The Theory and Practice of Econometrics

Second Edition

George G. Judge
University of Illinois

W. E. Griffiths
University of New England

R. Carter Hill
University of Georgia

Helmut Lütkepohl
Universität Osnabrück

Tsoung-Chao Lee
University of Connecticut

John Wiley and Sons

New York *Chichester* *Brisbane* *Toronto* *Singapore*

Library of Congress Cataloging in Publication Data:

The Theory and practice of econometrics.

Includes bibliographies and index.
1. Econometrics. I. Judge, George G.
HB139.T48 1985 330'.028 84-7254
ISBN 0-471-89530-X

Printed in the United States of America

10 9 8

22.9.5 Individual Exercises (Section 22.5)

Exercise 22.12
Adopt one of the rules outlined in Section 22.5.2b for selecting the biasing parameter k used in ridge regression and compute the ridge estimates for five samples. Compute the sample means and variances of the estimated coefficients and compare them with the least squares results.

Exercise 22.13
Use Strawderman's adaptive generalized ridge estimator (Section 22.5.2c) to estimate the model's coefficients for five samples. Compute the means and variances of these estimates and compare them to the least squares results.

22.9.6 Joint or Class Exercises (Section 22.5)

Exercise 22.14
Use the sample data from Exercise 22.3 and Strawderman's adaptive generalized ridge estimator to estimate the model's coefficients for 100 samples. Compute the sample means, variances, and squared error loss for these estimates and compare them to the least squares results.

22.9.7 Individual Exercises (Section 22.6)

Exercise 22.15
Use the James- and Stein-type minimax estimator (22.6.2) to estimate the model's parameters for five samples. Compute the sample means and variances of these estimates and compare them to the least squares results.

22.9.8 Joint or Class Exercises (Section 22.6)

Exercise 22.16
Use the sample data from Exercise 22.3 and the James- and Stein-type minimax estimator (22.6.2) to estimate the model's parameters for 100 samples. Compute the sample means, variances, and squared error loss of these estimates and compare them to the least squares results.

22.9.9 Joint or Class Exercises (Section 22.8)

Exercise 22.17
Contrast the results for the various estimators and diagnostic checks and discuss the best way to deal with this model and the corresponding data.

22.10 REFERENCES

Allen, D. M. (1974) "The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction," *Technometrics*, 16, 125–127.

Bacon, R. W. and J. A. Hausman (1974) "The Relationship Between Ridge Regression and the Minimum Mean Squared Error Estimator of Chipman," *Oxford Bulletin of Economics and Statistics,* 36, 115–124.

Baranchik, A. J. (1973) "Inadmissibility of Maximum Likelihood Estimators in Some Multiple Regression Problems with Three or More Independent Variables," *Annals of Statistics,* 1, 312–321.

Belsley, D., E. Kuh, and R. E. Welsh (1980) *Regression Diagnostics,* Wiley, New York.

Belsley, D. (1982a) "Centering, the Constant and Diagnosing Collinearity," Technical Report #33, Boston College.

Belsley, D. (1982b) "Assessing the Presence of Harmful Collinearity and other Forms of Weak Data through a Test for Signal to Noise," *Journal of Econometrics,* 20, 211–253.

Brooks, R. J. and T. Moore (1980) "On the Expected Length of the Least Squares Coefficient Vector," *Journal of Econometrics,* 12, 245–246.

Brown, W. G. and B. R. Beattie (1975) "Improving Estimates of Economic Parameters by Use of Ridge Regression with Production Function Applications," *American Journal of Agricultural Economics,* 57, 21–32.

Casella, G. (1977) Minimax Ridge Estimation, unpublished Ph.D. dissertation, Purdue University, Lafayette, Ind.

Conniffe, D. and J. Stone (1973) "A Critical View of Ridge Regression," *The Statistician,* 22, 181–187.

Dempster, A. P., M. Schatzoff, and M. Wermuth (1977) "A Simulation Study of Alternatives to Ordinary Least Squares," *Journal of the American Statistical Association,* 72, 77–104.

Dhrymes, P. J. (1974) *Econometrics: Statistical Foundations and Applications,* Springer-Verlag, New York.

Draper, N. and R. Van Nostrand (1979) "Ridge Regression and James and Stein Estimation: Review and Comments," *Technometrics,* 21, 451–466.

Efron, B. and C. Morris (1977) "Comment," *Journal of the American Statistical Association,* 72, 91–94.

Farebrother (1972) "Principal Component Estimators and Minimum Mean Square Error Criteria in Regression Analysis," *Review of Economics and Statistics,* 54, 332–336.

Farrar, D. E. and R. R. Glauber (1967) "Multicollinearity in Regression Analysis: The Problem Revisited," *Review of Economics and Statistics,* 49, 92–107.

Fomby, T. B. and S. R. Johnson (1977) "MSE Evaluation of Ridge Estimators Based on Stochastic Prior Information," *Communications in Statistics, A,* 6, 1245–1258.

Fomby, T. B. and R. C. Hill (1978a) "Multicollinearity and the Minimax Conditions for the Bock Stein-Like Estimator," *Econometrica,* 47, 211–212.

Fomby, T. B. and R. C. Hill (1978b) "Multicollinearity and the Value of a Priori Information," *Communications in Statistics, A,* 8, 477–486.

Fomby, T. B., R. C. Hill, and S. R. Johnson (1978) "An Optimality Property of Principal Components Regression," *Journal of the American Statistical Association,* 73, 191–193.

Fourgeaud, C., C. Gourieroux, and J. Pradel (1984) "Some Theoretical Results for Generalized Ridge Regression Estimator," *Journal of Econometrics,* forthcoming.

Friedmann, R. (1982) "Multicollinearity and Ridge Regression," *Allgemeines Statistisches Archiv,* 66, 120–128.

Gibbons, D. (1981) "A Simulation Study of Some Ridge Estimators," *Journal of the American Statistical Association,* 76, 131–139.

Goldstein, M. and A. F. M. Smith (1974) "Ridge Type Estimators for Regression Analysis," *Journal of the Royal Statistical Society B,* 36, 284–291.

Greenberg, E. (1975) "Minimum Variance Properties of Principal Components Regression," *Journal of the American Statistical Association,* 70, 194–197.

Hemmerle, W. J. (1975) "An Explicit Solution for Generalized Ridge Regression," *Technometrics,* 17, 309–314.

Hemmerle, W. J. and T. F. Brantle (1978) "Explicit and Constrained Generalized Ridge Regression," *Technometrics,* 20, 109–120.

Hill, R. C., and R. Ziemer (1983a) "Small Sample Performance of the Stein-Rule in Nonorthogonal Designs," *Economics Letters,* 10, 285–292.

Hill, R. C., and R. Ziemer (1984) "The Risk of Stein-Like Estimators in the Presence of Multicollinearity," *Journal of Econometrics,* forthcoming.

Hill, R. C., T. B. Fomby, and S. R. Johnson (1977) "Component Selection Norms for Principal Components Regression," *Communications in Statistics, A,* 6, 309–333.

Hocking, R. R., F. M. Speed, and M. J. Lynn (1976) "A Class of Biased Estimators in Linear Regression," *Technometrics,* 18, 425–437.

Hoerl, A. E. (1962) "Application of Ridge Analysis to Regression Problems," *Chemical Engineering Progress,* 58, 54–59.

Hoerl, A. E. (1964) "Ridge Analysis," Chemical Engineering Progress Symposium, Series 60, 67–77.

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

Hoerl, A. E. and R. W. Kennard (1970b) "Ridge Regression: Application to Nonorthogonal Problems," *Technometrics,* 12, 69–82.

Hoerl, A. E. and R. W. Kennard (1976) "Ridge Regression: Iterative Estimation of the Biasing Parameter," *Communications in Statistics, A,* 5, 77–88.

Hoerl, A. E. and R. W. Kennard (1979) "Ridge Regression-1979," unpublished mimeo.

Hoerl, A. E., R. W. Kennard, and K. F. Baldwin (1975) "Ridge Regression: Some Simulations," *Communications in Statistics, A,* 4, 105–123.

Johnson, S. R., S. C. Reimer, and T. P. Rothrock (1973) "Principal Components and the Problem of Multicollinearity," *Metroeconomica,* 25, 306–317.

Judge, G. G. and M. Bock (1983) "Biased Estimation," in *Handbook in Econometrics, Vol. I,* Z. Griliches and M. Intriligator, eds., North Holland, Amsterdam, 599–660.

Judge, G. G., R. Hill, W. Griffiths, H. Lütkepohl, and T. Lee (1982) *Introduction to the Theory and Practice of Econometrics,* Wiley, New York.

Judge, G. G. and M. E. Bock (1978) *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics,* North-Holland, Amsterdam.

Judge, G. G., T. A. Yancey, and M. E. Bock (1973) "Properties of Estimators After Preliminary Tests of Significance When Stochastic Restrictions Are Used in Regression," *Journal of Econometrics,* 1, 29–48.

King, N. (1972) "An Alternative for the Linear Regression Equation When the Predictor Variable Is Uncontrolled and the Sample Size Is Small," *Journal of the American Statistical Association,* 67, 217–219.

King, N. (1974) "An Alternative for Multiple Regression When the Prediction Variable Are Uncontrolled and the Sample Size Is Not So Small," unpublished manuscript.

Kumar, T. K. (1975) "Multicollinearity in Regression Analysis," *Review of Economics and Statistics,* 57, 365–366.

Lawless, J. F. and P. Wang (1976) "A Simulation Study of Ridge and Other Regression Estimators," *Communications in Statistics, A,* 5, 307–323.

Lawley, D. N. (1965) "Tests of Significance for the Latent Roots of Covariance and Correlation Matrices," *Biometrika,* 43, 128–136.

Leamer, E. (1983) "Model Choice and Specification Analysis," in *Handbook of Econometrics, Vol. I,* Z. Griliches and M. Intriligator, eds., North Holland, Amsterdam, 285–331.

Leamer, E. (1978) *Specification Searches: Ad Hoc Inference with Nonexperimental Data,* Wiley, New York.

Lin, K. and J. Kmenta (1982) "Ridge Regression Under Alternative Loss Criteria," *Review of Economics and Statistics,* 64, 488–494.

Lindley, D. V. and A. F. M. Smith (1972) "Bayes Estimates for the Linear Model," *Journal of the Royal Statistical Society, B,* 34, 1–41.

Lott, W. F. (1973) "The Optimal Set of Principal Component Restrictions on a Least Squares Regression," *Communications in Statistics, A,* 2, 449–464.

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

Marquardt, D. W. and R. D. Snee (1975) "Ridge Regression in Practice," *American Statistician,* 29, 3–20.

Mason, R. L., R. F. Gunst, and J. T. Webster (1975) "Regression Analysis and Problems of Multicollinearity," *Communications in Statistics,* 4, 277–292.

Massey, W. F. (1965) "Principal Components Regression Exploratory Statistic Research," *Journal of the American Statistical Association,* 60, 234–256.

McCallum, B. T. (1979) "Artificial Orthogonalization in Regression Analysis," *Review of Economics and Statistics,* 52, 110–113.

McDonald, G. C. and D. I. Galarneau (1975) "A Monte Carlo Evaluation of Some Ridge-Type Estimators," *Journal of the American Statistical Association,* 70, 407–416.

Mittelhammer, R. (1984) "Risk Comparisons of Restricted Least Squares, Pretest, OLS and Stein Estimators Under Model Misspecification," *Journal of Econometrics,* forthcoming.

Obenchain, R. L. (1975) "Ridge Regression Following a Preliminary Test of a Shrunken Hypothesis," *Technometrics,* 17, 431–441.

O'Hagen, J. and B. McCabe (1975) "Tests for the Severity of Multicollinearity in Regression Analysis: A Comment," *Review of Economics and Statistics,* 57, 368–370.

