01980602 & 0.00000001 & 0.05765988 & 1.32529504 & 0.02213744 & 0.00631444 \\ 0.04629599 & 0.04629442 & 0.00000000 & -0.00000004 & 0.03134646 & 0.45049457 & 0.00749867 & 0.00114495 \\ 0.01980602 & 0.00000000 & 0.01980564 & 0.00000000 & 0.01895061 & 0.19671280 & 0.00234865 & 0.00041155 \\ 0.00000001 & -0.00000004 & 0.00000000 & 0.00000000 & 0.00000000 & 0.00000000 & 0.00000000 & 0.00000002 \\ 0.05765988 & 0.03134646 & 0.01895061 & 0.00000000 & 0.05765900 & 0.57420212 & 0.00817570 & 0.00114276 \\ 1.32529504 & 0.45049457 & 0.19671280 & 0.00000000 & 0.57420212 & 12.89475788 & 0.21458995 & 0.06208332 \\ 0.02213744 & 0.00749867 & 0.00234865 & 0.00000000 & 0.00817570 & 0.21458995 & 0.00466134 & 0.00085111 \\ 0.00631444 & 0.00114495 & 0.00041155 & 0.00000002 & 0.00114276 & 0.06208332 & 0.00085111 & 0.00088991 \end{bmatrix}.$$

Consider first Figure 2, which provides a matrix of the bivariate cross-sectional contour plots for each pair of coefficients for the Florida data, again with contours at 0.1, 0.2, . . . , 0.9. As with the case of Texas, the problematic cross-sectional likelihood is for `Federal`, but in this case, the modes are not unique and so the Hessian matrix is not invertible. Except for this variable, the likelihoods are very well behaved. If we were forced to abandon the specification at this point, this is exactly the information that would be permanently lost. This is especially problematic when contrasted with the Texas case, for which the contours do not look a lot more informative.

A good data analyst using classical procedures with our data might reason as follows. The Texas data clearly suggest racial bias, but no results are available with the same specification in Florida. If we follow the textbook advice and respecify by dropping `Federal` and rerunning, we get the results in the second pair of columns in Table 2. The results for `Black` reveal a coefficient for Florida that is only a

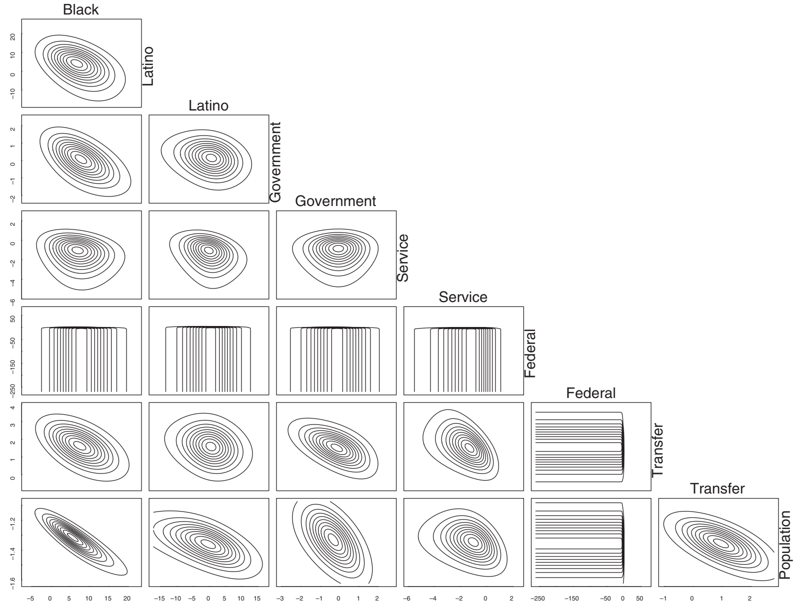

Figure 2: Contour Plot Matrix, Importance Sampling Model: Florida

third of the size as it was in Texas and only slightly larger than its standard error. The contrast is striking: Substantial racial bias in Texas and no evidence of such in Florida. Unfortunately, with classical methods, it is impossible to tell whether these interesting and divergent substantive results in Florida are due to omitted variable bias rather than political and economic differences between the states.

So what to do? One reasonable approach is to assume that the (unobservable) bias that resulted from omitting `Federal` in the Florida specification would be of the same degree and direction as the (observable) bias that would occur by omitting the variable in the Texas data. The advantage of this assumption is that we can easily estimate the bias in Texas by omitting `Federal`. We do this in the second pair of columns in Table 1. Those results suggest that there is no bias introduced since the results are nearly unchanged from the first two columns. Although this is a reasonable procedure (which most analysts have probably tried at one time or another), it is of course based on the unverifiable assumption that the biases are the same in the two states. In the present data, this assumption is false, as we now show.

We now recover some of the information lost in the Florida case by first applying our generalized inverse and generalized Cholesky procedures to the singular Hessian to create a pseudo-variance matrix. We then perform importance resampling using the multivariate normal, with the mode and pseudo-variance matrix, as the first approximation. We use a *t* distribution with three degrees of freedom as the approximation distribution to be conservative since we know from graphical evidence that one of the marginal distributions is problematic. The last two columns of Table 2 give the means and standard deviations of the marginal posterior for each parameter. We report ∞ for the standard error of `Federal` to emphasize the lack of information. Although the way the data and model were summarized contained no useful information about this parameter, the specification did control for `Federal`, and so any potentially useful information about the other parameters and their standard errors are revealed with our procedure without the potential for omitted variable bias that would occur by dropping the variable entirely.

