Queen's Economics Department Working Paper No. 304

On the Role of Jacobian Terms in Maximum Likelihood Estimation

James G. MacKinnon
Queen's University

Department of Economics
Queen's University
94 University Avenue
Kingston, Ontario, Canada
K7L 3N6

1978

On the Role of Jacobian Terms in Maximum Likelihood Estimation

James G. MacKinnon

Department of Economics
Queen's University
Kingston, Ontario, Canada
K7L 3N6

Abstract

Because of the presence of Jacobian terms, determinants which arose as a result of a transformation of variables, many common likelihood functions have singularities. This fact has several implications for maximum likelihood estimation. The most interesting of these is that singularities often correspond with economically meaningful restrictions, and they can be used to impose the latter. Several applications of this principle are presented. They suggest that maximum likelihood should be preferred to other estimation schemes not only because of its optimal large-sample statistical properties, but also because of its ability to incorporate certain *a priori* restrictions from economic theory.

I would like to thank Charles Beach, Alan Gelb, and Mark Gersovitz for valuable comments on earlier drafts, and Mike Peters for making an illuminating observation. All errors are mine.

August, 1978

1. Introduction

A likelihood function is simply the joint density of a set of sample observations, considered as a function of the parameters of the density. Suppose that $\boldsymbol{u} = [u_1, \ldots, u_k]^\top$ is a vector of k random variables with joint density $f(\boldsymbol{u})$, and $\boldsymbol{y} = [y_1, \ldots, y_k]^\top$ is a monotonic function of $\boldsymbol{u}$, with joint density $g(\boldsymbol{y})$. It is well known (Wilks, 1962, pp. 57–59) that

$$g(\boldsymbol{y}) = f\big(\boldsymbol{u}(\boldsymbol{y})\big)\left\|\frac{\partial u_i(\boldsymbol{y})}{\partial y_j}\right\|, \tag{1}$$

where the second term on the right-hand side is the absolute value of the Jacobian of the transformation. As a consequence of (1), many of the likelihood functions which are commonly encountered in econometrics include Jacobian terms, determinants which arose as a result of a transformation of variables. In many cases, these Jacobian terms can take on a value of zero for certain parameter values. Thus, when the likelihood function is graphed in parameter space, it will also take on a value of zero at certain points, which may be referred to as singularities, since the loglikelihood function tends to minus infinity as the parameter values approaches those points. The fact that many common loglikelihood functions have singularities has several implications for maximum likelihood estimation, which do not appear to be widely recognized. The purpose of this note is to draw attention to those implications.

It is obvious that the loglikelihood function can never achieve a maximum at a singularity, or in a sufficiently small neighborhood of one. This means that certain parameter values may be ruled out *a priori* when maximum likelihood estimation is employed. Moreover, a set of singularities may divide the parameter space into two or more regions, so that the loglikelihood function can be expected to have more than one local maximum. This may create difficulties of the obvious kind, but it may also be useful if the singularities correspond to economically meaningful restrictions. In the latter case, it may be possible to use the singularities in the loglikelihood function to bound the estimates within a region where those restrictions hold. Thus maximum likelihood estimation may be attractive not only because of its optimal large-sample statistical properties, but also because of its ability, in certain important cases, to incorporate *a priori* restrictions from economic theory.

In order to substantiate the above points, we shall consider three particular cases where singularities caused by the presence of Jacobian terms have interesting implications for maximum likelihood estimation. We shall first look at regression models with non-spherical disturbances, then at systems of linear equations, in particular models of demand and supply in a single market, and finally at systems of equations with truncated dependent variables.

2. Regression with Non-Spherical Disturbances

Consider the single-equation linear regression model,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{N}(0, \sigma^2\boldsymbol{V}), \tag{2}$$

where $\boldsymbol{V}$ is a positive definite matrix which is a function of one or more unknown parameters. The loglikelihood function for this model, concentrated with respect to σ^2, is

$$\ell(\boldsymbol{\beta}, \boldsymbol{V}) = \text{Const.} - \tfrac{1}{2} \log |\boldsymbol{V}| - \tfrac{n}{2} \log\bigl((\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\bigr). \tag{3}$$

Maximizing (3) conditional on $\boldsymbol{V}$ yields the familiar GLS estimates, which are maximum likelihood if $\boldsymbol{V}$ is known. But, in practice, $\boldsymbol{V}$ is almost never known. Econometricians then often employ pseudo-GLS procedures of various types. Such procedures are not maximum likelihood because they fail to take into account the Jacobian term, $-\frac{1}{2} \log |\boldsymbol{V}|$. This omission can be very serious, as the following two examples show.