Oman, S. (1982) "Contracting Towards Subspaces When Estimating the Mean of a Multivariate Normal Distribution," *Journal of Multivariate Analysis,* 12, 270–290.

Rao, C. R. (1973) *Linear Statistical Inference and Its Applications,* 2nd ed., Wiley, New York.

Schmidt, P. (1976) *Econometrics,* Dekker, New York.

Sclove, S. L., C. Morris, and R. Radhakrishnan (1972) "Nonoptimality of Preliminary Test Estimators for the Multinormal Mean, *Annals of Mathematical Statistics,* 42, 1481–1490.

Silvey, S. D. (1969) "Multicollinearity and Imprecise Estimation," *Journal of the Royal Statistical Society, Series B,* 35, 67–75.

Smith, G. (1974) "Multicollinearity and Forecasting," Cowles Foundation Discussion Paper No. 33.

Smith, G. and F. Campbell (1980) "A Critique of Some Ridge Regression Methods," *Journal of the American Statistical Association,* 75, 74–103.

Srivastava, V. and D. Giles (1982) "A Pre-Test General Ridge Regression Estimator: Exact Finite Sample Properties," Working paper No. 3/82, Department of Econometrics and Operations Research, Monash, University.

Stein, C. (1973) "Estimation of the Mean of a Multivariate Normal Distribution," Technical Report No. 48, Department of Statistics, Stanford University, Stanford, Calif.

Stein, C. (1981) "Estimation of the Mean of a Multivariate Normal Distribution," *Annals of Statistics,* 9, 1135–1151.

Stone, R. (1947) "On the Interpendence of Blocks of Transactions," Supplement to the *Journal of the Royal Statistical Society,* 9, 1–45.

Strawderman, W. E. (1978) "Minimax Adaptive Generalized Ridge Regression Estimators," *Journal of the American Statistical Association,* 73, 623–627.

Teekens, R. and P. M. C. de Boer (1977) "The Exact MSE-Efficiency of the General Ridge Estimator Relative to OLS," presented at the summer 1977 meeting of the Econometric Society.

Theil, H. (1963) "On the Use of Incomplete Prior Information in Regression Analysis," *Journal of the American Statistical Association,* 58, 401–414.

Theil, H. (1971) *Principles of Econometrics,* Wiley, New York.

Theobold, C. M. (1974) "Generalizations of Mean Square Error Applied to Ridge Regression," *Journal of the Royal Statistical Society, Series B,* 36, 103–106.

Thisted, R. (1977) Ridge Regression, Minimax Estimation, and Empirical Bayes Methods, unpublished Ph.D. dissertation, Stanford University, Stanford, Calif.

Thisted, R. (1978a) "Multicollinearity, Information and Ridge Regression," Technical Report No. 66, Department of Statistics, University of Chicago.

Thisted, R. (1978b) "On Generalized Ridge Regressions," Technical Report No. 57, Department of Statistics, University of Chicago.

Thisted, R. and C. Morris (1980) "Theoretical Results for Adaptive Ordinary Ridge Regression Estimators," Technical Report No. 94, Department of Statistics, University of Chicago.

Toro-Vizcarrondo, C. and T. D. Wallace (1968) "A Test of the Mean Square

Error Criterion for Restrictions in Linear Regression," *Journal of the American Statistical Association,* 63, 558–572.

Trenkler, G. (1984) "Some Further Remarks on Multicollinearity and the Minimax Conditions of the Bock Stein-Like Estimator," *Econometrica,* forthcoming.

Trivedi, P. K. (1978) "Estimation of a Distributed Lag Model Under Quadratic Loss," *Econometrica,* 46, 1181–1192.

Vinod, H. (1978) "A Ridge Estimator Whose MSE Dominates OLS," *International Economic Review,* 19, 727–737.

Vinod, H., A. Ullah, and K. Kadiyala, (1979) "Evaluation of the Mean Squared Error of Certain Generalized Ridge Estimators Using Confluent Hypergeometric Functions," Bell Laboratories Economic Discussion Paper 137, Murry Hill, N.J.

Vinod, H. D. (1976) "Application of New Ridge Regression Methods to a Study of Bell System Scale Economies," *Journal of the American Statistical Association,* 71, 835–841.

Vinod, H. D. (1978) "A Survey of Ridge Regression and Related Techniques for Improvements over Ordinary Least Squares," *Review of Economics and Statistics,* 60, 121–131.

Vinod, H. D. and A. Ullah (1981) *Recent Advances in Regression Methods,* Marcel Dekker, New York.

Wichern, D. and G. Churchill (1978) "A Comparison of Ridge Estimators," *Technometrics,* 20, 301–311.

Wichers, C. (1975) "The Detection of Multicollinearity: A Comment," *Review of Economics and Statistics,* 57, 366–368.

Willan, A. R. and D. G. Watts (1978) "Meaningful Multicollinearity Measures," *Technometrics,* 20, 407–411.

Zellner, A. (1971) *An Introduction to Bayesian Inference in Econometrics,* Wiley, New York.

Appendix B

Numerical Optimization Methods

B.1 INTRODUCTION

Estimating the parameters of a statistical model usually requires optimizing some kind of objective function. For instance, least squares estimates are obtained by minimizing a sum of squares, and maximum likelihood estimation is done by maximizing the likelihood function. In many situations it is not possible to give a closed form expression for the estimates as a function of the sample values. This occurs, for example, when the statistical model is nonlinear in the parameters, as in Chapter 6, or, more generally, when the likelihood function or sum of squares cannot be transformed so that the normal equations are linear.

In the following discussion we will assume that an objective function $H(\boldsymbol{\theta})$ is given that is to be *minimized* with respect to the $(K \times 1)$ parameter vector $\boldsymbol{\theta}$. This vector may of course contain variance-covariance parameters. The objective function is assumed to be sufficiently often differentiable so that all required derivatives exist. Note that *maximization* problems can be solved in this framework by simply minimizing the negative of the original objective function. In Section B.2 we will present some unconstrained optimization algorithms and in Section B.3 constrained minimization is discussed. Throughout we will give references for further details on the considered topics. For general reading and many more references the reader is referred to Bard (1974), Himmelblau (1972a), Künzi and Oettli (1969), Lootsma (1972a), Murray (1972a), Fletcher (1969), Fiacco and McCormick (1968), Powell (1982a), and Quandt (1983).

Computer software is available for the methods outlined in the following discussion. Some programs are listed in Bard (1974, Appendix G), and FORTRAN programs can be found in Himmelblau (1972a, Appendix B). For a discussion of FORTRAN subroutines and ALGOL 60 procedures, see Fletcher (1972b). Also, many computer packages contain easily used algorithms. For a general discussion of the available computer software see Powell (1982a, Part 6). In many programs slight modifications of the methods described here are used.

B.2 UNCONSTRAINED OPTIMIZATION

Most of the minimization methods discussed in the following are *iterative* and follow the general scheme depicted in Figure B.1. In this approach we try to

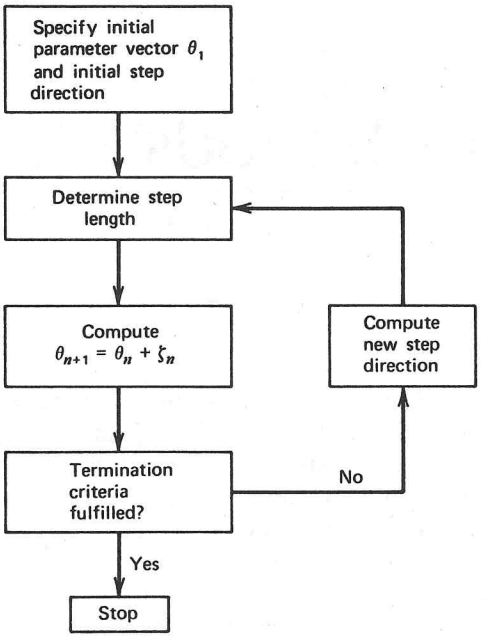

Figure B.1 Flow diagram for iterative optimization methods.

find a sequence $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N$ of vectors in the parameter space such that $\boldsymbol{\theta}_N$ minimizes $H(\boldsymbol{\theta})$ approximately. Starting with some *initial vector* $\boldsymbol{\theta}_1$, each of the following elements in the sequence is based on the preceding one in that we add a vector $\boldsymbol{\zeta}_n$, called a *step*, to $\boldsymbol{\theta}_n$ in order to determine $\boldsymbol{\theta}_{n+1}$. Thus

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n + \boldsymbol{\zeta}_n \tag{B.2.1}$$

Although this might not be the shortest way to the minimum of $H(\boldsymbol{\theta})$, we usually require $H(\boldsymbol{\theta}_n) > H(\boldsymbol{\theta}_{n+1})$. A step that meets this condition is called *acceptable*.

Ideally, the procedure should terminate when no further reduction of the objective function can be obtained. For practical purposes the iteration stops, for example, if for a prespecified small $\varepsilon > 0$:

1. $(\boldsymbol{\theta}_{n+\ell} - \boldsymbol{\theta}_n)'(\boldsymbol{\theta}_{n+\ell} - \boldsymbol{\theta}_n) < \varepsilon$.
2. $H(\boldsymbol{\theta}_n) - H(\boldsymbol{\theta}_{n+\ell}) < \varepsilon$ for a positive integer ℓ.
3. $\left[\left.\frac{\partial H}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}_n}\right]'\left[\left.\frac{\partial H}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}_n}\right] < \varepsilon$.
4. A prespecified upper bound for the number of iterations is attained.
5. A prespecified upper limit for the computation time is reached.

Usually, we do not wish to rely on a single one of these criteria, since (1) does not guarantee a termination if, for example, the minimum is not unique and (2)

and (3) may never occur if no minimum exists or the algorithm used does not approach the existing minimum for some reason. Conditions (4) and (5) are useful to stop a search in a region of the parameter space far away from a minimum. Different starting values may be tried to locate the minimum in a reasonable number of iterations. Let us now turn to a discussion of some of the possibilities in choosing a step.

B.2.1 Gradient Methods

Given a point $\boldsymbol{\theta}_n$ in the parameter space, we seek a direction $\boldsymbol{\delta}$ in which to go downhill—that is, a direction in which the objective function declines. However, we have to be careful not to step too far in this direction, since this may carry us beyond the trough and uphill again. On the other hand, too short a step will be inefficient. In other words, we have to choose an appropriate *step length* t and a *step direction* $\boldsymbol{\delta}$ such that

$$H(\boldsymbol{\theta}_n + t\boldsymbol{\delta}) < H(\boldsymbol{\theta}_n) \tag{B.2.2}$$

If $\boldsymbol{\delta}$ is a downhill direction, a small step in that direction will always decrease the objective function. Thus we are looking for $\boldsymbol{\delta}$ such that $H(\boldsymbol{\theta}_n + t\boldsymbol{\delta})$ is a decreasing function of t for t sufficiently close to zero. Consequently, for our $\boldsymbol{\delta}$,

$$\left.\frac{d[H(\boldsymbol{\theta}_n + t\boldsymbol{\delta})]}{dt}\right|_{t=0} = \left[\left.\frac{\partial H}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}_n}\right]' \left[\left.\frac{d(\boldsymbol{\theta}_n + t\boldsymbol{\delta})}{dt}\right|_{t=0}\right] = \left[\left.\frac{\partial H}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}_n}\right]' \boldsymbol{\delta} \tag{B.2.3}$$

has to be less than zero. Abbreviating the *gradient* of the objective function

$$\left.\frac{\partial H}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}_n}$$

by $\boldsymbol{\gamma}_n$, it is clear that we can choose

$$\boldsymbol{\delta} = -P_n\boldsymbol{\gamma}_n \tag{B.2.4}$$

where P_n is any positive definite matrix, that is, $\boldsymbol{\gamma}'P_n\boldsymbol{\gamma} > 0$ for all vectors $\boldsymbol{\gamma} \neq \mathbf{0}$ and, therefore, $\boldsymbol{\gamma}_n'\boldsymbol{\delta} = -\boldsymbol{\gamma}_n'P_n\boldsymbol{\gamma}_n < 0$ if $\boldsymbol{\gamma}_n \neq \mathbf{0}$. For $\boldsymbol{\gamma}_n = \mathbf{0}$, we can hope that we have reached the trough. As the general form of an iteration we get from (B.2.1)

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - t_nP_n\boldsymbol{\gamma}_n \tag{B.2.5}$$

where t_n is the step length in the nth iteration.

