EMPIRICAL LIKELIHOOD AND GENERALIZED PROJECTION PURSUIT

BY

ART B. OWEN

TECHNICAL REPORT NO. 393

APRIL 1992

PREPARED UNDER THE AUSPICES

OF

NATIONAL SCIENCE FOUNDATION GRANT DMS89-05874

DEPARTMENT OF STATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA
Empirical Likelihood and Generalized Projection Pursuit

By

Art B. Owen

Technical Report No. 393
April 1992

Prepared Under the Auspices
Of
National Science Foundation Grant DMS89-05874

Department of Statistics
Stanford University
Stanford, California
Empirical Likelihood and Generalized Projection Pursuit

Art B. Owen
Dept. of Statistics
Stanford University

Abstract

We present a class of models, the generalized projection pursuit models, and show how to draw inferences on their parameters using empirical likelihood. In a generalized projection pursuit model the observed response $Y$ has a parametric likelihood with parameter vector $\theta$. Each component of $\theta$ is of the form $g(\sum_e h_e(X^e\beta^e))$ for some functions $g$ and $h_e$, parameter column vectors $\beta^e$ and covariate row vectors $X^e$. The function $g$ plays the role of the inverse link function from generalized linear models. Inferential interest centers on functions of the $\beta^e$. In the neural network literature the model above for a component of $\theta$ is known as a feedforward network with a single hidden layer.

The model is illustrated by an example in which the log survival time of some plastic pipes has a mean and log standard deviation which are both piecewise planar in log temperature and log pressure. Empirical likelihood is used to make confidence statements about the life time of pipes at a particular pressure and temperature combination, to make a confidence band for the line along which the mean function bends, to test whether the mean and variance functions bend along parallel lines and to form a confidence band for the set of temperature pressure combinations consistent with a given mean lifetime.

A computational method for maximizing empirical likelihood subject to nonlinear equality constraints on the parameter set is described.
1. Introduction.

This paper introduces the generalized projection pursuit models which are formed by using feedforward neural networks with a single hidden layer for each parameter in a loglikelihood (or \( \psi \) function of an \( M \)-estimate). Empirical likelihood is used to draw inferences on the parameters of these models. The resulting inferences are asymptotically valid under mild assumptions on the sampling distribution of the data. Unlike inferences based directly on the parametric likelihood itself, the parametric likelihood model need not be correct for asymptotic confidence statements to hold.

The plan of this article is as follows: Section 2 defines generalized projection pursuit models, Section 3 describes empirical likelihood, Section 4 presents an example application to survival times of plastic pipes, and Section 5 discusses computational issues arising from maximizing empirical likelihood subject to nonlinear equality constraints on the parameters.


Consider a likelihood function \( f(Y, \theta) \) where \( Y = (Y^1, \ldots, Y^A) \) is an \( A \)-vector of observations, usually thought of as responses, and \( \theta = (\theta^1, \ldots, \theta^B)' \) is a \( B \)-vector of parameters.

A generalized projection pursuit model is one in which

\[
\theta^b = g_b \left( \sum_{c=1}^{C(b)} h_{bc}(X^{bc}\beta^b) \right), \quad b = 1, \ldots, B. \tag{2.1}
\]

The \( \beta^{bc} \) are parameter (column) vectors and the \( X^{bc} \) are covariate (row) vectors. We call \( g_b \) and \( h_{bc} \) activation functions, using the terminology from the artificial neural network literature. In the generalized linear model literature the inverse of \( g_b \) is called the link function.

The data consist of \( n \) observations \((Y_i, X_{i1}^{B1}, \ldots, X_i^{BC(B)}), i = 1, \ldots, n\). These may be i.i.d. or the \( X \)'s may have been chosen by a design. We use an unsuperscripted \( \beta \) to denote a vector containing all \( \sum_{1 \leq \beta \leq B} C(b) \) parameter vectors \( \beta^{bc} \).

The estimate \( \hat{\beta} \) of \( \beta \) is defined by maximizing

\[
I = \sum_{i=1}^{n} \log f(Y_i, \theta_i) \text{ where } \quad \theta_i^b = g_b \left( \sum_{c=1}^{C(b)} h_{bc}(X_i^{bc}\beta^{bc}) \right). \tag{2.2}
\]
In the regular cases to which we confine ourselves, \( \hat{\beta} \) is found by solving the likelihood equations obtained by setting to zero the gradient of \( l \) with respect to \( \beta \).

In the example of Section 4, \( Y \) is the log survival time of a plastic pipe. Predictors \( P \) and \( T \), are log pressure and log temperature respectively. The likelihood is the Gaussian one, namely

\[
 f = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(Y - \mu)^2}{2\sigma^2}\right) \text{ where } \theta = (\mu, \sigma)' \text{ and the model used has }
\]

\[
 \mu = \beta^{111} + \beta^{112}P + \beta^{113}T + h_{12}(\beta^{121} + \beta^{122}P + \beta^{123}T)
\]

and

\[
 \sigma = \exp(\beta^{211} + h_{22}(\beta^{221} + \beta^{222}P + \beta^{223}T)).
\]

Here each of \( g_1, h_{11} \) and \( h_{22} \) is the identity function \( I(z) = z \), \( g_2 \) is the exponential function, \( X^{21} = (1) \) and the other three of the \( X^{bc} \) vectors are \((1, P, T)\) row vectors. The functions \( h_{12}, h_{21}, \) chosen to make the arguments of the \( g_b \) piecewise planar, are discussed in Section 4. Piecewise planar regression models have been used by Breiman (1991) to fit nonparametric regressions that perform well in high dimensional problems. Meisel and Collins (1973) give some interesting case studies of piecewise planar regression.