First of all, suppose that the error terms follow a stationary first-order autoregressive error process

$$u_t = \rho u_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{N}(0, \sigma^2), \ \ |\rho| < 1. \tag{4}$$

The concentrated loglikelihood function for this model is

$$\ell(\boldsymbol{\beta}, \rho) = \text{Const.} + \tfrac{1}{2} \log |l - \rho^2| - \tfrac{n}{2} \log\bigl((\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{Q}\boldsymbol{Q}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\bigr), \tag{5}$$

where

$$\boldsymbol{Q} \equiv \begin{bmatrix} (1-\rho^2)^{1/2} & 0 & 0 & \cdots & 0 & 0 & 0 \\ -\rho & 1 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & -\rho & 1 \end{bmatrix}. \tag{6}$$

The first term in (5) is the logarithm of the absolute value of the determinant of $\boldsymbol{Q}$, which is the Jacobian of the transformation. It is clear that as $|\rho|$ approaches unity, this term tends to zero, so that there are singularities at $\rho = -1$ and $\rho = 1$. The loglikelihood function must therefore have at least one maximum in the interval $-1 < \rho < 1$, which is of course the stationarity region for the AR(1) process.

Thus the Jacobian term ensures that there will always exist estimates of ρ, and, under the usual conditions, of $\boldsymbol{\beta}$ as well, which maximize the likelihood function subject to the condition that the error process be stationary. This is not the case for two-step pseudo-GLS procedures, which may very well find estimates of ρ outside the stationarity region, nor is it the case for pseudo-ML procedures, such as iterated Cochrane-Orcutt, which drop the first observation and therefore incorrectly take the Jacobian of the transformation to be identically one.[1]

As a second example, suppose that $\boldsymbol{V}$ is a diagonal matrix with unity in the first $n-m$ diagonal elements and ω^2 in the remaining m diagonal elements. Thus the error terms are assumed to be independent but to have been generated by two different

[1] A simple technique for finding ML estimates for the linear regression model with AR(1) errors has recently been suggested by Beach and MacKinnon (1978a). For discussion of the AR(2) case, see Beach and MacKinnon (1978b).

regimes, with variance σ^2 for the first $n-m$ observations and $\omega^2\sigma^2$ for the remaining m observations. For this model, the concentrated log-likelihood function reduces to

$$\ell(\boldsymbol{\beta}, \omega) = \text{Const.} - \frac{n}{2}\log(\text{SSR}_1 + \text{SSR}_2/\omega^2), \tag{7}$$

where SSR_1 and SSR_2 are the sums of squared residuals for the first $n-m$ and last m observations, respectively, both of which are assumed to be greater than the number of elements in $\boldsymbol{\beta}$.

It is hard to imagine a reasonable pseudo-GLS procedure for this model. The problem is that one needs an estimate of ω^2. If that estimate is based on OLS estimates over the entire sample, one set of observations must have been given too much weight. If, on the other hand, it is based on estimates for the two regimes separately, the constraint that the $\boldsymbol{\beta}$ should be the same in both subsamples has been ignored. Finally, if one tries to estimate ω^2 by minimizing the generalized sum of squares, one will quickly discover that the minimum occurs at $\omega^2 = \infty$, since in that case the residuals from the last m observations will carry no weight.

