Department of Statistics
University of Wisconsin
1210 W. Dayton St.
Madision, Wisconsin

Technical Report 686

July, 1982

Multivariate Thin Plate Spline Estimates for the Posterior Probabilities in the Classification Problem

Miguel A. Villalobos [1]

University of Wisconsin

Grace Wahba [2]

University of Wisconsin

[1] Research supported in part by Consejo Nacional de Ciencia y Tecnologia, Mexico and in part by the National Cancer Institute grant No. 5-R01-CA18332-07.
[2] Research supported by Army Research Office grant No. DAAG29-80-K0042.

Multivariate Thin Plate Spline Estimates for the Posterior Probabilities in the Classification Problem

Miguel A. Villalobos

Grace Wahba

Abstract

A nonparametric estimate for the posterior probabilities in the classification problem using multivariate thin plate splines is proposed. This method presents a nonparametric alternative to logistic discrimination as well as to survival curve estimation. The degree of smoothnes of the estimate is determined from the data using generalized crossvalidation.

Key Words: Estimation of posterior probabilities; Discrimination; Maximum penalized log-likelihood; Thin plate spline; Generalized cross validation.

1. INTRODUCTION

Consider k populations $A_1,\ldots,A_k$ and a d-dimensional random vector $X=(X^1,\ldots,X^d)$. Assume that the probability distribution of X given that it comes from population A_j, j=1,...,k is absolutely continuous with respect to Lebesgue measure and let $f_j(x)$ denote the corresponding probability density function for j=1,...,k.

Suppose that a training sample $X_{ij}=x_{ij}$, $i=1,\ldots,n_j$, from the population A_j is available for each $j=1,\ldots,k$. Given these training samples and the prior probabilities q_j, $j=1,\ldots,k$ where $0 < q_j < 1$ for $j=1,\ldots,k$ and

$$\sum_{j=1}^{k} q_j = 1$$

we want to estimate the posterior probabilities

$$p_j(x) = q_j f_j(x) / \sum_{i=1}^{k} q_i f_i(x) = P(A_j \mid X=x) \quad j=1,\ldots,k.$$

The estimates of these posterior probabilities have a clear application in Bayes discriminant analysis and we believe that they will be useful for exploring properties of the data and for presenting the data in a way comprehensible to the layman.

In this paper we propose a class of optimization methods for estimating $p_1(x),\ldots,p_k(x)$, and for simplicity of notation we will consider the case where we have only two populations since the extension for more than two is straight-forward.

There is a large amount of literature in the area of discriminant analysis. Since we are proposing here a nonparametric method, in §2 we will present a brief review of the work in nonparametric discriminant analysis that is most closely related to our approach. In §3 we estimate the log likelihood ratio using thin plate splines. In §4 we propose a solution to the problem of estimating the posterior probabilities using smoothing thin plate splines. The results appear to provide a reasonable nonparametric alternative to logistic regression. In §5 we present some results of simulations in one and two dimensions and an application of the method to some real data.

Finally, in §6 we discuss some of the future work that needs to be done.

2. PREVIOUS WORK IN NONPARAMETRIC DISCRIMINANT ANALYSIS

Most of the work in discriminant analysis (for continuous variables) is based on Normality assumptions, usually with equal covariance matrices. For a summary of the work in discriminant analysis see Lachenbruch and Goldstein (1979). Here we will only be concerned with nonparametric discriminant analysis.

Fix and Hodges (1951) are, to our knowledge, the first to consider the nonparametric classification problem using a k-nearest neighbor approach. For further references related to this paper see Lachenbruch and Goldstein (1979).

During the last 10 years there has been a development of classification rules based on density estimates. These kind of rules are important because of the extensive research done in nonparametric density estimation. Another feature that makes these kind of methods attractive is a result by Glick (1972) that says that an estimate of the non-error rate of an arbitrary rule based on parametric or nonparametric density estimators is, in some sense asymptotically optimal provided that:

$$\hat{q}_i \hat{f}_i(x) \xrightarrow{p} q_i f_i(x)$$

pointwise for almost all x in R^d, i=1,...,k, and

$$\int_{R^d} \sum_{i=1}^{k} \hat{q}_i \hat{f}_i \xrightarrow{p} 1.$$

Kernel, maximum penalized likelihood and orthogonal series density estimates are among the most popular methods. All these density estimation methods involve the choice of a parameter that controls the degree of smoothing of the estimate. Several methods have been proposed to choose the smoothing parameter, among these there are three which are readily computable and objective. Two of these methods were suggested by Wahba (1977 and 1981a) and the third by Habbema, Hermans and Van den

Broek (1974). In this last paper the authors estimate the densities for each population using a kernel estimate. A complete description of kernel methods can be found in Tapia and Thompson (1978).

