CHAPTER 6

Numerical Issues Involved in Inverting Hessian Matrices

Jeff Gill and Gary King

6.1 INTRODUCTION

In the social sciences, researchers typically assume the accuracy of generalized linear models by using an asymptotic normal approximation to the likelihood function or, occasionally, by using the full posterior distribution. Thus, for standard maximum likelihood analyses, only point estimates and the variance at the maximum are normally seen as necessary. For Bayesian posterior analysis, the maximum and variance provide a useful first approximation (but see Chapter 4 for an alternative).

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 in slightly different context, German mathematician Ludwig Hesse) must be positive definite and hence invertible so as to compute the variance matrix, invertible Hessians do not exist for some combinations of datasets and models, 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 where the investigator is drawing data from a known statistical model, due to machine effects.

The Hessian can be noninvertible for both computational reasons and data reasons. Inaccurate implementation of the likelihood function (see Chapters 2 and 3), inaccurate derivative methods (see Chapter 8), or other inappropriate choices in optimization algorithms can yield noninvertible Hessians. Where these inaccuracies cause problems with Hessians, we recommend addressing these inaccuracies directly.

If these methods aren't feasible, or don't work, which often happens, we provide an innovative new library for doing generalized inverses . *Moreover, when a Hessian is not invertible for data reasons, no computational trick can make it invertible, given the model and data chosen, because the desired inverse does*

Numerical Issues in Statistical Computing for the Social Scientist, by Micah Altman, Jeff Gill, and Michael P. McDonald
ISBN 0-471-23633-0

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). For instance, in one of the best econometric textbooks, Davidson and MacKinnon (1993, pp. 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 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." The point of this chapter is to 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 appropriate only when the more complicated model is indeed of interest.

Respecification and reanalysis is important and appropriate advice in some applications of linear regression because a noninvertible Hessian has a clear substantive interpretation: It can only be caused by multicollinearity or 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 be aware of otherwise. It is also of interest in some nonlinear models, such as logistic regression, where the conditions of noninvertibility 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, because the same model specification may have worked in other contexts and really is the one from which the researcher wants estimates. 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).

The point developed in this chapter 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.

In situations where 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. Here, 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.[1] 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 normally much more involved to set up than calculating point estimates and asymptotic variance approximations to which social scientists have become accustomed, and so they have not been adopted widely. Our approach can be thought of as Bayesian, too, although informative prior distributions need not be specified; we focus only on methods that are relatively easy to apply. Although a sophisticated Bayesian analyst could figure out how to elicit information from a posterior with a noninvertible Hessian without our methods in particular instances, we hope that our proposals will make this information available to many more users and may even make it easier for those willing to do the detailed analysis of particular applications. In fact, the methods we discuss are appropriate even when the Hessian does invert and in many cases may be more appropriate than classical approaches. We begin in Section 6.2 by providing a summary of the posterior that can be calculated, even when the mode is uninteresting and the variance matrix is nonexistent. The road map to the rest of the chapter concludes that motivating section.

6.2 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, 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 1989].

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 the researcher 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 model specified 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

[1]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.

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 this mode?

In cases where 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, 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, the mean and its variance can be found, but a unique mode and its variance cannot. At least in these difficult cases, when the textbook suggestion of substantive respecification is not feasible or if it is not 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 do 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 that the posterior for a parameter θ is truncated normal with mean 0.5, standard deviation 10, and truncation is on the [0, 1] interval (cf. Gelman et al. 1995, p. 114, Prob. 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). In contrast, the mean posterior 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.

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. We thus hope to reduce the computational advantage of the mode over the mean by proposing some procedures for computing the mean and its variance.

6.3 DEVELOPING A SOLUTION USING BAYESIAN SIMULATION TOOLS

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 *pseudovariance 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 produce a pseudovariance 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 to our knowledge, the generalized Cholesky has not been used before for statistical purposes.) Surprisingly, the resulting matrix is not usually ill conditioned. In addition, although this is a "pseudo" rather than an "approximate" variance matrix (because the thing 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.[2]

Our solution is nothing more than a way to describe the difficult posterior form using importance sampling, which is a standard tool for Bayesians because they often end up with posterior forms that are difficult to describe analytically. This method of using a convenient candidate distribution and then accepting or rejecting values depending on their resemblance to those produced by the real posterior is supported by a large body of theoretical work starting with Ott (1979), Rubin (1987a), and Smith and Gelfand (1992). Recent discussions of the theoretical validity as well as properties of importance sampling are given by Geweke (1989), Gelman et al. (1995), Robert and Casella (1999), and Tanner (1996). Before continuing, it is also important to note that this proposed solution uses simulation but is not estimation based on Markov chain Monte Carlo analysis.

[2]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 (Riley 1955; Marquardt 1970; Searle 1971).

6.4 WHAT IS IT THAT BAYESIANS DO?

We are certainly "borrowing" from the Bayesian perspective: mean summaries and statistical summary through simulation. However, philosophically we are not requiring that one subscribe to the tenants of Bayesian inference: stipulation of prior distributions for unknown parameters, a belief that these parameters should be described distributionally conditional on the data observed and posteriors based on updating priors with likelihoods.

The essence of Bayesian inference is encapsulated in three general steps:

1. Specify a probability model for unknown parameter values that includes some prior knowledge about the parameters if available.
2. Update knowledge about the unknown parameters by conditioning this probability model on observed data.
3. Evaluate the fit of the model to the data and the sensitivity of the conclusions to the assumptions.

The second step constitutes the core of this process and is accomplished through Bayes' law:

$$\text{posterior probability} \propto \text{prior probability} \times \text{likelihood function}$$

$$\pi(\theta|\mathbf{D}) = \frac{p(\theta)L(\theta|\mathbf{D})}{\int_{\Theta} p(\theta)L(\theta|\mathbf{D})\,d\theta}$$

$$\propto p(\theta)L(\theta|\mathbf{D}),$$

where $\mathbf{D}$ is a generic symbol denoting the observed data at hand. A consequence is that $\pi(\theta|\mathbf{D})$ is a model summary that obviously retains its distributional sense. This is useful because it allows a more general look at what the model is asserting about parameter location and scale. It also pushes one away from simply describing this posterior with a point estimate and standard error for each parameter since this could miss some of the important features of the posterior shape. These additional features can include multimodality, skewness, and flat regions.

The Bayesian reporting mechanisms include the credible interval (computed exactly like the non-Bayesian confidence interval) and the highest posterior density (HPD) interval. The HPD interval contains the $100(1-\alpha)\%$ highest posterior density and therefore meets the criteria $C = \{\theta : \pi(\theta|\mathbf{x}) \geq k\}$, where k is the largest number assuring that $1 - \alpha = \int_{\theta:\pi(\theta|\mathbf{x})>k} \pi(\theta|\mathbf{x})\,d\theta$. This is the region where the probability that θ is in the region is maximized at $1-\alpha$, regardless of modality.

Bayesian statistical methods have some distinct advantages over conventional approaches in modeling social science data (Poirer 1988; Western 1998, 1999), including overt expression of model assumptions, an exclusive focus on probability-based statements, direct and systematic incorporation of prior knowledge, and the ability to "update" inferences as new data are observed. Standard Bayesian

statistical references include Box and Tiao (1973), Berger (1985), Bernardo and Smith (1994), and Robert (2001).

Our solution to the noninvertible Hessian problem is technically not at all Bayesian since there is no stipulation of priors and no treatment of posteriors as general conditional distributions in this Bayesian sense. We do, however, use this distributional treatment as an interim process since the importance sampling step samples from the difficult posterior as a complete distribution. Since the point estimate and subsequent standard errors are reported, it is essentially back to a likelihoodist result in summary. The key point from this discussion is that researchers do not need to subscribe to the Bayesian inference paradigm to find our techniques useful.

We next describe in substantive terms what is "wrong" with a Hessian that is noninvertible (Section 6.5), describe how we create a pseudovariance matrix (in Section 6.7), with algorithmic details and numerical examples, outline the concept of importance resampling to compute the mean and variance (in Section 6.9). We give our alternative procedure in Section 6.11.1, an empirical example (Section 6.10), and other possible approaches (in Section 6.11).