This model includes multivariate inference. For data pairs \( Y = (Y^1, Y^2) \), one might use the bivariate Gaussian density \( f \) and consider models of the form (2.1) for each component of \( \theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)' \).

Binomial responses, with \( s \) successes in \( t \) trials, fit into this framework by taking \( Y = (s, t), \theta = p \) and \( f = \binom{t}{s} p^s(1-p)^{t-s} \). Ordinarily \( t \) is thought of as a parameter, not a response, even though it is seldom estimated from the data. In general, parameters whose values are known for all data points may be subsumed into the response vector.

The gpp model generalizes the projection pursuit regression model of Friedman and Stuetzle (1981), in extending the conditional mean estimated there to other conditional parameters. This is not to say that their work can be subsumed in the gpp work reported here. Friedman and Stuetzle used nonparametric regression estimates for the \( h^{bc} \) functions whereas these functions are of specified parametric forms here. They also estimate the number of projections necessary. That
quantity is taken as fixed here, though it might be found through exploratory data analysis or in some cases through testing of nested models.

Gpp also extends generalized linear models (McCullagh and Nelder, 1989). In a typical generalized linear model the likelihood for $Y$ has a single parameter modeled as $\theta = g(X\beta)$. These models may be obtained from gpp by taking $g_1 = g$, $C(1) = 1$ and $h_{11} = I$. There are multi-parameter glms: see Pregibon (1984), and Chapter 10 of McCullagh and Nelder.

The gpp model is essentially one in which each parameter is a feedforward neural network with a single hidden layer. See Hertz, Krogh and Palmer (1991) for a discussion of neural networks. In a gpp model the $X$'s are the input layer, the $h$'s are the hidden layer and the $g$'s are the output layer. It is common in neural network modeling to take for all of the $h$ functions, the logistic function $L(x) = \exp(x)/(1 + \exp(x))$. The function $g$ is commonly the logistic function or the identity function depending on whether probabilities or means are being estimated. It is also common to have coefficients multiplying the $h$'s. Typically, for estimating means

$$\theta = \sum_c \eta_c L(X^c\beta^c)$$  \hspace{1cm} (2.3)

for coefficients $\eta_c$ and $\beta^c$. Parametrization (2.3) makes identifiability of $\beta^c$ difficult when $\eta_c$ is near zero. For the algorithms used here a parametrization based on (2.1) is more convenient; it makes it easier to keep certain Hessian matrices invertible. The usual neural network algorithms do not use second derivatives and hence have no problems with parametrization (2.3).

The emphasis in neural network modeling is primarily on models with a very large number of projections, and on methods of learning the parameters therein. The emphasis here is on smaller models that may be more interpretable and for which numerical methods based on second derivatives are feasible ways to estimate their coefficients.

Projection pursuit models are flexible enough to approximate any reasonably smooth function of the inputs $X$. This approximation can be made in parametrization (2.3) using only the logistic function (and coefficients $\eta_c$) provided that enough terms are included in the sum. Other families of functions, such as piecewise planar functions also have this property. To see this, approximate
$L(z)$ by a piecewise linear function.


Let $Z, Z_1, \ldots, Z_n$ be i.i.d. random vectors from a distribution $F_0$. The empirical distribution

$$F_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{z_i},$$

where $\delta_z$ denotes a point mass at $z$ is well known to be the nonparametric maximum likelihood estimate of $F_0$. The likelihood $F_n$ maximizes is of the form

$$L(F) = \prod_{i=1}^{n} F(z_i)$$

where $F(z_i)$ denotes $P_F(Z = z_i)$. It is clear that only distributions with an atom on each data point have nonzero likelihood. The nonparametric likelihood (3.1) has been used by Kaplan and Meier (1958) and by Vardi (1985) to define estimators of distribution functions in nonstandard sampling settings.

Confidence regions for a parameter $T(F)$ (with sample value $T(F_n)$) may be based on likelihood ratios

$$R(F) = \frac{L(F)}{L(F_n)}$$

of the likelihood in (3.1). Confidence sets take the form $\{T(F)|R(F) \geq c, F \ll F_n\}$ where $F \ll F_n$ means that $F = \sum_{i=1}^{n} w_i \delta_{z_i}$ with $0 \leq w_i \leq 1 = \sum w_i$. $F$ reweights the original sample, and the greater the departure of those weights from equality, the lower the likelihood ratio $R(F)$ is. Note that $\{T(F)|R(F) \geq c, F \ll F_n\}$ may contain the true value $T(F_0)$ of $T$ without $F_0$ itself belonging to $\{F|R(F) \geq c, F \ll F_n\}$.

Owen (1990) proves the following:

**Theorem** Let $X, X_1, X_2, \ldots$ be i.i.d. random vectors in $\mathbb{R}^p$, with $E(X) = \mu_0$ and $\text{var}(X) = \Sigma$ of rank $q > 0$. For positive $r < 1$ let $C_{r,n} = \{ X dF \mid F \ll F_n, R(F) \geq r \}$. Then $C_{r,n}$ is a convex set and

$$\lim_{n \to \infty} P(\mu_0 \in C_{r,n}) = P(\chi^2_q \leq -2 \log r).$$
Moreover if $E(\|X\|^4) < \infty$ then

$$|P(\mu \in C_{r,n}) - P(\chi^2_0 \leq -2\log r)| = O(n^{-1/2}).$$

The chisquare limit and the $O(n^{-1/2})$ rate are the same as one finds in parametric likelihood confidence regions. Theorem 1 is an empirical likelihood version of Wilks' (1938) theorem for parametric likelihoods. Like parametric likelihoods, empirical likelihoods generally give regions with coverage accuracy $O(n^{-1})$ in two sided problems, $O(n^{-1/2})$ in one sided problems and they may be Bartlett corrected or have their signed roots adjusted for greater accuracy. See DiCiccio, Hall and Romano (1991).

