Journal of Research of the National Bureau of Standards
Vol. 90, No. 6, November-December 1985

Statistical Properties of a Procedure for Analyzing Pulse Voltammetric Data

Thomas P. Lane

Massachusetts Institute of Technology, Cambridge, MA 02139

and

John J. O'Dea and Janet Osteryoung

State University of New York at Buffalo, Buffalo, NY 14214

Accepted: June 24, 1985

O'Dea et al. (1983, *J. Phys. Chem.* 97, 3911-3918) proposed an empirical procedure for obtaining estimates and confidence intervals for kinetic parameters in a model for pulse voltammetric data. Their goal was to find a procedure that would run in real time, not necessarily one that would have well-defined statistical properties. In this paper we investigate some of the statistical properties of their procedure. We show that their estimation method is equivalent to maximum likelihood estimation, and their confidence intervals, while related to likelihood ratio confidence regions, have a coverage probability that is not fixed and that is potentially quite large. We suggest modifications of their procedure that lead to more traditional confidence intervals. We examine the effect on their procedure of the presence of nuisance paramters. Finally we discuss the possibility of serially correlated errors.

Key words: autocorrelation; confidence intervals; estimation method; kinetic parameters; maximum likelihood estimation; serially correlated errors; statistical properties.

1. Introduction

O'Dea et al. (1983) proposed a nonlinear regression procedure for estimating, and obtaining confidence intervals for, kinetic parameters describing the reduction of Zn(II) at a stationary mercury electrode in aqueous solutions of $NaNO_3$. In this paper we examine the statistical properties of the procedure and suggest modifications to improve these properties.

In section 2 we describe O'Dea's procedure. In section 3 we show his estimation procedure to be equivalent to maximum likelihood estimation. In section 4 we show his interval estimation procedure produces intervals that are related to higher-dimensional confidence regions obtained by likelihood ratio theory, and we suggest modifications to the procedure that will produce confidence intervals with the desired coverage probability. In section 5 we question the assumption of

About the Authors: Thomas P. Lane is with the Statistics Center at MIT while John J. O'Dea and Janet Osteryoung are with the Department of Chemistry at the State University of New York at Buffalo.

independent errors and examine the effect of including another parameter in the model to describe the apparent autoregressive error structure.

The notation used here is similar to that used by O'Dea, but as is customary in literature on regression we use Y as the dependent variable.

2. Description of the Procedure

O'Dea models the observed response at time t_i to an arbitrary pulse sequence by

$$Y_i = a f_i + c + \epsilon_i\,,$$

where a and c are unknown constants that convey no kinetic information, $\{\epsilon_i\}$ is a sequence of errors that are assumed to be independent with mean 0 and unknown variance σ^2. The function $f_i = f(t_i, \alpha, k, E_{\frac{1}{2}})$ is the solution of an integral equation. It depends on unknown kinetic parameters α, k, and $E_{\frac{1}{2}}$, and it must be obtained by solving the integral equation numerically.

The kinetic parameters are of primary interest, so O'Dea uses a nonlinear optimization procedure to find the values of these parameters that maximize the correlation R between Y and $f(t, \alpha, k, E_{\frac{1}{2}})$. These values are taken as the estimates. Estimates of a and c can then be obtained by simple linear regression of Y on $f(t, \alpha, k, E_{\frac{1}{2}})$, and σ can be estimated by the standard deviation of the residuals from this regression.

With no error the correlation R calculated above would be equal to unity. O'Dea measures the deviation from unity by $\bar{R} = 1 - R$ and defines $\bar{R}_{\min}$ as the optimum value of $\bar{R}$. To measure the uncertainty in his estimate of α he fixes k and $E_{\frac{1}{2}}$ at their optimal values and finds the two values of α that give $R = 3\,\bar{R}_{\min}$. He calls the interval between these values a "confidence interval" for α, but he assigns no confidence level to the interval. He computes similar intervals for k and $E_{\frac{1}{2}}$.

3. Maximum Likelihood Estimation

A more traditional approach to a problem of this sort would be to write the likelihood or log likelihood function for the problem and maximize it as a function of the unknown parameters. For normally distributed errors this is equivalent to choosing parameter values that minimze the sum of squared residuals. In this section we show that the above estimation procedure is also equivalent to maximum likelihood estimation.