6.5 PROBLEM IN DETAIL: NONINVERTIBLE HESSIANS

Given a joint probability density $f(\mathbf{y}|\boldsymbol{\theta})$ for an $n \times 1$ observed data vector $\mathbf{y}$ and unknown $p \times 1$ parameter vector $\boldsymbol{\theta}$, denote the $n \times p$ matrix of first derivatives with respect to $\boldsymbol{\theta}$ as

$$g(\boldsymbol{\theta}|\mathbf{y}) = \partial \ln[f(\mathbf{y}|\boldsymbol{\theta})]/\partial\boldsymbol{\theta},$$

and the $p \times p$ matrix of second derivatives as

$$\mathbf{H} = \mathbf{H}(\boldsymbol{\theta}|\mathbf{y}) = \partial^2 \ln[f(\mathbf{y}|\boldsymbol{\theta})]/\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'.$$

Then the Hessian is $\mathbf{H}$, normally considered to be the estimate

$$E[g(\boldsymbol{\theta}|\mathbf{y})g(\boldsymbol{\theta}|\mathbf{y})'] = E[\mathbf{H}(\boldsymbol{\theta}|\mathbf{y})].$$

The standard maximum likelihood or maximum posterior estimate, which we denote as $\hat{\boldsymbol{\theta}}$, is obtained by setting $g(\boldsymbol{\theta}|\mathbf{y})$ 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. This occurs the vast majority of the time.

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

because 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, where the intuition is clearest but where 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 where 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, $1/\mathbf{H} \to \infty$ as $\mathbf{H} \to 0$).

The second problem occurs when the "mode" identified by the maximization algorithm is at the bottom of a valley instead of the top of a hill [$\mathbf{g}(\boldsymbol{\theta}|y)$ 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*, where 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 it 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 that 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 exist 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. Observe that $(-\mathbf{H})^{-1}$ is a positive definite matrix if for any nonzero $p \times 1$ vector $\mathbf{x}$, $\mathbf{x}'(-\mathbf{H})^{-1}\mathbf{x} > 0$. Nonpositive definiteness occurs in simple cases either because the variance is negative or the correlations are exactly -1 or 1.

6.6 GENERALIZED INVERSE/GENERALIZED CHOLESKY SOLUTION

The alternative developed here uses a generalized inverse, then a generalized Cholesky decomposition [if necessary when the generalized inverse of $(-\mathbf{H})$ is not positive definite], and subsequent refinement with importance sampling. The generalized inverse is produced by changing the parts of $-\mathbf{H}$ that get mapped to the variances so that they are no longer infinities. The generalized Cholesky adjusts inappropriate terms that would get mapped to the correlations (by slightly increasing variances in their denominator) to keep them within the required range of $[-1, 1]$. So the pseudovariance matrix is calculated as $\mathbf{V}'\mathbf{V}$, where $\mathbf{V} = \text{GCHOL}(\mathbf{H}^-)$, GCHOL$(\cdot)$ is the generalized Cholesky, and $\mathbf{H}^-$ is the generalized inverse of the Hessian.

The result of this process is a pseudovariance matrix that is in most cases well conditioned in that it is not nearly singular. Actually, this generalized inverse/generalized Cholesky approach is closely related to, but distinct from, the quasi-Newton *Davidson–Fletcher–Powell* (DFP) *method.* The difference is that the DFP method uses iterative differences to converge on an estimate of the negative inverse of a nonpositive definite Hessian. [See Greene (2003) for details.] However, the purpose of the DFP method is computational rather than statistical and therefore does not include our importance sampling step. Note that this method includes a default such that if the Hessian is really invertible, the pseudovariance matrix is the usual inverse of the negative Hessian.

6.7 GENERALIZED INVERSE

The literature on the theory and application of the generalized inverse is vast and spans several fields. Here we summarize some of the fundamental principles. [See Harville (1997) for further details.] The procedure begins with a generalized inverse procedure to address singularity in the $-\mathbf{H}$ matrix. This process resembles a standard matrix inversion to the greatest extent possible. The standard inverse $\mathbf{A}^{-1}$ of $\mathbf{A}$ meets five well-known conditions:

1. $\mathbf{H}\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}$
2. $\mathbf{A}^{-1}\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}$
3. $(\mathbf{A}\mathbf{A}^{-1})' = \mathbf{A}^{-1}\mathbf{A}$
4. $(\mathbf{A}^{-1}\mathbf{A}) = \mathbf{A}\mathbf{A}^{-1}$
5. $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$

(where conditions 1 to 4 are implied by condition 5). However, the *Moore–Penrose generalized inverse matrix*, $\mathbf{A}^-$ of $\mathbf{A}$, meets only the first four conditions listed above. Any matrix, $\mathbf{A}$, can be decomposed as

$$\underset{(p\times q)}{\mathbf{A}} = \underset{(p\times p)}{\mathbf{L}}\ \underset{(p\times q)}{\mathbb{D}}\ \underset{(q\times q)}{\mathbf{U}} \quad \text{where} \quad \mathbb{D} = \begin{bmatrix} \mathbf{D}_{r\times r} & 0 \\ 0 & 0 \end{bmatrix}, \tag{6.1}$$

and both $\mathbf{L}$ (lower triangular) and $\mathbf{U}$ (upper triangular) are nonsingular (even given a singular $\mathbf{A}$). The diagonal matrix $\mathbf{D}_{r\times r}$ has dimension and rank r corresponding to the rank of $\mathbf{A}$. When $\mathbf{A}$ is nonnegative definite and symmetric, the diagonals of $\mathbf{D}_{r\times r}$ are the eigenvalues of $\mathbf{A}$. If $\mathbf{A}$ is nonsingular, positive definite, and symmetric, as in the case of a proper invertible Hessian, $\mathbf{D}_{r\times r} = \mathbb{D}$ (i.e., $r = q$) and $\mathbf{A} = \mathbf{L}\mathbb{D}\mathbf{L}'$. The matrices $\mathbf{L}$, $\mathbb{D}$, and $\mathbf{U}$ are all nonunique unless $\mathbf{A}$ is nonsingular.

By rearranging (6.1) we can diagonalize any matrix as

$$\mathbb{D} = \mathbf{L}^{-1}\mathbf{A}\mathbf{U}^{-1} = \begin{bmatrix} \mathbf{D}_{r\times r} & 0 \\ 0 & 0 \end{bmatrix}. \tag{6.2}$$

Now define a new matrix, $\mathbb{D}^-$, created by taking the inverses of the nonzero (diagonal) elements of $\mathbb{D}$:

$$\mathbb{D}^- = \begin{bmatrix} \mathbf{D}_{r\times r}^- & 0 \\ 0 & 0 \end{bmatrix}. \tag{6.3}$$

If $\mathbb{D}\mathbb{D}^- = \mathbf{I}_{q\times q}$, we could say that $\mathbb{D}^-$ is *the* inverse of $\mathbb{D}$. However, this is not true:

$$\mathbb{D}\mathbb{D}^- = \begin{bmatrix} \mathbf{D}_{r\times r} & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \mathbf{D}_{r\times r}^- & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \mathbf{1} & 0 \\ 0 & 0 \end{bmatrix}.$$

Instead, we notice that

$$\mathbb{D}\mathbb{D}^-\mathbb{D} = \begin{bmatrix} \mathbf{1} & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \mathbf{D}_{r\times r} & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \mathbf{D}_{r\times r} & 0 \\ 0 & 0 \end{bmatrix} = \mathbb{D}.$$

So $\mathbb{D}^-$ is *a* generalized inverse of $\mathbb{D}$ because of the extra structure required. Note that this is *a* generalized inverse, not *the* generalized inverse, since the matrices on the right side of (6.1) are nonunique. By rearranging (6.1) and using (6.3) we can define a new $q \times p$ matrix: $\mathbf{G} = \mathbf{U}^{-1}\mathbb{D}^-\mathbf{L}^{-1}$. The importance of the *generalized inverse* matrix $\mathbf{G}$ is revealed in the following theorem.[3]

Theorem. (Moore 1920). $\mathbf{G}$ is a generalized inverse of $\mathbf{A}$ since $\mathbf{AGA} = \mathbf{A}$.