Empirical likelihood inferences extend to some very general parameters $T(F)$. Owen (1990) considers smooth functions of means (Theorem 2) and multidimensional M-estimates (Theorem 3). Owen (1991, Theorem 2) shows that identity of distribution is not required. When $T$ is the variance, the asymptotic coverage property holds provided the sampling distribution has a finite fourth moment. By comparison, normal theory confidence intervals for a variance have asymptotically correct coverage only for distributions having kurtosis zero. For distributions with heavier (lighter) than normal tails the normal theory intervals are asymptotically too narrow (wide) and have incorrect asymptotic coverage levels.

For robust statistics, like some $M$-estimates with nonzero breakdown, the restriction to $F \ll F_n$ need not be made. These statistics do not take extreme values when small amounts of probability are attached to infinite data values.

Owen (1991) extends empirical likelihood to regression problems in which the predictor variables may be fixed by a design. Kolaczyk (1992) makes an extension to generalized linear models. For generalized linear models, the parametric models are often violated by overdispersion. This occurs when the variance does not depend on the mean in the way stipulated by the parametric likelihood family. Empirical likelihood inferences do not require such stringent assumptions, although they typically require the true variances to be finite. Kolaczyk is able to make confidence
statements regarding the amount of overdispersion in a data set.

For statistics other than means, the resulting confidence regions may not be convex. The empirical likelihood method chooses the shape of the region from the data. There is no need to impose elliptical, rectangular or other shape constraints on the region. Hall (1990) finds that for statistics formed as smooth functions of vector means, the resulting confidence regions have asymptotically the correct size, shape and orientation but would be improved by a translation through a distance of $O(n^{-1})$.

The usual approach to computation of empirical likelihood confidence regions is through the profile likelihood ratio

$$\mathcal{R}(\tau) = \max\{R(F)|T(F) = \tau, F \ll F_n\}.$$  

Clearly $\tau \in \{T(F)|R(F) \geq c, F \ll F_n\}$ if and only if $\mathcal{R}(\tau) \geq c$. Computation is discussed further in Section 5.

If the likelihood model is not necessarily to be believed, one must take care in interpreting the parameters. In general the empirical likelihood inferences are appropriate for whatever the maximum likelihood estimators are really estimating, under mild conditions ensuring estimability and finiteness of certain moments. It is often the case that the maximum likelihood estimators also have a simple interpretation based for example, on moments or odds ratios. In that case the empirical likelihood inferences may be interpreted as relating to those moments or odds ratios. We conclude this section with a detailed example of how this lack of dependence on assumptions works. Some readers may prefer to go directly to the example in the next section.

The summary below is based on Owen (1991). Consider the example of a simple linear regression model in which $E(Y|X) = X\beta$, the data $(X_i, Y_i)$ are i.i.d., and the design matrix made of $n$ rows $X_i$ has full rank. One might entertain a model with the $N(X\beta, 1)$ likelihood for $Y$. The resulting estimating equations are the normal equations

$$0 = \sum_i X'_i(Y_i - X_i\hat{\beta}). \quad (3.3)$$

Suppose $E(Y|X) = X\beta$ and $0 < V(Y|X) = \sigma^2 < \infty$, and the $(X_i, Y_i)$ pairs are i.i.d., but $Y - X\beta$
is not necessarily normal. Then \( \hat{\beta} \) is still the minimum variance linear unbiased (Gauss-Markov) estimate of \( \beta \). Empirical likelihood confidence regions for \( \beta \) are asymptotically valid. The normal theory regions will be asymptotically valid provided that \( \sigma^2 = 1 \).

Standard parametric practice would of course be to use \( N(X\beta, \sigma^2) \) with \( \sigma^2 \) to be estimated jointly with \( \beta \). This would lead to asymptotically consistent parametric confidence statements provided that \( V(Y|X) \) is constant. The parametric inferences do not require normality of \( Y - X\beta \) for asymptotic correctness. If however, \( V(Y|X) \) is not constant, the pooled estimate of \( \sigma^2 \) will lead to misleading inferences in general, even asymptotically. By contrast, empirical likelihood inferences based on \( N(X\beta, \sigma^2) \) or even \( N(X\beta, 1) \) have asymptotically correct coverage levels, without requiring normality of \( Y - X\beta \) or constancy of \( V(Y|X) \) in \( X \).

With nonconstant \( V(Y|X) \), one would usually prefer the Gauss-Markov estimating equations

\[
0 = \sum_i X_i'(Y_i - X_i\hat{\beta})/V(Y|X_i) \tag{3.4}
\]

because the resulting estimate of \( \beta \) has smaller variance. The ordinary least squares estimate is inefficient. Empirical likelihood uses the same inefficient estimator, but at least gives asymptotically correct confidence statements for it. In some cases one might be able to guess the form of \( V(Y|X) \) up to a constant factor. Parametric and empirical likelihood methods would then both be based on the estimate (3.4) and might both be an improvement over ordinary least squares in terms of efficiency. But the asymptotic correctness of parametric inferences would depend, in general, on correctness of the guess for \( V \). Empirical likelihood inferences would not require a correct guess, just finiteness of certain moments involving \( X, Y \) and \( V \).

Now suppose that the mean of \( Y \) given \( X \) is not of the form \( X\beta \). Then \( \hat{\beta} \) may be thought of as an estimate of the minimizer of \( E((Y - X\beta)^2) \), that is, the best linear model in the least squares sense. Empirical likelihood inferences are asymptotically correct if viewed as statements about this \( \beta \). In this case estimates and intervals for \( \sigma^2 \) would be appropriate for the variance plus the mean squared bias.