If n is the number of observations, the log likelihood L is given by

$$L(a, c, \alpha, k, E_{\frac{1}{2}}, \sigma) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_i \left[\frac{Y_i - c - a f_i}{\sigma}\right]^2 .$$

The maximum is easily found by noting that for any value of σ, the expression is maximized by choosing a, c, α, k, and $E_{\frac{1}{2}}$ to minimize the sum of squared residuals $SS_R = \Sigma(Y_i - c - af_i)^2$. It is a simple matter to show that if $\bar{Y} + n^{-1}\Sigma Y_i$ and $SS_T = \Sigma(Y_i - \bar{Y})^2$, then $SS_R = (1 - R^2)SS_T$, so SS_R is a minimum when R is a maximum (in absolute value). Therefore O'Dea's estimates are the maximum likelihood estimates.

4. Confidence Intervals

Confidence regions for unknown parameters are often found by computing the maximum likelihood estimates and then finding other sets of parameter values for which the likelihood function, or an approximation to the likelihood function, is not much smaller. O'Dea's procedure is related to this approach.

Define $L(\alpha, k, E_{\frac{1}{2}})$ as the maximum over a, c, and σ of the log likelihood $L(a, c, \alpha, k, E_{\frac{1}{2}}, \sigma)$. Using the relationships between R, SS_R, and SS_T above and maximizing over σ gives

$$L(\alpha, k, E_{\frac{1}{2}}) = (-n/2)(1 + \log(2\pi) + \log SS_T + \log(1 - R^2) - \log n).$$

O'Dea's procedure involves finding six points—the two endpoints of the confidence intervals for each parameter—with the same correlation $R = 1 - 3\bar{R}_{\min}$. The above expression for $L(\alpha, k, E_{\frac{1}{2}})$ shows that these points have the same log likelihood as well.

Let $\hat{\alpha}, \hat{k}$, and $\hat{E}_{\frac{1}{2}}$ be the maximum likelihood estimates of α, k, and $E_{\frac{1}{2}}$. The quantity $\lambda = \exp(L(\hat{\alpha}, \hat{k}, \hat{E}_{\frac{1}{2}}) - L(\alpha, k, E_{\frac{1}{2}}))$ is called the likelihood ratio. It can be shown that $2 \log \lambda$ has an asymptotic chi-square distribution with 3 degrees of freedom if α, k, and $E_{\frac{1}{2}}$ are the true parameter values. Then $P[2 \log \lambda \leqslant 7.815] = 0.95$, so the parameter values for which $2 \log \lambda \leqslant 7.815$ form a 95% confidence region for the true values of the parameters. This region is bounded by the roughly ellipsoidal surface $\lambda(\alpha, k, E_{\frac{1}{2}}) = \exp(3.908)$, as shown in figure 1. In general, any surface of constant λ bounds some confidence region.

The confidence level of the region bounded by the surface containing O'Dea's points can be determined by computing λ. In his procedure

$$\begin{aligned} 2\log\lambda &= -n\left[\log(1 - (1 - \bar{R}_{\min})^2) - \log(1 - (1 - 3\bar{R}_{\min})^2)\right] \\ &\approx -n\log(2\bar{R}_{\min}/6\bar{R}_{\min}) = n\log 3, \end{aligned}$$

since $(\bar{R}_{\min})^2 << \bar{R}_{\min}$. Here $n = 81$, so $2\log\lambda \approx 89$. By comparison, the region bounded by the surface for

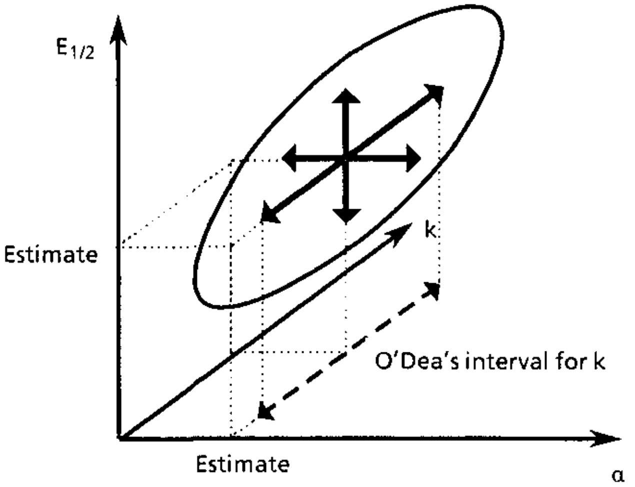

Figure 1–Relationship between O'Dea's intervals and a three-dimensional confidence ellipsoid. The ellipsoid passes through the endpoints of the solid line segments.