Maximum likelihood estimation, on the other hand, is quite feasible. The estimate of ω^2 must be finite, provided that the number of observations in each regime is large enough that neither SSR_1 nor SSR_2 can be made to equal zero by some choice of $\boldsymbol{\beta}$. The estimate of ω^2 cannot be arbitrarily large, because that would make the Jacobian term, $-m\log\omega$, tend to minus infinity. It also cannot be arbitrarily small, because as ω^2 tends to zero, SSR_2/ω^2 must become very large relative to SSR_1, so that the last term in (7) tends to $-(n/2)\log(\text{SSR2}/\omega^2)$, which can be rewritten as

$$-\frac{n}{2}\log\text{SSR}_2 + n\log\omega.$$

Clearly, $n\log\omega$ dominates $-m\log\omega$, the Jacobian term in (7), so that as ω^2 tends to zero, the log-likelihood function will tend to minus infinity. Thus, in this example, the Jacobian term makes maximum likelihood estimation feasible.

3. Systems of Linear Equations

We now turn to the familiar model of a system of simultaneous linear equations, which can be written in matrix notation as

$$\boldsymbol{Y\Gamma} + \boldsymbol{XB} = \boldsymbol{U}. \tag{8}$$

The concentrated log likelihood function for this model is

$$\ell(\boldsymbol{\Gamma}, \boldsymbol{B}) = \text{Const.} + n\log\|\boldsymbol{\Gamma}\| - \frac{n}{2}\log\left|(\boldsymbol{Y\Gamma} + \boldsymbol{XB})^\top(\boldsymbol{Y\Gamma} + \boldsymbol{XB})\right|. \tag{9}$$

The middle term is the logarithm of the absolute value of the Jacobian of the transformation. It is clear that when $\boldsymbol{\Gamma}$ is singular the value of $\ell(\boldsymbol{\Gamma}, \boldsymbol{B})$ is minus infinity, so that FIML estimates of $\boldsymbol{\Gamma}$ cannot be singular or nearly singular. This implies that FIML may be severely biased if the true $\boldsymbol{\Gamma}$ is almost singular. On the other hand, ruling

out singular estimates of $\boldsymbol{\Gamma}$ is surely desirable, since nature could not have generated unique observations on $\boldsymbol{Y}$ if $\boldsymbol{\Gamma}$ were singular.

More importantly, the equation

$$|\boldsymbol{\Gamma}| = 0 \tag{10}$$

divides the parameter space into more than one region. If, as is normally the case, each of the parameters in $\boldsymbol{\Gamma}$ which have to be estimated appear only once, then (10) is linear in every parameter individually, so that the parameter space is divided into exactly two parts, one in which $|\boldsymbol{\Gamma}| > 0$ and one in which $|\boldsymbol{\Gamma}| < 0$. The likelihood function can be expected to have at least one local maximum in each of them, as illustrated in Figure 1 for the case of only one parameter, θ.

The situation illustrated in Figure 1 clearly creates problems for numerical maximization techniques. The global maximum is at θ_1, but a maximization algorithm could easily converge to θ_2, either because it was started to the right of θ' and was unable to take steps large enough to enable it to cross the singularity, or because, although started to the left of θ', it took a large step across the singularity and was then unable to go back. Unless there is reason to believe in advance that the global maximum is to the left of θ', and the algorithm is constrained to stay in that region, there is always some danger that the algorithm may locate the "wrong" maximum.[2]

Multiple maxima created by singularities should not create as much trouble for the investigator as other types of multiple maxima. We know that, under the appropriate regularity conditions, one of the local maxima of the likelihood function yields consistent parameter estimates, and we normally choose the highest one. This involves finding all local maxima, which can be difficult and time-consuming. When multiple maxima are due to singularities, there may be some *a priori* reason to prefer the maximum which lies on one side of the singularity, regardless of whether or not it is actually the highest one. Such a case is illustrated below, and another one is presented in Section 3.

Finding a maximum conditional on, say, $|\boldsymbol{\Gamma}| > 0$, is very easy. With most algorithms, it is merely necessary to start in the appropriate region and assign an extremely bad value to the objective function whenever the condition is violated. The algorithm will then never jump across the singularity, and it must terminate in the interior of the desired region because the singularity prevents it from terminating at the boundary. Note that this technique could not be used for other types of restrictions, because there would be nothing to prevent the algorithm from terminating at the boundary.