When the \( X \)'s are fixed by design more care has to be taken in specifying what “asymptotic”
means. Mild conditions on moments of the design and response are needed to make empirical likelihood inferences asymptotically correct, provided that $E(Y|X) = X\beta$. Neither constant variance nor identity of distribution is required for the errors $Y - X\beta$.

When $E(Y|X)$ is not linear in $X$ and the $X$'s are fixed by design there is no natural population with respect to which one might minimize $E((Y - X\beta)^2)$. One could consider a sequence of $\beta^{(n)}$ of population values for $\beta$ which minimize

$$\frac{1}{n} \sum_{i}(E(Y_i|X_i = x_i) - x_i\beta)^2.$$ 

This $\beta^{(n)}$ minimizes $E((Y - X\beta)^2)$ for a distribution putting equal mass on each of the first $n$ points $x_i$ in the design. The sample error $\hat{\beta} - \beta^{(n)}$ converges strongly to zero under mild conditions on the design sequence and conditional distributions of $Y$ given $X$. Empirical likelihood confidence statements on this $\beta$ are not in general asymptotically correct. In general they are asymptotically conservative in that the confidence level claimed exceeds the true level.

4. Pipe strengths

A set of 208 plastic pipes was tested under varying conditions of pressure and temperature. The time until failure of each pipe was observed. The response of interest, $Y$, is the logarithm of this time. The predictors $P$ and $T$ are the logarithms of the pressure and temperature, respectively.

I thank Richard Gill for sending me the data.

The physical model suggested for these data has for $E(Y)$, a piecewise linear model in $X = (1, P, T)$. The reasoning is as follows: there are two failure modes, plastic and brittle. For each mode there is a failure time model in which $E(Y)$ is linear in $X$ and the smaller of these times is the one to win the race and be observed. See van de Geer (1987) for a discussion of this model as a generalization of the one dimensional change-point problem.

Figure 1 shows the design points at which data were obtained. The solid line in Figure 1 is an estimate (described below) of the dividing line between brittle and plastic failure. The area to the right of and below this line, with higher pressure and lower temperature, corresponds to brittle failure. Figure 2 shows $Y$ versus $P$. The piecewise behavior is readily apparent within the four
clumps of points corresponding to the four temperature settings shown in Figure 1. Because the slopes of the two regressions have quite different magnitude, it is not visually apparent from Figure 2 that a piecewise linear model is suitable. A plot, not shown here, of only the points in the brittle failure zone shows the predicted linear trend. It is also clear from Figure 2 that the variance of the response times may not be constant.

One way to parametrize this model is

$$E(Y|T,P) = X\beta^{11} - (X\beta^{12})_+$$  \hspace{1cm} (4.1)

where $z_+$ is the positive part of $z$, that is $z_+ = z$ if $z \geq 0$ and $z_+ = 0$ if $z \leq 0$. We prefer to approximate the function $z_+$ by a hyperbolic branch

$$z_\epsilon = \frac{z}{2} + (\epsilon^2 + \frac{z^2}{4})^{1/2}$$  \hspace{1cm} (4.2)

which approaches $z_+$ as $\epsilon$ approaches zero. This smooths out the likelihood surface, but for this data, using $\epsilon = 0.05$ doesn't change the fits much. With this smoothing for the mean, and a Gaussian likelihood for $Y$ our model is a parametric one for which standard asymptotic results can be expected. The asymptotics would be nonstandard if the true regression surface consisted of a single plane and we attempted to fit it by a model with an extraneous bend. See Davies (1987), Hinkley (1969) or Knowles and Siegmund (1989). Smoothing the model may not be necessary for standard asymptotic results to apply, as explained by Pollard (1985). It does give rise to a log likelihood that can be differentiated as many times as necessary, and this can be advantageous in computation.

If one takes the physical model very seriously, the regression surface should be smoothed out somewhat. Suppose we observe $Y = \min\{X\gamma_B + \epsilon_B, X\gamma_P + \epsilon_P\}$ instead of $Y = \min\{X\gamma_B, X\gamma_P\} + \epsilon$ and consider a point for which the regression planes for brittle and plastic failure coincide. The regression for $Y$ will be smaller than the common regression value by an amount equal to the expected minimum of the two error terms.

Parametrization (4.1) may strike some readers as unnatural. Here is why it was chosen: By parametrizing the piecewise linear function this way, the natural size for $\epsilon$ depends on the scale
of \( Y \), since quantities of the same dimension as \( Y \) are computed and then passed through the smoothed ramp function (4.2). In the alternative parametrization

\[
X\beta^{11} + \beta^{121}(\beta^{122} + \cos(\beta^{123})T + \sin(\beta^{123})P)\epsilon
\]  

(4.3)

the natural choice of \( \epsilon \) would appear to depend on the projection angle \( \beta^{123} \). Parametrization (4.1) requires one to specify whether a concave or convex model is desired, while parameterization (4.3) contains both models depending on the sign of \( \beta^{121} \). This gain may be illusory since a likelihood function based on parametrization (4.3) becomes degenerate as \( \beta^{121} \) approaches zero.

Upon fitting the model

\[
Y \sim N(X\beta^{11} - (X\beta^{12})._{.05}, 1)
\]  

(4.4)

by the Gauss Newton algorithm, and inspecting the residuals, it becomes apparent that the error variance is not constant. Figure 3 shows the residuals from the piecewise linear fit plotted against the crease projection \( X\beta^{12} \). Empirical likelihood inferences on the regression surface should still be trustworthy. But, by modeling the variance we can improve the estimate of the surface since the likelihood equations will then put greater weight on observations from regions with smaller error variance.