The results seem informative. They show that the effect of `Black` is indeed smaller in Florida than Texas, but the standard error for Florida is now almost a third of the size of the coefficient. Thus, the racial bias is clearly large in both states, although larger in Texas than Florida. This result thus precisely reverses the conclusion from the biased procedure of dropping the problematic `Federal` variable. Of course, without the generalized inverse/generalized Cholesky technique, there would be no results to evaluate for Florida at all.

All statistical results depend on assumptions, and most depend on a model of some kind. As a result, scholars are aware that classic statistical hypothesis tests and confidence intervals are all conditional on the veracity of the statistical model and so underestimate the degree of uncertainty they should have about their results. (Bayesian model averaging is one technique designed to also include some aspects of specification uncertainty.) That is, intervals and tests normally thus reflect *sampling uncertainty* but not *specification uncertainty*. Researchers may be less familiar with *summarization uncertainty*—the uncertainties that result from potentially inadequate summaries of the posterior distribution with small, low-dimensional statistics. However, this additional source of uncertainty can sometimes be substantial, particularly for the increasingly common complex nonlinear models. These include the risks of local minima, numerical instabilities, and non-unimodal posterior densities that are poorly summarized by point estimates.

For those doing empirical analyses that run into noninvertable Hessians, the additional summarization uncertainties involved in using the procedures we developed in this article and used in this section must also be recognized. Thus, although the standard errors did not systematically increase from the second to the last pair of columns in Table 1 and actually decreased in all cases in Table 2, the genuine level of uncertainty a researcher should have in these statistics must be considered a good deal higher than the numbers presented. We have tried to provide some graphical diagnostics to help reduce these uncertainties, but the preliminary nature of the methods should keep researchers on guard. Of course, the uncertainties a researcher should have in results that almost certainly have bias, such as by answering a question other than the one asked, would probably be higher still.

AN ALTERNATIVE PROCEDURE: DRAWING FROM THE SINGULAR NORMAL

We now describe a second procedure for drawing the random numbers from a different approximating density: the truncated singular normal. The basic idea is to try to draw directly from the singular multivariate density with a noninvertible Hessian. We think that the generalized Cholesky procedure will work better if the underlying model is identified, but numerical problems lead to apparent nonidentification. In contrast, we suspect that that this procedure will perform better when the underlying model would have a noninvertible Hessian even if we were able to run it on a computer with infinite precision. We offer no way to distinguish these two situations, but fortunately, it is relatively easy to run both.

To set up the procedure, again consider a Hessian matrix of second derivatives, $\mathbf{H}$, along with a $k \times 1$ associated vector of maximum likelihood estimates, $\hat{\boldsymbol{\theta}}$. The matrix $(-\mathbf{H})^{-1}$ does not exist due to either nonpositive definiteness or singularity ($r \leq k$). Suppose one can set some reasonable bounds on the posterior distribution of each of the k coefficient estimates in $\hat{\boldsymbol{\theta}}$. These bounds may be set according to empirical observation with similar models, as a Bayes-like prior assertion. Thus, we assume that $\boldsymbol{\theta} \in [\mathbf{g}, \mathbf{h}]$, where $\mathbf{g}$ is a $k \times 1$ vector of lower bounds and $\mathbf{h}$ is a $k \times 1$ vector of upper bounds.

The goal now is to draw samples from the distribution of $\hat{\boldsymbol{\theta}} : \hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, (-H)^{-1}) \propto e^{-T/2}$, truncated to be within $[\mathbf{g}, \mathbf{h}]$, and where $T = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\mathbf{H}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$. Note that the normal density does include an expression for the variance-covariance matrix—only the inverse (the negative of the Hessian), which exists and we have. We thus decompose T as follows:

$$\begin{aligned} T &= (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\mathbf{H}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \\ T &= (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\mathbf{U}'\mathbf{L}\mathbf{U}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}), \end{aligned} \tag{7.1}$$

where $\mathbf{U}'\mathbf{L}\mathbf{U}$ is the spectral decomposition of $\mathbf{H}$. Thus, $\text{rank}(\mathbf{H}) = r \leq k$, $\mathbf{H}$ has r eigenvalues (which we denote $d_1, \ldots, d_r$), $\mathbf{U}$ is $k \times k$ and orthogonal (and hence $(\mathbf{U})^{-1} = \mathbf{U}'$), and $\mathbf{L} = \text{diag}(\mathbf{L}_1, 0)$, where $\mathbf{L}_1 = \text{diag}(d_1, \ldots, d_r)$. Thus, the $\mathbf{L}$ matrix is a diagonal matrix with r leading values of eigenvalues and $n - r$ trailing zero values.

We now make the transformation $\mathbf{A} = \mathbf{U}(\hat{\boldsymbol{\theta}} - [\mathbf{h} + \mathbf{g}]/2)$, the density for which would normally be $\mathbf{A} \sim N(\mathbf{U}(\boldsymbol{\theta} - [\mathbf{h} + \mathbf{g}]/2), (-\mathbf{L})^{-1})$. This transformation centers the distribution of $\mathbf{A}$ at the middle of the bounds, and since $\mathbf{L}$ is diagonal, it factors into the product of independent densities. But this expression has two problems: (a) $(-\mathbf{L})^{-1}$ does not always exist, and (b) $\mathbf{A}$ has complicated multivariate support (a hypercube not necessarily parallel with the axes of the elements of $\mathbf{A}$), which is difficult to draw random numbers from. We now address these two problems.