Clearly, there are many downhill directions in these mountains, at least as many as there are positive definite matrices. Since our limited horizon in most cases does not allow an a priori optimal choice, many different ways to the

trough are proposed in the literature. In some algorithms the step length is determined simultaneously with the direction, whereas other methods specify the direction only and allow various possibilities for the step length selection. These can range from using the first trial of t that fulfills (B.2.2) to an optimization of the step length in the given direction. For a description of some possible procedures, see Flanagan, Vitale, and Mendelsohn (1969). Determination of the optimal step length in each iteration is likely to decrease the number of steps to the trough but increases the computational cost for each iteration. Bard (1970) and Flanagan, Vitale, and Mendelsohn (1969) found that for their test problems it did not pay to use the most expensive procedures for the step length selection and according to Dennis (1973, p. 161) efforts to improve the gradient methods focus on modifications of the direction rather than the step length.

Usually, the gradient methods are differentiated by the step direction, or since almost all use the basic formula (B.2.5), by the *direction matrix* P_n. In the following section we will discuss possible choices, some of which are summarized in Table B.1.

TABLE B.1 *SOME GRADIENT METHODS*

Method	Direction Matrix in the nth Iteration P_n	Section
Steepest descent	I_K	B.2.2
Newton–Raphson	$\left[\left.\frac{\partial^2 H}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\right\|_{\boldsymbol{\theta}_n}\right]^{-1}$	B.2.3
Rank one correction	$P_{n-1} + \frac{\boldsymbol{\eta}_{n-1}\boldsymbol{\eta}_{n-1}'}{\boldsymbol{\eta}_{n-1}'(\boldsymbol{\gamma}_n - \boldsymbol{\gamma}_{n-1})}$	B.2.4
Davidon–Fletcher–Powell	$P_{n-1} + \frac{\boldsymbol{\zeta}_{n-1}\boldsymbol{\zeta}_{n-1}'}{\boldsymbol{\zeta}_{n-1}'(\boldsymbol{\gamma}_n - \boldsymbol{\gamma}_{n-1})} - \frac{P_{n-1}(\boldsymbol{\gamma}_n - \boldsymbol{\gamma}_{n-1})(\boldsymbol{\gamma}_n - \boldsymbol{\gamma}_{n-1})'P_{n-1}}{(\boldsymbol{\gamma}_n - \boldsymbol{\gamma}_{n-1})'P_{n-1}(\boldsymbol{\gamma}_n - \boldsymbol{\gamma}_{n-1})}$	B.2.4
Gauss	$[Z(\boldsymbol{\theta}_n)'Z(\boldsymbol{\theta}_n)]^{-1}$	B.2.5
Method of scoring	$-\left[E\left.\frac{\partial^2 \ln \ell}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\right\|_{\boldsymbol{\theta}_n}\right]^{-1}$	B.2.5
Brown–Dennis	$\left[Z(\boldsymbol{\theta}_n)'Z(\boldsymbol{\theta}_n) - \sum_{t,t'=1}^{T}[y_t - f_t(\boldsymbol{\theta}_n)]F_{t',n}\right]^{-1}$	B.2.6
Marquardt	$[Z(\boldsymbol{\theta}_n)'Z(\boldsymbol{\theta}_n) + \lambda_n I_K]^{-1}$	B.2.7
Quadratic hill climbing	$\left[\left.\frac{\partial^2 H}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\right\|_{\boldsymbol{\theta}_n} + \lambda_n I_K\right]^{-1}$	B.2.7

B.2.2 Method of Steepest Descent

It can be shown that the initially steepest descent is obtained if we choose

$$P_n = I_K \tag{B.2.6}$$

in all iterations, where K is the dimension of the parameter space as before. Although this method is very simple, its use cannot be recommended in most cases, since it may converge very slowly if the minimum is in a long and narrow valley, that is, if the objective function is ill-conditioned. It is clear that using the same direction matrix P_n in each iteration does not allow a flexible adjustment to different shapes of the objective function surface. However, the steepest descent method can be valuable if it is combined with other algorithms as discussed in Section B.2.7.

B.2.3 Newton–Raphson Method

The Newton–Raphson or simply Newton algorithm uses the inverse of the Hessian matrix to specify the step direction in each iteration; that is,

$$P_n = \left[\frac{\partial^2 H}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}_n}\right]^{-1} \tag{B.2.7}$$

In the sequel the Hessian of $H(\boldsymbol{\theta})$ in $\boldsymbol{\theta}_n$ will be denoted by $\mathcal{H}_n$. To see why this direction matrix is chosen, we approximate $H(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_n$ by its Taylor series expansion up to the quadratic terms:

$$H(\boldsymbol{\theta}) \simeq H(\boldsymbol{\theta}_n) + \boldsymbol{\gamma}_n'(\boldsymbol{\theta} - \boldsymbol{\theta}_n) + \tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_n)'\mathcal{H}_n(\boldsymbol{\theta} - \boldsymbol{\theta}_n) \tag{B.2.8}$$

First-order conditions for a minimum of the right-hand side are

$$\boldsymbol{\gamma}_n + \mathcal{H}_n(\boldsymbol{\theta} - \boldsymbol{\theta}_n) = \mathbf{0} \tag{B.2.9}$$

or

$$\boldsymbol{\theta} = \boldsymbol{\theta}_n - \mathcal{H}_n^{-1}\boldsymbol{\gamma}_n \tag{B.2.10}$$

Hence, if $H(\boldsymbol{\theta})$ is quadratic, that is, we have an exact equality in (B.2.8), we reach the minimum in one step of length one. In general, however, the Hessian may not be positive definite outside a small neighborhood of the minimum, and thus iterations of the form (B.2.10) may carry us to a maximum or saddle point. Consequently, a step (B.2.10) may not be acceptable. Other disadvantages are that first and second partial derivatives have to be determined analytically, which places a heavy burden on the user of this algorithm. Whereas the analytic determination of the ingredients of the gradient is acceptable in most

cases, analytic computation of higher-order partial derivatives is considered to be at least a possible error source. The gradient methods described in the following subsections can be viewed as efforts to overcome the disadvantage that analytic determination of second-order partial derivatives is necessary, without giving up the advantage of fast local convergence.

To illustrate the Newton algorithm we use an example that is discussed in more detail in Judge et al. (1982, Chapter 24). We consider the nonlinear statistical model

$$y_t = \theta_1^* + \theta_2^* x_{t2} + \theta_2^{*2} x_{t3} + e_t, \qquad t = 1, \ldots, 20 \tag{B.2.11}$$

$$\mathbf{y} = \mathbf{f}(\boldsymbol{\theta}^*) + \mathbf{e} \tag{B.2.12}$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_{20})'$, $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*)'$ is the true parameter vector, $\mathbf{e} = (e_1, e_2, \ldots, e_{20})'$ and

$$\mathbf{f}(\boldsymbol{\theta}) = \begin{bmatrix} \theta_1 + \theta_2 x_{12} + \theta_2^2 x_{13} \\ \theta_1 + \theta_2 x_{22} + \theta_2^2 x_{23} \\ \vdots \\ \theta_1 + \theta_2 x_{20,2} + \theta_2^2 x_{20,3} \end{bmatrix}$$

TABLE B.2 *DATA FOR EXAMPLE MODEL*

t	y_t	x_{t1}	x_{t2}	x_{t3}
1	4.284	1.000	0.286	0.645
2	4.149	1.000	0.973	0.585
3	3.877	1.000	0.384	0.310
4	0.533	1.000	0.276	0.058
5	2.211	1.000	0.973	0.455
6	2.389	1.000	0.543	0.779
7	2.145	1.000	0.957	0.259
8	3.231	1.000	0.948	0.202
9	1.998	1.000	0.543	0.028
10	1.379	1.000	0.797	0.099
11	2.106	1.000	0.936	0.142
12	1.428	1.000	0.889	0.296
13	1.011	1.000	0.006	0.175
14	2.179	1.000	0.828	0.180
15	2.858	1.000	0.399	0.842
16	1.388	1.000	0.617	0.039
17	1.651	1.000	0.939	0.103
18	1.593	1.000	0.784	0.620
19	1.046	1.000	0.072	0.158
20	2.152	1.000	0.889	0.704

The data used for this example are given in Table B.2. The x_{t2} and x_{t3} are pseudo-random numbers from a uniform distribution on the unit interval, and the y_t are computed by adding a normal pseudo-random number to $\theta_1^* + \theta_2^* x_{t2} + \theta_2^{*2} x_{t3}$, where the true parameters θ_1^* and θ_2^* were chosen to be $\theta_1^* = \theta_2^* = 1$. That is, we have actually added the random error to $1 + x_{t2} + x_{t3}$.

To determine the least squares estimate of $\boldsymbol{\theta}^*$ we have to minimize

$$H(\boldsymbol{\theta}) = [\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})]'[\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})] \tag{B.2.13}$$

Loci in the parameter space with constant $H(\boldsymbol{\theta})$ are depicted in Figure B.2. Using the iterations

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \mathcal{H}_n^{-1}\boldsymbol{\gamma}_n$$

and three different starting values, we obtained the results given in Table B.3. The algorithm terminates at two different points in the parameter space. This outcome is not surprising since $H(\boldsymbol{\theta})$ has two different local minima (see Figure B.2). The example shows that an optimization algorithm does not necessarily converge to the global minimum. We will return to this problem later on.

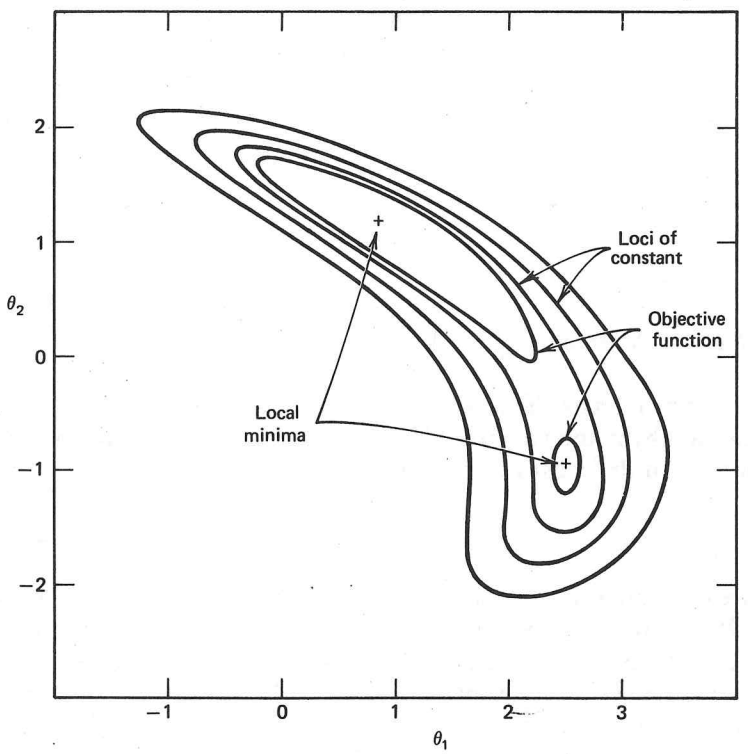

Figure B.2 Loci of constant objective function $H(\boldsymbol{\theta})$.