The error standard deviation appears to be approximately of the form

\[
V(Y|T, P)^{1/2} = \exp (X\beta^{21} + (X\beta^{22})_{+})
\]

with the log standard deviation crease line \( X\beta^{22} = 0 \) close to the mean crease line \( X\beta^{12} = 0 \). The maximum empirical likelihood estimates from the model

\[
Y \sim N(X\beta^{11} - (X\beta^{12})._{.05}, (X\beta^{21} + (X\beta^{22})._{.05})^{2})
\]  

(4.5)

are:

\[
\beta^{11} = (-14.68263, -1106.669, 6795.153)', \\
\beta^{12} = (50.18946, 12550.52, -28483.48)', \\
\beta^{21} = (-3.059234, -174.333, 562.5376)', \\
\beta^{22} = (17.2454, 3984.993, -9184.009)'.
\]
These estimates are given with more significant digits than would naturally be used to study this set of pipes, in order to allow other workers to make comparisons. The parameters themselves are not readily interpreted, but certain functions of them are, and inferences are drawn on them below.

The dashed line in Figure 1 indicates the crease along which \( X\beta^{22} = 0 \). The variance changes along a line nearly parallel to the one separating plastic from brittle failure. The change takes place within the plastic failure zone.

If one maximizes the empirical likelihood ratio, subject to \( \beta^{212} = \beta^{213} = 0 \) the maximum is 0.661. Referring \(-2 \log(0.661) = 0.828\) to a \( \chi^2_{(2)} \) distribution, the result is clearly not significant. We therefore simplify the model by replacing \( X\beta^{21} \) by the constant \( \beta^{211} \). In this model, the log failure time has constant variance for pipes well inside the plastic failure zone. The resulting estimates for the parameters are:

\[
\beta^{11} = (-14.5434, -1084.55, 6731.54),
\beta^{12} = (50.1318, 12521.7, -28432.1),
\beta^{21} = (-1.74319),
\beta^{22} = (15.611, 3723.37, -8430.28).
\]

The lines on which the mean and variance models fold appear to be parallel. We can test whether this might be true. If one maximizes the empirical likelihood ratio subject to the constraint that \( \beta^{122}/\beta^{123} = \beta^{222}/\beta^{223} \) the result is 0.668 and referring \(-2 \log(0.668) = 0.808\) to a \( \chi^2_{(1)} \) distribution the data clearly do not reject the hypothesis that the mean and variance lines are parallel. If we were convinced, on scientific grounds, that these lines had to be parallel we could impose this constraint to reduce the size of the model by one parameter and improve our inferences.

Figure 4 shows contours of the estimated surface for \( E(Y) \). The design points are superimposed. It is clear that the lifetime of these pipes plunges dramatically along a ridge. Suppose we are interested in the lifetimes of pipes for which \( P = P_0 = 0.003 \) and \( T = T_0 = .003084433 \). This combination of pressure and temperature lies on the crease between failure modes, and is roughly in the middle of the data. According to the model, the mean log survival time for a pipe under
such conditions is

\[ S(X_0, \beta) = X_0 \beta^{11} - (X_0 \beta^{12})_{.05}, \]

where \( X_0 = (1, P_0, T_0) \). The empirical maximum likelihood estimate of this mean is

\[ S(X_0, \hat{\beta}) = 2.92. \]

We would expect a confidence interval for \( S(X_0, \beta) \) to be negatively skewed because uncertainty about the position of the crease could cause \( X_0 \) to lie in the brittle failure zone and have a much lower survival time, or cause \( X_0 \) to lie in the plastic failure zone and have only a slightly higher survival time. Figure 5 shows the profile likelihood of \( S(X_0, \beta) \). It exhibits the expected skewness and has a small second mode, which was not anticipated. Figure 6 shows the profile loglikelihood of \( S(X_0, \beta) \) with reference lines appropriate to asymptotic .95 and .99 confidence levels.

It would be interesting to know where the transition between plastic and brittle failure occurs. A point \((P, T)\) is on the transition line if and only if \((1, P, T)\beta^{12} = 0\). This can be tested by maximizing the empirical likelihood subject to the constraint \((1, P, T)\beta^{12} = 0\). Let \( R_{tfm}(P, T) \) be this maximum, where the subscript “tfm” is mnemonic for the transition between failure modes. One could imagine computing \( R_{tfm} \) at all points in the \((P, T)\) plane and forming a set of those points for which \( R_{tfm}(P, T) > c \).

How could one choose the value of \( c \)? A choice which turns out to be conservative, is to use the chisquare distribution on 3 degrees of freedom. There are three parameters in \( \beta^{12} \), and the set of points \((P, T)\) for which \( R_{tfm}(P, T) > c \) is determined by the set of values of \( \beta^{12} \) with an empirical likelihood ratio as large as \( c \). In this special case however, only two degrees of freedom are really used. For example, assuming \( \beta^{123} \neq 0 \), the ratios \( \beta^{121}/\beta^{123} \) and \( \beta^{122}/\beta^{123} \) determine whether a given point is on the crease or not. This encoding corresponds to the slope and intercept of the crease line. The third degree of freedom in \( \beta^{12} \) governs the angle between the two planes at the crease, not where the crease is.

We expect the confidence set for the transition line to be shaped roughly like the region between the two branches of an hyperbola. This inside region is traced out by all the lines in the
$P,T$ plane for which $\beta^{12}$ takes values with large enough empirical likelihood. Similarly shaped regions commonly arise as simultaneous confidence sets for linear regressions.