First, in place of $\mathbf{L}$, we use $\mathbf{L}^*$, defined such that $L_i^* = L_i$ if $L_i > 0$ and L_i^* equals some small positive value otherwise (and where the subscript refers to the row and column of the diagonal element). Except for the specification of the support of A that we consider next, this transforms the density into

$$\begin{aligned} \mathbf{A} &\sim N(\mathbf{U}'\boldsymbol{\theta}, (-\mathbf{L}^*)^{-1}) \\ &= \prod_i N(U_i(\theta_i - [h_i + g_i]/2, -1/L_i^*), \end{aligned} \tag{7.2}$$

Second, instead of trying to draw directly from the support of $\mathbf{A}$, we draw from a truncated density with support that is easy to compute and encompasses the support of $\mathbf{A}$ (but is larger than it), transform back via $\hat{\boldsymbol{\theta}} = \mathbf{U}'\mathbf{A} + (\mathbf{h} + \mathbf{g})/2$, and accept the draw only if $\hat{\boldsymbol{\theta}}$ falls within its (easy-to-verify) support, $[\mathbf{g}, \mathbf{h}]$. The encompassing support we use for each element in the vector $\mathbf{A}$ is the hypercube $[-Q, Q]$, where the scalar Q is the maximum Euclidean distance from $\boldsymbol{\theta}$ to any of the 2^k corners of the hyperrectangle defined by the bounds. Since by definition $\boldsymbol{\theta} \in [\mathbf{g}, \mathbf{h}]$, we should normally avoid the sometimes common pitfall of rejection sampling—having to do an infeasible number of draws from $\mathbf{A}$ to accept each draw of $\hat{\boldsymbol{\theta}}$.

The principle of rejection sampling is satisfied here—that we can sample from any space (in our case, using support for $\mathbf{A}$ larger than its support) so long as it fully encompasses the target space and the accept/reject algorithm operates appropriately. If $-\mathbf{H}$ were positive definite, this algorithm would return random draws from a truncated normal distribution. When $-\mathbf{H}$ is not positive definite, it returns draws from a singular normal, truncated as indicated.

So now we have draws of $\hat{\boldsymbol{\theta}}$ from a singular normal distribution. We then repeat the procedure m, which serves as draws from the enveloping distribution that is used in the importance sampling procedure. That is, we take these simulations of $\hat{\boldsymbol{\theta}}$ and accept or reject according to the importance ratio. We keep going until we have enough simulated values.

OTHER APPROACHES

The problem of computational misspecification is well studied in statistics, particularly in the case of the linear model. McCullagh and Nelder (1989) discuss this problem in the context of generalized linear models in which specifications that introduce overlapping subspaces due to redundant information in the factors produce *intrinsic aliasing*. This occurs when a linear combination of the factors reduces to fewer terms than the number of specified parameters. McCullagh and Nelder solve the aliasing problem by introducing "suitable constraints," which are linear restrictions that increase the dimension of the subspace created by the specified factors. A problem with this approach is that the "suitable constraints" are necessarily an arbitrary and possibly atheoretical imposition. In addition, it is often difficult to determine a minimally affecting yet sufficient set of constraints.

McCullagh and Nelder (1989) also identify *extrinsic aliasing*, which produces the same modeling problem but as a result of data values. The subspace is reduced below the number of factors because of redundant case-level information in the data. This is only a problem, however, in very low-sample problems atypical of social science applications.

Another well-known approach to this problem in linear modeling is ridge regression. Ridge regression essentially trades the multicollinearity problem for introduced bias. Suppose that the $\mathbf{X}'\mathbf{X}$ matrix is singular or nearly singular. Then specify the smallest scalar possible, ζ, that can be added to the characteristic roots of $\mathbf{X}'\mathbf{X}$ to make this matrix nonsingular. The linear estimator is now defined as

$$\hat{\boldsymbol{\beta}}(\theta) = (\mathbf{X}'\mathbf{X} + \zeta\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}.$$

There are two well-known problems with this approach. First, the coefficient estimate is, by definition, biased, and there currently exists no theoretical approach that guarantees some minimum degree of bias. Some approaches have been suggested that provide reasonably small values of ζ based on graphical methods (Hoerl and Kennard 1970a, 1970b), empirical Bayes (Efron and Morris 1972; Amemiya 1985), or generalized ridge estimators based on decision theoretical considerations (James and Stein 1961; Berger 1976; Strawderman 1978). Second, because ζ is calculated with respect to the smallest eigenvalue of $\mathbf{X}'\mathbf{X}$, it must be added to every diagonal of the matrix: $\mathbf{X}'\mathbf{X} + \zeta\mathbf{I}$. So, by definition, the matrix is changed more than necessary (cf. Appendix 10.1). For a very informative discussion (and one that seems to have mortally wounded ridge regression), see Smith and Campbell (1980) along with the comments that follow.