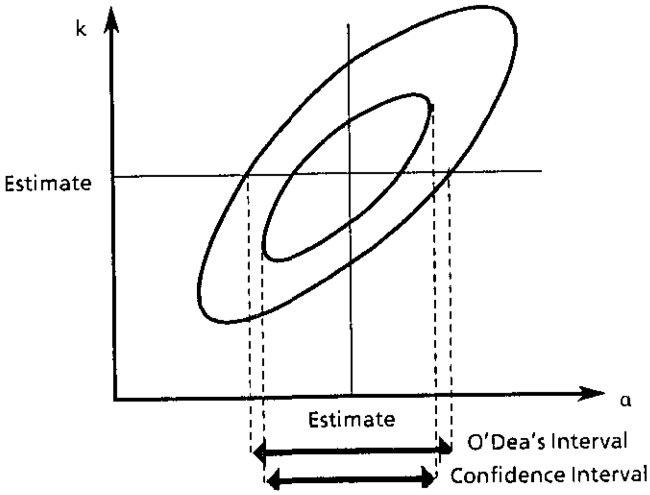

Figure 2–Comparison of confidence interval with interval computed by O'Dea's procedure in two dimensions.

which $2\log\lambda=16.268$ has a confidence level of 99.9%, so a three-dimensional confidence region found using $2\log\lambda=89$ would be very conservative.

A more customary confidence level is 95%. Since the likelihood ratio can be written as a function of the correlation, it is possible to use a modification of O'Dea's procedure to find points on the boundary of a 95% confidence region. Rather than increasing $\bar{R}$ by a factor of 3, the appropriate factor is the value of b for which 81 log $b=7.815$, or $b\approx 1.10$. For example, in a sample data set that does not appear in O'Dea's original paper, $\alpha=.22522$. Increasing $\bar{R}$ by a factor of 3 produces the interval [.22139, .22916], while the factor 1.10 leads to the interval [.22435, .22609].

On the other hand O'Dea's goal was not to find points in a confidence region for all three parameters, but to find separate confidence intervals for each parameter. In order to use the distribution of the likelihood ratio, O'Dea's procedure must be modified so that in computing the endpoints of the confidence interval for one parameter, the likelihood is maximized over the other two parameters. Twice the log of this likelihood ratio has an asymptotic chi-square distribution with one degree of freedom.

In the case of α, for example, this is done by comparing $2\log\lambda=2[L(\hat{\alpha},\hat{k},\hat{E}_{\frac{1}{2}})-L(\alpha,\tilde{k}(\alpha),\tilde{E}_{\frac{1}{2}}(\alpha)]$ to a chi-square distribution with one degree of freedom, where $\tilde{k}(s)$ and $\tilde{E}_{\frac{1}{2}}(s)$ are the values of k and $E_{\frac{1}{2}}$ that maximize $L(\alpha,k,E_{\frac{1}{2}})$ subject to the restriction $\alpha=s$. The 95% point of this distribution is 3.841, so the values of α for which $2\log\lambda\leqslant 3.841$ form a 95% confidence interval for the true parameter value.

This is best illustrated in two dimensions, as in figure 2. Here approximately elliptical contours of constant λ are plotted as a function of α and k for constant $E_{\frac{1}{2}}$. (The complete contours are ellipsoids in three dimensions.) The inner ellipse has $2\log\lambda=3.841$, while the outer ellipse has $2\log\lambda=7.815$. The endpoints of the confidence interval for α are the points on the contour that have tangents perpendicular to the α axis. This interval can be compared with the interval found by O'Dea's procedure, which is that portion of the $k=\hat{k}$ line that is within the outer ellipse.

Which is larger? If the two ellipses have major axes parallel to the coordinate axes, O'Dea's intervals are longer and their coverage probabilities exceed 95%. While this is not desirable, it increases the probability that his interval will contain the true parameters. But if the major axes are not parallel to the coordinate axes and if the lengths of the minor axes are small, O'Dea's intervals are shorter and have a coverage probability less than 95%. Unfortunately it is not possible to determine which is the case by looking only at the points examined in his procedure.

