PERCENTILE MATCHING: AN ALTERNATIVE METHOD OF CURVE FITTING

James H. Smith, Lockheed Engineering and Management Services Company, Inc., Houston, Texas

ABSTRACT

Conventional methods of curve fitting, particularly least squares, often have the undesirable property of destroying the a priori density of the predicted variable. That is, given a set of ordered pairs \( \{(x, y)\} \) and an estimated function of \( x, \hat{f} \), the density of \( \hat{f}(x) \) will in many cases differ markedly from the density of \( y \). The method of curve fitting proposed in this paper is designed to force the density of \( \hat{f}(x) \) to be nearly identical to that of \( y \), in cases where the expectation of \( y \) given \( x \) is monotonic in \( x \). An example of the application of this technique, using several SAS procedures, is provided. In developing a simulator of the Landsat data collection system as a part of the Inventory Technology Development project of the Agriculture and Resources Inventory Surveys Through Aerospace Remote Sensing program, the proportion of sky cover recorded by ground observers is modeled as a function of radiometric readings from National Oceanic and Atmospheric Administration polar-orbiting meteorological satellites.

INTRODUCTION

When a model is fit to a set of ordered pairs, \( \{(x, y)\} \), the resulting set of predicted \( y \)'s, say \( \{ \hat{y} \} \), frequently possesses a distribution different from that of the original \( y \)'s. In most applications, this disparity causes little concern, since the usual goal of modeling is to produce predictions with minimum expected error (with respect to some norm, such as \( L_1 \) or \( L_2 \)).

This paper describes a problem which arose recently in a simulation modeling context and which required the distribution of the predicted \( y \)'s to closely resemble the distribution of the original data. The problem was solved by a technique which we refer to as percentile matching. That is, the predicted value of \( y \) for an observed \( x \) is found by selecting \( y \) so that its percentile in the cumulative distribution function (cdf) of \( y \) matches the percentile of \( x \) in the cdf of \( X \). While this method requires the assumption that \( E(y|x) \) is monotonic in \( x \), it does not assume a functional form to express the relationship between \( X \) and \( Y \). Although the procedure bears a resemblance to rank regression methods (see Iman and Conover, 1979), it differs fundamentally in that the latter depends on the Spearman rank correlation coefficient while percentile matching depends only on the sign of the correlation between \( X \) and \( Y \). Before describing the technique in detail, we turn our attention to the problem which motivated the development.

THE PROBLEM

A persistent problem encountered in Earth resource monitoring by remote-sensing satellites such as Landsat (see Landsat Data User's Handbook, 1979) is the data loss that occurs because of cloud cover. A set of simulation software has recently been developed (Smith et al., 1982; Ramey et al., 1982) as a part of the research conducted through the Agriculture and Resources Inventory Surveys Through Aerospace Remote Sensing (AgRISTARS) program. This study involves modeling the process through which Landsat data are collected, processed, and aggregated to regional crop acreage estimates (see Erickson et al., 1982; MacDonald and Hall, 1980).

The design of the simulator required the fitting of a relation between radiometric readings from the environmental satellite NOAA-4 and cloud cover proportions recorded by ground observers. The predictor variable, \( X \), was calculated by a principal component transform (using the SAS procedure PRINCOMP) on the NOAA-4 data. The dependent variable, \( Y \), consisted of ground-based observations of sky cover from the North Central United States. For each day of the 1976 growing season, both the ground observations and the principal component transformed satellite data were interpolated to fill a 0.5 degree latitude-longitude grid covering the region, and 25 grid points were randomly sampled from the grid on each day, providing a data set consisting of some 3,300 observations of \( X \) and \( Y \). Figure 1 is a plot of 300 observations sampled from this data set. A major difficulty in fitting a model to the \( y \)'s is apparent: the density of \( Y \) is a mixture of continuous and discrete densities (sky cover proportion is 0 or 1 with nonzero probability).

FIGURE 1—PLOT OF CLOUD COVER AS A FUNCTION OF \( X \)
Figure 2 is a plot of sample means of the y's, taken on intervals (in X) of unit length. It is apparent from the figure that the conditional expectation of Y, given X, increases monotonically with X. Figure 3 is a histogram of the 3,300 y values, which shows the U-shape of the cloud cover density (see Barrett and Grant, 1976). The design of the simulator required that the prediction equation produce a set of predictions with a density similar to that seen in the figure, but standard least squares regression using any of several models failed to produce acceptable densities. Figure 4, for example, shows a histogram of the y's predicted with a cubic regression model, where any predicted values less than zero or greater than one have been set to 0 or 1, respectively. The shape of the density has obviously been severely distorted.