Another alternative was proposed by Rao and Mitra (1971). Define $\delta\boldsymbol{\theta}$ as an unknown correction that has an invertible Hessian. Then (ignoring higher order terms in a Taylor series expansion of $\delta\boldsymbol{\theta}$), $f(\mathbf{x}|\boldsymbol{\theta}) = H(\boldsymbol{\theta})\delta\boldsymbol{\theta}$. Since $H(\boldsymbol{\theta})$ is singular, a solution is available only by the generalized inverse, $\delta\boldsymbol{\theta} = H(\boldsymbol{\theta})^{-}f(\mathbf{x}|\boldsymbol{\theta})$. When there exists a parametric function of $\boldsymbol{\theta}$ that is estimable and whose first derivative is in the column space of $H(\boldsymbol{\theta})$, then there exists a unique, maximum likelihood estimate of this function, $\phi(\hat{\boldsymbol{\theta}})$, with asymptotic variance-covariance matrix $\phi(\hat{\boldsymbol{\theta}})H(\boldsymbol{\theta}_0)^{-}\phi(\hat{\boldsymbol{\theta}})$. The difficulty with this procedure, of course, is finding a substantively reasonable version of $\phi(\hat{\boldsymbol{\theta}})$. Rao and Mitra's point is nevertheless quite useful since it points out that any generalized inverse has a first derivative in the column space of $H(\boldsymbol{\theta})$.

An additional approach is to apply bootstrapping to the regression procedure to produce empirical estimates of the coefficients that can then be used to obtain subsequent values for the standard errors. The basic procedure (Davidson and MacKinnon 1993:330-31; Efron and Tibshirani 1993:111-12) is to bootstrap from the residuals of a model in which coefficients estimates are obtained but the associated measures of uncertainty are unavailable or unreliable. The steps for the linear model are as follows (Freedman 1981): (1) for the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, obtain $\hat{\boldsymbol{\beta}}$ and the centered residuals $\boldsymbol{\epsilon}^*$; (2) sample size n with replacement m times from $\boldsymbol{\epsilon}^*$ and calculate m replicates of the outcome variable by $\mathbf{y}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\epsilon}^*$; (3) regress the m iterates of the $\mathbf{y}^*$ vector on $\mathbf{X}$ to obtain m iterates of $\hat{\boldsymbol{\beta}}$; and (4) summarize the coefficient estimates with the mean and standard deviation of these bootstrap samples. The generalized linear model case is only slightly more involved since it is necessary to incorporate the link function, and the (Pearson) residuals need to be adjusted (see Shao and Tu 1995:341-43).

Applying this bootstrap procedure to our Florida data in which the coefficient estimates are produced and the Hessian fails, we obtain the standard error vector: [9.41, 9.08, 1.4, 2.35, 25.83, 1.43,11.86, 6.32] (in the same order as Table 2). These are essentially the same standard errors as those in the model dropping `Federal`, except that the uncertainty for `Population` is much higher. This bootstrapping procedure does not work well in non-iid settings (it assumes that the error between $\mathbf{y}$ and $\mathbf{X}\hat{\boldsymbol{\beta}}$ is independent of $\mathbf{X}$), and it is possible that spatial correlation that is likely to be present in FIPS-level population data is responsible for this one discrepency. An alternative bootstrapping procedure, the paired bootstrap, generates m samples of size n directly from (y_j, x_j) together to produce $\mathbf{y}^*$, $\mathbf{X}^*$ and then generates $\hat{\boldsymbol{\beta}}$ values. While the paired bootstrap is less sensitive to non-iid data, it can produce simulated data sets (the $\mathbf{y}^*$, $\mathbf{X}^*$) that are very different from the original data (Hinkley 1988).

By far, the most common way of recovering from computational problems resulting from collinearity is respecification. Virtually every basic and intermediate text on linear and nonlinear regression techniques gives this advice. The respecification process can vary from ad hoc trial error strategies to more sophisticated approaches based on principal components analysis (Greene 1993:271-73). While these approaches often "work," they force the user to change their research question due to technical concerns. As the example in Section 6 shows, we should not be forced to alter our thinking about a research question as a result of computational issues.

CONCLUDING REMARKS

In this article, we seek to rescue analyses with noninvertable Hessians that might otherwise be left in the trash bin. Although the likelihood estimated may have certain problems associated with it, the data in these problems may still contain revealing information about the question at hand. We therefore help researchers avoid giving up the question they posed originally and instead extract at least some of the remaining available information. The methods we offer that are intended to accomplish these tasks are hardly infallable. As such, considerable care should go into using them, and much further research needs to be conducted to help extract the additional information available and to understand in precisely what circumstances the ideas in this article can be applied.

Finally, we note that an opportunity exists for advancement in the field of linear algebra that could help work in ours. In particular, we apply the generalized inverse and the generalized Cholesky sequentially because theoretical developments in the two areas have been developed separately and apparently independently. We conjecture that theoretical, or at least computational, efficiencies can be found by combining the two procedures. In addition, it may also be possible to produce an even better result by using the information that the Hessian is not merely a symmetric matrix but that it was formed as a matrix of second derivatives. We thus encourage future linear algebra researchers to find a way to begin with a Hessian matrix and to produce the "nearest" possible well-conditioned, positive-definite (and hence nonsingular) pseudo-variance matrix. This procedure would have many important applications.

APPENDIXES

THE GENERALIZED CHOLESKY

We now describe the classic Cholesky decomposition and recent generalizations designed to handle non-positive-definite matrices. A matrix $\mathbf{C}$ is positive definite if, for any $\mathbf{x}$ vector except $\mathbf{x} = \mathbf{0}$, $\mathbf{x}'\mathbf{C}\mathbf{x} > 0$, or in other words if $\mathbf{C}$ has all positive eigenvalues. Symmetric positive-definite matrices are nonsingular, have only positive numbers on the diagonal, and have positive determinants for all principle-leading submatrices. The Cholesky matrix is defined as $\mathbf{V}$ in the decomposition $\mathbf{C} = \mathbf{V}'\mathbf{V}$. We thus construct our pseudo-variance matrix as $\mathbf{V}'\mathbf{V}$, where $\mathbf{V} = \text{GCHOL}(\mathbf{H}^{-})$, GCHOL$(\cdot)$ is the generalized Cholesky described below, and $\mathbf{H}^{-}$ is the Moore-Penrose generalized inverse of the Hessian matrix.