Figure 7 plots the estimated crease along with contours of $\mathcal{R}_{12}$ corresponding to confidence levels .99 and .95. This plot was formed by computing $\mathcal{R}_{12}$ at many points in the $(P,T)$ plane, interpolating the results to a grid and contouring the grid values. The confidence bounds have small wiggles in them that are due not to the empirical likelihood function, but are artifacts of the interpolation and contouring methods used. These are the interp and contour functions in S (Becker, Chambers and Wilks 1989), which, as general purpose functions, had no way of knowing that the contours should be convex envelopes.

A striking feature of Figure 7 is that there is very little uncertainty in the angle of the transition line. That is $\beta^{122}/\beta^{123}$ is well determined by the data. There is some uncertainty allowing this line to be shifted parallel to itself, with the result that the confidence region for the points $(P,T)$ on the transition line is very nearly the region between two parallel lines. Also apparent from Figure 7 is that shifts of the crease towards the plastic failure region have greater plausibility than shifts towards the brittle direction. Shifts into the brittle failure region are more plausible at high temperatures; this is probably because the data points from highest temperature do not appear to straddle the crease line. Instead it appears that they are all in the brittle failure region.

Suppose that an expected log lifetime of 3 time units were critical to success of the application using these pipes. The contour on which the expected log lifetime is predicted to be 3 units appears in Figure 4. This contour passes through both the plastic and brittle failure regions. For a piecewise planar regression this contour would be a broken line, but here it is somewhat smoothed. To get a confidence band for the contour $S(X,\beta) = 3.0$ we maximize empirical likelihood subject to the constraint $S(X,\beta) = 3.0$ for various $X = (1,P,T)$ and compute the set of $(P,T)$ values for which this likelihood is sufficiently large. Four degrees of freedom are appropriate here: two to describe one of the linear parts of the contour $S = 3$, a third to describe where on that line the second linear part intersects, and a fourth to describe the angle at the intersection. The resulting confidence set is the union of a set of bent lines in the same way that the confidence set for the transition line is
the union of a set of lines.

At a sample of points clustered around the contour \( S(X, \hat{\beta}) = 3 \) in the \((P, T)\) plane, the empirical likelihood was maximized subject to the constraint \( S(X, \beta) = 3 \). Figure 8 shows those points at which \( S(X, \beta) = 3 \) cannot be rejected at the .95 confidence level based on the empirical likelihood. The outer limits of these points give an impression of the region in the \((P, T)\) plane that could have a mean log survival time of 3. Also shown is the sample contour, \( S(X, \hat{\beta}) = 3 \). There is a fairly narrow bent band in which a mean log survival time of 3 units is reasonable. To the left and above this band the mean is greater than 3 units and to the right and below, the mean is less than 3 units. The part of the band within the brittle failure region is narrower as might be expected because of the greater steepness of the regression surface there. Interpolation and contouring of \( R(S(X, \beta) = 3) \), similar to that done for Figure 7, produced an unsuitable plot in which the 95% contour consisted of a number of closed curves in the 95% region.

5. Computation.

In this section we consider how to maximize the empirical likelihood for simple hypotheses that specify all of the parameters. Then we consider "rectangular" composite hypotheses in which some of the parameters are fixed and others are free to vary. Finally we consider composite hypotheses expressed in terms of a list of equality constraints on the parameters.

The parameter estimate \( \hat{\beta} \) is defined by the estimating equations

\[
\sum_{i=1}^{n} Z_i(\hat{\beta}) = 0 \tag{5.1}
\]

where \( Z_i = \nabla_{\beta} \log f(Y_i, \theta_i) \) is the gradient with respect to \( \beta \) of the contribution to the parametric log likelihood from the \( i \)'th observation. An expression for \( Z_i \) may easily be worked out in terms of the functions \( g_0 \) and \( h_{bc} \), the covariates \( X^{bc} \) and the components of \( \beta \) by applying the chain rule.

Two derivatives of \( Z_i \) with respect to \( \beta \) are needed. Thus to implement a new log likelihood one needs to provide routines for the first three derivatives of the log likelihood with respect to its parameters. These are of course the score function and its first two derivatives. Also required are
the activation functions \( g_b \) and \( h_b \) and their first two derivatives. We do not write out expressions for the repeated derivatives of the log likelihood since they become tedious.

In this section we treat \( \beta \) as a vector of \( p \) components.

To test a simple hypothesis \( \beta = \beta_0 \), one computes \( Z_i = Z_i(\beta_0) \) and maximizes \( \prod_{i=1}^n w_i \) subject to \( 0 \leq w_i \leq 1 = \sum w_i \) and \( \sum w_i Z_i = 0 \). This computation is described in detail in Owen (1990). Here we sketch the solution. Some simple Lagrange multiplier calculations lead to \( w_i = w_i(\lambda) = n^{-1}(1 + \lambda' Z_i)^{-1} \) where the multiplier \( \lambda \) is a column vector of length \( p \), for which \( \sum w_i(\lambda) Z_i = 0 \). The solution \( \lambda \) may be found by minimizing

\[
G_0(\lambda) = -\sum_{i=1}^n \log(1 + \lambda' Z_i)
\]

over \( \lambda \). Maximization of the log empirical likelihood in \( n \) parameters with \( p+1 \) equality constraints has been replaced by an unconstrained minimization over \( p \) parameters. This minimization over the Lagrangian vector is the convex dual problem to the original maximization. This dual problem is an easy one numerically since \( G_0 \) is convex in \( \lambda \).