A SOLUTION

In the above application, the most crucial shortcoming of any of the usual modeling techniques was the failure to produce proportions of zeros and ones which were reasonably close to the observed proportions. We can assure the correct proportion of zeros (say P₀) in the y's by drawing a discriminant boundary along the x-axis so that 100P₀ percent of the observed x's lie to the left of the boundary. Then for any value of X to the left of this boundary, we set the predicted value of Y to be zero. Similarly, we can assure the correct proportion of ones (say P₁) by first drawing a discriminant boundary so that 100P₁ percent of the x's lie to the right of this boundary and then setting the predicted value of Y to be one. Note that the first discriminant boundary is the 100P₀th percentile of X and the second is the 100(1-P₁)th. Of course, any two disjoint sets of intervals containing, respectively, 100P₀ and 100P₁ percent of the x's could be used to ensure that the correct percentages of zeros and ones are obtained, but the monotonicity of the expected value of Y in X makes the choices above best for the problem at hand.

Once the above discrimination has been performed, we could then proceed to fit the remaining data with some standard technique. However, it is appealing to extend the idea of percentile matching introduced above to fit the entire set of data. The method is described in the next section.

THE METHOD

The following discussion first assumes that the distributions of the variables are known and then treats the case where the distributions are estimated via empirical cdf's. Let F(X) denote the (true) cumulative distribution of X (i.e., the independent variable) and G(Y) denote the cdf of Y, and assume that E(Y|X) increases monotonically in X. Then to find the predicted value of Y, say \( \hat{y}_0 \), corresponding to a given value of X, say \( x_0 \), first find \( F(x_0) \) and call this value \( C_0 \). Note that \( x_0 \) is the 100C₀th percentile of X. The predicted value of Y, \( \hat{y}_0 \), is then simply the 100C₀th percentile of Y. That is,

\[
\hat{y}_0 = G^{-1}(C_0) = G^{-1}(F(x_0))
\]
Assume that X and Y are from a bivariate normal population with mean vector $\mu$ and variance-covariance matrix $\Sigma$,

$$
\mu = \begin{bmatrix}
\mu_1 \\
\mu_2
\end{bmatrix},
\Sigma = \begin{bmatrix}
\sigma_1^2 & \rho \sigma_1 \sigma_2 \\
\rho \sigma_1 \sigma_2 & \sigma_2^2
\end{bmatrix}
$$

The least squares predictor of $Y$ is the conditional expectation of $Y$, given that $X = x$:

$$
y_0 = \mu_2 + \rho \frac{\sigma_2}{\sigma_1} (x - \mu_1) \quad (3)
$$

It is easy to show that the predictor derived using percentile matching, assuming it is known that $p > 0$, is

$$
y_P = \mu_2 + \rho \frac{\sigma_2}{\sigma_1} (x - \mu_1) \quad (4)
$$

and it is easily seen that the ratio of the standard errors (SE's) of these two predictors is

$$
\frac{\text{SE of } y_1}{\text{SE of } y_0} = \frac{1 + \rho}{2} \quad (5)
$$

We next consider the case where the true distributions of $X$ and $Y$ are unknown. By properly defining the range and domain of the empirical cdf of $Y$, it is possible to define an inverse of this function and apply equation (1), replacing the true distributions by the empirical cdfs. When no ties occur among either the $x$'s or the $y$'s, percentile matching is then equivalent to rank matching: if the rank of $x_0$ in the set of $x$'s is $R$, then predict $Y$ by the observed $y$ which has rank $R$ among the set of $y$'s. Matching ranks fails to preserve the density of $Y$, however, if there are ties among the $x$'s and each of the tied $x$ values is assigned a common rank (e.g., the average rank), since several distinct $y$ values may correspond to this rank. If the tied $x$'s are assigned random ranks (Conover, 1980), then the density of the $y$'s will match exactly that of the $y$'s regardless of the presence of any ties among the $y$'s. The obvious drawback of a rank-matching approach is that it requires lists containing the original data, and these lists can be quite lengthy. The next section describes a means by which this requirement can be avoided.

**IMPLEMENTATION**

In cases where the original data sets are very large, it is inconvenient to maintain files containing them. A reasonable alternative might be to fit the empirical cdfs with some parametric models such as nth-degree polynomials. Since the inverse of the cdf of $Y$ is the function actually used in the calculations, this function should be fit rather than the cdf. This approach has the disadvantage of requiring the assumption of a model form; but if the number of observations is large, the shape of each cdf is usually quite smooth and allows very good approximations by polynomial models of moderate degree. If there are discontinuities in the cdf of $Y$, as in the case of the cloud-cover data, then a piecewise approach like that described in the next example may be required. The following example describes the fitting of the cloud-cover data set using percentile matching (the empirical cdfs referred to were found using the SAS FREQ procedure).