The new matrix $\mathbf{G}$ necessarily has rank r since the product rule states that the result has rank less than or equal to the minimum of the rank of the factors, and $\mathbf{AGA} = \mathbf{A}$ requires that $\mathbf{A}$ must have rank less than or equal to the lowest rank of itself or $\mathbf{G}$. Although $\mathbf{G}$ has infinitely many definitions that satisfy the Theorem, any one of them will do for our purposes: for example, in linear regression, the

[3]The generalized inverse is also sometimes referred to as the *conditional inverse, pseudo inverse,* and *g-inverse.*

fitted values, defined as $\mathbf{XGX'Y}$, with $\mathbf{G}$ as the generalized inverse of $\mathbf{X'X}$, $\mathbf{X}$ as a matrix of explanatory variables, and $\mathbf{Y}$ as the outcome variable, are invariant to the definition of $\mathbf{G}$. In addition, we use our pseudovariance only as a first approximation to the surface of the true posterior, and we will improve it in our importance resampling stage. Note, in addition, that $\mathbf{AG}$ is always idempotent [$\mathbf{GAGA} = \mathbf{G(AGA)} = \mathbf{GA}$], and $\text{rank}(\mathbf{AG}) = \text{rank}(\mathbf{A})$. These results hold whether or not $\mathbf{A}$ is singular.

Moore (1920) and (apparently unaware of Moore's work) Penrose (1955) reduced the infinity of generalized inverses to the one unique solution given above by imposing four reasonable algebraic constraints, all met by the standard inverse. This G matrix is unique if the following hold:

1. *General condition:* $\mathbf{AGA} = \mathbf{A}$
2. *Reflexive condition:* $\mathbf{GAG} = \mathbf{G}$
3. *Normalized condition:* $(\mathbf{AG})' = \mathbf{GA}$
4. *Reverse normalized condition:* $(\mathbf{GA})' = \mathbf{AG}$

The proof is lengthy, and we refer the interested reader to Penrose (1955). There is a vast literature on generalized inverses that meet some subset of the Moore–Penrose condition. A matrix that satisfies the first two conditions is called a *reflexive* or *weak generalized inverse* and is order dependent. A matrix that satisfies the first three conditions is called a *normalized generalized inverse*. A matrix that satisfies the first and fourth conditions is called a *minimum norm generalized inverse*.

Because the properties of the Moore–Penrose generalized inverse are intuitively desirable, and because of the invariance of important statistical results to the choice of generalized inverse, we follow standard statistical practice by using this form from now on. The implementations of the generalized inverse in Gauss and Splus are both the Moore–Penrose version.

The Moore–Penrose generalized inverse is also easy to calculate using QR factorization. QR factorization takes the input matrix, $\mathbf{A}$, and factors it into the product of an orthogonal matrix, $\mathbf{Q}$, and a matrix, $\mathbf{R}$, which has a triangular leading square matrix ($\mathbf{r}$) followed by rows of zeros corresponding to the difference in rank and dimension in $\mathbf{A}$:

$$\mathbf{A} = \begin{bmatrix} \mathbf{r} \\ \mathbf{0} \end{bmatrix}.$$

This factorization is implemented in virtually every professional-level statistical package. The Moore–Penrose generalized inverse is produced by

$$\mathbf{G} = \begin{bmatrix} \mathbf{r}^{-1}\mathbf{0} \end{bmatrix} \mathbf{Q}',$$

where $\mathbf{0}$ is the transpose of the zeros' portion of the $\mathbf{R}$ matrix required for conformability.

6.7.1 Numerical Examples of the Generalized Inverse

As a means of motivating a simple numerical example of how the generalized inverse works, we develop a brief application to the linear model where the $\mathbf{X}'\mathbf{X}$ matrix is noninvertible because $\mathbf{X}$ is singular. In this context, the generalized inverse provides a solution to the normal equations (Campbell and Meyer 1979, p. 94), and both the fitted values of $\mathbf{Y}$ and the residual error variance are invariant to the choice of $\mathbf{G}$ (Searle 1971, pp. 169–71). We use the Moore–Penrose generalized inverse.

Let

$$\mathbf{X} = \begin{bmatrix} 5 & 2 & 5 \\ 2 & 1 & 2 \\ 3 & 2 & 3 \\ 2.95 & 1 & 3 \end{bmatrix} \qquad \mathbf{Y} = \begin{bmatrix} 9 \\ 11 \\ -5 \\ -2 \end{bmatrix}$$

(Our omission of the constant term makes the numerical calculations cleaner but is not material to our points.) Applying the least squares model to these data ($\mathbf{X}$ is of full rank) yields the coefficient vector

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = (222.22, -11.89, -215.22)',$$

fitted values,

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{b}} = (11.22, 2.11, -2.78, -2.00)',$$

and variance matrix

$$\sum = \begin{bmatrix} 57283.95 & -1580.25 & -56395.06 \\ -1580.25 & 187.65 & 1491.36 \\ -56395.06 & 1491.36 & 55550.62 \end{bmatrix}.$$

What we call the *standardized correlation matrix*, a correlation matrix with standard deviations on the diagonal, is then

$$\mathbf{C}_s = \begin{bmatrix} 239.34 & -0.48 & -0.99 \\ -0.48 & 13.69 & 0.46 \\ -0.99 & 0.46 & 235.69 \end{bmatrix}.$$

Now suppose that we have a matrix of explanatory effects that is identical to $\mathbf{X}$ except that we have changed the bottom left number from 2.95 to 2.99:

$$\mathbf{X}_2 = \begin{bmatrix} 5 & 2 & 5 \\ 2 & 1 & 2 \\ 3 & 2 & 3 \\ 2.99 & 1 & 3 \end{bmatrix}.$$

Using the same $\mathbf{Y}$ outcome vector and applying the same least squares calculation now gives

$$\hat{\mathbf{b}}_2 = (1111.11, -11.89, -1104.11)'$$

and

$$\hat{\mathbf{Y}} = (11.22, 2.11, -2.78, -2.00)'.$$

However, the variance–covariance matrix reacts sharply to the movement toward singularity as seen in the standardized correlation matrix:

$$\mathbf{C}_s = \begin{bmatrix} 1196.70 & -0.48 & -0.99 \\ -0.48 & 13.70 & 0.48 \\ -0.99 & 0.48 & 1193.00 \end{bmatrix}.$$

Indeed, if $\mathbf{X}_3 = 2.999$, $\mathbf{X}'\mathbf{X}$ is singular (with regard to precision in Gauss and Splus) and we must use the generalized inverse. This produces

$$\tilde{\mathbf{b}}_3 = \mathbf{G}\mathbf{X}'\mathbf{Y} = (1.774866, -5.762093, 1.778596)'$$

and

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{Y} = (11111.11, -11.89, -11104.11)'.$$

The resulting pseudovariance matrix (calculated now from $\mathbf{G}\sigma^2$) produces larger standard deviations for the first and third explanatory variables, reflecting greater uncertainty, again displayed as a standardized correlation matrix:

$$\mathbf{C}_s = \begin{bmatrix} 11967.0327987 & -0.4822391 & -0.9999999 \\ -0.4822391 & 13.698 & 0.4818444 \\ -0.9999999 & 0.4818444 & 11963.3201730 \end{bmatrix}.$$

6.8 GENERALIZED CHOLESKY DECOMPOSITION

We now describe the classic Cholesky decomposition and recent generalizations designed to handle nonpositive 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 principal leading submatrices. The Cholesky matrix is defined as $\mathbf{V}$ in the decomposition $\mathbf{C} = \mathbf{V}'\mathbf{V}$. We thus construct our pseudovariance matrix as $\mathbf{V}'\mathbf{V}$, where $\mathbf{V} = \text{GCHOL}(\mathbf{H}^-)$, $\text{GCHOL}(\cdot)$ is the generalized Cholesky described below, and $\mathbf{H}^-$ is the Moore–Penrose generalized inverse of the Hessian.