The kernel estimate used in Habbema, Hermans and Van den Broek (1974) is of the form:

$$f_j(x)=(n_j \sigma_j^d)^{-1} \sum_{i=1}^{n_j} K\left\{\frac{(x-x_{ij})}{\sigma_j}\right\} \qquad (2.1)$$

for j=1,...,k, where, as before d is the dimension of the vector x and k is the number of populations. K is a multivariate normal kernel and the smoothing parameters σ_j are estimated by maximizing what might be called the "cross-validation likelihood function":

$$V(\sigma_j)=\prod_{i=1}^{n_j} f_j^{[k]}(x_{ji})$$

where $f_j^{[k]}$ is an estimate of f_j computed as in (2.1) but leaving out the point x_{jk}. For a detailed description of the algorithm to carry out this kernel discriminant analysis see Hermans and Habbema (1976).

Hermans and Habbema (1975) compare five methods for estimating posterior probabilities using some medical data for which the true posterior probability function is unknown. Four of these five methods are parametric and the fifth one is the kernel method described above. The four parametric methods involve:

(1) Multinormal distributions, equal covariance matrices, estimated parameters.

(2) Multinormal distributions, equal covariance matrices, bayesian or predictive approach.

(3) Multinormal distributions, unequal covariance matrices, estimated parameters.

(4) Multinormal distributions, unequal covariance matrices, bayesian or predictive approach.

The nonparametric method is:

(5) Direct estimation of the density functions using a kernel method.

Later, Remme, Habbema and Hermans (1980) carry out a simulation study to compare the performances of methods 1,3 and 5 above. Their simulations show that the performance of the kernel method was either better or as good as the performance of other methods, except in the simulations with multinormal distributions with equal covariance matrices. It performed increasingly well with increasing sample sizes, however, the improvement was very slow in samples simulated from lognormal distributions.

Another nonparametric classification method is given by Chi and Van Ryzin (1977). Their procedure is based upon the idea of a histogram density estimator but bypasses the direct density estimation calculations.

In most of the references listed above the approach has been to estimate each density separately and from this form an estimate of the posterior probabilities. By the Neyman-Pearson lemma, we know that if we want to classify an object as coming from one of two populations with densities f_1 and f_2, we should base the classification on the likelihood ratio f_1/f_2 and hence it would be attractive to have a method to estimate the likelihood ratio directly. Silverman (1978) considers the direct estimation of the log likelihood ratio for one dimensional data. He assumes that $h=log(f_1/f_2)$ is in $C_2(I)$, where I is some interval containing all the observations. He finds the conditional log-likelihood of h and penalizes it according to the smoothness of h using $\int_I (h'')^2$ as the smoothing penalty functional. He estimates h by maximizing the penalized log likelihood and shows that the estimate is a cubic spline. However, he does not give a data-based method to choose the smoothing parameter.

In §3 we extend the result in Silverman (1978) to the d-dimensional case and show how the search for a data-based method to choose the smoothing parameter motivated us to consider the problem of estimating the posterior probabilities directly instead of

the likelihood ratio.

Anderson and Blair (1982) introduce penalized maximum likelihood estimates in the context of logistic regression and discrimination. They obtain estimates of the logistic parameters and a nonparametric spline estimate of the marginal distribution of the regressor x.

3. MAXIMUM PENALIZED LOG LIKELIHOOD ESTIMATION

In this section we extend the result given by Silverman (1978) to the d-dimensional case and describe a natural generalization to d dimensions of the one dimensional penalty functional $\int \left[f^{(m)}(x)\right]^2 dx$.

