“Hessian Not Invertible” Help!

Jeff Gill
Department of Political Science
California Polytechnic

Gary King
Department of Government
Harvard University[1]

PARTIAL DRAFT, September 30, 1998

1 Introduction

Most statistical applications in the social sciences are based on a posterior distribution or likelihood function. In the vast majority of these applications, scholars assume the accuracy of the usual asymptotic normal approximation and thus require only the vector of maximum likelihood (or maximum posterior) estimates and the variance matrix of the estimates at the maximum. Although the negative of the Hessian (the matrix of second derivatives of the posterior with respect to the parameters) must be positive definite and hence invertible in order to compute the variance matrix, invertible Hessians do not exist for some data sets, and so statistical procedures often fail for this reason before completion. Indeed, receiving a computer-generated "Hessian not invertible" message (because of singularity or non-positive definiteness) rather than a set of statistical results is a frustrating, but common, occurrence in quantitative research. It even occurs with regularity during most Monte Carlo experiments where the investigator is drawing data from the assumed statistical model.

When a Hessian is not invertible, no computational trick can make it invertible, given the model and data chosen, since the needed inverse does not exist. The textbook advice in this situation is to rethink one's model, respecify it, and rerun the analysis. 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 a data set with more explanatory variables than observations (although even in the simple case, this is not the end of the story; see Searle, 1971). As such, a noninvertible Hessian may indicate a substantive problem that a researcher would not otherwise be aware of. In nonlinear models, however, noninvertible Hessians are traceable to the shape of the posterior density, but how to connect the problem to the substantive question being analyzed is normally far from obvious. As such, even though the textbook strategy may be worth trying in some cases, it is often difficult or impossible to implement.

In addition, for many applications, the textbook advice is disconcerting, if not downright wrong, since the model specification may have worked in other contexts and really is the one the researcher wants estimates from. In fact, although a noninvertible Hessian means the desired variance matrix does not exist, the likelihood function may still contain considerable information about the substantive 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 country or population subgroup), dropping only those cases with noninvertible Hessians, as is commonly done, can easily cause selection bias in the ultimate conclusions drawn from the set of analyses. Similarly, Monte Carlo studies that evaluate estimators risk severe bias if conclusions are based (as usual) on only those iterations with invertible Hessians.

In this paper, we provide a means of eliciting information in a convenient format from a likelihood or posterior distribution when the Hessian does not invert. Our approach is appropriate even when the Hessian does invert, and in many cases may be more appropriate than traditional reporting methods even when the Hessian is fine. We begin in Section 2 by providing an useful summary of the likelihood that can be calculated, even when the mode is uninteresting and the variance matrix nonexistent. The roadmap to the rest of the paper concludes that motivating section.

2 Means vs Modes

When a likelihood function or posterior distribution contains information but the variance matrix cannot be computed, all hope should not be considered lost. In low dimensional problems, plotting the likelihood or posterior is an obvious solution that would reveal all relevant information. 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). Thus, 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 change their *summary* of the likelihood function so that useful information can still be elicited without changing their substantive questions, statistical specification, assumptions, data, or model.

In statistical analyses, researchers collect data, specify a model, and form the likelihood or posterior. They then summarize this information, essentially by posing a question about the likelihood surface. The question answered by the standard maximum likelihood (or maximum posterior) estimates is

> *What is the mode of the posterior density and the curvature at that mode?*

In cases where the mode is on a plateau or at a boundary, or the posterior's surface has ridges or saddlepoints, the curvature will produce a noninvertible Hessian. In these cases, the unusable 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 about their likelihood function (with flat prior) or posterior distribution:

> *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. 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 the mode and its variance cannot. At least in these hard 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 the posterior (or likelihood) for θ is truncated normal with mean 0.5, standard deviation 10, and truncation at the $[0, 1]$ interval. In this case, the likelihood, based on a sample of data, will be a small segment of the normal curve. Except when the unit interval captures the mode of the normal (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 do 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 if one was specified. The mean will also not be a very useful summary for a bimodal likelihood function, 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 capture 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 likelihood function. We think 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 a procedure for computing the mean and its variance for any likelihood or posterior. This procedure is easy to apply in most cases, even when the Hessian fails.

In brief, we use a generalized inverse (when necessary, to avoid singularity) and generalized Cholesky decomposition (when necessary, to guarantee positive definiteness), which together almost always produce a "pseudo-variance matrix" for the mode that is a reasonable summary of the curvature of the likelihood surface. (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 normally ill conditioned. In addition, although this is a "pseudo" rather than "approximate" variance matrix (since 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 sampling. This procedure starts with a normal approximation, with mean and variance set at our mode and pseudo-variance, and then uses a probabilistic rejection algorithm to draw from 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.