In order to find out more about the characteristics of the likelihood function (9), data were generated from the following model:

$$\begin{aligned} y_1 + \gamma_1 y_2 &= x_1 + u_1 \\ \gamma_2 y_1 + y_2 &= x_2 + u_2. \end{aligned} \tag{11}$$

[2] For a good discussion of numerical maximization techniques, see Bard (1974).

The coefficients of the exogenous variables were fixed at unity and treated as known; x_1 was a trend variable, and x_2 was sinusoidal. The error terms were normally distributed with variance unity and covariance zero. The values of γ_1 and γ_2 were chosen to be 0.9, so that $|\boldsymbol{\Gamma}| = 1 - \gamma_1\gamma_2$ is fairly close to singular.

Only one set of data, consisting of fifty observations, was generated from this model, since the object was not to perform a sampling experiment, but rather to examine the characteristics of the likelihood function. This function is graphed, for $\hat{\gamma}_1$ and $\hat{\gamma}_2$ between 0.7 and 1.5, in Figure 2. The global maximum, labelled M in the figure, is at $\hat{\gamma}_1 = 0.918$ and $\hat{\gamma}_2 = 0.878$; at that point, $\ell(\hat{\gamma}_1, \hat{\gamma}_2) = -273.49$. The contour line numbered "1" corresponds to $\ell = -445.02$, the line numbered "10" corresponds to $\ell = -278.96$, and the lines in between are equally spaced. No contours are shown for values of ℓ less than -445.02, so that Figure 2 is slightly misleading; the blank space between the two contours labelled "1" should actually contain an infinite number of contour lines, since $\ell(\gamma_1, \gamma_2)$ goes to minus infinity in that region.

The behavior of the likelihood function on the "wrong" side of the singularity is very strange. There are apparently two local maxima in addition to the global maximum. These are at $\gamma_1 = \infty$, $\gamma_2 = 0.786$, where $\ell = -338.22$, and $\gamma_1 = 1.243$, $\gamma_2 = \infty$, where $\ell = -322.72$. Why such absurd estimates should be associated with rather "qood" values of the likelihood function is not at all clear.

Unless one has *a priori* information on the sign of $|\boldsymbol{\Gamma}|$, one should clearly attempt to find all local maxima on both sides of the singularity. In many cases, however, such information may be available. Consider the simple case of a two-equation system explaining demand and supply in a single market. The model may be written as

$$Q_t^d = \alpha P_t + \boldsymbol{X}_t\boldsymbol{\beta} + u_{1t} \tag{12}$$

$$Q_t^s = \gamma P_t + \boldsymbol{Z}_t\boldsymbol{\delta} + u_{2t} \tag{13}$$

where Q_t^d and Q_t^s are quantity demanded and supplied in period t (it is assumed that $Q_t^d = Q_t^s$), P_t is the price, and $\boldsymbol{X}_t$ and $\boldsymbol{Z}_t$ are vectors of exogenous variables with enough exclusions to allow identification. In this case, the Jacobian of the transformation is

$$\begin{vmatrix} 1 & -\alpha \\ 1 & -\gamma \end{vmatrix} = \alpha - \gamma. \tag{14}$$

The condition that this Jacobian be negative is simply the condition that the demand curve should cut the supply curve from above. It would therefore be very reasonable to impose that condition as an *a priori* restriction.

4. Equation Systems with Truncated Dependent Variables

Amemiya (1974) has recently introduced a simultaneous equations model with truncated dependent variables, sometimes referred to as the simultaneous Tobit model. In the two-equation case, the model may be written as

$$\gamma_{11} y_{1t} + \gamma_{12} y_{2t} \geq \boldsymbol{X}_{1t}\boldsymbol{\beta}_1 + u_{1t} \tag{15}$$

$$\gamma_{21} y_{1t} + \gamma_{22} y_{2t} \geq \boldsymbol{X}_{2t}\boldsymbol{\beta}_2 + u_{2t} \tag{16}$$

$$y_{1t},\ \ y_{2t} \geq 0, \tag{17}$$

where equations (15) and (16) hold as equalities if y_{1t} and y_{2t} are positive, respectively. The error terms u_{1t} and u_{2t} are assumed to have a bivariate normal distribution with means zero and covariance matrix $\boldsymbol{V}$, which may be written as $f(u_{1t}, u_{2t}; \boldsymbol{V})$.