Let $Y_1,\ldots,Y_n$, $n=n_1+n_2$ denote the combined samples from the two populations and define

$$Z_i = \begin{cases} 1 & \text{if } Y_i \in A_1 \\ 0 & \text{if } Y_i \in A_2. \end{cases} \tag{3.1}$$

As before, let q_1 and q_2 be the prior probabilities and consider the estimation of

$$h = \log(q_1 f_1 / q_2 f_2)$$

Following Silverman (1978) it can be shown that the conditional log likelihood of h is given by

$$ln(h) = \sum_{i=1}^{n} \left[z_i h(y_i) - \log\left\{1 + exp(h(y_i))\right\}\right] \tag{3.2}$$

where z_i and y_i are the observed values of Z_i and Y_i, so that a "maximum likelihood" estimator of h would be obtained by maximizing (3.2). To avoid the undesirable solution:

$$h(y_i) = \begin{cases} \infty & \text{if } Z_i = 1 \\ -\infty & \text{if } Z_i = 0 \end{cases}$$

we should use the underlying assumption that h is, in some sense, not too rough. Therefore we must penalize the likelihood according to the roughness of h.

We will assume that h is in a reproducing kernel Hilbert space H of real valued "smooth" functions which map $\mathbf{R}^d$ into R. More precisely, (see Wahba and Wendelberger, 1980), H is the space of all Schwartz distributions for which all the partial derivatives in the distributional sense, of total order m, are square integrable. A "maximum penalized log likelihood" estimator of h, is the function $h_{n\lambda} \in H$ that minimizes

$$L(h) = -ln(h) + \lambda J_m(h) \tag{3.3}$$

where J_m is given by

$$J_m(h) = \sum_{|\alpha|=m} \left[\frac{m!}{\alpha_1! \cdots \alpha_d!} \right] \|D^\alpha h\|^2_{L_2(\mathbf{R}^d)} \tag{3.4}$$

where, $\alpha=(\alpha_1, \ldots, \alpha_d)$, $|\alpha| = \sum_{j=1}^{d} \alpha_j$ and $D^\alpha = \prod_{j=1}^{d} \frac{\partial^{\alpha_j}}{(\partial x^j)^{\alpha_j}}$. For the special case d=2, m=2,

$$J_2(h) = \int\int \left\{ \left[\frac{\partial^2 h}{(\partial x^1)^2} \right]^2 + 2 \left[\frac{\partial^2 h}{(\partial x^1)(\partial x^2)} \right]^2 + \left[\frac{\partial^2 h}{(\partial x^2)^2} \right]^2 \right\} dx^1 dx^2$$

In Theorems 3.1 and 3.2 below we establish conditions for the existence and uniqueness of the minimizer of (3.3) and characterize the solution as a thin plate spline (polynomial spline when d=1). Because the proofs of these theorems are lengthy they will not be given, but will appear in Villalobos (1982). The proof of Theorem 3.2 follows Wahba and Wendelberger(1980).

Theorem 3.1: The minimizer of (3.3) in H exists and is unique provided there is no level curve of a polynomial of degree less than m that completely separates the samples from the two populations.

Silverman (personal communication) has previously conjectured Theorem 3.1 and has also noted a rather elegant property of the minimizer of (3.3) as $\lambda \to \infty$. As $\lambda \to \infty$, $h_{n\lambda}$ tends to an element of the null space of J_m, so that for m=2 the estimated log likelihood ratio will be linear and for m=3 it will be quadratic in x. Thus the parametric

estimate for multivariate normal densities is included as a limiting case. (Compare Wahba, 1978).

Theorem 3.2: The minimizer $h_{n\lambda}$ of (3.3) in H, if it exists, is of the form:

$$h_{n\lambda} = \sum_{i=1}^{n} c_i E_m(y, y_i) + \sum_{j=1}^{M} d_j \varphi_j(y)$$

where the function E_m is given by

$$E_m(s, t) = E(|s - t|)$$

where

$$E(|\tau|) = \begin{cases} \vartheta_m |\tau|^{2m-d} \ln|\tau|, & d \text{ even} \\ \vartheta_m |\tau|^{2m-d}, & d \text{ odd} \end{cases}$$

and

$$\vartheta_m = \begin{cases} \dfrac{(-1)^{d/2+1+m}}{2^{2m-1}\pi^{d/2}(m-1)!(m-d/2)!}, & d \text{ even} \\ \dfrac{\Gamma(d/2-m)}{2^{2m}\pi^{d/2}(m-1)!}, & d \text{ odd.} \end{cases}$$