In this paper, we describe in substantive terms what is "wrong" with a Hessian that is noninvertible (Section 3), and summarize existing knowledge about the generalized inverse (Section 4) and the generalized Cholesky (Section 5). We describe how we use these techniques with importance sampling to compute the mean and variance in Section 6. We then give three examples, two constructed from real data used in prior empirical research and one based on Monte Carlo evidence (Section 7).

3 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 or mass function, $f(y|\boldsymbol{\theta})$, for n-dimensional observed data vector y and unknown p-dimensional parameter vector $\boldsymbol{\theta}$, denote the $n \times p$ matrix of first derivatives with respect to $\boldsymbol{\theta}$ as $\mathbf{g}(\boldsymbol{\theta}|y) = \partial \ln[f(y|\boldsymbol{\theta})]/\partial\boldsymbol{\theta}$, 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 $E[\mathbf{g}(\boldsymbol{\theta}|y)g(\boldsymbol{\theta}|y)'] = E[\mathbf{H}(\boldsymbol{\theta}|y)]$. The maximum likelihood or maximum posterior estimate, which we denote $\hat{\boldsymbol{\theta}}$ is obtained by setting $\mathbf{g}(\boldsymbol{\theta}|y)$ equal to zero and solving, analytically or numerically. When $\mathbf{H}$ is non-positive definite in the neighborhood of $\hat{\boldsymbol{\theta}}$ then the theory is well known and no problems arise.

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 *non-positive definiteness*, which means that $(-\mathbf{H})^{-1}$ may exist but its contents do not make sense as a variance matrix. Statistical software normally describe both problems as "noninvertibility" since their inversion algorithms take computational advantage of the fact that the Hessian must be 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. We then move to more typical multiple parameter problems, which are more complicated but where we can help even more.

In one dimension, the Hessian is a single number measuring the degree to which the likelihood curves downward on either side of the maximum. When all is well, the $\mathbf{H} < \mathbf{0}$, 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, $-\mathbf{1}/\mathbf{H}$, which of course is a positive number as a variance must be.

The first problem, singularity, occurs in the one dimensional case when the likelihood is flat around the mode. This means that the curvature is zero 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, $\mathbf{1}/\mathbf{H} \to \infty$ as $\mathbf{H} \to \mathbf{0}$).

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 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 $-\mathbf{1}/\mathbf{H}$ is negative (or in other words is not positive definite) even though it exists, 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 defined as 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. Intuitively, singularity indicates that the variances to be calculated would be (essentially) infinite. When $(-\mathbf{H})^{-1}$ exists, it is only a valid variance matrix when the result is positive definite. Non-positive definiteness occurs either because the variance is negative or the correlations are not in the $(-1, 1)$ interval.

Below, we use a generalized inverse procedure to address singularity in the $\mathbf{H}$ matrix, and a generalized Cholesky to address cases where the inverse or generalized inverse of $(-\mathbf{H})$ is not positive definite. More specifically, the generalized inverse is primarily changing whatever is in $\mathbf{H}$ that gets 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 $(-1, 1)$. The interesting result, that we now proceed to describe is a pseudo-variance matrix that is well behaved and not nearly singular.

4 The Generalized Inverse

This section briefly reviews the mathematical and statistical considerations behind the 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.

4.1 The Many Generalized Inverse Matrices

Any matrix, $\mathbf{A}$, can be decomposed as

$$\underset{(p\times q)}{\mathbf{A}} = \underset{(p\times p)}{\mathbf{L}}\underset{(p\times q)}{\mathbb{D}}\underset{(q\times q)}{\mathbf{U}} \qquad \text{where,} \qquad \mathbb{D} = \begin{bmatrix} \mathbf{D}_{r\times r} & 0 \\ 0 & 0 \end{bmatrix}, \tag{4.1}$$

and both $\mathbf{L}$ (unit lower triangular) and $\mathbf{U}$ (unit upper triangular) are non-singular (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 non-negative definite and symmetric, then the diagonals of $\mathbf{D}_{r\times r}$ are the eigenvalues of $\mathbf{A}$. If $\mathbf{A}$ is non-singular, positive definite, and symmetric, as in the case of a proper invertible Hessian, then $\mathbf{D}_{r\times r} = \mathbb{D}$ (i.e. $r = q$) and $\mathbf{A} = \mathbf{L}\mathbb{D}\mathbf{L}'$. The matrices $\mathbf{L}$, $\mathbb{D}$, and $\mathbf{U}$ are all non-unique unless $\mathbf{A}$ is nonsingular.