The Classic Algorithm

The classic Cholesky algorithm assumes a positive-definite matrix and symmetric variance matrix ($\mathbf{C}$). It then proceeds via the following decomposition:

$$\underset{(k\times k)}{\mathbf{C}} = \underset{(k\times k)}{\mathbf{L}}\,\underset{(k\times k)}{\mathbf{D}}\,\underset{(k\times k)}{\mathbf{L}'}. \tag{A.1}$$

The basic Cholesky procedure is a one-pass algorithm that generates two output matrices that can then be combined for the desired “square root” matrix. The algorithm moves down the main diagonal of the input matrix determining diagonal values of $\mathbf{D}$ and triangular values of $\mathbf{L}$ from the current column of $\mathbf{C}$ and previously calculated components of $\mathbf{L}$ and $\mathbf{C}$. Thus, the procedure is necessarily sensitive to values in the original matrix and previously calculated values in the $\mathbf{D}$ and $\mathbf{L}$ matrices. There are k stages in the algorithm corresponding to the k-dimensionality of the input matrix. The jth step ($1 \leq j \leq k$) is characterized by two operations:

$$\mathbf{D}_{j,j} = \mathbf{C}_{j,j} - \sum_{\ell=1}^{j-1} \mathbf{L}_{j,\ell}^2 \mathbf{D}_{\ell,\ell}, \quad \text{and} \tag{A.2}$$

$$\mathbf{L}_{i,j} = \left[\mathbf{C}_{i,j} - \sum_{\ell=1}^{j-1} \mathbf{L}_{j,\ell}\mathbf{L}_{i,\ell}\mathbf{D}_{\ell,\ell}\right] / \mathbf{D}_{j,j}, \quad i = j+1, \ldots, k, \tag{A.3}$$

where $\mathbf{D}$ is a positive diagonal matrix so that upon the completion of the algorithm, the square root of it is multiplied by $\mathbf{L}$ to give the Cholesky decomposition. From this algorithm, it is easy to see why the Cholesky algorithm cannot tolerate singular or non-positive-definite input matrices. Singular matrices cause a divide-by-zero problem in (A.3), and non-positive-definite matrices cause the sum in (A.2) to be greater than $\mathbf{C}_{j,j}$, thus causing negative diagonal values. Furthermore, these problems exist in other variations of the Cholesky algorithm, including those based on svd and qr decomposition. Arbitrary fixes have been tried to preserve the mathematical requirements of the algorithm, but they do not produce a useful result (Fiacco and McCormick 1968; Gill, Golub, Murray, and Sanders 1974; Matthews and Davies 1971).

The Gill/Murray Cholesky Factorization

Gill and Murray (1974) introduced and Gill, Murray, and Wright (1981) refined an algorithm to find a nonnegative diagonal matrix, $\mathbf{E}$, such that $\mathbf{C} + \mathbf{E}$ is positive definite and the diagonal values of $\mathbf{E}$ are as small as possible. This could easily be done by taking the greatest negative eigenvalue of $\mathbf{C}$, λ_1, and assigning $\mathbf{E} = -(\lambda_1 + \epsilon)\mathbf{I}$, where ϵ is some small positive increment. However, this approach (implemented in various computer programs, such as the Gauss "maxlik" module) produces $\mathbf{E}$ values that are much larger than required, and therefore the $\mathbf{C} + \mathbf{E}$ matrix is much less like $\mathbf{C}$ than it could be.

To see Gill et al.'s (1981) approach, we rewrite the Cholesky algorithm provided as (A.2) and (A.3) in matrix notation. The jth submatrix of its application at the jth step is

$$\mathbf{C}_j = \begin{bmatrix} c_{j,j} & \mathbf{c}'_j \\ \mathbf{c}_j & \mathbf{C}_{j+1} \end{bmatrix}, \tag{A.4}$$

where $c_{j,j}$ is the jth pivot diagonal; $\mathbf{c}'{}_j$ is the row vector to the right of $c_{j,j}$, which is the transpose of the $\mathbf{c}_j$ column vector beneath $c_{j,j}$; and $\mathbf{C}_{j+1}$ is the $(j+1)$th submatrix. The jth row of the $\mathbf{L}$ matrix is calculated by $L_{j,j} = \sqrt{c_{j,j}}$, and $\mathbf{L}_{(j+1):k,j} = \mathbf{c}_{(j+1):k,j}/L_{j,j}$. The $(j+1)$th submatrix is then updated by

$$\mathbf{C}^*_{j+1} = \mathbf{C}_{j+1} - \frac{\mathbf{c}_j \mathbf{c}'{}_j}{L^2_{j,j}}. \tag{A.5}$$

Suppose that at each iteration, we defined $L_{j,j} = \sqrt{c_{j,j} + \delta_j}$, where δ_j is a small positive integer sufficiently large so that $\mathbf{C}_{j+1} > \mathbf{c}_j \mathbf{c}'/L^2_{j,j}$. This would obviously ensure that each of the j iterations does not produce a negative diagonal value or divide-by-zero operation. However, the size of δ_j is difficult to determine and involves tradeoffs between satisfaction with the current iteration and satisfaction with future iterations. If δ_j is picked such that the new jth diagonal is just barely bigger than zero, then subsequent diagonal values are greatly increased through the operation of (A.5). Conversely we do not want to be adding large δ_j values on any given iteration.