Here $M = \binom{m+d-1}{d}$ is the dimension of the space of polynomials of degree less than m in d-dimensional space, and $\{\varphi_1, \ldots, \varphi_M\}$ span this space. For example, if d=2, m=3, then M=6 and $\varphi_1(x^1,x^2) = 1$, $\varphi_2(x^1,x^2) = x^1$, $\varphi_3(x^1,x^2) = x^2$, $\varphi_4(x^1,x^2) = x^1x^2$, $\varphi_5(x^1,x^2) = (x^1)^2$ and $\varphi_6(x^1,x^2) = (x^2)^2$.

The vectors $c=(c_1,\ldots,c_n)'$ and $d=(d_1,\ldots,d_M)'$ are the solution to the optimization problem:

minimize

$$\sum_{i=1}^{n} \left[\log\left\{1 + \exp\left[\sum_{j=1}^{n} c_j E_m(y_i, y_j) + \sum_{j=1}^{M} d_j \varphi_j(y_i)\right]\right\} - z_i\left\{\sum_{j=1}^{n} c_j E_m(y_i, y_j) + \sum_{j=1}^{M} d_j \varphi_j(y_i)\right\}\right] + \lambda c'Kc$$

subject to T'c = 0, where K is the nxn matrix with (i,j) entry $E_m(y_i, y_j)$ and T is the nxM

matrix with (i,j) entry $\varphi_j(y_i)$.

We computed the estimate using four one dimensional simulated samples and in some cases the estimation turned out to be very expensive. There is still the problem of choosing the value of the smoothing parameter. Since the conditional distribution of Z_i given $Y_i=y_i$ is $binomial(1,p_1)$, where

$$p_1=P(Z_i=1 \mid Y_i=y_i)$$

then

$$E\left[Z_i \mid Y_i=y_i\right] = \frac{\exp\left\{h(y_i)\right\}}{1+\exp\left\{h(y_i)\right\}}.$$

so that an (ordinary) cross validation (or "leaving out one") estimate of λ would be the value that minimizes

$$\frac{1}{n}\sum_{k=1}^{n}\left[\frac{\exp\left\{h_{n\lambda}^{[k]}(y_k)\right\}}{1+\exp\left\{h_{n\lambda}^{[k]}(y_k)\right\}} - z_k\right]^2 \qquad (3.5)$$

where $h_{n\lambda}^{[k]}$ is the estimate of h given λ, obtained by leaving out the k^{th} observation. Obviously this would be prohibitive to compute. But expression (3.5) suggests a different way of approaching the problem, in fact, this is what motivated us to consider the direct estimation of the posterior probability instead of the likelihood ratio.

4. SPLINE ESTIMATE FOR THE POSTERIOR PROBABILITY

In this section we propose a class of optimization methods to estimate the posterior probabilities p_1 and p_2. Since $p_1+p_2=1$ we will only estimate $p=p_1$ and the estimate for p_2 will be

$$\hat{p}_2 = 1-\hat{p} = 1-\hat{p}_1.$$

Let

$$w_j = n_j/n \quad j=1,2$$

where $n=n_1+n_2$ and let $Z_j, j=1,\ldots,n$ be defined as in (3.1). We propose to estimate

$$h(y) = h_1(y) = \frac{w_1 f_1(y)}{\sum_{j=1}^{2} w_j f_j(y)}$$

We will assume that either h is in the reproducing kernel Hilbert space H or can be well approximated by some function in H.

If $\hat{h}$ is an estimate of h then we can obtain an estimate of $p=p_1$ by

$$\hat{p}(y) = \frac{(q_1/w_1)\hat{h}(y)}{\sum_{j=1}^{2}(q_j/w_j)\hat{h}_j(y)}$$

where $\hat{h}_1=\hat{h}$ and $\hat{h}_2=1-\hat{h}$.