6.8.1 Standard Algorithm

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

$$\underset{(k\times k)}{\mathbf{C}} = \underset{(k\times k)}{\mathbf{L}}\ \underset{(k\times k)}{\mathbf{D}}\ \underset{(k\times k)}{\mathbf{L}'}. \tag{6.4}$$

The basic Cholesky procedure is a one-pass algorithm that generates two output matrices which 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:

$$\mathbf{D}_{j,j} = \mathbf{C}_{j,j} - \sum_{\ell=1}^{j-1} \mathbf{L}_{j,\ell}^2 \mathbf{D}_{\ell,\ell} \tag{6.5}$$

and

$$\mathbf{L}_{i,j} = \left[\mathbf{C}_{i,j} - \sum_{\ell=1}^{j-1} \mathbf{L}_{j,\ell}\mathbf{L}_{i,\ell}\mathbf{D}_{\ell,\ell} \right] \Big/ \mathbf{D}_{j,j}, \qquad i = j+1, \ldots, k, \tag{6.6}$$

where $\mathbf{D}$ is a positive diagonal matrix so that on completion of the algorithm, its square root 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 nonpositive definite input matrices. Singular matrices cause a divide-by-zero problem in (6.6), and nonpositive definite matrices cause the sum in (6.5) to be greater than $\mathbf{C}_{j,j}$, 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, Matthews and Davies 1971; Gill et al. 1974).

6.8.2 Gill–Murray Cholesky Factorization

Gill and Murray (1974) introduced, and Gill et al. (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 $\mathbf{E} = -(\lambda_1 + \epsilon)I$, where ϵ is a small positive increment. However, this approach (implemented in various computer programs, such as the Gauss "maxlike" 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 (6.5) and (6.6) in matrix notation. The jth submatrix of its application at the jth step is

$$\mathbf{C}_j = \begin{bmatrix} c_{j,j} & \mathbf{c}'_j \\ \mathbf{c}_j & \mathbf{C}_{j+1} \end{bmatrix}, \tag{6.7}$$

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

$$\mathbf{C}^*_{j+1} = \mathbf{C}_{j+1} - \frac{\mathbf{c}_j \mathbf{c}'_j}{L^2_{j,j}}. \tag{6.8}$$

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 trade-offs 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, subsequent diagonal values are greatly increased through the operation of (6.8). Conversely, we don't 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

$$\|\mathbf{c}_j\|_\infty \beta^{-2} - c_{j,j} \leqslant \delta_j \tag{6.9}$$

where

$$\beta = \max \begin{cases} \max(\text{diag}(\mathbf{C})) \\ \max(not\text{diag}(\mathbf{C}))\sqrt{k^2 - 1} \\ \epsilon_m, \end{cases}$$

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*) (see Chapter 4). This algorithm always produces a factorization and has the advantage of not modifying already positive definite $\mathbf{C}$ matrices. However, the bounds in (6.9) have been shown to be nonoptimal and thus provide $\mathbf{C}+\mathbf{E}$ that is again farther from $\mathbf{C}$ than necessary.

6.8.3 Schnabel–Eskow Cholesky Factorization

Schnabel and Eskow (1990) improve on the $\mathbf{C}+\mathbf{E}$ procedure of Gill and Murray 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. Suppose that $\mathbf{C} \in \mathbb{R}^k$ with eigenvalues $\lambda_1, \ldots, \lambda_k$, and define the ith Gerschgorin bound as

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

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

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

$$\delta_j = \max\left(\epsilon_{\mathrm{m}}, \max_i(\mathsf{G}_i(\text{lower}))\right). \tag{6.10}$$

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, so a smaller quantity is subtracted in (6.8). Adding this condition to (6.10) and protecting the algorithm from problems associated with the machine epsilon produces the following determination of the additional amount in $L_{j,j} = \sqrt{c_{j,j} + \delta_j}$:

$$\delta_j = \max\left(\epsilon_{\mathrm{m}}, -\mathbf{C}_{j,j} + \max(\|\mathbf{a}_j\|, (\epsilon_{\mathrm{m}})^{1/3} \max(\mathrm{diag}(\mathbf{C})), \mathbf{E}_{j-1,j-1}\right). \tag{6.11}$$

The algorithm follows the same steps as that of Gill–Murray except that the determination of δ_j is done by (6.11). 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 $\|\mathbf{E}\|_\infty$.

6.8.4 Numerical Examples of the Generalized Cholesky Decomposition

Suppose that we have the positive definite matrix

$$\Sigma_1 = \begin{bmatrix} 2 & 0 & 2.4 \\ 0 & 2 & 0 \\ 2.4 & 0 & 3 \end{bmatrix}.$$

This matrix has the Cholesky decomposition:

$$\text{chol}(\Sigma_1) = \begin{bmatrix} 1.41 & 0 & 1.70 \\ 0 & 1.41 & 0 \\ 0 & 0 & 0.35 \end{bmatrix}.$$

Now suppose that we have a very similar but nonpositive definite matrix that requires the generalized Cholesky algorithm. The only change from the input matrix above is that the values on the corners have been changed from 2.4 to 2.5:

$$\Sigma_2 = \begin{bmatrix} 2 & 0 & 2.5 \\ 0 & 2 & 0 \\ 2.5 & 0 & 3 \end{bmatrix}.$$

This matrix has the generalized Cholesky decomposition

$$\text{GCHOL}(\Sigma_2) = \begin{bmatrix} 1.41 & 0 & 1.768 \\ 0 & 1.41 & 0 \\ 0 & 0 & 0.004 \end{bmatrix}$$

So the generalized Cholesky produces a very small change here so as to obtain a positive definite input matrix. This reflects the fact that this nonpositive definite matrix is actually very close to being positive definite. Now suppose that we create a matrix that is deliberately very far from positive definite status:

$$\Sigma_3 = \begin{bmatrix} 2 & 0 & 10 \\ 0 & 2 & 0 \\ 10 & 0 & 3 \end{bmatrix}$$

This matrix has the Cholesky decomposition

$$\text{GCHOL}(\Sigma_3) = \begin{bmatrix} 1.41 & 0 & 7.071 \\ 0 & 1.41 & 0 \\ 0 & 0 & 0.008 \end{bmatrix}$$

The effects are particularly evident when we square the Cholesky result:

$$\text{GCHOL}(\Sigma_3)'\text{GCHOL}(\Sigma_3) = \begin{bmatrix} 2 & 0 & 10 \\ 0 & 2 & 0 \\ 10 & 0 & 50 \end{bmatrix},$$

so the diagonal of the **E** matrix is very large: [8, 6, 11].

6.9 IMPORTANCE SAMPLING AND SAMPLING IMPORTANCE RESAMPLING

The algorithm called *sampling importance resampling* (SIR) or simply *importance resampling* is a Monte Carlo simulation technique used to draw random numbers directly from an exact (finite sample) posterior distribution. The original idea comes from Rubin (1987a, pp. 192–94), but see also Wei and Tanner (1990), Tanner (1996), and Gill (2002). For social science applications, see King (1997) and King et al. (1998). The primary requirement for effective implementation of the algorithm is the specification of a reasonable approximation to the exact (but inconvenient) posterior. If this requirement is not met, the procedure can take excessively long to be practical or can miss features of the posterior distribution. Also, while the approximating distribution is required, it need not be normalized. So there is a lot of flexibility in this choice.

A common choice for the approximation distribution, based on flexibility and convenience, is the multivariate normal distribution. Sometimes the multivariate t distribution is substituted when the sample size is small or there is general concern about the tails. Using the normal or t-distribution should be relatively uncontroversial for our purposes here, since the algorithm in applied cases 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 in the original matrix calculation. So this first approximation retains as many of the assumptions of the original model as possible. However, other distributions can easily be used if that seems necessary.

Using either the normal or t-distribution, the mean is set at $\hat{\boldsymbol{\theta}}$, the vector of maximum likelihood or maximum posterior estimates. Recall that this vector of point estimates was reported by the computer program that failed before it failed the variance calculation. For the normal this is simple: Set the variance equal to our pseudovariance matrix. For the t, the pseudovariance is required that there be an adjustment by the degrees of freedom to yield the appropriate scatter matrix.