TABLE B.3 *ITERATIONS OF THE NEWTON ALGORITHM*

n	$\theta_{n,1}$	$\theta_{n,2}$	$H(\boldsymbol{\theta}_n)$
1	3.000000	2.000000	264.3918
2	−0.084033	1.811210	20.6328
3	0.625029	1.423940	16.5105
4	0.817259	1.272776	16.0961
5	0.862590	1.237516	16.0818
6	0.864782	1.235753	16.0817
7	0.864787	1.235748	16.0817
8	0.864787	1.235748	16.0817
1	0.000000	2.000000	29.2758
2	0.334936	1.600435	17.7382
3	0.735040	1.336953	16.1955
4	0.849677	1.247743	16.0832
5	0.864541	1.235946	16.0817
6	0.864787	1.235749	16.0817
7	0.864787	1.235748	16.0817
8	0.864787	1.235748	16.0817
1	1.500000	0.500000	20.2951
2	2.256853	0.007135	20.7735
3	2.467047	−0.436460	21.0312
4	2.316982	−0.202435	20.9467
5	2.359743	−0.320579	20.9809
6	2.354457	−0.319153	20.9805
7	2.354471	−0.319186	20.9805
8	2.354471	−0.319186	20.9805

B.2.4 Quasi-Newton Methods

Some algorithms approximate the inverse of the Hessian of the objective function in each iteration by adding a *correction matrix* to the approximate inverse of the Hessian used in the most recent step,

$$P_{n+1} = P_n + M_n \tag{B.2.14}$$

where M_n is the correction matrix and P_n is an approximation to $\mathcal{H}_n^{-1}$. In the $(n + 1)$st step P_{n+1} is used as the direction matrix. These methods are sometimes called *variable metric methods*. Starting with some symmetric matrix P_1 and proceeding inductively, a reasonable choice of P_{n+1} results from a first-order approximation of the gradient

$$\boldsymbol{\gamma}_n \simeq \boldsymbol{\gamma}_{n+1} + \mathcal{H}_{n+1}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}) \tag{B.2.15}$$

which leads to

$$\mathcal{H}_{n+1}^{-1}(\gamma_{n+1} - \gamma_n) \simeq (\theta_{n+1} - \theta_n) \tag{B.2.16}$$

if the Hessian $\mathcal{H}_{n+1}$ is nonsingular. Replacing $\mathcal{H}_{n+1}^{-1}$ in this equation by $P_{n+1} = P_n + M_n$, we get

$$M_n(\gamma_{n+1} - \gamma_n) = \eta_n \tag{B.2.17}$$

where $\eta_n = (\theta_{n+1} - \theta_n) - P_n(\gamma_{n+1} - \gamma_n)$. Of course, M_n is required to be symmetric in order to obtain the symmetry of P_{n+1}. But even with this restriction the $K(K + 1)/2$ possibly different elements of M_n are not uniquely determined by the K equations (B.2.17) if $K > 1$. Thus different correction matrices fulfilling (B.2.17) are suggested in the literature; for example, one alternative is

$$M_n = \frac{\eta_n \eta_n'}{\eta_n'(\gamma_{n+1} - \gamma_n)} \tag{B.2.18}$$

This is the only symmetric matrix of rank one which meets the requirements of (B.2.17) [Broyden (1965, 1967)]. The resulting algorithm is therefore called the *rank one correction* (ROC) method. This choice does not necessarily lead to an acceptable step, since $P_n + M_n$ may not be positive definite.

Another famous member of this family of algorithms is the *Davidon–Fletcher–Powell* (DFP) method [Davidon (1959), Fletcher and Powell (1963)] for which

$$M_n = \frac{\zeta_n \zeta_n'}{\zeta_n'(\gamma_{n+1} - \gamma_n)} - \frac{P_n(\gamma_{n+1} - \gamma_n)(\gamma_{n+1} - \gamma_n)'P_n}{(\gamma_{n+1} - \gamma_n)'P_n(\gamma_{n+1} - \gamma_n)} \tag{B.2.19}$$

As shown by Fletcher and Powell (1963), if the step length t_n for each step $\zeta_n = -t_n P_n \gamma_n$ is selected so as to minimize $H(\theta_n + \zeta_n)$ for the given θ_n, P_n, and γ_n, then $P_{n+1} = P_n + M_n$ will always be positive definite. Therefore, choosing P_n as the step direction in the nth iteration guarantees an acceptable step. The identity matrix I_K can be used as P_1. In practice, however, the necessary computations may not be precise enough to obtain a positive definite P_n in each iteration [Bard (1968)]. Nevertheless, the DFP method is very popular and has proved very efficient in many applications.

Clearly, the choice of the step size is crucial in this algorithm, and many procedures have been developed to efficiently minimize the objective function in the direction $-P_n\gamma_n$ in the nth iteration. Some methods are discussed in Himmelblau (1972a, Section 2.6) and Dixon (1972).

Many other possible formulae have been suggested for the correction matrix M_n. Some are presented in Himmelblau (1972a, Table 3.5–1), and Huang (1970) has given a classification of many of them. See also Huang and Levy (1970), Dennis and Moré (1977), and Brodlie (1977).

Sometimes the inverse Hessian of the objective function is used to construct approximate confidence intervals for the parameter estimates. If a quasi-Newton algorithm is started with $P_1 = I_K$ and converges after only a few iterations, the direction matrix obtained in the last step may not be a good approximation to the inverse of the Hessian of the objective function. In that case this matrix should not be used to compute asymptotic confidence intervals for the parameter estimates. It is suggested to restart the algorithm and do some more iterations.

Instead of its inverse, the Hessian can also be approximated directly. In that case the inverse of the approximating matrix is used as the direction matrix. Dual formulae for many of the above-mentioned algorithms exist.

B.2.5 Gauss Method

The Gauss method, sometimes called the *Gauss-Newton method,* is based on another possibility to approximate the Hessian if the objective function is of a special form. Suppose, for instance, that we have a statistical model of the type

$$\mathbf{y} = \mathbf{f}(X,\boldsymbol{\theta}^*) + \mathbf{e} = \mathbf{f}(\boldsymbol{\theta}^*) + \mathbf{e} \tag{B.2.20}$$

where $\mathbf{y} = (y_1, \ldots, y_T)'$, $\mathbf{f}(\boldsymbol{\theta}) = [f_1(\boldsymbol{\theta}), \ldots, f_T(\boldsymbol{\theta})]'$, and $\mathbf{e} = (e_1, \ldots, e_T)'$, as discussed in Chapter 6, and the objective function is the sum of squared errors,

$$H(\boldsymbol{\theta}) = [\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})]'[\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})] = \mathbf{e}(\boldsymbol{\theta})'\mathbf{e}(\boldsymbol{\theta}) \tag{B.2.21}$$

Then the Hessian of $H(\boldsymbol{\theta})$ is

$$\mathcal{H}(\boldsymbol{\theta}) = 2Z(\boldsymbol{\theta})'Z(\boldsymbol{\theta}) - 2\sum_{t,t'=1}^{T}[y_t - f_t(\boldsymbol{\theta})]\left[\frac{\partial^2 f_{t'}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}}\right] \tag{B.2.22}$$

where $Z(\boldsymbol{\theta}) = [\partial\mathbf{f}/\partial\boldsymbol{\theta}'|_{\boldsymbol{\theta}}]$. Since the mean of $e_t = y_t - f_t(\boldsymbol{\theta}^*)$ is assumed to be zero the second term on the right-hand side is deleted and the first term is taken as an approximation of $H(\boldsymbol{\theta})$. Consequently,

$$P_n = [2Z(\boldsymbol{\theta}_n)'Z(\boldsymbol{\theta}_n)]^{-1} \tag{B.2.23}$$

and, since $\boldsymbol{\gamma}_n = -2Z(\boldsymbol{\theta}_n)'[\mathbf{y} - \mathbf{f}(\boldsymbol{\theta}_n)]$, we get with step length 1,

$$\begin{aligned}\boldsymbol{\theta}_{n+1} &= \boldsymbol{\theta}_n + [Z(\boldsymbol{\theta}_n)'Z(\boldsymbol{\theta}_n)]^{-1}Z(\boldsymbol{\theta}_n)'[\mathbf{y} - \mathbf{f}(\boldsymbol{\theta}_n)] \\ &= [Z(\boldsymbol{\theta}_n)'Z(\boldsymbol{\theta}_n)]^{-1}Z(\boldsymbol{\theta}_n)'[\mathbf{y} - \mathbf{f}(\boldsymbol{\theta}_n) + Z(\boldsymbol{\theta}_n)\boldsymbol{\theta}_n]\end{aligned} \tag{B.2.24a}$$

which is the least squares estimator for the model

$$\tilde{\mathbf{y}}(\boldsymbol{\theta}_n) = Z(\boldsymbol{\theta}_n)\boldsymbol{\theta} + \mathbf{e} \tag{B.2.25}$$

where

$$\bar{\mathbf{y}}(\boldsymbol{\theta}_n) = \mathbf{y} - \mathbf{f}(\boldsymbol{\theta}_n) + Z(\boldsymbol{\theta}_n)\boldsymbol{\theta}_n \tag{B.2.26}$$

This shows that the Gauss algorithm can be viewed as a sequence of linear regressions. In each step we compute the LS estimator for a linear approximation of the nonlinear model. It is sometimes useful to write (B.2.24a) as

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \left(\left[\frac{\partial \mathbf{e}}{\partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}_n}\right]'\left[\frac{\partial \mathbf{e}}{\partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}_n}\right]\right)^{-1}\left[\frac{\partial \mathbf{e}}{\partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}_n}\right]' \mathbf{e}(\boldsymbol{\theta}_n) \tag{B.2.24b}$$

where

$$\frac{\partial \mathbf{e}}{\partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}} = -Z(\boldsymbol{\theta})$$

has been used.

Another interesting feature of the Gauss algorithm is its relationship to the *method of scoring* [e.g., Maddala (1977)], which can be used for maximum likelihood estimation. For this algorithm the direction matrix is given by

$$P_n = -\left[E\,\frac{\partial^2 \ln \ell}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}_n}\right]^{-1} \tag{B.2.27}$$

where ℓ is the likelihood function. Thus, using the negative log likelihood function as the objective function, the Hessian is approximated by its expected value. Assuming independently, identically, normally distributed errors and deleting the variance σ^2 in the usual way from the minimization procedure, $[Z(\boldsymbol{\theta}_n)'Z(\boldsymbol{\theta}_n)]^{-1}$ can be used as direction matrix instead of (B.2.27). For a generalization of this method see Berndt, Hall, Hall, and Hausman (1974).

The Gauss method is also applicable for objective functions different from (B.2.21). [See Bard (1974, Sections 5-9 to 5-11)]. The simplicity of this algorithm and its good local convergence properties [Dennis (1973)] make it an attractive method where applicable. However, in general, P_n is singular and thus not positive definite if $\boldsymbol{\theta}_n$ is not close to $\boldsymbol{\theta}_{\min}$, which minimizes $H(\boldsymbol{\theta})$ and, if the linear approximation of the nonlinear model is poor or the residuals are large, convergence can be slow. For the example model (B.2.11) iterations of the Gauss algorithm are given in Table B.4.