There are two sensible remedies to this problem. The first, the likelihood ratio method, is similar in spirit to O'Dea's original procedure. This method involves finding the confidence interval as described above by finding those values of the first parameter that produce the proper likelihood ratio when the likelihood is maximized over the other two parameters. In this problem, though, it is time consuming to calculate f_i and its derivatives do not have simple analytic expressions, so repeated maximization of the likelihood may be too computationally burdensome.

The other method, an asymptotic normal approximation, is the one we use here. This involves assuming that the maximum likelihood estimates have a multivariate normal distribution with a mean vector equal to the true parameter values and with covariance matrix equal to minus the inverse of the second derivative of the log likelihood L. (This is equivalent to the likelihood ratio method applied to a quadratic approximation to the log likelihood.) Since there is no analytic expression for the second derivative in this problem, we use a numerical approximation.

In the example given above, the estimated covariance matrix is

$$S = \begin{bmatrix} 1.0334 & -.6908 & .0396 \\ -.6908 & 8.6984 & -.7509 \\ .0396 & -.7509 & .1384 \end{bmatrix} \times 10^{-7}.$$

A 95% confidence interval for α is given by $\hat{\alpha}\pm 1.96(S_{11})^{\frac{1}{2}}$, or [.22459, .22585]. This is narrower than the interval obtained above using O'Dea's procedure with a factor of 1.10.

5. Residual Autocorrelation

The above derivations are valid if the errors $\{\epsilon_i\}$ are independent normal random variables with a common variance. In practice this assumption must be checked. This is especially true when, as in this case, measurements are taken over time. It is often reasonable to suspect that measurements at neighboring time points may be correlated.

The errors $\{\epsilon_i\}$ are not observed, but they can be estimated by the residuals, or the differences between the observed Y_i and the fitted values $\hat{Y}_i=\hat{c}+\hat{a}\,f(t_i,\hat{\alpha},\hat{k},\hat{E}_{\frac{1}{2}})$. Figure 3 is a plot of Y and $\hat{Y}$ as a function of time. Figure 4 is a plot of the residuals $e_i=Y_i-\hat{Y}_i$ over time. If the residuals were independent we would not expect to find any pattern here, but in fact there is a pronounced tendency for residuals at neighboring time points to have the same sign.

There are three possible causes for this phenomenon. First, it is possible that this is an artifact of the fitting procedure. Even when the errors are independent, fitting an ordinary linear regression produces residuals that have some correlation (for example, they sum to zero). It is possible that minimizing the sum of squared residuals in this more complicated model produces residuals with some autocorrelation. However experiments with the model do not support this hypothesis.

A second possible cause is model inadequacy. The model relates an imposed voltage to an observed current, and the voltage is highly correlated with time. If the equation used here is not the true relationship between the current and the voltage, there may be correlation between the current and the residuals, and this dependence could be masquerading as time dependence. The cure for this difficulty is to propose alternative models that better fit the data.