6.9.1 Algorithm Details

The basic 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 for those farther away. The main difficulty is in determining an approximation distribution that somewhat resembles the difficult posterior. So we use normal or t-distributions centered at the posterior mean and the pseudovariance matrix 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.

Denote $\tilde{\boldsymbol{\theta}}$ as one random draw of $\boldsymbol{\theta}$ from the approximating distribution, and use it to compute the *importance ratio*: the ratio of the posterior $P(\cdot)$ to the normal approximation, where both are evaluated at $\tilde{\boldsymbol{\theta}}$: $P(\tilde{\boldsymbol{\theta}}|y)/N(\tilde{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}}, \mathbf{V}'\mathbf{V})$. Then keep $\tilde{\boldsymbol{\theta}}$, as if it where a random draw from the posterior, with probability

proportional to this ratio. The procedure is repeated until the desired (generally large) number of simulations have been accepted.

Suppose that we wish to obtain the marginal distribution for some parameter θ_1 from a joint distribution: $f(\theta_1, \theta_2|\mathbf{X})$. If we actually knew the parametric form for this joint distribution, it would be straightforward to integrate out the second parameter analytically over its support as shown in basic texts:

$$f(\theta_1|\mathbf{X}) = \int f(\theta_1, \theta_2|\mathbf{X})\, d\theta_2. \tag{6.12}$$

However, in many settings this is not possible, and more involved numerical approximations are required. Suppose that we could posit a *normalized* conditional posterior approximation density of θ_2, $\hat{f}(\theta_2|\theta_1, \mathbf{X})$, that would often be given a normal or t form, as mentioned above. The trick that this approximation gives 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

$$\begin{aligned} f(\theta_1|\mathbf{X}) &= \int f(\theta_1, \theta_2|\mathbf{X})\, d\theta_2 \\ &= \int \frac{f(\theta_1, \theta_2|\mathbf{X})}{\hat{f}(\theta_2|\theta_1, \mathbf{X})} \hat{f}(\theta_2|\theta_1, \mathbf{X})\, d\theta_2 \\ &= E_{\theta_2}\left[\frac{f(\theta_1, \theta_2|\mathbf{X})}{\hat{f}(\theta_2|\theta_1, \mathbf{X})}\right]. \end{aligned} \tag{6.13}$$

The fraction

$$\frac{f(\theta_1, \theta_2|\mathbf{X})}{\hat{f}(\theta_2|\theta_1, \mathbf{X})}, \tag{6.14}$$

called the *importance weight*, determines the probability of accepting sampled values of θ_2. This setup provides a rather simple procedure to obtain the estimate of $f(\theta_1|\mathbf{X})$. The steps are summarized as follows:

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 (6.13) to obtain $f(\theta_1^{(i)}|\mathbf{X})$ by taking the means of the N ratios just calculated.

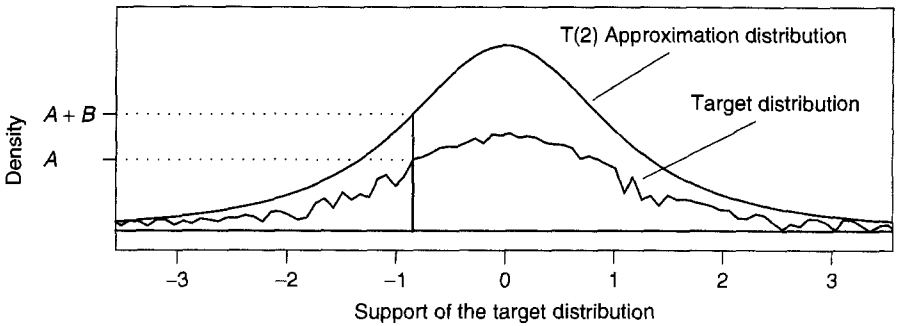

Fig. 6.1 Importance sampling illustration.

The user controls the level of accuracy of this estimate by increasing the granularity of the grid and the number of draws per position on that grid. In addition, this procedure can also be used to perform standard numerical integration, provided that a suitable normalized approximation function can be found (albeit somewhat less efficiently than standard algorithms; see Gill 2002, Chap. 8). These considerations make importance sampling a very useful and very common tool in applied mathematics.

The importance sampling algorithm is illustrated in Figure 6.1, where the importance ratio calculation is shown for an arbitrary point along the x-axis. The approximation distribution is t with two degrees of freedom and the target distribution is a contrived problematic form. The point indicated is accepted into the sample with probability $A/(A + B)$, which can be viewed as the quality of the approximation at this point.

6.9.2 SIR Output

The resulting simulations can easily be displayed with a histogram to give the full marginal distribution of a quantity interest (see Tanner 1996; King et al. 2000) or just a parameter of the model. Taking the sample average and sample standard deviation of the simulations can be used to compute the mean and standard error or full variance matrix of the parameters if these common summaries are desired. The computed variance matrix of the means will almost always be positive definite, as long as enough simulations are drawn such that there are sufficient elements of the mean vector and variance matrix (normally, one would want at least one order of magnitude more than that number). It is also possible, however, 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 it is already on the variance–covariance metric (inverted), and the only problem is that some of the correlations will be exactly 1 or -1, which can actually be very informative substantively, and standard errors, for example, will still be available.

One diagnostic often used to detect a failure of importance resampling is when many candidate values of $\tilde{\theta}$ are rejected due to low values of the importance ratio. In this case the procedure will take a very long time, and to be useful a better approximation is certainly needed. Here, the long run time indicates a problem, and letting it run longer may eventually yield sufficient sample size. However, this can be very frustrating and time consuming from a practical point of view. There is a danger here, though: if the approximation distribution entirely misses a range of values of θ that have 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. Therefore, one check is to compute a very large number of simulations with an artificially large variance matrix, such as the pseudovariance matrix multiplied by a positive factor, which we label F. This works since obviously the coverage is more diffuse. Like all related simulation procedures, it is impossible to cover the full continuum of values that θ can take, and the procedure can miss subtle features such as pinholes in the surface, very sharp ridges, or other eccentricities.

6.9.3 Relevance to the Generalized Process

The importance sampling procedure cannot be relied on *completely* in our case, since we know that the likelihood surface is nonstandard by definition of the problem. The normal approximation requires an invertible Hessian. The key to extracting at least some information from the Hessian via the derived pseudovariance 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 that they are localized, some parameters effectively have infinite standard errors, or pairs of parameters have perfect correlations. The suggestion here 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 profile 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.

To make the normal or t-approximation work more efficiently, it is generally advisable to reparameterize so that the parameters are unbounded and approximately symmetric. This strategy is pretty standard in this literature and normally makes the maximization routine work better. This can be broadly used; for example, instead of estimating $\sigma^2 > 0$ as a variance parameter directly, one could estimate γ, where $\sigma^2 = e^{\gamma}$, since γ can take on any real number.

6.10 PUBLIC POLICY ANALYSIS EXAMPLE

This real-data example looks at public policy data focused on poverty and its potential causes, measured by state at the county level (FIPS). The data highlight a common and disturbing problem in empirical model fitting. Suppose that a researcher seeks to apply a given model specification to multiple datasets for the purpose of comparison: comparing models across 50 U.S. states, 25 OECD

countries, 15 EU countries, or even the same unit in some time series. Normally, if the Hessian fails to invert for a small number of the cases, generally the researcher respecifies the model for nonsubstantive, technical reasons, even though some other specification may be preferred for substantive reasons. If the researcher respecifies only the problem cases, differences among the results are contaminated by investigator-induced omitted variable bias. Otherwise, all equations are respecified in an effort to get comparable results, in which case the statistical analyses differs from the original substantive question posed. Obviously, neither approach is satisfactory from a substantive research perspective.

It is important to note, prior to giving the empirical example, *that we do not extract, fabricate, or simulate information from the likelihood function that does not exist*. That is, the culpable dimension will be given an infinite variance posterior, reflecting a complete lack of information about its form. What the algorithm does accomplish is the recovery of information on the other dimensions that otherwise would not be available to researchers. Therefore, a model that would have been dismissed as nonidentified for purely data reasons can now be partially recovered.