By rearranging 4.1 we can diagonalize any matrix as

$$\mathbb{D} = \mathbf{L}^{-1}\mathbf{A}\mathbf{U}^{-1} = \begin{bmatrix} \mathbf{D_{r\times r}} & 0 \\ 0 & 0 \end{bmatrix}. \tag{4.2}$$

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

$$\mathbb{D}^{-} = \begin{bmatrix} \mathbf{D^{-}_{r\times r}} & 0 \\ 0 & 0 \end{bmatrix}. \tag{4.3}$$

If $\mathbb{D}\mathbb{D}^{-} = \mathbf{I_{q\times q}}$ then 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 (4.1) are non-unique. By rearranging (4.1) and using (4.3) 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.[1]

Theorem 4.1 *(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 Theorem 4.1, any one of them will do for our purposes. For example, in linear regression, the fitted values — defined as $X\mathbf{G}X'y$, with $\mathbf{G}$ as the generalized inverse of $X'X$, X as a matrix of explanatory variables and y as the outcome variable — are invariant to the definition of $\mathbf{G}$. In addition, we only use our pseudo-variance as a first approximation to the surface of the true posterior, and we will improve it in our importance sampling stage.[2]

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

[2] Note in addition that $\mathbf{AG}$ is always idempotent ($\mathbf{GAGA} = \mathbf{G}(\mathbf{AGA}) = \mathbf{GA}$), and $\text{rank}(\mathbf{AG}) = \text{rank}(\mathbf{A})$. These results hold whether $\mathbf{A}$ is singular or not.

4.2 The Unique Moore-Penrose Generalized Inverse Matrix

Moore (1920) and (unaware of Moore's work) Penrose (1955) reduced the infinity of generalized inverses to one unique solution by imposing four reasonable algebraic constraints, all met by the standard inverse. If

1. general condition: $\mathbf{AGA} = \mathbf{A}$,
2. reflexive condition: $\mathbf{GAG} = \mathbf{G}$,
3. normalized condition: $(\mathbf{AG})' = \mathbf{GA}$, and
4. reverse normalized condition: $(\mathbf{GA})' = \mathbf{AG}$

then this $\mathbf{G}$ matrix is unique. 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:

$$G = \left[\mathbf{r}^{-1}\mathbf{0}\right]\mathbf{Q}'$$

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

5 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 $\mathbf{V} = \text{Gchol}(\mathbf{H}^{-})$, $\text{Gchol}(\cdot)$ is the generalized Cholesky described below, and $\mathbf{H}^{-}$ is the generalized inverse of the Hessian matrix.

5.1 The Classic Algorithm

The classic Cholesky algorithm exploits the symmetry of covariance ($\mathbf{C}$) matrices by changing the decomposition from (4.1) to:

$$\underset{(k\times k)}{\mathbf{C}} = \underset{(k\times k)}{\mathbf{L}}\underset{(k\times k)}{\mathbf{D}}\underset{(k\times k)}{\mathbf{L}'}. \tag{5.1}$$

The basic Cholesky procedure is a one pass algorithm which 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 j^{th} step $(1 \leq j \leq k)$ is characterized by two operations:

$$\mathbf{D}_{j,j} = \mathbf{C}_{j,j} - \sum_{\ell=1}^{j-1} \mathbf{L}_{j,\ell}^2 \mathbf{D}_{\ell,\ell}, \qquad \text{and} \tag{5.2}$$

$$\mathbf{L}_{i,j} = \left[\mathbf{C}_{i,j} - \sum_{\ell=1}^{j-1} \mathbf{L}_{j,\ell}\mathbf{L}_{i,\ell}\mathbf{D}_{\ell,\ell}\right] / \mathbf{D}_{j,j}, \qquad i = j+1, \ldots, k, \tag{5.3}$$

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 (5.3), and non-positive definite matrices cause the sum in (5.2) 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 singular value decomposition 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 & McCormick 1968, Gill, Golub, Murray, & Saunders 1974, Matthews and Davies 1971).

5.2 The Gill/Murray Cholesky Factorization

Gill and Murray (1974) introduced, and Gill, Murray and Wright (1981) refined, an algorithm to find a non-negative 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 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 $\mathbf{C}+\mathbf{E}$ matrix is much less like $\mathbf{C}$ than it could be.

To see Gill, Murray and Wright's approach, we rewrite the Cholesky algorithm provided as (5.2) and (5.3) in matrix notation. The j^{th} submatrix of its application at the j^{th} step is

$$\mathbf{C}_j = \begin{bmatrix} c_{j,j} & \mathbf{c}'_j \\ \mathbf{c}_j & \mathbf{C}_{j+1} \end{bmatrix} \tag{5.4}$$

where $c_{j,j}$ is the j^{th} 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 j^{th} 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:

$$\mathbf{C}^{*}_{j+1} = \mathbf{C}_{j+1} - \frac{\mathbf{c}_j \mathbf{c'}_j}{L^2_{j,j}} \tag{5.5}$$

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 trade-offs between satisfaction with the current iteration and satisfaction with future iterations. If δ_j is picked such that the new j^{th} diagonal is just barely bigger than zero, then subsequent diagonal values are greatly increased through the operation of (5.5). Conversely we don't want to be adding large δ_j values on any given iteration.