We can think of the vector $Z=(Z_1,\ldots,Z_n)'$ of zeroes and ones as noisy observations on the values $h(y_1),\ldots,h(y_n)$. To see this note that, if we draw an observation Y from the density f_j with probability w_j, $j=1,2$, and Z is the random variable which is 1 or 0 according as j is 1 or 2, then

$$E(Z \mid Y=y) = h(y).$$

Finally, following the approach of Wahba (1980 and 1981b), we suggest estimating h by minimizing:

$$\frac{1}{n}\sum_{i=1}^{n}\left\{h(y_i) - z_i\right\}^2 + \lambda J_m(h) \qquad (4.1)$$

subject to

$$0 \le h(s_i) \le 1, \quad i=1,\ldots,L \tag{4.2}$$

where J_m is given by (3.4), and $s_1,\ldots,s_L$ is a fine regular grid of points in $\mathbf{R}^d$, chosen so that a smooth function which satisfies all the constraints at this points will appear to satisfy them over all of S, where S is a subset of $\mathbf{R}^d$ such that

$$\sum_{j=1}^{2} w_j f_j(y) > \varepsilon > 0$$

for every $y \in S$.

The smoothing parameter lambda will be chosen by the method of generalized cross-validation for constrained problems as described by Wahba (1980,1981b)

It can be shown that the quadratic form (4.1) is strongly convex for any z provided that the n by M matrix T_1 with (i,j) entry $\varphi_j(y_i)$ is of rank M. Then (4.1) will have a minimizer in any closed convex set in H, in particular, the minimizer of (4.1) exists in the set:

$$C_L = \left\{ h \in H: 0 \le h(s_i) \le 1, \ i=1,\ldots,L \right\}.$$

Note that to extend this to more than two populations we also need to enforce the constraint that

$$\sum_{j=1}^{k} h_j = 1.$$

Following Kimeldorf and Wahba (1971) and Wahba and Wendelberger (1980) the minimizer of (4.1) can be shown to be of the form:

$$h_{n\lambda}(y) = \sum_{i=1}^{n} c_i E_m(y,y_i) + \sum_{k=1}^{L} b_k E_m(y,s_k) + \sum_{j=1}^{M} d_j \varphi_j(y)$$

where the vectors $c=(c_1,\ldots,c_n)'$, $b=(b_1,\ldots,b_L)'$ and $d=(d_1,\ldots,d_M)'$ are determined as the solution to the following optimization problem:

minimize

$$\frac{1}{n}\|z - K\begin{bmatrix}c\\b\end{bmatrix} - T_1 d\|_n^2 + \lambda[c' : b'] \begin{bmatrix}K_{11} & K_{12}\\ K_{21} & K_{22}\end{bmatrix}\begin{bmatrix}c\\b\end{bmatrix} \tag{4.3}$$

subject to

$$\begin{bmatrix}K_{21} & K_{22} & T_2\\ -K_{21} & -K_{22} & -T_2\end{bmatrix}\begin{bmatrix}c\\b\\d\end{bmatrix} \le \begin{bmatrix}1\\0\end{bmatrix} \tag{4.4}$$

and

$$T\begin{bmatrix}c\\b\end{bmatrix} = 0 \tag{4.5}$$

where $K=\left[K_{11}:K_{12}\right]$, $T = \left[T_1' : T_2'\right]$ and the description of K_{11}, K_{12}, K_{22}, T_1 and T_2 is given in table 4.1.

In order to be able to use a quadratic programming routine to solve the problem for a fixed value of λ we should eliminate the equality constraint (4.5). To do this we want to find a matrix S of dimension n+L by n+L-M such that for some n+L-M dimensional vector e we have

$$\begin{bmatrix}c\\b\end{bmatrix} = Se.$$

To do this consider the Q-R decomposition of T'. There exist matrices $Q_1\ (n+L \times M)$ and $Q_2\ (n+L \times n+L-M)$ such that

$$\begin{bmatrix}Q_1'\\Q_2'\end{bmatrix} T = \begin{bmatrix}R\\0\end{bmatrix} \tag{4.6}$$

where R is an M by M upper triangular matrix. Now, condition (4.5) implies that e'S'T=0