Gill et al. (1981) note the effect of the $\mathbf{c}_j$ vector on subsequent iterations and suggest that minimizing the summed effect of δ_j is equivalent to minimizing the effect of the vector maximum norm of $\mathbf{c}_j$, $\|\mathbf{c}_j\|_\infty$ at each iteration. This is done at the jth step by making δ_j the smallest nonnegative value satisfying

$$\begin{aligned} &\|\mathbf{c}_j\|_\infty \beta^{-2} - c_{j,j} \leqslant \delta_j \\ &\text{where } \beta = \max \begin{cases} \max(\text{diag}(\mathbf{C})) \\ \max(\textit{not}\,\text{diag}(\mathbf{C}))\sqrt{k^2 - 1} \\ \epsilon_{\mathrm{m}} \end{cases} \end{aligned} \tag{A.6}$$

where ϵ_{m} is the smallest positive number that can be represented on the computer used to implement the algorithm (normally called the machine epsilon). This algorithm always produces a factorization and has the advantage of not modifying already positive-definite $\mathbf{C}$ matrices. However, the bounds in (A.6) have been shown to be nonoptimal and thus provide $\mathbf{C} + \mathbf{E}$ that is further from $\mathbf{C}$ than necessary.

The Schnabel/Eskow Cholesky Factorization

Schnabel and Eskow (1990) improve on the $\mathbf{C} + \mathbf{E}$ procedure of Gill and Murray (1974) by applying the Gerschgorin circle theorem to reduce the infinity norm of the $\mathbf{E}$ matrix. The strategy is to calculate δ_j values that reduce the *overall* difference between $\mathbf{C}$ and $\mathbf{C} + \mathbf{E}$. Their approach is based on the following theorem (stated in the context of our problem).

THEOREM A.1. *Suppose* $\mathbf{C} \in \mathbb{R}^k$ *with eigenvalues* $\lambda_1, \ldots, \lambda_k$. *Define the* i*th Gerschgorin bound as*

$$\mathbf{G}_i\,(\mathit{lower,upper}) = \left[\mathbf{C}_{i,i} - \sum_{\substack{j=1\\ j\neq i}}^{n} |C_{i,j}|, \mathbf{C}_{i,i} + \sum_{\substack{j=1\\ j\neq i}}^{n} |C_{i,j}| \right].$$

Then, $\lambda_i \in \left[\mathbf{G}_1 \cup \mathbf{G}_2 \cup \cdots \cup \mathbf{G}_k\right] \; \forall \lambda_{1\leq i\leq k}$.

But we know that λ_1 is the largest negative amount that must be corrected, so this simplifies to the following decision rule:

$$\delta_j = \max\left(\epsilon_{\mathrm{m}}, \max_i(\mathbf{G}_i\,(\mathrm{lower}))\right). \tag{A.7}$$

In addition, we do not want any δ_j to be less than δ_{j-1} since this would cause subsequent submatrices to have unnecessarily large eigenvalues, and so a smaller quantity is subtracted in (A.5). Adding this condition to (A.7) and protecting the algorithm from problems associated with machine epsilon produces the following determination of the additional amount in $L_{j,j} = \sqrt{c_{j,j} + \delta_j}$:

$$\delta_j = \max(\epsilon_{\mathrm{m}}, -\mathbf{C}_{j,j} + \max(\|\mathbf{a}_j\|, (\epsilon_{\mathrm{m}})^{\frac{1}{3}}\max(\mathrm{diag}(\mathbf{C}))), \mathbf{E}_{j-1,j-1}). \tag{A.8}$$

The algorithm follows the same steps as that of Gill and Murray (1974) except that the determination of δ_j is done by (A.8). The Gerschgorin bounds, however, provide an order of magnitude improvement in $\|\mathbf{E}\|_\infty$. We refer to this Cholesky algorithm based on Gerschgorin bounds as the generalized Cholesky since it improves the common procedure, accommodates a more general class of input matrices, and represents the "state of the art" with regard to minimizing $\|\mathbf{E}\|_\infty$.

A Review of Importance Sampling

Importance sampling is a general technique that uses ratios of densities from repeated draws to estimate integrals and to obtain marginal distributions. For example, suppose we wished to obtain the marginal distribution for some parameter θ_1 from a joint distribution: $f(\theta_1, \theta_2|\mathbf{X})$. If we knew the parametric form for this joint distribution, it is often straightforward to analytically integrate out the second parameter over its support, as shown in basic mathematical statistics texts: $f(\theta_1|\mathbf{X}) = \int f(\theta_1, \theta_2|\mathbf{X})d\theta_2$. However, in a great many circumstances, this is not possible, and numerical approximations are required. Suppose we could posit a *normalized* conditional posterior approximation density of $\theta_2 : \hat{f}(\theta_2|\theta_1, \mathbf{X})$, which can often be given a normal or t form. The trick that this approximation allows is that an expected value formulation can be substituted for the integral and repeated draws used for numerical averaging. Specifically, the form for the marginal distribution is developed as

$$\begin{aligned} f(\theta_1|\mathbf{X}) &= \int f(\theta_1, \theta_2|\mathbf{X})d\theta_2 \\ &= \int \frac{f(\theta_1, \theta_2|\mathbf{X})}{\hat{f}(\theta_2|\theta_1, \mathbf{X})} \hat{f}(\theta_2|\theta_1, \mathbf{X})d\theta_2 \\ &= E_{\theta_2}\left[\frac{f(\theta_1, \theta_2|\mathbf{X})}{\hat{f}(\theta_2|\theta_1, \mathbf{X})}\right]. \end{aligned} \tag{A.9}$$

The fraction $\frac{f(\theta_1,\theta_2|\mathbf{X})}{\hat{f}(\theta_2|\theta_1,\mathbf{X})}$ is called the importance weight and determines the probability of accepting sampled values of θ_2. This setup allows the following rather simple procedure to obtain the estimate of $f(\theta_1|\mathbf{X})$.