Gill, Murray, and Wright 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 j^{th} step by making δ_j the smallest non-negative value satisfying:

$$\|\mathbf{c}_j\|_\infty \beta^{-2} - c_{j,j} \leqslant \delta_j \tag{5.6}$$

$$\text{where: } \beta = \max \begin{cases} \max(\text{diag}(\mathbf{C})) \\ \max(not\text{diag}(\mathbf{C}))\sqrt{k^2-1} \\ \epsilon_{\mathtt{m}} \end{cases}$$

where $\epsilon_{\mathtt{m}}$ is the smallest positive number that can be represented on the utilized computer (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 (5.6) have been shown to be non-optimal and thus provide $\mathbf{C}+\mathbf{E}$ that is further from $\mathbf{C}$ than necessary.

5.3 The Schnabel/Eskow Cholesky Factorization

Schnabel and Eskow 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 5.1 *Suppose $\mathbf{C} \in \mathbb{R}^k$ with eigenvalues $\lambda_1, \ldots, \lambda_k$. Define the i^{th} Gerschgorin bound as:*

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

Then $\lambda_i \in [\mathsf{G}_1 \cup \mathsf{G}_2 \cup \ldots \cup \mathsf{G}_k] \qquad \forall \lambda_{1 \leq i \leq k}$.

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

$$\delta_j = \max\left(\epsilon_{\mathtt{m}}, \max_i(\mathsf{G}_i(\text{lower}))\right) \tag{5.7}$$

In addition, we don't 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 (5.5). Adding this condition to (5.7) and protecting the algorithm from problems associated with machine epsilon produces the following determination of the additional amount in $L_{j,j} = \sqrt{c_{j,j} + \delta_j}$:

$$\delta_j = \max\left(\epsilon_{\mathtt{m}}, -\mathbf{C}_{j,j} + \max(\|\mathbf{a}_j\|, (\epsilon_{\mathtt{m}})^{\frac{1}{3}} \max(\mathrm{diag}(\mathbf{C}))), \mathbf{E}_{j-1,j-1}\right) \tag{5.8}$$

The algorithm follows the same steps as that of Gill/Murray except that the determination of δ_j is done by (5.8). 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 $\|\mathbf{E}\|_\infty$.

6 Importance Sampling

We now describe importance sampling and how to use it, the generalized inverse, and the generalized Cholesky to compute the mean and its variance. A key point is that the procedures we describe in this section work whether or not the Hessian is invertible.

Importance sampling, also called "sampling importance resampling" , is an iterative simulation technique used to draw random numbers from an exact (finite sample) posterior distribution.[3] This algorithm is not based on Markov Chains, and so does not have the same problems with assessing convergence typically associated with Monte Carlo Markov Chain (MCMC) methods, and it is normally much faster. Other methods can be used to compute the mean instead of a mode of a distribution, but importance sampling is probably the simplest method that works without modification across a wide range of applications. That is, it does not require special mathematical analysis for each application, so it is fairly easy to implement computationally.

The chief requirement for a successful implementation of importance sampling is simulations from a distribution that is a reasonable approximation to the true posterior. If this requirement is not met, the procedure can take too long to be practical or can miss features of the posterior.

In our case, we use the multivariate normal distribution as our approximation in general, or the multivariate t distribution when the sample size is small. This decision should be is 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 instead, but are not normally 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 (calculated as $\mathbf{V}'\mathbf{V}$, $\mathbf{V} = \mathrm{Gchol}(\mathbf{H}^{-})$, $\mathrm{Gchol}(\cdot)$ is the generalized Cholesky, and $\mathbf{H}^{-}$ is the Moore-Penrose generalized inverse of the Hessian matrix).[4] For the t, the pseudo-variance is adjusted by the degrees of freedom to yield the scatter matrix.

[3] For political science applications of importance sampling, see King (1997) and King, Honaker, Joseph, and Scheve (1998). For technical references, see Rubin, 1987: 192–4; Tanner, 1996; Gelman et al., 1996; Wei and Tanner, 1990.

[4] 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, p.350), but its purpose is computational rather than statistical and so the importance sampling step is omitted as well.

The idea of importance sampling 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 more frequently than those farther away. To be more precise, denote $\tilde{\boldsymbol{\theta}}$ as one random draw of $\boldsymbol{\theta}$ from the normal distribution (for example), and use it to compute the importance ratio, the ratio of the true likelihood or 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 used directly to compute a quantity of interest (see King, Tomz, and Wittenberg, 1998) or they can be used to compute a mean and variance matrix. To do the latter, we merely compute the sample average and variance of a vector of simulations. The variance matrix will normally be positive definite, so long as more simulations than elements of the mean vector and variance matrix were drawn (and normally, one would want at least 10 times that number). It is conceivable, 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 very unusual case, singularity in the variance matrix (as opposed to the Hessian) is not normally a problem, since all that will happen is that some of the correlations will be exactly 1 or -1, which can be very informative substantively.