6.10.1 Texas

The example here uses data from the 1989 county-level economic and demographic survey for all 2276 nonmetropolitan U.S. counties ("ERS Typology") organized hierarchically by state such that each state is a separate unit of analysis with counties as cases. The U.S. Bureau of the Census, U.S. Department of Agriculture, and state agencies collect these data to provide policy-oriented information about conditions leading to high levels of rural poverty. The dichotomous outcome variable indicates whether 20% or more of the county's residents live in poverty (a standard measure in this field). The specification includes the following explanatory variables:

- `Govt`: a dichotomous factor indicating whether various government activities contributed a weighted annual average of 25% or more labor and proprietor income over the three preceding years.
- `Service`: a dichotomous factor indicating whether service-sector activities contributed a weighted annual average of 50% or more labor and proprietor income over the three preceding years.
- `Federal`: a dichotomous factor indicating whether federally owned lands make up 30% 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% or more of total personal income over the preceding three years.
- `Population`: the log of the county population total for 1989.
- `Black`: the proportion of black residents in the county.
- `Latino`: the proportion of Latino residents in the county.

This model provides the results given in Table 6.1.

Table 6.1 Logit Regression Model: Nonsingular Hessian, Texas

	Standard Results		Without `Federal`		Importance Sampling	
Parameter	Est.	Std. Err.	Est.	Std. Err.	Est.	Std. Err.
`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

A key substantive question is whether the black fraction 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 they are succeeding in a racially fair manner or whether there is more poverty in counties with larger fractions of African Americans. That is, whether the hypothesized effect 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 standard logistic regression model, so $P(Y_i = 1|X_i) = [1 + \exp(X_i\beta)]^{-1}$, where X_i is a vector of all our explanatory variables for case i. Using this specification, 43 of the U.S. states produce invertible Hessians and therefore 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 state models. From this 43:7 dichotomy, a matched pair of similar states is chosen for discussion here, where one case produces a (barely) invertible Hessian with the model specification (Texas) and the other is noninvertible (Florida). 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 the black fraction predicts poverty in similar fashions.

The logit model for Texas counties ($n = 196$) produces the results in the first pair of columns in Table 6.1. The coefficient on the black fraction is very large, and statistically reliable, thus supporting the racial bias hypothesis. It turns out that the variable `Federal` is problematic in these models and as noted below, actually prevents estimation using the Florida data. The second pair of columns reestimates the Texas model without the `Federal` variable, and the results for the black fraction (and the other variables) are mostly 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 without the need for a pseudovariance matrix calculation. The means here are very close to the modes, and the standard errors

in the two cases are very close as well, so the importance resampling in this (invertible) case did not generate important differences.

Below is the Hessian from this estimation, which supports the claim that the variable `Federal` is a problematic component of the model. Note the zeros and very small values in the fourth row and column of H.

$$\mathbf{H} = \begin{vmatrix} 0.13907100 & 0.00971597 & 0.01565632 & 0.00000000 \\ 0.00971597 & 0.00971643 & 0.00000000 & 0.00000000 \\ 0.01565632 & 0.00000000 & 0.01594209 & 0.00000000 \\ 0.00000000 & 0.00000000 & 0.00000000 & 0.00000003 \\ 0.01165964 & 0.00022741 & 0.00305369 & 0.00000000 \\ 1.27113747 & 0.09510282 & 0.14976776 & 0.00000044 \\ 0.01021141 & 0.00128841 & 0.00170421 & -0.00000001 \\ 0.03364064 & 0.00211645 & 0.00246767 & 0.00000000 \end{vmatrix}$$

$$\begin{vmatrix} 0.01165964 & 1.27113747 & 0.01021141 & 0.03364064 \\ 0.00022741 & 0.09510282 & 0.00128841 & 0.00211645 \\ 0.00305369 & 0.14976776 & 0.00170421 & 0.00246767 \\ 0.00000000 & 0.00000044 & -0.00000001 & 0.00000000 \\ 0.01166205 & 0.10681518 & 0.00136332 & 0.00152559 \\ 0.10681518 & 11.77556446 & 0.09904505 & 0.30399224 \\ 0.00136332 & 0.09904505 & 0.00161142 & 0.00131032 \\ 0.00152559 & 0.30399224 & 0.00131032 & 0.01222711 \end{vmatrix}$$

To see this near singularity implied by this Hessian, Figure 6.2 provides a matrix of the bivariate profile contour plots for each pair of coefficients from the Texas data, with contours at 0.05, 0.15, ... , 0.95 where the 0.05 contour line bounds approximately 0.95 of the data, holding all other parameters constant at their maxima. These easy-to-compute profile 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 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 the maximum value in each profile plot, thus allowing estimation.

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 typically the case, the graphed profile contours reveal more information. In particular, the plot indicates that distribution 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 6.1, cannot reveal this additional information. In contrast, the importance resampling results reveal the richer set of information. For example, to compute the entries in the last two columns of Table 6.1, we first took many random draws of the parameters from

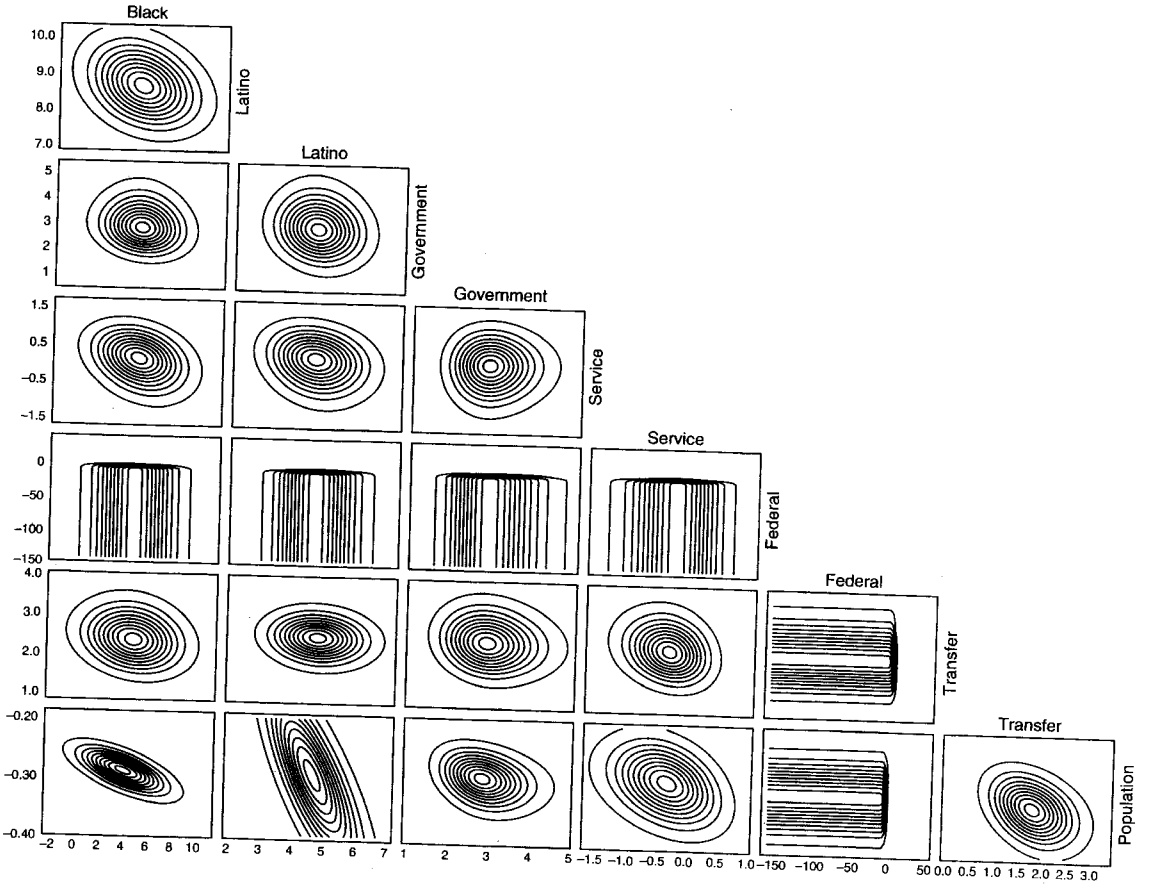

Fig. 6.2 Contourplot matrix, Texas data.

their exact posterior distribution. If instead of summarizing this information with their means and standard deviations, as in Table 6.1, we presented univariate or bivariate histograms of the draws, we would reveal all the information in Figure 6.2. In fact, the histograms would give the exact marginal distributions of interest (the full posterior, with other parameters integrated out) rather than merely the profile contours as shown in the figures, so the potential information revealed, even in this case where 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 dataset.