B.2.6 Combining Gauss and Quasi-Newton Methods

Since the performance of the Gauss method depends on the size of the residuals in a particular model or more precisely on

$$\sum_{t,t'=1}^{T} [y_t - f_t(\boldsymbol{\theta})]\left[\frac{\partial^2 f_{t'}}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}}\right] \tag{B.2.28}$$

TABLE B.4 *ITERATIONS OF THE GAUSS ALGORITHM*

n	$\theta_{n,1}$	$\theta_{n,2}$	$H(\boldsymbol{\theta}_n)$
1	3.000000	2.000000	264.3918
2	0.723481	1.404965	16.6635
3	0.837007	1.259230	16.0880
4	0.861002	1.238408	16.0818
5	0.864359	1.236040	16.0817
6	0.864740	1.235780	16.0817
7	0.864782	1.235752	16.0817
8	0.864787	1.235749	16.0817
9	0.864787	1.235749	16.0817
1	3.000000	−1.000000	25.5156
2	2.498561	−0.989894	20.4856
3	2.498566	−0.985678	20.4824
4	2.498571	−0.983903	20.4823
5	2.498574	−0.983154	20.4823
6	2.498575	−0.982837	20.4823
7	2.498576	−0.982703	20.4823
8	2.498576	−0.982646	20.4823
9	2.498576	−0.982623	20.4823
10	2.498576	−0.982612	20.4823
11	2.498576	−0.982607	20.4823
12	2.498576	−0.982605	20.4823
13	2.498576	−0.982605	20.4823
1	1.500000	0.500000	20.2951
2	1.067414	1.213585	16.6646
3	0.868351	1.233424	16.0818
4	0.865161	1.235496	16.0817
5	0.864828	1.235721	16.0817
6	0.864792	1.235746	16.0817
7	0.864788	1.235748	16.0817
8	0.864787	1.235748	16.0817
9	0.864787	1.235748	16.0817

Brown and Dennis (1971) suggest the possibility of combining the Gauss algorithm with a quasi-Newton method. Instead of the inverse Hessian of the objective function, we approximate the Hessian of $f_t(\boldsymbol{\theta})$ iteratively; that is, we choose

$$F_{t,n+1} = F_{t,n} + M_{t,n} \tag{B.2.29}$$

where $F_{t,n}$ is an approximation to

$$\left.\frac{\partial^2 f_t}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}_n} \tag{B.2.30}$$

From

$$\left.\frac{\partial f_t}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}_n} \simeq \left.\frac{\partial f_t}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}_{n+1}} + \left[\left.\frac{\partial^2 f_t}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}_{n+1}}\right] (\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}) \tag{B.2.31}$$

it follows that $M_{t,n}$ should satisfy

$$M_{t,n}(\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n) = \boldsymbol{\mu}_{t,n} \tag{B.2.32}$$

with

$$\boldsymbol{\mu}_{t,n} = \left.\frac{\partial f_t}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}_{n+1}} - \left.\frac{\partial f_t}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}_n} - F_{t,n}(\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n) \tag{B.2.33}$$

[compare (B.2.15) to (B.2.17)]. For instance, a correction matrix of rank one such as

$$M_{t,n} = \frac{\boldsymbol{\mu}_{t,n}(\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n)'}{(\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n)'(\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n)} \tag{B.2.34}$$

could be used.

The direction matrix for this algorithm is

$$P_n = \{Z(\boldsymbol{\theta}_n)'Z(\boldsymbol{\theta}_n) - \sum_{t,t'=1}^{T} [y_t - f_t(\boldsymbol{\theta}_n)]F_{t',n}\}^{-1} \tag{B.2.35}$$

As one alternative the unit matrix could be substituted for $F_{t,1}$ for all t. For another possible choice see Dennis (1973, p. 173). Brown and Dennis (1971) found a good performance of this algorithm in a comparison with other methods. However, the computer storage requirements for this procedure are relatively high.

B.2.7 Modifications

B.2.7a Marquardt's Method

The Marquardt algorithm [Marquardt (1963)], sometimes referred to as the *Marquardt–Levenberg method,* can be used to modify procedures that do not guarantee a positive definite direction matrix P_n. This algorithm utilizes the fact that

$$P_n + \lambda_n \bar{P}_n \tag{B.2.36}$$

is always positive definite if $\bar{P}_n$ is positive definite and the scalar λ_n is sufficiently large. A possible choice for $\bar{P}_n$ is the identity matrix. Typically, this method is used in combination with the Gauss algorithm and $Z(\boldsymbol{\theta}_n)'Z(\boldsymbol{\theta}_n)$ is

modified rather than its inverse. Thus the new direction matrix is given by

$$P_n = [Z(\theta_n)'Z(\theta_n) + \lambda_n \bar{P}_n]^{-1} \tag{B.2.37}$$

where I_K can be used as $\bar{P}_n$. For a λ_n close to zero, this method is equivalent to the Gauss algorithm. On the other hand, for increasing λ_n, the steepest descent method is approached. Since the Gauss algorithm performs very well in a neighborhood of the minimum, we start with a small λ_1 and decrease $\lambda_n > 0$ in each iteration unless this results in an unacceptable step. Note that the step length is determined simultaneously with the step direction. For the typical step (B.2.5) t_n always takes the value one.

Similarly, we could modify the Hessian of the objective function and use

$$P_n = [\mathcal{H}_n + \lambda_n I_K]^{-1} \tag{B.2.38}$$

as the direction matrix. This algorithm is usually referred to as the *quadratic hill-climbing method,* since it was introduced in a maximization context [Goldfeld, Quandt, and Trotter (1966)]. The performance of these algorithms is not invariant under transformations of the parameter space and thus an improvement may be obtained by using a matrix $\bar{P}_n$ that is different from I_K [Marquardt (1963), Goldfeld and Quandt (1972, Chapter 1)]. Marquardt's method appears to perform very well in practice even if the initial parameter vector θ_1 is not close to the minimum of the objective function.

B.2.7b Some Other Modifications

Some other possibilities to enforce acceptable steps are based upon the eigenvalue decomposition of the direction matrix

$$P_n = U_n' \Lambda_n U_n \tag{B.2.39}$$

where U_n is an orthogonal matrix and Λ_n a diagonal matrix with diagonal elements consisting of the characteristic roots $\lambda_j, j = 1, 2, \ldots, K$, of P_n. If P_n is not positive definite, then some λ_j are negative or zero and can be replaced by

$$\lambda_j^* = \max[|\lambda_j|, \mu] \tag{B.2.40}$$

[Greenstadt (1967)], or

$$\bar{\lambda}_j = \begin{cases} \max[|\lambda_j|, \mu] & \text{if } |\lambda_j|^{-1} > \varepsilon \\ \eta & \text{if } |\lambda_j|^{-1} \leq \varepsilon \end{cases} \tag{B.2.41}$$

[Bard (1974), p. 93]. Here 0^{-1} is defined to be infinity and ε, μ, and η are suitable positive constants. In these proposals a positive lower bound for the absolute value of the characteristic roots is given, since in practical computations a very small number is treated as zero.

Another possibility is to replace a nonpositive definite matrix P_n by the identity matrix and perform a step in the direction of steepest descent. This modification is sometimes used in conjunction with quasi-Newton methods and can be interpreted as a restarting of the algorithm.

B.2.8 Conjugate Gradient Algorithm

This algorithm was suggested by Fletcher and Reeves (1964) and is based on an idea that is different from the other gradient methods. Suppose the objective function $H(\boldsymbol{\theta})$ is quadratic and thus the Hessian $\mathcal{H}$ is constant. Two directions $\boldsymbol{\delta}$ and $\bar{\boldsymbol{\delta}}$ in the K-dimensional $\boldsymbol{\theta}$-space are called *conjugate* if $\boldsymbol{\delta}'\mathcal{H}\bar{\boldsymbol{\delta}} = 0$. For a given $\boldsymbol{\delta} \neq \mathbf{0}$ there are $K - 1$ conjugate directions in the parameter space. Starting at an initial vector $\boldsymbol{\theta}_1$ with a step in the steepest descent direction and then performing steps in all conjugate directions will lead to the global minimum, provided that $H(\boldsymbol{\theta})$ is minimized in the given direction at each stage, that is, the optimal step length has to be chosen in each iteration. It turns out that the conjugate directions can be computed inductively by the formula

$$\boldsymbol{\delta}_{n+1} = -\boldsymbol{\gamma}_n + \frac{\boldsymbol{\gamma}_n'\boldsymbol{\gamma}_n}{\boldsymbol{\gamma}_{n-1}'\boldsymbol{\gamma}_{n-1}}\boldsymbol{\delta}_n \tag{B.2.42}$$

using $\boldsymbol{\delta}_1 = -\boldsymbol{\gamma}_1$, the direction of steepest descent. As before, $\boldsymbol{\gamma}_n$ denotes the gradient at $\boldsymbol{\theta}_n$. In the conjugate gradient algorithm a step is thus chosen to be of the form

$$\boldsymbol{\zeta}_n = t_n\boldsymbol{\delta}_n \tag{B.2.43}$$

where t_n is selected so as to minimize $H(\boldsymbol{\theta}_n + t_n\boldsymbol{\delta}_n)$. This guarantees a nonincreasing objective function in each step, but it does not guarantee an arrival at the minimum in K steps if $H(\boldsymbol{\theta})$ is not quadratic. Therefore, the algorithm has to be restarted periodically in the direction of the negative gradient.

An obvious advantage of this method is the simple nature of the steps. In fact, it is only necessary to store the gradient and the step direction vector at each iteration stage whereas the storage of the gradient and the direction matrix is required for other gradient algorithms. Note that the DFP method can also be interpreted as a conjugate gradient procedure. For a further discussion see Polak (1971) and Fletcher (1972a).

B.2.9 Jacobi's Method and the Gauss-Seidel Algorithm

From (B.2.5) it follows that the above-mentioned algorithms consist of iterations of the general form

$$\boldsymbol{\theta}_{n+1} = \mathbf{F}(\boldsymbol{\theta}_n) \tag{B.2.44}$$

where $\mathbf{F}(\boldsymbol{\theta}) = [F_1(\boldsymbol{\theta}), \ldots, F_K(\boldsymbol{\theta})]'$ and the optimum is reached when $\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n$. In other words, the algorithms determine a *fixed point* of the function $\mathbf{F}(\cdot)$. If the original problem is formulated in this way, the iteration (B.2.44) can be used directly. This procedure is called *Jacobi method*. It does not converge in general.

The *Gauss-Seidel algorithm* follows similar ideas. It is applicable if the original problem is set up so that, in the optimum,

$$\theta_i = G_i(\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_K), \qquad i = 1, 2, \ldots, K \tag{B.2.45}$$

In this case iterations similar to (B.2.44) can be carried out. For further details and references see Quandt (1983).

B.2.10 Derivative Free Methods

One disadvantage of the gradient methods is that they require analytic computation of at least first partial derivatives to find the step direction. Actually, it is possible to let the computer remove this burden from the user, since approximately

$$\left.\frac{\partial H}{\partial \theta_i}\right|_{\boldsymbol{\theta}} \simeq [H(\theta_1, \ldots, \theta_{i-1}, \theta_i + \Delta\theta_i, \theta_{i+1}, \ldots, \theta_K) - H(\theta_1, \ldots, \theta_{i-1}, \theta_i - \Delta\theta_i, \theta_{i+1}, \ldots, \theta_K)]/2\Delta\theta_i \tag{B.2.46}$$

if $\Delta\theta_i$ is sufficiently small. Also, a similar one-sided approximation could be used. But it is clear that this convenience will not only increase the computational cost but also the inaccuracy of the calculations. Of course, it is possible to compute other necessary derivatives in a similar way. Details are given in Bard (1974, Section 5-18) and Quandt (1983).