In our experience, this procedure works well for even quite complicated likelihoods, as long as the routinely assumed asymptotic normal approximation is plausible in cases without the Hessian difficulties. A diagnostic for it not working is if too many candidate values of $\tilde{\boldsymbol{\theta}}$ are rejected. In this case the procedure will take an unreasonably long time, and to generate a useful result a better approximation is needed. That is, a long run time indicates a problem, but letting it run longer is a reasonable solution from a statistical perspective, although not necessarily from a practical one. The only danger of using the procedure is if the approximation distribution entirely misses a range of values of $\boldsymbol{\theta}$ that have posterior density systematically different from the rest. Since the normal covers all of $(-\infty, \infty)$, 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). Of course, it is impossible to cover all the values that $\boldsymbol{\theta}$ can take, and the procedure can therefore miss features like pinholes in the surface or other eccentricities.

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 make the maximization routine work better as well.) 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.[5]

[5] An alternative to our procedure 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, $\delta\boldsymbol{\theta} = H(\boldsymbol{\theta})^- f(\mathbf{x}|\boldsymbol{\theta})$. Rao and Mitra assert that 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 $H(\boldsymbol{\theta})$.

7 Numerical Examples

This section provides examples in which model specifications that lead to singular Hessian matrices, and thus requiring our techniques, are compared with similarly specified but not problematic models. We emphasize the substantive importance that comes from understanding what the generalizations provide to the results. The noninvertibility in Section 7.1 is generated in obvious fashion by feeding it explanatory variables that are perfectly correlated. In Section 7.2, the problem arises more naturally and its source is not obvious, but our approach seems to work as well.

7.1 Logit Regression Model

The first example data come from Martin (1992), but are greatly simplified to emphasize the statistical discussion. The dichotomous outcome variable is the presence of an international institution. The initial explanatory variables are WW, a dichotomous factor indicating World War I or II, and NUM, the number of senders. These data are analyzed with a simple logit model ($P(Y_i = 1|X_i) = \pi_i = [1 + e^{-X_i\beta}]^{-1}$), in which all but two of the explanatory variables are dropped from the specification. This model produces the following Hessian and Covariance matrix:

$$\mathbf{H} = \begin{bmatrix} 0.0715 & 0 & 0.4277 \\ 0 & 0.000001 & 0 \\ 0.4277 & 0 & 4.9132 \end{bmatrix} \qquad \mathbf{\Sigma} = \begin{bmatrix} 0.3740 & 0 & -0.0326 \\ 0 & 2135691.5 & 0 \\ -0.0326 & 0 & 0.0054 \end{bmatrix}$$

where the covariance matrix is the inverse of the Hessian (divided by $n = 78$). We display the results as a traditional regression summary, along with what we call a *standardized correlation matrix*, which is a correlation matrix with standard deviations (standard errors) on the diagonal:

Simple Logit Model, No Hessian Problem

Parameter	Estimate	Standard Error	Est/SE
Intercept	-3.1389	0.6155	-5.133
WW	-14.3642	1461.4005	-0.010
NUM	0.1958	0.0738	2.654

Standardized Correlation Matrix:

$$\begin{bmatrix} 0.615 & 0.000 & -0.722 \\ 0.000 & 1461.401 & 0.000 \\ -0.722 & 0.000 & 2.654 \end{bmatrix}$$

Now consider a new explanatory variable in the model specification. In order to make the example as forceful as possible, at the cost of authenticity, we will construct a third explanatory variable as a linear function of the other two: COMBO = WW + NUM/100. This is done to illustrate the effects and subsequent treatment of a difficult Hessian matrix. Once again a logit specification is used, and the following singular Hessian results:

$$\mathbf{H} = \begin{bmatrix} 0.07152074 & 0.00000005 & 0.00427760 & 0.42774826 \\ 0.00000005 & 0.00000002 & 0.00000000 & 0.00000000 \\ 0.00427760 & 0.00000000 & 0.00049129 & 0.04913012 \\ 0.42774826 & 0.00000000 & 0.04913012 & 4.91290180 \end{bmatrix}$$

where greater displayed precision is required. Obviously this matrix cannot be inverted in the standard fashion. Furthermore, even the generalized inverse of this matrix is not positive definite. Our solution is to apply the Moore-Penrose Generalized Inverse followed by the Schnabel-Eskow generalized Cholesky factorization to get our pseudo-variance matrix,

$$\tilde{\mathbf{S}} = \begin{bmatrix} 1.79965037 & -0.47559405 & 0.15926313 & -0.01968730 \\ -0.47559405 & 730.80902976 & -0.00079386 & 0.00009813 \\ 0.15926313 & -0.00079386 & 72.89110580 & 72.89149681 \\ -0.01968730 & 0.00009813 & 72.89149681 & 728.90629058 \end{bmatrix}$$

where the $\tilde{S}$ notation is used to distinguish this result from the conventional variance matrix. Using these estimates of uncertainty, the model results are summarized as:

Simple Logit Model, Singular Hessian

Parameter	Estimate	Standard Error	Est/SE
Intercept	-3.138906	1.341511	-2.339830
WW	-7.183151	27.03348	-0.265713
COMBO	-7.180483	8.537629	-0.841039
NUM	0.267585	26.99827	0.009911

Standardized Correlation Matrix:

$$\begin{bmatrix} 1.34151048 & -0.01311415 & 0.01390541 & -0.00054357 \\ -0.01311415 & 27.03347979 & -0.00000343 & 0.00000013 \\ 0.01390541 & -0.00000343 & 8.53762882 & 0.31623050 \\ -0.00054357 & 0.00000013 & 0.31623050 & 26.99826458 \end{bmatrix}$$

Several interesting and important observations come from comparing the two regression summaries. First, the generalized procedure had very little effect on the intercept term except to increase its standard error. Second, the new term, COMBO, takes coefficient magnitude away from the two explanatory variable that produced a reasonable outcome. This dilution of the original coefficient magnitudes into three is seen by noting that the sum of the three explanatory coefficients in the singular model is almost exactly the the sum of the two coefficients in the non-singular model.

****[add importance sampling]****

7.2 Negative Binomial Regression Model

The data for this example provide murder counts in Detroit along with thirteen potential explanatory variables. These data were collected by J.C. Fisher (1976) and used by Miller (1990, Appendix A) as an example where stepwise regression fails. The initial model, which produces a satisfactory Hessian, specifies the following explanatory variables: death rate from accidents (ACC), percent of homicides cleared by arrests (CLR), and percent unemployed (UMP). Because the assumption of independence is highly suspect, and because the data demonstrate overdispersion, the Poisson regression model is replaced with the more flexible negative binomial specification, with the variance specified as a function of UMP.

The negative binomial specification provides the following Hessian and variance matrix:

$$\mathbf{H} = \begin{bmatrix} 41.3216 & 1822.4184 & 2923.9282 & -0.4206 & -1.6797 \\ 1822.4184 & 80884.9140 & 129225.2100 & -19.9051 & -75.7069 \\ 2923.9282 & 129225.2100 & 213584.8600 & -34.7928 & -132.2470 \\ -0.4206 & -19.9051 & -34.7928 & 0.3836 & 2.0766 \\ -1.6797 & -75.7070 & -132.2470 & 2.0766 & 12.5601 \end{bmatrix}$$

$$\mathbf{\Sigma} = \begin{bmatrix} 0.345706 & -0.006794 & -0.000636 & -0.224345 & 0.035670 \\ -0.006794 & 0.000162 & -0.000004 & 0.004093 & -0.000659 \\ -0.000637 & -0.000005 & 0.000012 & 0.000724 & -0.000107 \\ -0.224345 & 0.004093 & 0.000723 & 2.099769 & -0.344879 \\ 0.035670 & -0.000659 & -0.000107 & -0.344880 & 0.062818 \end{bmatrix}$$

where again the variance matrix is the inverse of the Hessian (divided by $n = 13$). This produces the following model summary:

Negative Binomial, No Hessian Problem

Parameter	Estimate	Standard Error	Est/SE
Intercept	7.5881	0.5880	12.906
ACC	0.0403	0.0127	3.162
CLR	-0.0505	0.0035	-14.557
Var. Incpt.	4.5446	1.4491	3.136
UMP	-0.4383	0.2506	-1.749

Standardized Correlation Matrix:

$$\begin{bmatrix} 0.588 & -0.908 & -0.312 & -0.263 & 0.242 \\ -0.908 & 0.013 & -0.109 & 0.222 & -0.206 \\ -0.312 & -0.109 & 0.004 & 0.144 & -0.123 \\ -0.263 & 0.222 & 0.144 & 1.449 & -0.950 \\ 0.242 & -0.206 & -0.123 & -0.950 & 0.251 \end{bmatrix}$$

We now introduce a new variable, average hourly earnings (AHE), with collinear effects such that the Hessian is no longer invertible using conventional means. This produces the following non-singular, but non-positive definite Hessian.

$$\mathbf{H} = \begin{bmatrix} 4.427758 & -0.080856 & -0.011726 & -4.332436 & 0.762877 & 0.182450 \\ -0.080856 & 0.001731 & 0.000061 & 0.103885 & -0.033836 & 0.006207 \\ -0.011726 & 0.000061 & 0.000128 & -0.001358 & 0.009660 & -0.006309 \\ -4.332437 & 0.103885 & -0.001358 & 22.822484 & 4.707084 & -7.038142 \\ 0.762878 & -0.033836 & 0.009660 & 4.707084 & -3.430507 & 1.582113 \\ 0.182450 & 0.006208 & -0.006309 & -7.038142 & 1.582113 & 0.161358 \end{bmatrix}$$