1. Divide the support of θ_1 into a grid with the desired level of granularity determined by $k : \theta_1^{(1)}, \theta_1^{(2)}, \ldots, \theta_1^{(k)}$.
2. For each of the $\theta_1^{(i)}$ values along the k-length grid, determine the density estimate at that point by performing the following steps:
   (a) Simulate N values of $\hat{\theta_2}$ from $\hat{f}(\theta_2|\theta_1^{(i)}, \mathbf{X})$.
   (b) Calculate $f(\theta_1^{(i)}, \hat{\theta}_{2n}|\mathbf{X}) / \hat{f}(\hat{\theta}_{2n}|\theta_1^{(i)}, \mathbf{X})$ for $i = 1$ to N.
   (c) Use (A.9) to obtain $f(\theta_1^{(i)}|\mathbf{X})$ by taking the means of the N ratios just calculated.

The user can fix the level of accuracy of this estimate by changing the granularity of the grid and the number of draws per position on that grid. Obviously, this procedure can also be used to perform numerical integration, provided a suitable normalized approximation function

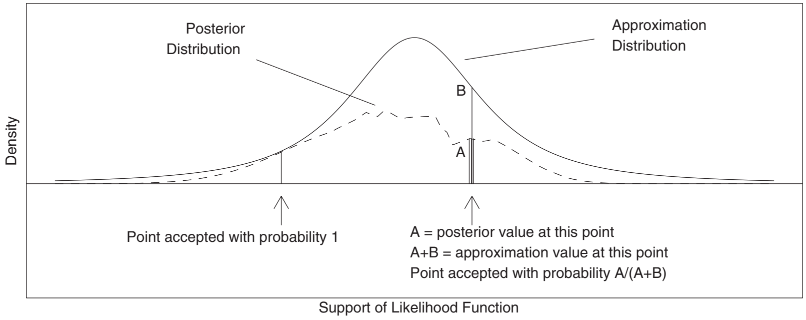
Posterior Distribution
Approximation Distribution
B
A
Density
Point accepted with probability 1
A = posterior value at this point
A+B = approximation value at this point
Point accepted with probability A/(A+B)
Support of Likelihood Function

Figure 3: Importance Sampling Illustration

can be found. This makes importance sampling a very useful tool in applied mathematics.

Importance sampling is depicted in Figure 3, where two points along the support of the posterior density are indicated. At the first point, the approximation density and the posterior density provide identical values, so this point is accepted with probability 1. The second point is accepted with probability equal to the ratio $A/(A+B)$ (i.e., the quality of the approximation).

NOTES

1. In one of the best econometric textbooks, Davidson and MacKinnon (1993:185-86) write, “There are basically two options: Get more data, or estimate a less demanding model . . . . If it is not feasible to obtain more data, then one must accept the fact that the data one has contain a limited amount of information and must simplify the model accordingly. Trying to estimate models that are too complicated is one of the most common mistakes among inexperienced applied econometricians.” We provide an alternative to simplifying or changing the model, but the wisdom of Davidson and MacKinnon’s advice is worth emphasizing in that our approach is only appropriate when the more complicated model is indeed of interest.

2. For simplicity, we refer to the objective function as the posterior distribution from here on, although most of our applications will involve flat priors, in which case, of course, the posterior is equivalent to a likelihood function.

3. In the case of multimodal posterior forms, $\pi(\theta|\mathbf{X})$, the most informative reporting tool, is the Bayesian highest posterior density (HPD) interval. The $100(1-\alpha)$ percent HPD interval is the region of the parameter support for the coefficient θ that meets the following criteria: $C = \{\theta : \pi(\theta|\mathbf{X}) \geq k\}$, where k is the largest number ensuring that $1-\alpha = \int_{\theta:\pi(\theta|\mathbf{X})\geq k} \pi(\theta|\mathbf{x})d\theta$.

This region can be noncontiguous for multimodal forms since it describes the highest probability area of the support, regardless of location.

4. This part of our method is what most separates it from previous procedures in the literature that sought to find a working solution based on the generalized inverse alone (Marquardt 1970; Riley 1955; Searle 1971).

5. In general, $(-\mathbf{H})^{-1}$ is positive definite if, for any nonzero $p \times 1$ vector $\mathbf{x}$, $\mathbf{x}'(-\mathbf{H})^{-1}\mathbf{x} > 0$.

6. The generalized inverse/generalized Cholesky approach is related to the quasi-Newton DFP (Davidon-Fletcher-Powell) method. The difference is that DFP uses iterative differences to converge on an estimate of the negative inverse of a non-positive-definite Hessian (Greene 1993:350), but its purpose is computational rather than statistical, and so the importance sampling step is omitted as well.

7. See Rubin (1987:192-94), Tanner (1996), Gelman et al. (1995), and Wei and Tanner (1990). For applications, see King (1997) and King, Honaker, Joseph, and Scheve (1998).