A major problem with this model, as Amemiya points out, is that it may have either no solutions or more than one solution; that is, for certain values of the parameters, the independent variables, and the error terms, there may be no values for y_{1t} and y_{2t}, or more than one set of values, which satisfy (15)–(17). In his Theorem 3, Amemiya shows that the model will have a unique solution if and only if every principal minor of $\boldsymbol{\Gamma}$ is positive; in the two-equation case here, $\boldsymbol{\Gamma}$ is the 2×2 matrix with elements γ_{ij}. Amemiya goes on to propose a consistent but inefficient estimator which is computationally less burdensome than maximum likelihood. He fails to point out that, if maximum likelihood is employed, the Jacobian terms in the loglikelihood function can be used to impose the principal minors condition on the γ_{ij}. This is a remarkable property of the maximum likelihood estimator, and a remarkably useful one.

The likelihood function is quite complicated. First, divide the set of observations into four subsets:

$$\begin{aligned} S_1 &= \{t | y_{1t} > 0,\ y_{2t} > 0\}, \\ S_2 &= \{t | y_{1t} > 0,\ y_{2t} = 0\}, \\ S_3 &= \{t | y_{1t} = 0,\ y_{2t} > 0\}, \\ S_4 &= \{t | y_{1t} = 0,\ y_{2t} = 0\} \end{aligned} .$$

Then the likelihood function, of which we would actually maximize the logarithm, is the product of four factors:

$$\begin{aligned} L = (\gamma_{11}\gamma_{22} - \gamma_{12}\gamma_{21}) \prod_{S_1} & f(\gamma_{11} y_{1t} + \gamma_{12} y_{2t} - \boldsymbol{X}_{1t}\boldsymbol{\beta}_1, \gamma_{21} y_{1t} + \gamma_{22} y_{2t} - \boldsymbol{X}_{2t}\boldsymbol{\beta}_2; \boldsymbol{V}) \\ \times \gamma_{11} \prod_{S_2} & \int_{-\infty}^{\gamma_{21} y_{1t} - \boldsymbol{X}_{2t}\boldsymbol{\beta}_2} f(\gamma_{11} y_{1t} - \boldsymbol{X}_{1t}\boldsymbol{\beta}_1, u_2; \boldsymbol{V}) du_2 \\ \times \gamma_{22} \prod_{S_3} & \int_{-\infty}^{\gamma_{12} y_{2t} - \boldsymbol{X}_{1t}\boldsymbol{\beta}_1} f(\gamma_{22} y_{2t} - \boldsymbol{X}_{2t}\boldsymbol{\beta}_2, u_1; \boldsymbol{V}) du_1 \\ \times \prod_{S_4} & \int_{-\infty}^{-\boldsymbol{X}_{2t}\boldsymbol{\beta}_2} \int_{-\infty}^{-\boldsymbol{X}_{1t}\boldsymbol{\beta}_1} f(u_1, u_2; \boldsymbol{V}) du_1 du_2. \end{aligned} \tag{18}$$

Here each factor corresponds to one of the subsets, and each of the products is taken over all observations in the indicated subset.

It may be observed that all of the principal minors of $\boldsymbol{\Gamma}$ appear in this likelihood function. They are just the Jacobians of the various transformations, where different y_{it} are identically zero. There should really be absolute value signs around these expressions, namely, $(\gamma_{11}\gamma_{22} - \gamma_{12}\gamma_{21})$, γ_{11}, and γ_{22}, but they were omitted because we will want to constrain the expressions to be positive anyway. These constraints can easily be implemented in the manner proposed in the last section when the logarithm of (18) is maximized with the aid of a numerical maximization routine. Some of the constraints may be rendered redundant by normalization; for example, it would be usual to normalize γ_{11} and γ_{22} to equal unity.

The ability to constrain ML estimates of the simultaneous Tobit model to satisfy the conditions for existence and uniqueness is an important one. In a simulation study of this model, Warner (1976) found that Amemiya's consistent estimator produced estimates which violated these conditions a substantial fraction of the time, while the maximum likelihood estimates, even without being explicitly constrained to satisfy them, always did so.