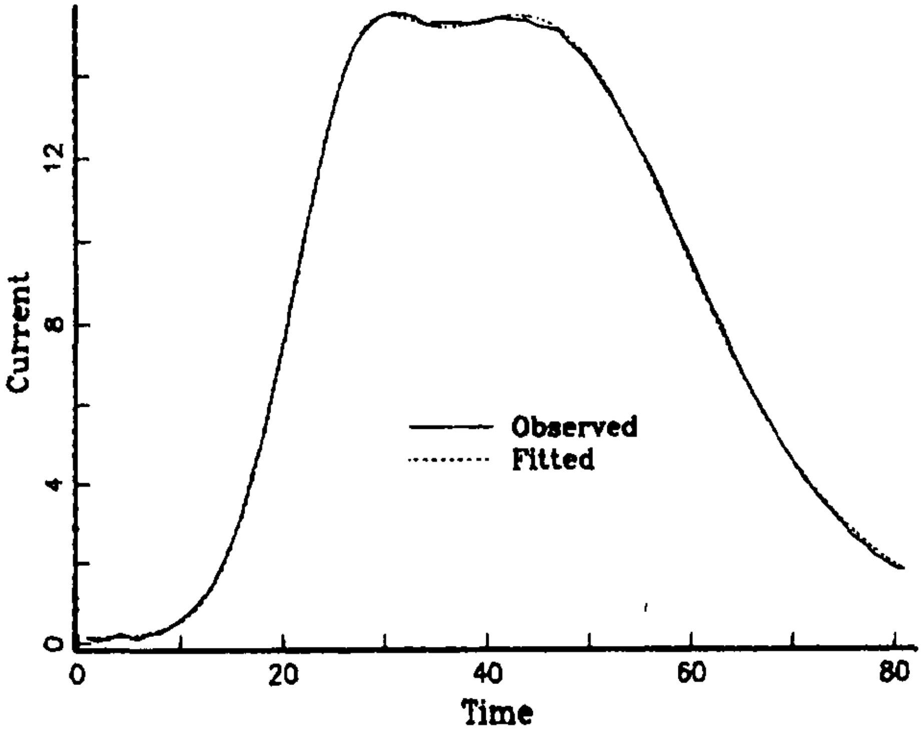

Figure 3–Observed and Fitted Current.

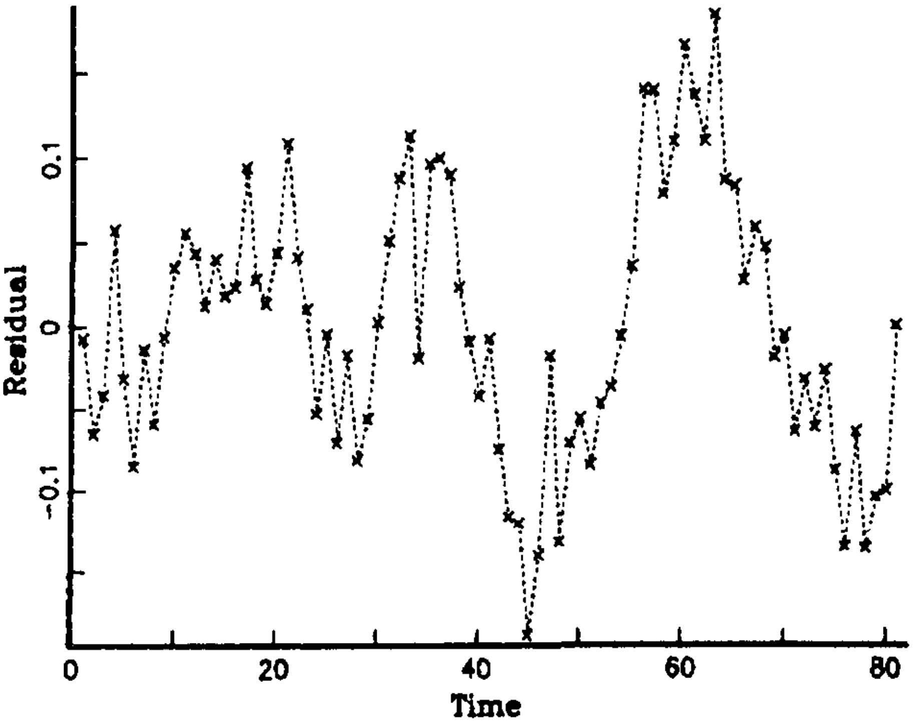

Figure 4–Residuals as a Function of Time.

The third possible cause is actual autocorrelation in the errors, and this autocorrelation can be modeled as well. We proceed under the assumption that the true errors have time-dependent correlation.

Two common models for time series are the first order autoregressive model

$$\epsilon_i=\rho\ \epsilon_{i-1}+u_i$$

and the first order moving average model

$$\epsilon_i=u_i+\rho\ u_{i-1},$$

where in both cases $\{u_i\}$ is a sequence of independent normal random variables with mean 0 and common unknown variance, and ρ is an unknown parameter between -1 and 1. Other possible models are the higher order models, where terms from earlier time points are used, and mixed models, where ϵ_i is modeled as a linear combination of $\epsilon_{i-1},\ldots,\epsilon_{i-p}$ and $u_i,\ldots,u_{i-q}$.

Two tools useful for identifying a good model are the autocorrelation function and the partial autocorrelation function. These appear in figures 5 and 6. The sample autocorrelation function is simply the correlation of e_i and e_{i-k} plotted as a function of k. For a moving average process of order q the true autocorrelation function is 0 for $k>q$. For autoregressive processes and mixed processes the true autocorrelation function approaches 0 as $k\rightarrow\infty$, but it is not identically 0 for all k beyond some finite value. The sample autocorrelation function in figure 5 seems to be more consistent with that of the autoregressive and mixed models, since there does not seem to be a sharp cutoff.