**Example 2**

As has been mentioned before, the ground-observed cloud proportions ($Y$'s) contained large proportions of zeros ($P_0$) and ones ($P_1$). The independent variable, $X$, was thus partitioned into three intervals:

1. $X \leq P_0$,
2. $P_0 < X < 1 - P_1$,
3. $X \geq 1 - P_1$.

where $F$ now represents the empirical cdf of $X$. The prediction for $Y$ when $x$ lies in interval (1) is then zero, and is one when $x$ lies in interval (3). For $x$ in the interval (2), the prediction of $Y$ was found by first fitting a fifth-degree polynomial to the set of ordered pairs {$(x, F(x))$} with $P_0 < F(x) < 1 - P_1$, using the SAS GLM procedure, and then by fitting a fifth-degree polynomial to the set of ordered pairs {$(G(y), y)$} with $0 < y < 1$, where $G$ denotes the empirical cdf of $Y$. The first curve was constrained through the points $(X_0, P_0)$ and $(X_1, 1 - P_1)$, where $X_0$ is the $(100P_0)$th percentile of $X$ and $X_1$ is the $100(1 - P_1)$th percentile. The second curve was constrained through $(P_0, 0)$ and $(1 - P_1, 1)$. These constraints were included to make the fitted curves continuous at the points where the distribution of $Y$ changes from continuous to discrete (at $Y = 0$ and $Y = 1$). Now the value of $y$ is found from the following equation, where $G_1$ is used to denote the polynomial fit to the cdf of $X$ and $G_1^{-1}$ is used to denote the polynomial fit to the inverse cdf of $y$:

$$
y = \begin{cases} 
0, & \text{if } F(x) < P_0 \\
G_1^{-1}(F_1(x)), & \text{if } P_0 < F(x) < P_1 \\
1, & \text{if } P_1 < F(x)
\end{cases} \quad (6)
$$

Figure 5 shows a plot of $\hat{y}$ as a function of $x$. Figure 6 is a histogram of the $y$'s, and a comparison with Figure 3 shows the similarity of the two densities. Note that the
density of $y$ is skewed slightly to the right, because of ties among the $x$'s which are caused by rounding to three decimal places.

**FIGURE 5.— PREDICTED CLOUD COVER AS A FUNCTION OF $X$**

![Figure 5](image)

**FIGURE 6.— DISTRIBUTION OF PREDICTED CLOUD COVER**

![Figure 6](image)

SUMMARY

The method of percentile matching not only proves to be straightforward but also has the advantage of not requiring the assumption of a model for the relation between $X$ and $Y$. In cases where the model form is known, and where the distribution of the deviations from this model is known to be Gaussian, the method of least squares is always preferable for predicting $Y$. If the appropriate model form is unknown or if one is concerned with preserving the distribution of $Y$, then the method of percentile matching may be more appropriate, provided that $E(Y|X)$ is monotone. As an example of the effectiveness of the proposed method, consider the Spearman rank correlation coefficient (calculated using the PROC CORR procedure) between the predicted and observed cloud proportions. The cubic model discussed earlier gives a coefficient of .79834, while the predictor derived in the example above gives a coefficient of .79873. Thus, our predictor is as good as at least one reasonable least squares model, and it has the desirable property of preserving the distribution of the dependent variable.

Additional information can be obtained from James H. Smith (Mail Code C09), Lockheed Engineering and Management Services Company, Inc., 1830 NASA Road 1, Houston, Texas 77258.

ENDNOTES

1 We adopt the convention of using uppercase letters to denote random variables and lowercase letters to denote observations of these random variables. References to the distribution of a lowercase named variable indicate the empirical distribution.

2 AgrISTARS is a multiyear program of research, development, evaluation, and application of aerospace remote sensing for agricultural resources, which began in fiscal year 1980. This program is a cooperative effort of the U.S. Department of Agriculture, the National Aeronautics and Space Administration, the National Oceanic and Atmospheric Administration (U.S. Department of Commerce), the Agency for International Development (U.S. Department of State), and the U.S. Department of the Interior. The work which is the subject of this document was performed by the Earth Resources Applications Division, Space and Life Sciences Directorate, Lyndon B. Johnson Space Center, National Aeronautics and Space Administration and Lockheed Engineering and Management Services Company, Inc. The tasks performed by Lockheed Engineering and Management Services Company, Inc., were accomplished under Contract NAS 9-18800.

3 Courtesy of the NOAA Operational Heat Budget Archive, NOAA Environmental Data Information Services, Washington, D.C.

REFERENCES