This property of maximum likelihood estimates extends to truncated dependent variable models with n equations. If there are n equations, the number of factors in the likelihood function is 2^n, and $2^n - 1$ of these are multiplied by the $2^n - 1$ principal minors of the $\boldsymbol{\Gamma}$ matrix. Estimation is of course likely to become impractical for large n, however.

5. Conclusion

Econometricians should pay more attention to those properties of likelihood functions associated with the presence of Jacobian terms. These terms can be bothersome, because they can create singularities and hence multiple maxima. But singularities created by Jacobian terms can also be useful, since they can be used to impose economically meaningful constraints without altering the properties of maximum likelihood estimation. In the case of the AR(1) model, the singularity enforces the stationarity constraint. In the case of the two-equation linear market model, the singularity enforces the constraint that the demand curve should cut the supply curve from above. And in the case of the simultaneous Tobit model, a whole set of singularities enforce the conditions for the existence and uniqueness of a solution to the estimated model. Alternative estimation procedures, such as generalized least squares and Amemiya's consistent estimator for the generalized Tobit model, do not incorporate these constraints. This provides a powerful reason for preferring maximum likelihood to other estimation techniques in these and similar situations.

References

Amemiya, Takeshi (1974). "Multivariate regression and simultaneous equation models when the dependent variables are truncated normal," *Econometrica*, 42, 999–1012.

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

Beach, Charles M., and James G. MacKinnon (1978a). "A maximum likelihood procedure for regression with autocorrelated errors," *Econometrica*, 46, 51–58.

Beach, Charles M., and James G. MacKinnon (1978b). "Full maximum likelihood estimation of second-order autoregressive error models," *Journal of Econometrics*, 7, 187–198.

Warner, Dennis L. (1976). "A Monte Carlo study of limited dependent variable estimation," in S. M. Goldfeld and R. E. Quandt, ed., *Studies in Nonlinear Estimation*. New York, Ballinger.

Wilks, Samuel S. (1962). *Mathematical Statistics*. New York, John Wiley and Sons.