B.2.10a Direct Search Methods

Another possible way of avoiding partial derivatives is to apply a *direct search method*. These procedures are useful if the first partial derivatives of the objective function do not exist or are difficult to compute. For an example see Goldfeld and Quandt (1972, Chapter 5). Starting from some initial $(K \times 1)$ parameter vector $\boldsymbol{\theta}_1$, a search is performed in K directions $\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \ldots, \boldsymbol{\delta}_K$, which are at least linearly independent but often orthogonal. A typical iteration is

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n + t_n \boldsymbol{\delta}_j \tag{B.2.47}$$

The step length t_n is chosen such that $H(\boldsymbol{\theta}_{n+1}) \leq H(\boldsymbol{\theta}_n)$. The methods differ in how they select step length and direction and how and when they apply a new set of direction vectors. An algorithm suggested by Powell (1964) has proved quite successful in some applications. As in the conjugate gradient method, a search is performed along conjugate directions. For details see Powell's article or Himmelblau (1972a, Chapter 4), where other derivative free methods are

also described. A short review and comparison of some procedures is given by Fletcher (1965). Other references include Rosenbrock (1960), Hooke and Jeeves (1961), Powell (1965), and Swann (1972).

B.2.10b Simplex Algorithm

A search method based on yet another idea is the so-called *simplex algorithm* suggested by Spendley, Hext, and Himsworth (1962), not to be confused with the simplex method for the solution of linear programming problems. A simplex is spanned by $K + 1$ vectors $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{K+1}$ in the K-dimensional space. These vectors are the vertices of the simplex. For example, in the plane the two-dimensional simplexes are triangles and in three-space tetrahedrons are simplexes of maximum dimension. Suppose that m is such that

$$H(\boldsymbol{\theta}_m) = \max_{j = 1, 2, \ldots, K+1} H(\boldsymbol{\theta}_j)$$

Then $\boldsymbol{\theta}_m$ is replaced by $\bar{\boldsymbol{\theta}}$, say, a point on the ray from $\boldsymbol{\theta}_m$ through the centroid of the remaining points. The procedure is repeated, always replacing the vertex that leads to a maximum value of the objective function as illustrated in Figure B.3. Usually, modifications of this basic step will be necessary in order to guarantee satisfactory progress in locating the minimum of $H(\boldsymbol{\theta})$. A detailed

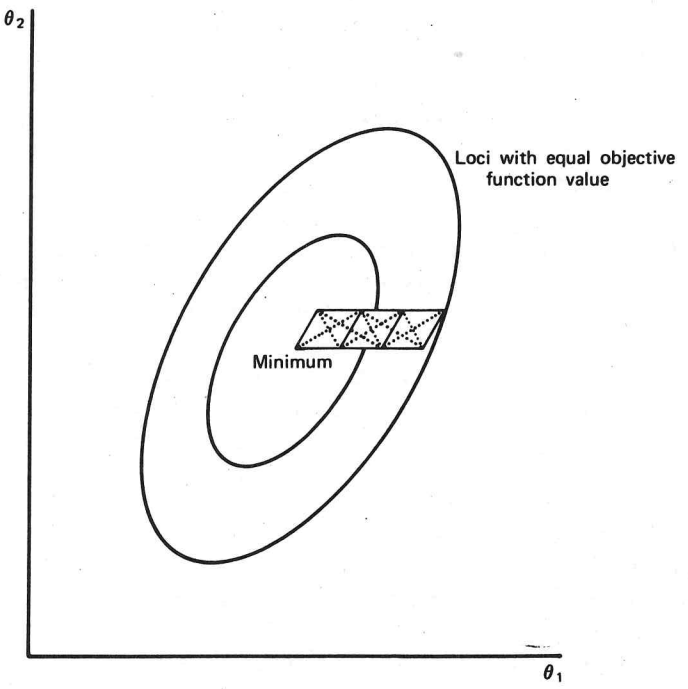

Figure B.3 Iterations of the simplex algorithm.

description of the algorithm is given by Nelder and Mead (1965); for some modifications see Parkinson and Hutchinson (1972a, 1972b). The method seems to be robust and has been successfully applied for problems with a high-dimensional parameter space. Computer programs for the simplex method and also for Powell's algorithm are given in Himmelblau (1972a, Appendix B).

B.2.10c Grid Search

If the statistical model is nonlinear in only one parameter, and if the feasible range of this parameter is small as, for example, in the discounted polynomial lag model (Chapter 10), then sometimes a simple grid search over the range of this parameter is used. The optimal values of the linear parameters are computed conditionally on the given values of the nonlinear parameter at each grid point by applying linear estimation techniques.

A grid search is also recommended to find reasonable starting values for gradient or direct search methods. If the feasible region in the parameter space is large, a search over randomly selected points can be carried out to keep the cost of this procedure in acceptable limits. Some possible strategies for a random search are discussed in Schrack and Borowski (1972).

B.2.11 Concluding Remarks and Critique

Since for applied researchers the actual performance of an algorithm for their particular problems is of major interest, we have not given theoretical convergence results. These can be found in the references [e.g., Ortega and Rheinboldt (1970) and Dennis (1977)]. Rather, we compare the methods on the basis of some performance comparisons reported in the literature. It should be clear that the wide variety of possible objective functions does not permit an evaluation that is correct for any given problem and also the differences in the evaluation criteria used in different studies hamper generally valid statements. Criteria such as (1) robustness measured, for example, by the number of times the algorithm under consideration has converged to the minimum for a set of trial problems, (2) precision in the solution, (3) the number of function evaluations, and (4) execution time, are often used [Himmelblau (1972b)].

Bard (1970) found that for his test problems Gauss-type methods including the Marquardt algorithm were superior to quasi-Newton procedures, and Himmelblau's (1972b) results seem to indicate that quasi-Newton methods perform better than the conjugate gradient and derivative-free algorithms. This ranking, however, depends on the procedure used for step length selection. In Himmelblau's study, Powell's direct search method could compete with some variable metric algorithms. Box (1966) reports similar results. Comparing direct search methods, Parkinson and Hutchinson (1972a) found that a modification of the simplex algorithm has some advantages in terms of robustness and computer storage space over Powell's method for test problems with high-dimensional parameter spaces. Although the conjugate gradient method appears to be less efficient than quasi-Newton procedures in most comparisons, Sargent and

Sebastian (1972) mention its good performance for some functions involving many parameters. As we have seen, this algorithm needs less storage space than most other gradient methods.

Summarizing, there seems to be evidence that the Gauss-type methods are preferable if applicable, but a general ranking of other algorithms appears to be difficult and possibly for the applied worker not too useful. A combination of Gauss and quasi-Newton methods might be the first choice if the residuals are expected to be relatively large. However, this procedure is more costly in terms of computer storage space than other algorithms. An application of the Newton–Raphson algorithm or a modification thereof is often hampered by the inconvenience of providing analytic second partial derivatives. To find the optimal algorithm in the first trial requires some experience and even then it will be difficult or impossible. Bard (1974, pp. 116–117) remarks "that the state of the art of nonlinear optimization is such that one cannot as yet write a computer program that will produce the correct answer to every parameter estimation problem in a single computer run." This, however, should not discourage the applied researcher from using the described procedures, since even for a few trials the cost will usually be in acceptable limits if the number of parameters is relatively small. Moreover, the application of the methods is very easy if the available computer packages are used. For a discussion of computer related problems see Murray (1972b).

Generally, all possible simplifications of the minimization procedure should be utilized. A reparameterization of ill-conditioned objective functions, for instance, can sometimes simplify the minimization problem. An example is given in Draper and Smith (1981, Chapter 10). Unfortunately, the difficulty of finding a good reparameterization is one limitation of this method. If the model under consideration is linear in some parameters, a nonlinear procedure can sometimes be combined profitably with the methods to solve a linear estimation problem. At each stage of the iteration procedure the optimal values of the linear parameters are computed conditional on the given values for the nonlinear parameters.

Finally, we remind the reader that the described algorithms usually converge only to local minima. To find the global minimum, it can be helpful to apply different methods and starting values (see Section B.2.3). The reader is referred to Quandt (1983) and Dixon and Szegö (1975) for a discussion of more sophisticated, global optimization techniques.

B.3 CONSTRAINED OPTIMIZATION

In Section B.2 we have assumed that a priori any point in the parameter space qualifies as a solution of the optimization problem. In practice, nonsample information about the parameters is often available in the form of equality and/or inequality constraints. Knowledge of this kind can help to find good starting values for an optimization algorithm and also reduces the region in which we have to search for the minimum of the objective function. Moreover, it can aid

in picking the correct optimizing vector if the optimization problem has no unique solution.

On the other hand, the algorithms described in Section B.2 are designed for unconstrained problems and therefore either the algorithms have to be modified or the objective function has to be changed in such a way that the solution of an unconstrained optimization fulfills the constraints. In the following we will first discuss equality constraints and then consider inequality constraints.

B.3.1 Equality Constraints

Suppose that the true parameters are known to fulfill equality constraints given in the form

$$\mathbf{q}(\boldsymbol{\theta}) = \mathbf{0} \tag{B.3.1}$$

where $\mathbf{q}(\boldsymbol{\theta})$ is a differentiable function with values in the J-dimensional Euclidean space. The differentiability assumption will be met in almost all practical situations and covers, for instance, the case of linear constraints $R\boldsymbol{\theta} = \mathbf{r}$, where R is a $(J \times K)$ matrix and $\mathbf{r}$ a $(J \times 1)$ vector. In this case $\mathbf{q}(\boldsymbol{\theta}) = R\boldsymbol{\theta} - \mathbf{r}$. In the following section we will discuss some possible modifications of the objective function $H(\boldsymbol{\theta})$.

B.3.1a Reparameterization

Sometimes equality restrictions can be introduced by reducing the dimension of the parameter space. For instance, if $\boldsymbol{\theta} = (\theta_1,\theta_2)'$ and a constraint is $\theta_2 = \theta_1^2$, then we can define

$$\mathbf{g}(\theta_1) = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$$

and minimize $H_R(\theta_1) = H(\mathbf{g}(\theta_1))$ with respect to θ_1 only. Generally, if the restrictions can be expressed in the form

$$\boldsymbol{\theta} = \mathbf{g}(\boldsymbol{\alpha}) \tag{B.3.2}$$

where $\boldsymbol{\alpha}$ is a $(J \times 1)$ vector, the constrained optimum of $H(\boldsymbol{\theta})$ can be found by an unconstrained minimization of $H_R(\boldsymbol{\alpha}) = H(\mathbf{g}(\boldsymbol{\alpha}))$.

B.3.1b Lagrange Multipliers

If a reparameterization of the above-described form is not possible, a vector $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_J)'$ of Lagrange multipliers can be introduced and the constraints are incorporated in the minimization procedure by defining

$$\hat{H}_R(\boldsymbol{\theta},\boldsymbol{\lambda}) = H(\boldsymbol{\theta}) + \boldsymbol{\lambda}'\mathbf{q}(\boldsymbol{\theta}) \tag{B.3.3}$$

It can be shown that a constrained minimum of $H(\boldsymbol{\theta})$ is obtained at a stationary point of $\hat{H}_R(\boldsymbol{\theta},\boldsymbol{\lambda})$. Thus we have to find a solution to $(\partial\hat{H}_R/\partial\boldsymbol{\nu}|_{\nu}) = \mathbf{0}$, where $\boldsymbol{\nu}' = (\boldsymbol{\theta}',\boldsymbol{\lambda}')$. This can be done by minimizing

$$H_R(\boldsymbol{\nu}) = \left[\frac{\partial\hat{H}_R}{\partial\boldsymbol{\nu}}\bigg|_{\nu}\right]'\left[\frac{\partial\hat{H}_R}{\partial\boldsymbol{\nu}}\bigg|_{\nu}\right]$$

But notice that this method increases the number of parameters in the objective function and thereby may increase the minimization difficulties, whereas a reduction of the parameter space as described above is likely to reduce the computational burden. Furthermore, $(\partial\hat{H}_R/\partial\boldsymbol{\nu}|_{\nu}) = \mathbf{0}$ is only a necessary condition for a constrained minimum of $H(\boldsymbol{\theta})$.