8. Although logit is known to have a globally concave likelihood surface in theory, actual estimates need not be strictly concave due to numerical imprecision. In the present data, the fact that the Hessian is just barely invertible makes it sensitive to numerical imprecision, and as it turns out, there are at least two local maxima on the marginal likelihood for `Federal`. The statistical package Gauss found a solution at -11.69 and the package R at -5.78 (reported). This discrepancy is typical of software solutions to poorly behaved likelihood functions as algorithmic differences in the applied numerical procedures have different intermediate step locations. The solution difference is not particularly troubling as no reasonable analyst would place faith in either coefficient estimate, given the large reported standard error.

REFERENCES

Amemiya, Takeshi. 1985. *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

Berger, James O. 1976. "Admissible Minimax Estimation of a Multivariate Normal Mean With Arbitrary Quadratic Loss." *Annals of Statistics* 4:223–26.

Davidson, Russell and James G. MacKinnon. 1993. *Estimation and Inference in Econometrics*. New York: Oxford University Press.

Efron, Bradley and C. Morris. 1972. "Limiting the Risk of Bayes and Empirical Bayes Estimators—Part II: The Empirical Bayes Case." *Journal of the American Statistical Association* 67:130–39.

Efron, Bradley and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Fiacco, A. V. and G. P. McCormick. 1968. *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. New York: John Wiley.

Freedman, D. A. 1981. "Bootstrapping Regression Models." *Annals of Statistics* 9:1218–28.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 1995. *Bayesian Data Analysis*. New York: Chapman & Hall.

Gill, Phillip E., G. H. Golub, Walter Murray, and M. A. Sanders. 1974. "Methods for Modifying Matrix Factorizations." *Mathematics of Computation* 28:505–35.

Gill, Phillip E. and Walter Murray. 1974. "Newton-Type Methods for Unconstrained and Linearly Constrained Optimization." *Mathematical Programming* 7:311–50.

Gill, Phillip E., Walter Murray, and M. H. Wright. 1981. *Practical Optimization*. London: Academic Press.

Greene, William H. 1993. *Econometric Analysis*. 2d ed. New York: Macmillan.

Hinkley, David V. 1988. "Bootstrap Methods." *Journal of the Royal Statistical Society, Series B* 50:321–37.

Hoerl, A. E. and R. W. Kennard. 1970a. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics* 12:55–67.

———. 1970b. "Ridge Regression: Applications to Nonorthogonal Problems." *Technometrics* 12:69–82.

James, W. and C. Stein. 1961. "Estimation With Quadratic Loss." In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, edited by Jerzy Neyman. Berkeley: University of California Press.

King, Gary. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior From Aggregate Data*. Princeton, NJ: Princeton University Press.

———. 1998. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. Ann Arbor: University of Michigan Press.

King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 1998. "Listwise Deletion Is Evil: What to Do About Missing Data in Political Science." Paper presented at the annual meetings of the American Political Science Association, Boston.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44:347–61.

Leamer, Edward E. 1973. "Multicollinearity: A Bayesian Interpretation." *Review of Economics and Statistics* 55(3):371–80.

Marquardt, D. W. 1970. "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation." *Technometrics* 12:591–612.

Matthews, A. and D. Davies. 1971. "A Comparison of Modified Newton Methods for Unconstrained Optimization." *Computer Journal* 14:213–94.

McCullagh, P. and J. A. Nelder. 1989. *Generalized Linear Models*. New York: Chapman & Hall.

Moore, E. H. 1920. "On the Reciprocal of the General Algebraic Matrix (Abstract)." *Bulletin of the American Mathematical Society* 26:394–95.

Penrose, R. A. 1955. "A Generalized Inverse for Matrices." *Proceedings of the Cambridge Philosophical Society* 51:406–13.

Rao, C. Radhakrishna and Sujit Kumar Mitra. 1971. *Generalized Inverse of Matrices and Its Applications*. New York: Johnn Wiley.

Riley, James. 1955. "Solving Systems of Linear Equations With a Positive Definite, Symmetric but Possibly Ill-Conditioned Matrix." *Mathematical Tables and Other Aides to Computation* 9:96–101.

Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.

Schnabel, Robert B. and Elizabeth Eskow. 1990. "A New Modified Cholesky Factorization." *SIAM Journal of Scientific Statistical Computing* 11(6):1136–58.

Searle, S. R. 1971. *Linear Models*. New York: John Wiley.

Shao, Jun and Dongsheng Tu. 1995. *The Jackknife and Bootstrap*. New York: Springer.

Smith, Gary and Frank Campbell. 1980. "A Critique of Some Ridge Regression Methods (With Comments)." *Journal of the American Statistical Association* 75:74–103.

Strawderman, W. E. 1978. "Minimax Adaptive Generalized Ridge Regression Estimators." *Journal of the American Statistical Association* 72:890–91.

Tanner, M. A. 1996. *Tools for Statistic Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. New York: Springer.

Wei, G. C. G. and M. A. Tanner. 1990. "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithm." *Journal of the American Statistical Association* 85:699–704.

Jeff Gill is an associate professor of political science at the University of California, Davis. His primary research applies Bayesian modeling and data analysis to substantive questions in public policy, budgeting, bureaucracy, and Congress. His homepage can be found at http://psfaculty.ucdavis.edu/jgill/.

Gary King is the David Florence Professor of Government at Harvard University. He also serves as director of the Harvard-MIT Data Center. His homepage can be found at http://gking.harvard.edu.