There is usually more interest in composite hypotheses, which specify a set \( B \) of values for \( \beta \). For example to test whether \( \beta_j = 0 \) one maximizes the empirical likelihood over possible values for \( \beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p \) while holding \( \beta_j = 0 \). Each maximization is in fact a minimization over \( \lambda \). Letting

\[
G(\lambda, \beta) = -\sum_{i=1}^n \log(1 + \lambda' Z_i(\beta))
\]

we compute

\[
\max_{\beta \in B} \min_{\lambda} G(\lambda, \beta), \quad (5.2)
\]

with \( B = \{\beta | \beta_j = 0\} \). The inner minimization in (5.2) is of a convex function as noted above. The outer maximization in (5.2) is, at least for large \( n \), of a concave function (Owen, 1990).

The problem (5.2) is one of finding a saddle point, or maximin value. It can be approached as in Owen (1990) by using a nested algorithm in which an optimization routine at the outer level calls a function at the inner level that minimizes with respect to \( \lambda \). A disadvantage of this approach
is that at the early stages much of the work done in optimizing over \( \lambda \) is expended for values of \( \beta \) that are not near the ultimate solution.

For the work reported here, the saddle point was computed by applying Newton's method to \( G \) as a function of \( \lambda \) and \( \beta \). This amounts to Newton's method for solving nonlinear equations applied to

\[
\nabla_\beta G = 0, \nabla_\lambda G = 0
\]

(5.3)

where \( \nabla_\beta \) denotes the gradient with respect to those components of \( \beta \) that are not fixed by the composite hypothesis.

Another way to derive this algorithm is to approximate \( G \) by a function \( Q \), quadratic in \( \lambda \) and those components of \( \beta \) not fixed by the hypothesis. So long as \( Q_{\lambda \lambda} = \nabla_\lambda' \nabla_\lambda Q \) is positive definite one can find \( \lambda(\beta) \) to minimize \( Q \) over \( \lambda \) with \( \beta \) fixed. One substitutes this \( \lambda(\beta) \) into \( Q \) to get a quadratic function of \( \beta \) alone. If \( Q_{\beta \beta} - Q_{\beta \lambda} Q_{\lambda \lambda}^{-1} Q_{\lambda \beta} \) is negative definite then the maximum over \( \beta \) is easily found. The resulting step for \( \lambda, \beta \) is the same as the Newton step.

The solution to (5.3) is approached with the quadratic convergence typical of Newton's method, provided that the starting values are close enough to the solution. That Newton's method can converge to a saddlepoint is well known as a pitfall of minimization routines. Here it is the desired behavior.

In practice Newton's method must be used with care, since it can be unreliable when the starting point is not close enough to the solution. A standard remedy is to halve the Newton step if it does not result in progress towards the desired solution. In maximization (respectively minimization) problems there is a natural indicator of progress: was the objective function increased (reduced)? In this problem we accept a step if the norm of the gradient

\[
\|\nabla_\beta G\|^2 + \|\nabla_\lambda G\|^2
\]

(5.4)

has been reduced. This is not completely satisfactory. The Hessian of (5.4) may be shown to be \( 2Q^2 \) where \( Q \) is the Hessian of \( G \) from (5.2). Thus if \( Q \) is badly conditioned, the criterion used to accept a step will be worse conditioned still.
Problems of ill conditioning may arise when the likelihood is overparametrized, or when very extreme values of the likelihood are desired. In Figures 5 and 6 this ill conditioning was reached at $\log R = -22.7$ and $-20.3$ at the upper and lower ends of the profile respectively. These end points correspond to confidence limits more extreme than one would need in practice. Figures 7 and 8 are based on empirical likelihood maximizations attempted at a large number of points in the $P, T$ plane. The simple step halving strategy often failed to reduce the gradient to negligibility. A tolerance of $10^{-9}$ was used, but when the method failed to converge the gradient could be as large as $10^{-2}$. Failure of the algorithm is more likely for points far from the sample crease in Figure 7 or sample contour in Figure 8.

A substantial improvement in speed is gained by replacing the original empirical likelihood maximization problem by (5.2). Owen (1991) and Kolaczyk (1992) both use the sequential quadratic programming package NPSOL on a problem with $n + p$ parameters $w_1, \ldots, w_n, \beta_1, \ldots, \beta_p$. Solving the equations for a step direction takes $O((n + p)^3)$ work. To find the Newton step in solving (5.3) takes $O(np^2)$ to form the Hessian and $O(p^3)$ work to solve the equations for the Newton step. There is no part of the cost that grows as $n^3$ in the Newton method. Kolaczyk reports, in a personal communication, that one empirical likelihood profile for a logistic regression using the Kyphosis data ($n = 81, p = 4$) described in Becker, Chambers and Wilks (1988) took at least 6 hours of elapsed time to compute. This can’t be directly translated into a cpu time, but cpu time was probably a substantial fraction of the elapsed time. On the same cpu (DECstation 5000/200) the Newton method computes profiles for all four parameters in under 5 minutes of cpu time.

The solution to (5.3) may be used to test whether a parameter or set of parameters could plausibly be zero. This test is often of interest because it suggests a smaller model might be appropriate. When a parameter $\beta_j$ has an important interpretation, then one should consider profiling it. This is done by repeated solution of (5.3) with $B = \{\beta|\beta_j = \beta_j^*\}$ and varying $\beta_j^*$. It is convenient to start from $\beta_j^* = \hat{\beta}_j$ and then to increase $\beta_j^*$ by small amounts so that each application of Newton’s method need only make small changes to $\beta$ and $\lambda$. This produces half of the desired profile empirical likelihood. The other half is computed by reducing $\beta_j^*$ from $\hat{\beta}_j$. Owen

By fixing none of the components of $\beta$ one can find the solution to (5.1) from reasonable starting values. In many settings the solution to (5.1) is easy to find by some other method (such as linear or logistic regression). For the gpp model, the mle is not generally available from other sources, and solving (5.3) with all components of $\beta$ varying provides a way of solving (5.1). Each problem may require its own methods for finding starting values; this can be true of direct methods of solving (5.1) and hence cannot be less true of solving (5.3). For the pipe data, estimates of the parameters in model (4.4) (using Gauss Newton) were used as starting values for $\beta^{11}$ and $\beta^{12}$ in model (4.5). The residuals from (4.4) were analyzed to get starting values for $\beta^{21}$ and $\beta^{22}$. The Gauss Newton method itself requires starting values; these were based on examination of the residuals from a simple linear model relating $Y$ to $P$ and $T$. The starting values for $\lambda$ may be taken to be zero.