B.3.1c Penalty Functions

Since using Lagrange multipliers has the above-mentioned disadvantages, other methods, for instance adding a *penalty function*, $A(\boldsymbol{\theta})$ say, to $H(\boldsymbol{\theta})$ are proposed. The penalty function $A(\boldsymbol{\theta})$ has to be chosen such that it is zero or almost zero for feasible $\boldsymbol{\theta}$ and very large outside the feasible region, for example,

$$A(\boldsymbol{\theta}) = d\mathbf{q}(\boldsymbol{\theta})'\mathbf{q}(\boldsymbol{\theta}) \tag{B.3.4}$$

where d is a sufficiently large constant to guarantee a minimum in the feasible region. The modified objective function is

$$H_R(\boldsymbol{\theta}) = H(\boldsymbol{\theta}) + A(\boldsymbol{\theta}) \tag{B.3.5}$$

For more details and other possible penalty functions, see Fletcher (1977) and Lootsma (1972b).

If the constraints are simple enough to permit an easy computation of second partial derivatives of $A(\boldsymbol{\theta})$, Gauss-type methods can still be applied to minimize the modified objective function if $H(\boldsymbol{\theta})$ has the form discussed in Section B.2.5. The direction matrix for the nth iteration becomes

$$\bar{P}_n = \left[P_n^{-1} + \frac{\partial^2 A}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}_n}\right]^{-1} \tag{B.3.6}$$

where P_n is the direction matrix used in the Gauss or modified Gauss algorithm.

B.3.1d Augmented Lagrangian Method

The augmented Lagrangian method combines (B.3.3) and (B.3.5) giving an objective function

$$H_R(\boldsymbol{\theta},\boldsymbol{\lambda}) = H(\boldsymbol{\theta}) + \boldsymbol{\lambda}'\mathbf{q}(\boldsymbol{\theta}) + d\mathbf{q}(\boldsymbol{\theta})'\mathbf{q}(\boldsymbol{\theta}) \tag{B.3.7}$$

Using this specification the λ's can be given fixed nonzero values in a first round and $H_R(\boldsymbol{\theta},\boldsymbol{\lambda})$ is minimized with respect to $\boldsymbol{\theta}$ without constraints. The resulting $\boldsymbol{\theta}$ vector can be used in (B.3.3) to determine a new $\boldsymbol{\lambda}$ that is used as a fixed vector in (B.3.7) in the next round. This procedure is repeated until convergence. For details see Fletcher (1975) and Powell (1978). Powell (1982b) recommends the method in conjunction with the conjugate gradient algorithm when the number of parameters is large and the number of restrictions is small.

B.3.2 Inequality Constraints

Suppose now that constraints for the parameters are given in the form of inequalities

$$\mathbf{q}(\boldsymbol{\theta}) \geq \mathbf{0} \tag{B.3.8}$$

where $\mathbf{q}(\boldsymbol{\theta})$ is a differentiable function with values in the J-dimensional Euclidean space. Let us first look at possible modifications of the objective functions that transform the optimization problem into an unconstrained one.

B.3.2a Modifications of the Objective Function

A *barrier function* $B(\boldsymbol{\theta})$, which is almost zero for all $\boldsymbol{\theta}$ satisfying the constraints and very large at the boundary of the feasible region, can have the form

$$B_n(\boldsymbol{\theta}) = c_n \sum_{j=1}^{J} [1/q_j(\boldsymbol{\theta})] \tag{B.3.9}$$

[Carroll (1961)] or

$$B_n(\boldsymbol{\theta}) = -c_n \sum_{j=1}^{J} \ln[q_j(\boldsymbol{\theta})] \tag{B.3.10}$$

[Lootsma (1967)], where the c_n are small positive constants with $c_1, c_2, c_3, \ldots$, approaching zero. In order to find the minimum of the original objective function in the feasible region

$$H_n(\boldsymbol{\theta}) = H(\boldsymbol{\theta}) + B_n(\boldsymbol{\theta}) \tag{B.3.11}$$

has to be minimized starting with a feasible initial vector $\boldsymbol{\theta}_1$, for consecutive n until $\min H_{n-1}(\boldsymbol{\theta}) = \min H_n(\boldsymbol{\theta})$ approximately. A method of this type is called an *interior point method*. The general form of the barrier function is

$$B(\boldsymbol{\theta}) = c \sum_{j=1}^{J} \rho[q_j(\boldsymbol{\theta})] \tag{B.3.12}$$

where $\rho[\cdot]$ is a continuously differentiable function defined on the positive real numbers, with $\rho[q_j(\boldsymbol{\theta})] \to \infty$ if $q_j(\boldsymbol{\theta}) \to 0$. This formulation restricts the parame-

ters to the interior of the feasible region, and problems can arise if the minimum is obtained on or near the boundary. In this case an *exterior point method* may be appropriate. A penalty function of the general form

$$A(\boldsymbol{\theta}) = d \sum_{j=1}^{J} \eta[q_j(\boldsymbol{\theta})] \tag{B.3.13}$$

is added to the original objective function. In (B.3.13), $\eta[\cdot]$ is a continuously differentiable function defined on the real number line and satisfying

$$\eta[q_j(\boldsymbol{\theta})] \begin{cases} = 0 & \text{for } q_j(\boldsymbol{\theta}) \geq 0 \\ > 0 & \text{for } q_j(\boldsymbol{\theta}) < 0 \end{cases} \tag{B.3.14}$$

A possible choice is

$$\eta[q_j(\boldsymbol{\theta})] = \{\min[0, q_j(\boldsymbol{\theta})]\}^2 \tag{B.3.15}$$

If we add $A(\boldsymbol{\theta})$ to $H(\boldsymbol{\theta})$, the penalty for leaving the feasible region grows with d. Thus an increasing series of d-values may have to be tried in order to obtain a feasible minimum of the objective function. Note that the Hessian of the new objective function does not exist at the boundary of the feasible region. For more details about the use of barrier and penalty functions, as well as other methods and references, see Lootsma (1972b), Davies and Swann (1969), Powell (1969), Fiacco and McCormick (1968), Fletcher (1977), and Murray (1969). The Gauss method can be modified to allow for a penalty or barrier function term in the objective function as in (B.3.6).

Sometimes inequality constraints can be eliminated by a parameter transformation. For instance, if θ_i is restricted to be greater than zero it can be replaced by $e^{\tilde{\theta}_i}$. Since we are dealing with nonlinear models anyway, such a transformation may not increase the computational burden. Other possible transformations are listed in Box (1966) and Powell (1972).

B.3.2b Modifications of the Optimization Algorithm

In this subsection we discuss methods of how to modify the optimization procedure rather than the objective function. The *gradient projection method* proposed by Rosen (1960, 1961) can be applied combined with any gradient method. An unconstrained minimization is carried out as long as all steps remain inside the feasible region. If the nth iteration, say, leads to a $\boldsymbol{\theta}_{n+1}$ outside the feasible region, the step length t_n is reduced such that $\boldsymbol{\theta}_{n+1}$ is on the boundary and consequently fulfills some of the inequality constraints, called *active constraints,* with an equality sign. If the next unconstrained iteration results in a $\boldsymbol{\theta}_{n+1}$ outside the feasible region, we treat the active constraints for $\boldsymbol{\theta}_n$ as equality constraints and perform the next step along the boundary, and so forth. If we have strict inequalities, the gradient projection may not be applicable, since for the boundary values $H(\boldsymbol{\theta})$ may simply not be defined. For example, this occurs if the restriction is $\theta_i > 0$ and this parameter appears as the

argument of a logarithm in the objective function. In this case the application of the gradient projection method requires a modification of the constraint of the form $\theta_i - \varepsilon \geq 0$, where ε is a small positive number.

The same applies for the so-called *projection method*. Any iterative algorithm is used for an unconstrained minimization, but each infeasible $\boldsymbol{\theta}_n$ is projected on $\bar{\boldsymbol{\theta}}_n$, a vector on the boundary of the feasible region, and the next step is started from $\bar{\boldsymbol{\theta}}_n$. This method is very easy if the constraints are of a simple type, for instance, $\theta_{ni} \geq 0$. Then if θ_{ni}, the ith coordinate of $\boldsymbol{\theta}_n$, is less than zero and thus $\boldsymbol{\theta}_n$ is infeasible, θ_{ni} is simply replaced by zero, and all other coordinates are left unchanged. Unfortunately, this method does not guarantee a constrained minimum if the procedure terminates at the boundary of the feasible region [Jennrich and Sampson (1968)]. But its simplicity makes it worth trying especially if the minimum of the restricted objective function is expected to be in the interior of the feasible region.

B.3.3 Joint Equality and Inequality Constraints

Sometimes both equality and inequality constraints for the parameter vector $\boldsymbol{\theta}$ are present. In this case the methods discussed in Sections B.3.1 and B.3.2 can be combined. Suppose the restrictions are given as

$$\mathbf{q}_1(\boldsymbol{\theta}) = \mathbf{0} \qquad \text{and} \qquad \mathbf{q}_2(\boldsymbol{\theta}) \geq \mathbf{0} \tag{B.3.16}$$

where $\mathbf{q}_1(\cdot)$ is $(J_1 \times 1)$ and $\mathbf{q}_2(\cdot)$ is $(J_2 \times 1)$. Combining (B.3.4) and (B.3.15) we get a new objective function,

$$H_R(\boldsymbol{\theta}) = H(\boldsymbol{\theta}) + d\{\mathbf{q}_1(\boldsymbol{\theta})'\mathbf{q}_1(\boldsymbol{\theta}) + \sum_{j=1}^{J_2} (\min[0, q_{2j}(\boldsymbol{\theta})])^2\} \tag{B.3.17}$$

where $q_{2j}(\boldsymbol{\theta})$ is the jth coordinate of $\mathbf{q}_2(\boldsymbol{\theta})$.

B.3.4 Some Comments

Having presented a range of different possibilities for constrained and unconstrained optimization, the question arises as to which method to use with each inequality-constrained problem, and how this method should be combined with an optimization algorithm. This, of course, depends on the model under consideration. Generally, problems can arise when quasi-Newton methods and penalty or barrier functions are applied, since these methods iteratively approximate the Hessian of the objective function, which is increasingly ill-conditioned if the sequence of weighting parameters c_n for an interior point barrier function approaches zero; and does not even exist at the boundary of the feasible region if the exterior point penalty function is used. Moreover, the performance of the quasi-Newton methods depends on the unidimensional search. Therefore, relatively sophisticated, unidimensional search procedures, which are in many

cases not adequate in the presence of penalty or barrier functions, are usually applied in ready-for-use computer programs. It is possible that a long trial step is carried out beyond the barrier at the boundary of the feasible region. This can cause trouble especially if the original objective function is not defined outside the feasible region. Also, the use of the simple projection method together with a quasi-Newton method cannot be recommended, since frequent interruptions of the unconstrained iteration process hinder a successful approximation of the Hessian.

Consequently, Gauss-type algorithms should be used in the presence of inequality constraints, if possible. In a first-round minimization the projection method can be applied to guarantee a feasible estimate. If the minimization procedure terminates at the boundary of the feasible region, other methods should be tried. If the structure of the objective function prohibits the use of Gauss-type methods, then a quasi-Newton method with a special unidimensional search procedure [Lasdon, Fox, and Ratner (1973)] or an algorithm designed particularly for the minimization of objective functions with penalty or barrier function terms could be applied [Lasdon (1972)].

In general it is desirable to utilize any possible simplifications that may for instance result from a reparameterization of the objective function. Also, if the objective functions or restrictions have a particular form, simplifications may be possible. For a discussion of linear constraints see Powell (1982b) and the references given in that article. Of course, at the extreme end, where the normal equations and restrictions are linear, the above methods are not required (see Chapters 2 and 3).