Applying the generalized inverse/generalized Cholesky (the Moore-Penrose generalized inverse provides the regular inverse when the matrix is non-singular), gives the following pseudo-variance matrix:

$$\tilde{\mathbf{S}} = \begin{bmatrix} 0.642731 & 0.009677 & 0.001403 & 0.518513 & -0.091302 & -0.021836 \\ 0.009677 & 0.868105 & -0.000021 & -0.014985 & 0.004016 & -0.000307 \\ 0.001403 & -0.000021 & 0.868235 & -0.000718 & -0.000708 & 0.000594 \\ 0.518513 & -0.014985 & -0.000718 & 0.796491 & -0.395086 & 0.693935 \\ -0.091302 & 0.004016 & -0.000708 & -0.395086 & 1.579433 & 0.095269 \\ -0.021836 & -0.000307 & 0.000594 & 0.693935 & 0.095269 & 1.379880 \end{bmatrix}$$

If we apply this estimate of uncertainty to the incomplete model results, the following table results:

Negative Binomial, Singular Hessian

Parameter	Estimate	Standard Error	Est/SE
Intercept	7.575934	0.80170522	9.44977505
ACC	0.041288	0.93172162	0.04431367
CLR	-0.050944	0.93179112	-0.05467320
Var. Incpt.	4.167551	0.89246363	4.66971522
AHE	0.138427	1.25675503	0.11014637
UMP	-0.468935	1.17468307	-0.39920129

Standardized Correlation Matrix:

$$\begin{bmatrix} 0.80170522 & 0.01295496 & 0.00187861 & 0.72469319 & -0.09061843 & -0.02318655 \\ 0.01295496 & 0.93172162 & -0.00002423 & -0.01802135 & 0.00342970 & -0.00028015 \\ 0.00187861 & -0.00002423 & 0.93179112 & -0.00086358 & -0.00060471 & 0.00054293 \\ 0.72469319 & -0.01802135 & -0.00086358 & 0.89246363 & -0.35224927 & 0.66192359 \\ -0.09061843 & 0.00342970 & -0.00060471 & -0.35224927 & 1.25675503 & 0.06453272 \\ -0.02318655 & -0.00028015 & 0.00054293 & 0.66192359 & 0.06453272 & -0.39920129 \end{bmatrix}$$

This example provides a more realistic and a much more useful result. Note that despite the introduction of a problematic new variable which precludes the development of a Covariance matrix, none of the original coefficient estimates are substantially changed from their original values in the previous healthy model. The introduction of this new variable and the subsequent treatment by the generalized technique does, however, cause near uniform enlargement of the standard errors (the exception being the constant in the variance term). This demonstrates that under common circumstances we can create nearby starting values for importance sampling algorithms even though such estimates would have been completely inaccessible with the generalized technique.

****[add importance sampling]****

7.3 Monte Carlo Example

8 Concluding Remarks

The dominant statistical procedure in empirical social science is to specify a common distributional form and obtain estimates of the modes of the parameters with maximum likelihood. In this procedure a variance matrix is calculated from the curvature of the likelihood function at the mode. Unfortunately, many aspects of the multidimensional density blanket can cause the variance generating procedure to fail. In these cases an estimate of the parameter uncertainty at the mode simply doesn't exist, and researchers typically give up and change the substantive question they are asking.

In this paper we argue that the standard maximum likelihood approach is somewhat myopic. The exclusive focus on the mode and associated variance of the likelihood surface means that some models will not provide useful information in certain, and often very interesting, situations. The mode is not a very good or very complete summary of the posterior in many cases, and a better summary is achieved by drawing from the posterior directly and computing the mean or variance (or other quantities of interest).

Importance sampling is a well-established procedure for determining full information posterior densities in which we make no assumptions other than those that the researcher specified initially. However, getting the initial parameter estimates to begin importance sampling was previously unavailable to all but the most resourceful applied statisticians. Our generalized inverse/generalized Cholesky procedure changes that. With no additional distributional assumptions, asymptotic or otherwise, and a relatively simple set of calculations, the user can generate a meaningful posterior distribution for any parameters of interest and easily compute the mean and its variance. Since for many if not most applications, researchers would have chosen the mean over the mode had there been an easy and reliable way to generate estimates of each, our summary of the posterior distribution may be as or more useful than the mode and variance if they were available. Thus, our point goes well beyond solving the annoying and diverting "Hessian not invertible" message spewed out on occasion by most computer programs.