For the pipe data, we were interested in non-rectangular constraints. A linear constraint on the parameters was of interest in drawing inferences on the location of the transition between failure modes. A nonlinear constraint was used to test whether the creases in the models for mean and log standard deviation are parallel. Another nonlinear constraint was used to test whether the mean log survival at a particular point in the plane might take a given value.

Suppose the composite hypothesis of interest is of the form

$$\beta \in B = \{ \beta \mid C_j(\beta) = 0, j = 1, \ldots, J \},$$  \hspace{1cm} (5.5)

where the $C_j$ are twice differentiable functions of $\beta$. Equation (5.3) uses a partition of the components of $\beta$ into those that are fixed and those that are unconstrained. The constraints in (5.5) do not generally fit this pattern. The maximization in (5.2) must become a constrained one.

From a purely formal point of view the constraints in (5.5) constitute $J$ additional estimating equations, and they could be handled as such, by adding them to the list of the components of the $Z$. Therefore, in principle, methods based on (5.3) can be adapted to hypotheses (5.5). But since the $C_j(\beta)$ do not depend on the observations, is better to handle them separately. At the very least
one should not recompute them each time through the loop over the observations used to compute the \( Z_i \).

Let
\[
G_c(\lambda, \beta, \eta) = G(\lambda, \beta) + \sum_{j=1}^{J} \eta_j c_j(\beta). \tag{5.6}
\]

\( G_c \) incorporates into (5.2) a Lagrange multiplier for each constraint. Then (5.3) is replaced by
\[
\nabla_\beta G_c = 0, \; \nabla_\lambda G_c = 0, \; \nabla_\eta G_c = 0. \tag{5.7}
\]

In the pipe example the constraint functions were themselves parameterized. For example a constraint of the form
\[
C(\beta) = \gamma' \beta - \gamma_0 \tag{5.8}
\]
forces \( \gamma' \beta = \gamma_0 \). To construct Figures 5, 6 and 8 a constraint of the form
\[
\gamma_1' \beta - (\gamma_2' \beta)_{.05} - \gamma_0
\]
was used. For Figure 8 the parameters \( \gamma_1 \) and \( \gamma_2 \) were varied to describe different points in the \( P, T \) plane. The computation for Figures 5 and 6 was done by profiling \( \gamma_0 \) using the method described above for profiling a component of \( \beta \).

6. Acknowledgements.

I thank Richard Gill for providing the pipe data. I am grateful to Eric Kolaczyk, John Nelder and Daryl Pregibon for helpful discussions. Special thanks go to Peter Hall who suggested computing confidence contours around a contour of the mean log survival function.

A research version, Nel2.0, of a C program for computing empirical likelihoods is available from the author. It includes code for gpp, glm and some other estimating equations and for several constraints including linear \( (\sum \gamma_j \beta_j - \gamma_0 = 0) \), conic \( (\sum \gamma_j \beta_j^2 - \gamma_0 = 0) \), glm \( (g(\sum \gamma_j \beta_j) - \gamma_0 = 0) \) and gpp \( (g(\sum h_c(\sum \gamma_j \beta_j)) - \gamma_0 = 0) \). (This availability is subject to change without notice; present plans are to make the code available until at least the end of 1992.)
REFERENCES


Davies, R.B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 74, 33–43.


CAPTIONS

Figure 1: Log temperature and pressure for 208 plastic pipes tested until failure. The solid line is an estimate of the border between regions for brittle and plastic failure of the pipes. The dashed line is an estimate of the line along which the model for the variance of log survival times changes.

Figure 2: Log survival time versus log pressure for 208 plastic pipes.

Figure 3: The vertical axis shows the residuals from a piecewise planar regression of log survival plotted versus log temperature and pressure. The horizontal axis shows log temperature and pressure projected on the normal to line at which the regression planes meet.

Figure 4: Contours of piecewise planar regression surface relating log survival time to log temperature and pressure. Experimental points are marked with an x.

Figure 5: Profile empirical likelihood for mean log survival of pipes at log pressure $P_0 = 0.003$ and log temperature $T_0 = .003084433$. The underlying models for mean log survival time and for log standard deviation of log survival time are both piecewise planar as described in the text.

Figure 6: Profile log empirical likelihood corresponding to empirical likelihood plotted in Figure 6. Horizontal lines are reference lines for 95% and 99% confidence sets for mean log survival time.

Figure 7: The dotted line is the estimate of the line separating plastic and brittle failure regions. Surrounding it are 95% and 99% confidence envelopes for this line.

Figure 8: The bent line is the estimate of the contour on which mean log survival time is 3 units. The points shown are interior to the 95% confidence region for this contour.
Design Points for Pipe Experiment
Raw Data from Pipe Experiment
Residuals versus Crease Projection

Figure 3
Figure 4

Contours of Mean Log Survival
Mean Survival at $P=0.003 \ T=0.00308$
Mean Survival at $P=0.003 \ T=0.00308$

Predicted Mean Log Survival
With .95 and .99 confidence limits
Confidence Sets For Failure Mode Transition

Log Pressure
at .95 and .99 levels, mle is dashed line
95% Confidence Set For Log Mean Survival = 3.0

Figure 8