B.4 REFERENCES

Bard, Y. (1968) "On a Numerical Instability of Davidon-Like Methods," *Mathematics of Computation,* 22, 665–666.

Bard, Y. (1970) "Comparison of Gradient Methods for the Solution of Nonlinear Parameter Estimation Problems," *SIAM Journal of Numerical Analysis,* 7, 157–186.

Bard, Y. (1974) *Nonlinear Parameter Estimation,* Academic, New York.

Berndt, E. R., B. H. Hall, R. E. Hall, and J. A. Hausman (1974) "Estimation and Inference in Nonlinear Structural Models," *Annals of Economic and Social Measurement,* 3, 653–665.

Box, M. J. (1966) "A Comparison of Several Current Optimization Methods, and the Use of Transformations in Constrained Problems," *The Computer Journal,* 9, 67–77.

Brodlie, K. W. (1977) "Unconstrained Minimization," in D. Jacobs, ed., *The State of the Art in Numerical Analysis,* Academic, London, 229–268.

Brown, G. G. (1974) "Nonlinear Statistical Estimation with Numerical Maximum Likelihood," Working Paper No. 222, University of California, Los Angeles.

Brown, K. M. and J. E. Dennis, Jr. (1971) "A New Algorithm for Nonlinear

Least-Squares Curve Fitting," in J. R. Rice, ed., *Mathematical Software,* Academic, New York, 391–396.

Broyden, C. G. (1965) "A Class of Methods for Solving Nonlinear Simultaneous Equations," *Mathematics of Computation,* 19, 577–593.

Broyden, C. G. (1967) "Quasi-Newton Methods and Their Application to Function Minimization," *Mathematics of Computation,* 21, 368–381.

Carroll, C. W. (1961) "The Created Response Surface Technique for Optimizing Nonlinear, Restrained Systems," *Operations Research,* 9, 169–184.

Davidon, W. C. (1959) "Variable Metric Method for Minimization," A.E.C. Research and Development Report, ANL-5990 (revised).

Davies, D. and W. H. Swann (1969) "Review of Constrained Optimization," in R. Fletcher, ed., *Optimization,* Academic, London, 187–202.

Dennis, J. E. Jr. (1973) "Some Computational Techniques for the Nonlinear Least Squares Problem," in G. D. Bryne and C. A. Hall, eds., *Numerical Solutions of Systems of Nonlinear Algebraic Equations,* Academic, New York, 157–183.

Dennis, J. E., Jr. (1977) "Non-linear Least Squares and Equations," in D. Jacobs, ed., *The State of the Art in Numerical Analysis,* Academic, London, 269–312.

Dennis, J. E. and J. J. Moré (1977) "Quasi-Newtón Methods, Motivation and Theory," *SIAM Review,* 9, 46–89.

Dixon, L. C. W. (1972) "The Choice of Step Length, a Crucial Factor in the Performance of Variable Metric Algorithms," in F. R. Lootsma, ed., *Numerical Methods for Non-linear Optimization,* Academic, London, 149–170.

Dixon, L. C. W. and G. P. Szegö (1975) *Towards Global Optimization,* North-Holland, Amsterdam.

Draper, N. R. and H. Smith (1981) *Applied Regression Analysis,* 2nd ed., Wiley, New York.

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

Flanagan, P. D., P. A. Vitale, and J. Mendelsohn (1969) "A Numerical Investigation of Several One-Dimensional Search Procedures in Nonlinear Regression Problems," *Technometrics,* 11, 265–284.

Fletcher, R. (1965) "Function Minimization Without Evaluating Derivatives—A Review," *The Computer Journal,* 8, 33–41.

Fletcher, R., ed. (1969) *Optimization,* Academic, London.

Fletcher, R. (1972a) "Conjugate Direction Methods," in W. Murray, ed., *Numerical Methods for Unconstrained Optimization,* Academic, London, 73–86.

Fletcher, R. (1972b) "A Survey of Algorithms for Unconstrained Optimization," in W. Murray, ed., *Numerical Methods for Unconstrained Optimization,* Academic, London, 123–129.

Fletcher, R. (1975) "An Ideal Penalty Function for Constrained Optimization," in O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, eds., *Nonlinear Programming 2,* Academic, New York, 121–163.

Fletcher, R. (1977) "Methods for Solving Non-Linearly Constrained Optimiza-

tion Problems," in D. Jacobs, ed., *The State of the Art in Numerical Analysis,* Academic, London, 365–407.

Fletcher, R. and M. J. D. Powell (1963) "A Rapidly Convergent Descent Method for Minimization," *The Computer Journal,* 6, 163–168.

Fletcher, R. and C. M. Reeves (1964) "Function Minimization by Conjugate Gradients," *The Computer Journal,* 7, 149–154.

Goldfeld, S. M. and R. E. Quandt (1972) *Nonlinear Methods in Econometrics,* North-Holland, Amsterdam.

Goldfeld, S. M., R. E. Quandt, and H. F. Trotter (1966) "Maximization by Quadratic Hill-Climbing," *Econometrica,* 34, 541–551.

Greenstadt, J. (1967) "On the Relative Efficiencies of Gradient Methods," *Mathematics of Computation,* 21, 360–367.

Himmelblau, D. M. (1972a) *Applied Nonlinear Programming,* McGraw-Hill, New York.

Himmelblau, D. M. (1972b) "A Uniform Evaluation of Unconstrained Optimization Techniques," in F. R. Lootsma, ed., *Numerical Methods for Nonlinear Optimization,* Academic, London, 69–97.

Hooke, R. and T. A. Jeeves (1961) "Direct Search Solution of Numerical and Statistical Problems," *Journal of the Association for Computing Machinery,* 8, 212–229.

Huang, H. Y. (1970) "A Unified Approach to Quadratically Convergent Algorithms for Function Minimization," *Journal of Optimization Theory and Applications,* 5, 405–423.

Huang, H. Y. and A. V. Levy (1970) "Numerical Experiments on Quadratically Convergent Algorithms for Function Minimization," *Journal of Optimization Theory and Applications,* 6, 269–282.

Jennrich, R. I. and P. F. Sampson (1968) "Application of Stepwise Regression to Non-Linear Estimation," *Technometrics,* 10, 63–72.

Judge, G. G., R. C. Hill, W. E. Griffiths, H. Lütkepohl, and T. C. Lee (1982) *Introduction to the Theory and Practice of Econometrics,* Wiley, New York.

Künzi, H. P. and W. Oettli (1969) *Nichtlineare Optimierung: Neuere Verfahren Bibliographie,* Lecture Notes in Operations Research and Mathematical Systems, 16, Springer, Berlin.

Lasdon, L. S. (1972) "An Efficient Algorithm for Minimizing Barrier and Penalty Functions," *Mathematical Programming,* 2, 65–106.

Lasdon, L. S., R. L. Fox, and M. W. Ratner (1973) "An Efficient One-Dimensional Search Procedure for Barrier Functions," *Mathematical Programming,* 4, 279–296.

Lootsma, F. R. (1967) "Logarithmic Programming: A Method of Solving Nonlinear-Programming Problems," *Philips Research Reports,* 22, 329–344.

Lootsma, F. R., ed. (1972a) *Numerical Methods for Non-linear Optimization,* Academic, London.

Lootsma, F. R. (1972b) "A Survey of Methods for Solving Constrained Minimization Problems via Unconstrained Minimization," in F. R. Lootsma, ed., *Numerical Methods for Non-linear Optimization,* Academic, London, 313–347.

Maddala, G. S. (1977) *Econometrics,* McGraw-Hill, New York.

Marquardt, D. W. (1963) "An Algorithm for Least Squares Estimation of Nonlinear Parameters," *Journal of the Society for Industrial and Applied Mathematics,* 11, 431–441.

Murray, W. (1969) "An Algorithm for Unconstrained Minimization," in R. Fletcher, ed., *Optimization,* Academic, London, 247–258.

Murray, W., ed. (1972a) *Numerical Methods for Unconstrained Optimization,* Academic, London.

Murray, W. (1972b) "Failure, the Causes and Cures," in W. Murray, ed., *Numerical Methods for Unconstrained Optimization,* Academic, London, 107–122.

Nelder, J. A. and R. Mead (1965) "A Simplex Method for Function Minimization," *The Computer Journal,* 7, 308–313.

Ortega, J. M. and W. C. Rheinboldt (1970) *Iterative Solution of Nonlinear Equations in Several Variables,* Academic, New York.

Parkinson, J. M. and D. Hutchinson (1972a) "A Consideration of Non-gradient Algorithms for the Unconstrained Optimization of Functions of High Dimensionality," in F. R. Lootsma, ed., *Numerical Methods for Non-linear Optimization,* Academic, London, 99–113.

Parkinson, J. M. and D. Hutchinson (1972b) "An Investigation into the Efficiency of Variants on the Simplex Method," in F. R. Lootsma, ed., *Numerical Methods for Non-linear Optimization,* Academic, London, 115–135.

Polak, E. (1971) *Computational Methods in Optimization,* Academic, New York.

Powell, M. J. D. (1964) "An Efficient Method for Finding the Minimum of a Function of Several Variables Without Calculating Derivatives," *The Computer Journal,* 7, 155–162.

Powell, M. J. D. (1965) "A Method for Minimizing a Sum of Squares of Nonlinear Functions Without Calculating Derivatives," *The Computer Journal,* 7, 303–307.

Powell, M. J. D. (1969) "A Method for Nonlinear Constraints in Minimization Problems," in R. Fletcher, ed., *Optimization,* Academic, London, 283–298.

Powell, M. J. D. (1972) "Problems Related to Unconstrained Optimization," in W. Murray, ed., *Numerical Methods for Unconstrained Optimization,* Academic, London, 29–55.

Powell, M. J. D. (1978) "Algorithms for Nonlinear Constraints that Use Lagrangian Functions," *Mathematical Programming,* 14, 224–248.

Powell, M. J. D., ed. (1982a) *Nonlinear Optimization 1981,* Academic, London.

Powell, M. J. D. (1982b) "Algorithms for Constrained and Unconstrained Optimization Calculations," in M. Hazewinkel and A. H. G. Rinnooy Kan, eds., *Current Developments in the Interface: Economics, Econometrics, Mathematics,* Reidel, Dordrecht, 293–312.

Quandt, R. E. (1983) "Computational Problems and Methods," in Z. Griliches and M. Intriligator, eds., *Handbook of Econometrics,* Vol. I, North-Holland, Amsterdam.

Rosen, J. B. (1960) "The Gradient Projection Method for Nonlinear Programming: I. Linear Constraints," *SIAM Journal,* 8, 181–217.

Rosen, J. B. (1961) "The Gradient Projection Method for Nonlinear Programming: II. Nonlinear Constraints," *SIAM Journal,* 9, 514–532.

Rosenbrock, H. H. (1960) "An Automatic Method for Finding the Greatest or Least Value of a Function," *The Computer Journal,* 3, 175–184.

Sargent, R. W. H. and D. J. Sebastian (1972) "Numerical Experience with Algorithms for Unconstrained Minimization," in F. R. Lootsma, ed., *Numerical Methods for Non-linear Optimization,* Academic, London, 45–68.

Schrack, G. and N. Borowski (1972) "An Experimental Comparison of Three Random Searches," in F. R. Lootsma, ed., *Numerical Methods for Nonlinear Optimization,* Academic, London, 137–147.

Spendley, W., G. R. Hext, and F. R. Himsworth (1962) "Sequential Application of Simplex Designs in Optimization and Evolutionary Operation," *Technometrics,* 4, 441–461.

Swann, W. H. (1972) "Direct Search Methods," in W. Murray, ed., *Numerical Methods for Unconstrained Optimization,* Academic, London, 13–28.