Note that although logit is known to have a globally concave likelihood surface in theory, actual estimates are not strictly concave, due to numerical imprecision. In the present data, the Hessian is barely invertible, making the likelihood surface sensitive to numerical imprecision. As it turns out, there are at least two local maxima on the marginal likelihood for `Federal` and thus potential attractors. 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 difference in the results here is not particularly troubling, as no reasonable analyst would place faith in either coefficient estimate for `Federal`, given the large reported standard error. Note also that `Govt` and `Service` fall below conventional significance threshold levels as well. Our primary concern with `Federal` is that it alone prevents the Florida model (Section 6.10.2) from producing conventional results.

6.10.2 Florida

We ran the same specification used in Texas for Florida (33 counties), providing the maximum likelihood parameter estimates in Table 6.2 and the following Hessian, *which is now noninvertible.* The standard errors are represented in the table with question marks since standard estimates are not available.

Table 6.2 Logit Regression Model: Singular Hessian, Florida

	Standard Results		Without `Federal`		Importance Sampling	
Parameter	Est.	Std. Err.	Est.	Std. Err.	Est.	Std. Err.
`Black`	5.86	???	5.58	5.34	5.56	2.66
`Latino`	4.08	???	3.21	8.10	3.97	2.73
`Government`	−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

$$\mathbf{H} = \begin{vmatrix} 0.13680004 & 0.04629599 & 0.01980602 & 0.00000001 \\ 0.04629599 & 0.04629442 & 0.00000000 & -0.00000004 \\ 0.01980602 & 0.00000000 & 0.01980564 & 0.00000000 \\ 0.00000001 & -0.00000004 & 0.00000000 & 0.00000000 \\ 0.05765988 & 0.03134646 & 0.01895061 & 0.00000000 \\ 1.32529504 & 0.45049457 & 0.19671280 & 0.00000000 \\ 0.02213744 & 0.00749867 & 0.00234865 & 0.00000000 \\ 0.00631444 & 0.00114495 & 0.00041155 & 0.00000002 \end{vmatrix}$$

$$\begin{vmatrix} 0.05765988 & 1.32529504 & 0.02213744 & 0.00631444 \\ 0.03134646 & 0.45049457 & 0.00749867 & 0.00114495 \\ 0.01895061 & 0.19671280 & 0.00234865 & 0.00041155 \\ 0.00000000 & 0.00000000 & 0.00000000 & 0.00000002 \\ 0.05765900 & 0.57420212 & 0.00817570 & 0.00114276 \\ 0.57420212 & 12.89475788 & 0.21458995 & 0.06208332 \\ 0.00817570 & 0.21458995 & 0.00466134 & 0.00085111 \\ 0.00114276 & 0.06208332 & 0.00085111 & 0.00088991 \end{vmatrix}$$

Consider first Figure 6.3, which provides the same type of matrix of the bivariate profile plots for each pair of coefficients for the Florida data, like Texas with contours at $0.1, 0.2, \ldots, 0.9$. The problematic profile likelihood is clearly for `Federal`, but in this case the modes are not unique, so the Hessian is not invertible. Interestingly, except for this variable, the posteriors are very well behaved and easy to summarize. If one was forced to abandon the specification at this point, this is exactly the information that would be lost forever. The loss is especially problematic when contrasted with the Texas case, for which the contours do not look a lot more informative, but we were barely able to get an estimate.

Here is the key trap. A diligent data analyst using classical procedures with our data might reason as follows:

- The Texas data clearly suggest racial bias, but no results are available in Florida with the same specification.
- Follow the textbook advice and respecify by dropping `Federal` and rerunning the model for both Texas and Florida (these results are in both Tables 6.1 and 6.2).
- Note that the new results for `black` reveal a coefficient for Florida that is only a third of the size it was in Texas and only slightly larger than its standard error.

Now the contrast with the previous results is striking: a substantial racial bias in Texas and no evidence of such in Florida. However, with this approach it is impossible to tell whether these interesting and divergent substantive results in Florida are due to omitted variable bias rather than true political and economic differences between the states.

What can our analysis 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

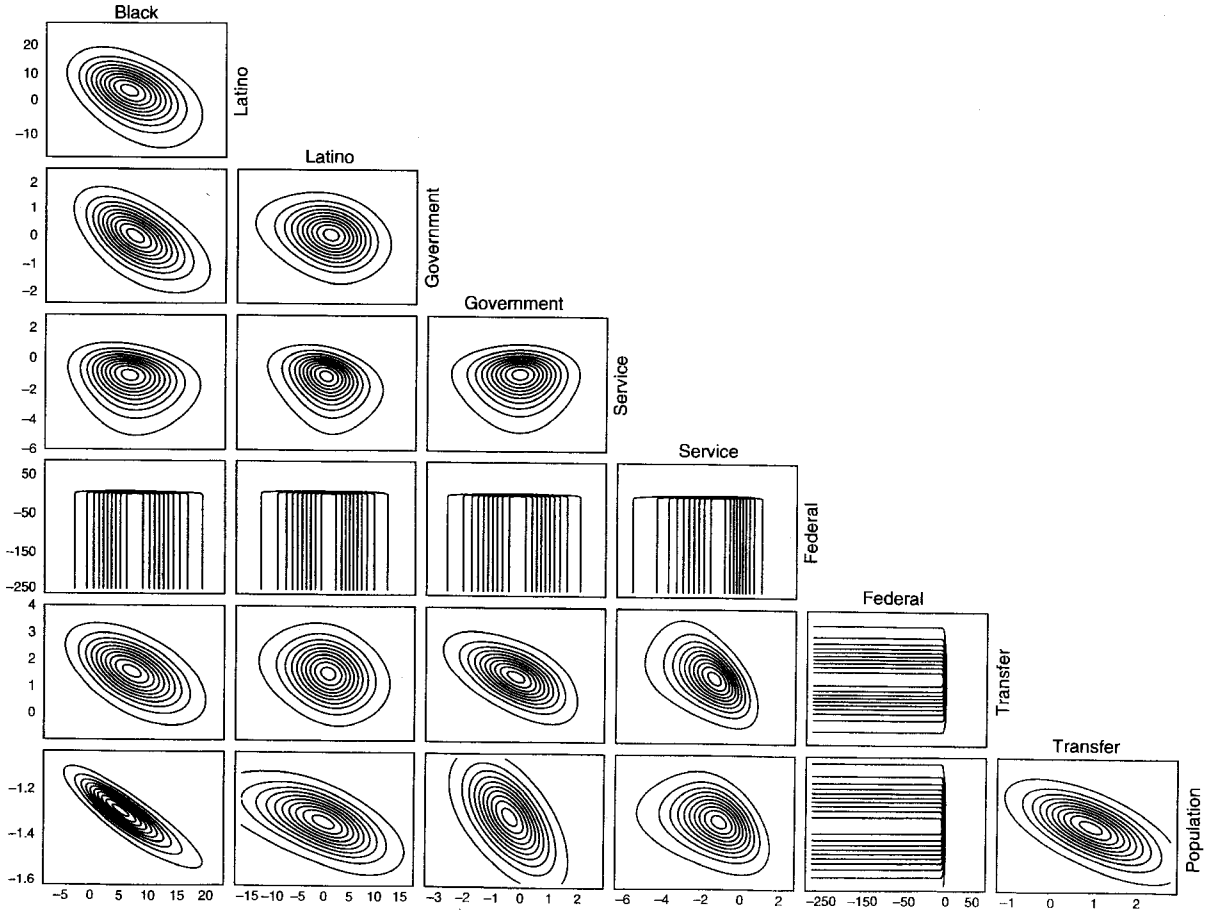

Fig. 6.3 Contourplot matrix, Florida data.

would occur by omitting the variable in the Texas data. Therefore, one can easily estimate the bias in Texas by omitting `Federal`. This is done in the second pair of columns in Table 6.1, and the results suggest that there is no bias introduced since the results are nearly unchanged from the first two columns. Although this seems like a reasonable procedure (and one that most analysts have no doubt tried at one time or another), it is of course based on the completely unverifiable assumption that the biases are the same in the two states. With the present data, this assumption is false, as our procedure now shows.