The partial autocorrelation function is more complicated, but its interpretation is quite simple. It is the "dual" of the autocorrelation function, in that it is 0 for all $k>p$ for an autoregressive process of order p, and it approaches 0 as $k\rightarrow\infty$ but it does not vanish for moving average and mixed processes. The sample partial autocorrelation function in figure 6 shows a large value at $k=1$ and smaller values for $k>1$. It is never exactly 0, but for most k values the sample partial autocorrelation falls inside the boundary that marks the values that are significantly different from 0. The function seems to be consistent with what might be expected from a first order autoregressive process.

The new model

$$Y_i=af_i+c+\epsilon_i \qquad \text{with} \qquad \epsilon_i=\rho\epsilon_{i-1}+u_i$$

is equivalent to

$$Y_i=\rho Y_{i-1}+a(f_i-\rho f_{i-1})+c(1-\rho)+u_i,$$

which is a nonlinear regression model with independent errors.

There are now four parameters to be estimated, in addition to a, c, and σ. But if only the original three parameters are of interest, it is possible to treat ρ as one of the nuisance parameters by using a variant of the Cochrane-Orcutt procedure, as follows:

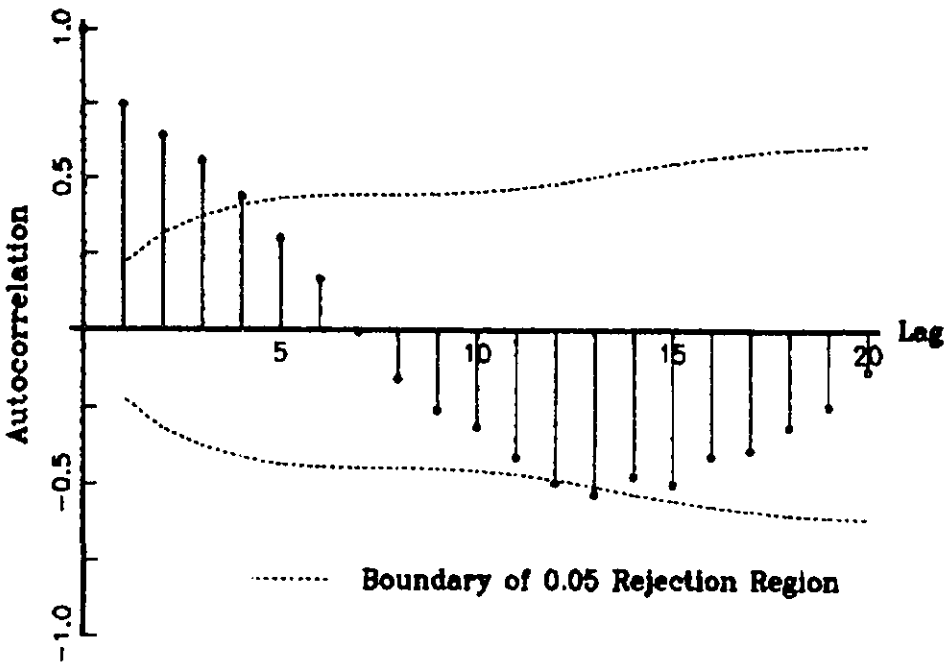

Figure 5-Residual Autocorrelation Function.