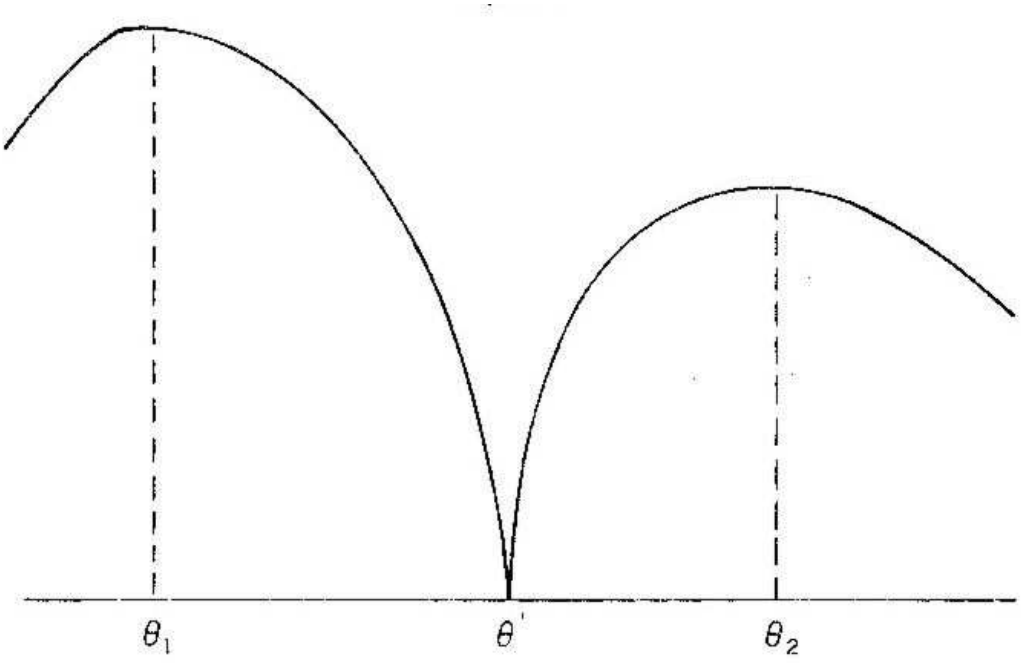

Figure 1

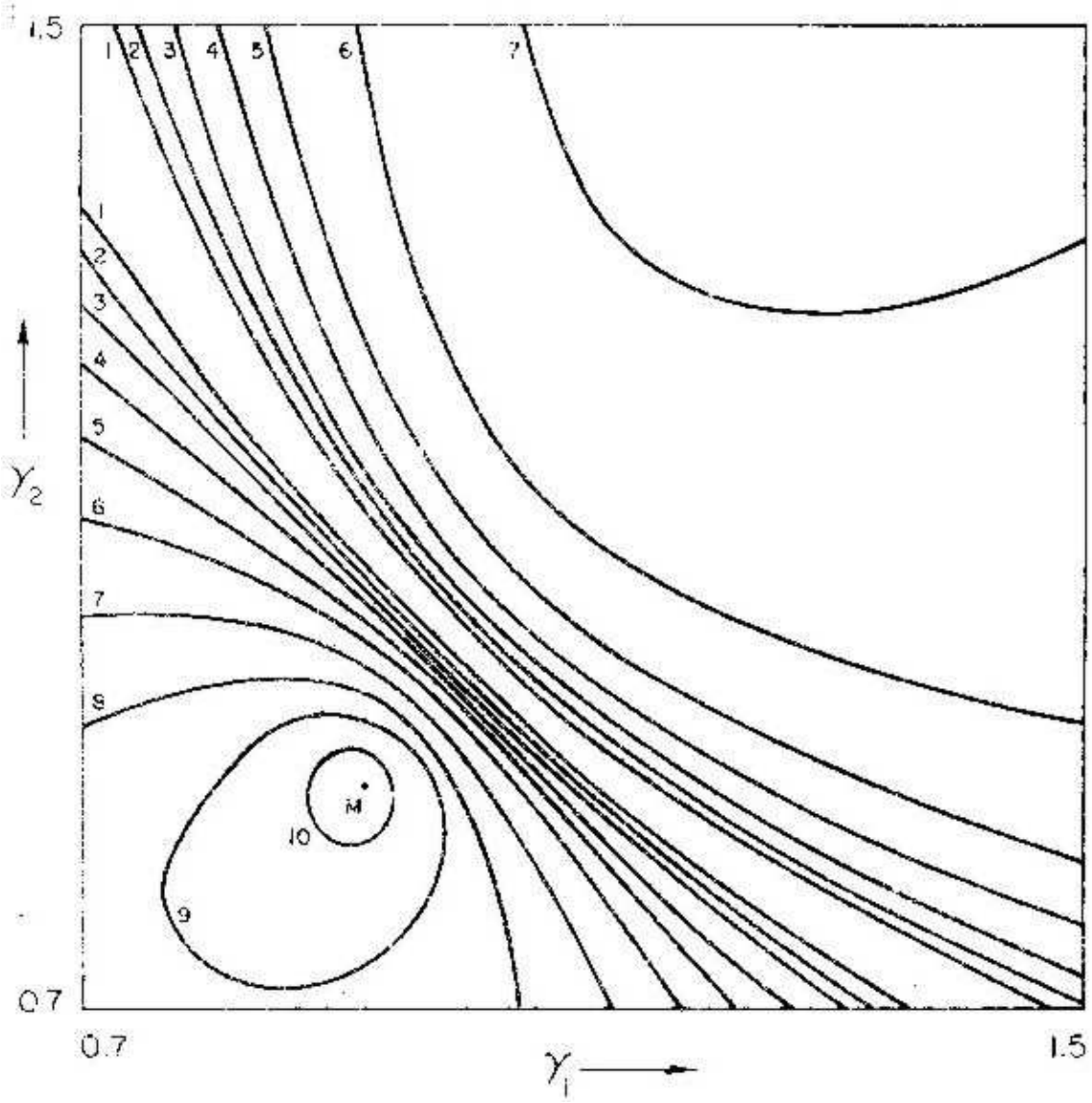

Figure 2