Scholars of linear algebra have made many advances in recent years, some of which we build on in this paper. However, we believe an opportunity for further advancement in that field that could help work in our's. 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. It may also be possible produce an even better result by using the information that the Hessian is not merely a symmetric matrix, but 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 positive definite (and hence nonsingular) pseudo-variance matrix that is not ill conditioned.

Appendices

A Generalized Inverse Numerical Examples

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 non-invertible 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, p.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

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

fitted values,

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

and variance matrix

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

The standardized correlation matrix 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 we have a matrix of explanatory affects 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 LS model now gives

$$\hat{\mathbf{b}}_\mathbf{2} = (1111.11, -11.89, -1104.11)',$$

and

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

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

$$\mathbf{C}_s = \begin{bmatrix} 1758.79 & -0.48 & -0.99 \\ -0.48 & 20.12 & 0.48 \\ -0.99 & 0.48 & 1753.35 \end{bmatrix}$$

Indeed, if $\mathbf{X_3} = 2.999$, then $\mathbf{X}'\mathbf{X}$ is singular (with regard to precision in Gauss and Splus) and we must use the generalized inverse. This produces

$$\tilde{\mathbf{b}}_\mathbf{3} = \mathbf{G}\mathbf{X}'\mathbf{Y} = (1.774866, -5.762093, 1.778596)',$$

and

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{Y} = (6.2431253, 1.3448314, -0.8637996, 4.8965190)'.$$

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

$$\mathbf{C}_s = \begin{bmatrix} 16328.7257311 & -0.4822391 & -0.9999999 \\ -0.4822391 & 18.6815417 & 0.4818444 \\ -0.9999999 & 0.4818444 & 16323.6599450 \end{bmatrix}$$

B Generalized Cholesky Numerical Examples

Suppose we have the positive definite matrix

(Example B.1)
$$\Sigma_1 = \begin{bmatrix} 2 & 0 & 2.4 \\ 0 & 2 & 0 \\ 2.4 & 0 & 3 \end{bmatrix}.$$

This matrix has the Cholesky decomposition:

$$\text{chol}(\Sigma_1) = \begin{bmatrix} 1.41 & 0 & 1.69 \\ 0 & 1.41 & 0 \\ 0 & 0 & 0.35 \end{bmatrix}.$$

Now suppose we have a very similar, but non-positive definite matrix that requires the generalized Cholesky algorithm. The only change from the input matrix in Example B is that the values on the corners have been changed from 2.4 to 2.5:

(Example B.2)
$$\Sigma_3 = \begin{bmatrix} 2 & 0 & 2.5 \\ 0 & 2 & 0 \\ 2.5 & 0 & 3 \end{bmatrix}.$$

This matrix has the generalized Cholesky decomposition:

$$\text{Gchol}(\Sigma_2) = \begin{bmatrix} 1.58 & 0 & 1.58 \\ 0 & 1.41 & 0 \\ 0 & 0 & 0.71 \end{bmatrix}$$

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

(Example B.3)
$$\Sigma_2 = \begin{bmatrix} 2 & 0 & 10 \\ 0 & 2 & 0 \\ 10 & 0 & 3 \end{bmatrix}$$

This matrix has the Cholesky decomposition:

$$\text{Gchol}(\Sigma_3) = \begin{bmatrix} 3.16 & 0 & 3.16 \\ 0 & 2.82 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

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

$$\text{Gchol}(\Sigma_3)'\text{Gchol}(\Sigma_3) = \begin{bmatrix} 10 & 0 & 10 \\ 0 & 8 & 0 \\ 10 & 0 & 14 \end{bmatrix}$$

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

References

Albert, Arthur. 1973 "The Gauss-Markov Theorem for Regression Models with Possible Singular Covariances." *SIAM Journal on Applied Mathematics* 24, No.2 (March), 182-7.

Campbell, S. L. and C. D. Meyer, Jr. 1979. *Generalized Inverses of Linear Transformations.* New York: Dover Publications.

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

Fisher, J.C. 1976. "Homicide in Detroit: The Role of Firearms." *Criminology* 14, 387-400.

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.* Second Edition. New York: MacMillan.

Harville, David A. *Matrix Algebra From a Statistician's Perspective.* New York: Springer.

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

King, Gary. 1989. *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 to the annual meetings of the American Political Science Association, Boston.

King, Gary; Michael Tomz; and Jason Wittenberg. 1998. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation," paper presented to the annual meetings of the American Political Science Association, Boston.

Martin, Lisa L. 1992. *Coercive Cooperation: Explaining Multilateral Economic Sanctions.* Princeton, N.J.: Princeton University Press.

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

Miller, Alan J. 1990. *Subset selection in regression.* New York: Chapman and Hall.

Moore, E. H. 1920. "On the Reciprocal of the General Algebraic Matrix. (Abstract)" *Bulletin of the American Mathematical Society* 26, 394-5.

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: Wiley & Sons.

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

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