We now recover the information lost in the Florida case by first applying our generalized inverse and generalized Cholesky procedures to the singular Hessian to create a pseudovariance matrix. We then perform importance resampling using the multivariate normal, with the mode and pseudovariance matrix, as the first approximation. We use a t-distribution with three degrees of freedom as the approximation distribution so as to be as conservative as possible since we know from graphical evidence that one of the marginal distributions is problematic. The last two columns of Table 6.2 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 data and model contain no useful information about this parameter, the specification did control for `Federal`, 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 are indeed quite informative. They show that the effect of `Black` is indeed smaller in Florida than in 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.

6.11 ALTERNATIVE METHODS

6.11.1 Drawing from the Singular Normal

In this section we describe another procedure for drawing the random numbers from a different approximating density: the truncated singular normal. The key idea is to draw directly from the singular multivariate density with a noninvertible Hessian. It should be true that the generalized Cholesky procedure will work better if the underlying model is identified, but numerical problems lead to apparent nonidentification. However, the singular normal procedure will perform better when the underlying model would have a noninvertible Hessian even if one were able to run it on a computer with infinite precision.

Again consider the matrix of second derivatives, $\mathbf{H}$, along with a $k \times 1$ associated vector of maximum likelihood estimates, $\hat{\boldsymbol{\theta}}$. Again, the matrix $(-\mathbf{H})^{-1}$ does

not exist due either to nonpositive definiteness or to singularity ($r \leq k$). Suppose that one can actually 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 (Hathaway 1985; O'Leary and Rust 1986; McCullagh and Nelder 1989; Geyer 1991; Wolak 1991; Geyer and Thompson 1992; Dhrymes 1994, Sec. 5.11). 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 not include an expression for the variance–covariance matrix—only the inverse (i.e., the negative of the Hessian), which exists here. We thus decompose T as follows:

$$\begin{aligned} T &= (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\mathbf{H}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \\ &= (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\mathbf{U}'\mathbf{L}\mathbf{U}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}), \end{aligned} \tag{6.15}$$

where $\mathbf{U}'\mathbf{L}\mathbf{U}$ is the *spectral decomposition* of $\mathbf{H}$; $\text{rank}(\mathbf{H}) = r \leq k$; $\mathbf{H}$ has r non-zero eigenvalues, denoted $d_1, \ldots, d_r$; $\mathbf{U}$ 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.

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:

- $(-\mathbf{L})^{-1}$ does not always exist.
- $\mathbf{A}$ 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^* is equal to some small positive value otherwise (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

$$\begin{aligned} \mathbf{A} &\sim N(\mathbf{U}'\boldsymbol{\theta}, (-\mathbf{L}^*)^{-1}) \\ &= \prod_i N(U_i(\theta_i - [h_i + g_i]/2, -1/L_i^*). \end{aligned} \tag{6.16}$$

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 $\hat{\theta} = \mathbf{U}'\mathbf{A} + (\mathbf{h}+\mathbf{g})/2$, and accept the draw only if $\hat{\theta}$ 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 θ to any of the 2^k corners of the hyperrectangle defined by the bounds. Since by definition $\theta \in [\mathbf{g}, \mathbf{h}]$, 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 $\hat{\theta}$.

Now 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) as long as it fully encompasses the target space and the standard 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, but truncated as indicated.

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

6.11.2 Aliasing

The problem of computational misspecification and covariance calculation is well studied in the context of generalized linear models, particularly in the case of the linear model (Albert 1973). McCullagh and Nelder (1989) discuss this problem in the context of generalized linear models where specifications that introduce overlapping subspaces due to redundant information in the factors produce *intrinsic aliasing*. This occurs when a linear combination of the factors is reduced to fewer terms than the number of parameters specified. 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 factors specified. 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 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 political science applications.

6.11.3 Ridge Regression

Another well-known approach to this problem in linear modeling is *ridge regression*, which 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 very 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,b), empirical Bayes (Efron and Morris 1972; Amemiya 1985), minimax considerations (Casella 1980, 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 (in contrast to the Schnabel–Eskow method). For a very important critique, see Smith and Campbell (1980) along with the comments that follow.

6.11.4 Derivative Approach

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 $\delta\boldsymbol{\theta}$)

$$f(\mathbf{x}|\boldsymbol{\theta}) = H(\boldsymbol{\theta})\,\delta\boldsymbol{\theta}. \tag{6.17}$$

Since $H(\boldsymbol{\theta})$ is singular, a solution is available only by the generalized inverse:

$$\delta\boldsymbol{\theta} = H(\boldsymbol{\theta})^{-} f(\mathbf{x}|\boldsymbol{\theta}). \tag{6.18}$$

When there exists a parametric function of $\boldsymbol{\theta}$ that is estimable and whose first derivative is in the column space of $H(\boldsymbol{\theta})$, there exists a unique, maximum likelihood estimate of this function, $\phi(\hat{\boldsymbol{\theta}})$, with asymptotic variance–covariance matrix:

$$\phi(\hat{\boldsymbol{\theta}})H(\boldsymbol{\theta}_0)^{-}\phi(\hat{\boldsymbol{\theta}}). \tag{6.19}$$

The difficulty with this procedure 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 $H(\boldsymbol{\theta})$.

6.11.5 Bootstrapping

An additional approach is to apply bootstrapping to the regression procedure so as to produce empirical estimates of the coefficients, which can then be used to

obtain subsequent values for the standard errors. The basic procedure (Davidson and MacKinnon 1993, pp. 330–31; Efron and Tibshirani 1993, pp. 111–12) is to bootstrap from the residuals of a model where coefficients estimates are obtained but where the associated measures of uncertainty are unavailable or unreliable.

The steps for the linear model are given by 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 $\hat{\boldsymbol{\beta}}$.
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, pp. 341–43).

Applying this bootstrap procedure to the problematic Florida data where the coefficient estimates are available but 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 6.2). 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 discrepancy.

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 datasets (the $\mathbf{y}^*$, $\mathbf{X}^*$) that are very different from the original data (Hinkley 1988).

6.11.6 Respecification (Redux)

Far and away the most common way of recovering from computational problems resulting from forms of collinearity is respecification. Virtually every basic and intermediate textbook 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 (Krzanowski 1988). Although these approaches often work, they force the user to change their research question due to technical concerns. As the example in Section 6.10 shows, we should not be forced to alter our thinking about a research question as a result of computational issues.

6.12 CONCLUDING REMARKS

The purpose of this chapter is twofold. The first objective is to illuminate the central role of the Hessian in standard likelihood-based estimation. The second objective is to provide a working solution to noninvertible Hessian problems that might otherwise cause researchers to discard their substantive objectives. This method is based on some established theories, but is new as a complete method. Currently, the generalized inverse/generalized Cholesky procedure is implemented in King's EI software (<http://gking.harvard.edu/stats.shtml>), and the R and Gauss procedures are freely available at <http://www.hmdc.harvard.edu/numerical_issues/>.

So what can a frustrated practitioner do? We have given several alternatives to the standard (albeit eminently practical) recommendation to respecify. Our new method is intended to provide results even in circumstances where it is not usually possible to invert the Hessian and obtain coefficient standard errors. The usual result is that the problematic coefficient has huge posterior variance, indicating statistical unreliability. This is exactly what should happen: A model is produced and poor contributors are identified.

Although the likelihood estimation process from a given dataset may have imposing problems, the data may still contain revealing information about the question at hand. The point here is therefore to help researchers avoid giving up the question they posed originally and instead, to extract at least some of the remaining information available. The primary method we offer here is certainly not infallible, nor are the listed alternatives. Therefore, considerable care should go into their use and interpretation.