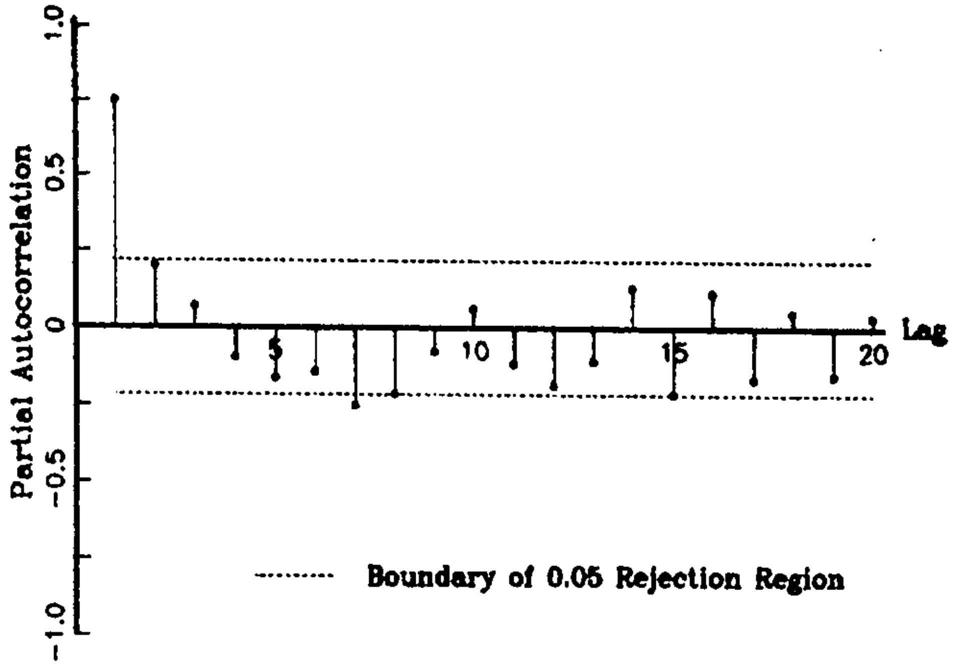

Figure 6-Residual Partial Autocorrelation Function.

1) For any given α, k, and $E_{\frac{1}{2}}$, compute $\{f_i\}$.
2) Estimate a and c by linear regression to get $\{e_i\}$.
3) Estimate ρ by the sample correlation of the $\{e_i\}$.
4) Regress $Y_i - \rho Y_{i-1}$ on $f_i - \rho f_{i-1}$ to get new estimates of a and c; and new residuals $\{e_i\}$.
5) Repeat steps 3 and 4 until convergence.
6) Compute the sum of squares $\Sigma u_i^2 = \Sigma(e_i - \rho e_{i-1})^2$.

The computer time needed for these steps is much less than that needed to compute $\{f_i\}$, so the estimation is much faster if the nonlinear optimization program searches only in the three-dimensional space of $(\alpha, k, E_{\frac{1}{2}})$. For each set of trial parameter values the above steps can be performed to minimize the residual sum of squares over the nuisance parameters. The resulting estimate of α is .22473.

The other calculation can also be repeated for this new model. The estimated covariance matrix is

$$S = \begin{bmatrix} 4.8344 & -5.6256 & .4766 \\ -5.6256 & 37.5982 & -2.6193 \\ .4766 & -2.6193 & .6513 \end{bmatrix} \times 10^{-7}.$$

This is roughly four times the previous covariance matrix, so the length of the new confidence interval is about twice that of the previous confidence interval. The new interval is [.22336, .22600].

The four intervals around α are compared in figure 7. The procedure used in O'Dea's original paper produces an interval obtained from a three dimensional confidence ellipsoid with a very large confidence level, and it is quite long. The interval is shortened by using a 95% confidence ellipsoid, but it still does not have a 95% coverage probability. The 95% confidence interval is still shorter. Taking into account the apparent autoregressive error structure leads to a confidence interval that is about twice as long as the one in the independence model, but still only a third as long as the interval obtained by O'Dea's procedure.

Thanks to Herman Chernoff for numerous discussions. Thanks also to IBM for providing graphics hardware and software.

References

[1] O'Dea, J. J.; J. Osteryoung and R. A. Osteryoung, Square wave voltammetry and other pulse techniques for the determination of kinetic parameters. The reduction of zinc (II) at mercury electrodes, *J. Phys. Chem.* **87**, 3911-3918 (1983).

[2] Morrison, D. F. *Multivariate Statistical Mehods*, McGraw Hill: New York (1976).

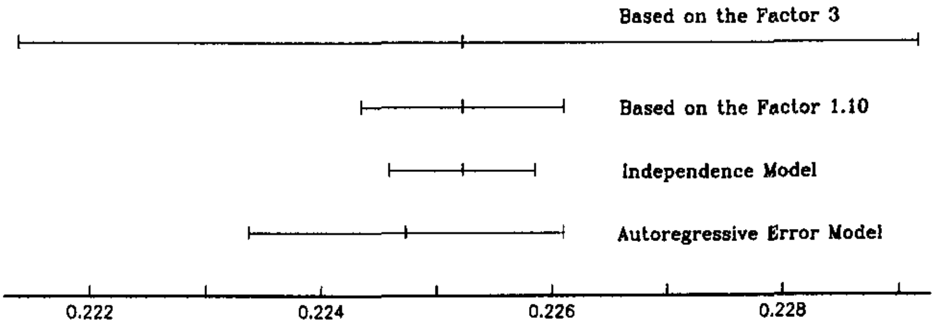

Figure 7-Confidence Intervals for Alpha.