Table 4.1

Matrix	Dimension	(i,j)element	
K_{11}	n x n	$E_m(y_i,y_j)$	i=1,...,n j=1,...,n
K_{12}	n x L	$E_m(y_i,s_j)$	i=1,...,n j=1,...,L
K_{22}	L x L	$E_m(s_i,s_j)$	i=1,...,L j=1,...,L
T_1	n x M	$\varphi_j(y_i)$	i=1,...,n j=1,...,M
T_2	L x M	$\varphi_j(s_i)$	i=1,...,L j=1,...,M

so that by (4.6) we should take $S=Q_2$, then minimizing (4.3) subject to (4.4) and (4.5) is equivalent to minimizing

$$(1/n) \parallel z - KSe - T_1 d \parallel^2 + e'SKSe$$

subject to

$$\begin{bmatrix} [K_{21} \mid K_{22}]S & T_2 \\ -[K_{21} \mid K_{22}]S & -T_2 \end{bmatrix} \begin{bmatrix} e \\ d \end{bmatrix} \leq \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

A program to compute this estimate choosing the smoothing parameter by generalized cross-validation is not available yet. However, a program to find $h_{n\lambda}$ to minimize

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\left(h(y_i) - z_i\right)^2}{\sigma_i^2} + \lambda J_m(h), \tag{4.7}$$

where σ_i's i=1,...,n are specified weights, and choose λ by generalized cross validation has been developed by J. Wendelberger (1981) and is available through the Madison Academic Computing Center (1981). In the next section we use this program to give some examples without the constraints (4.2).

5. EXAMPLES

Two spline estimates are presented. The first estimate (E.V., for equal variances), is the minimizer $\hat{h}$ of (4.7) with $\sigma_i^2 = 1$, and $\lambda=\hat{\lambda}$, the generalized cross-validation estimate of λ. The second estimate (U.V., for unequal variances), is the minimizer $\tilde{h}$ of (4.7) where the weights σ_i^2 are estimates of the variances of the Z_i's. These variance estimates are

$$\sigma_i^2 = Var\left[Z_i \mid x_i\right] = \hat{h}(x_i)\left[1-\hat{h}(x_i)\right], \quad i=1,\ldots,n.$$

In the U. V. estimate of h, the generalized cross-validation estimate of λ is obtained by viewing z_i/σ_i as the data (more details may be found in Wendelberger, 1981). Assuming the usual notation for the univariate Normal distribution and using $N_2(\mu_1,\mu_2,\sigma_{11},\sigma_{22},\sigma_{12})$ for the bivariate Normal, six simulated samples were generated as

shown in table 5.1.

In figures 1-6 we give the data, the true h and the spline (E.V.) estimates. For samples 1-3 the Normal theory h is also given and for samples 1,4,5 and 6 the spline (U.V.) estimates are shown. In each of these examples m was taken as 2. The normal theory estimates were obtained using the population sample means and variances.

In figure 3, the constraint $h(x) \leq 1$ is grossly violated. However, if one imagines the spline (E.V.) curve as a thin rod and one pushes down on it so that the constraint is satisfied, one can imagine that a rather good estimate will result when the constraints are imposed. Although the variances of the Z_i's are actually unequal it is apparent from the plots shown that we do not get a significant improvement by estimating the variances first. In fact, in some cases, for example, for samples 1 and 6 the weighted estimate looks worse than the unweighted one. This might be because in weighting the observations, we are forcing the spline (U.V.) estimate to be very close to the data where h is very close to zero or one, and this happens at regions where the observations are very sparse and far from their corresponding groups.

However, judging from the plots, the estimate using equal weights does a decent job in most cases. It is close to the true function in the regions where the actual posterior probabilities are close to 0.5. It does not do well (nor is it expected to) in the regions where there is almost no data.

Table 5.1

Sample	f_1	f_2	n_1	n_2
1	$N(0,1)$	$N(1.5,1)$	60	72
2	$.5N(-1,1)+.5N(1,1)$	$N(0,1)$	60	72
3	$.4N(0,0.25)+.6Cauchy$	$N(0,0.25)$	60	72
4	$N_2(0,0,0.25,0.25,0)$	$N_2(1,1,0.25,0.25,0)$	80	80
5	$N_2(0,0,0.25,0.25,0.01)$	$N_2(1.5,1.5,0.5,0.5,0.25)$	80	80
6	$.6N_2(-1,-1,0.25,0.25,0)$ $+.4N_2(1,1,1,0.5,0.5)$	$N_2(0,0,0,0.25,0.25,0)$	175	175

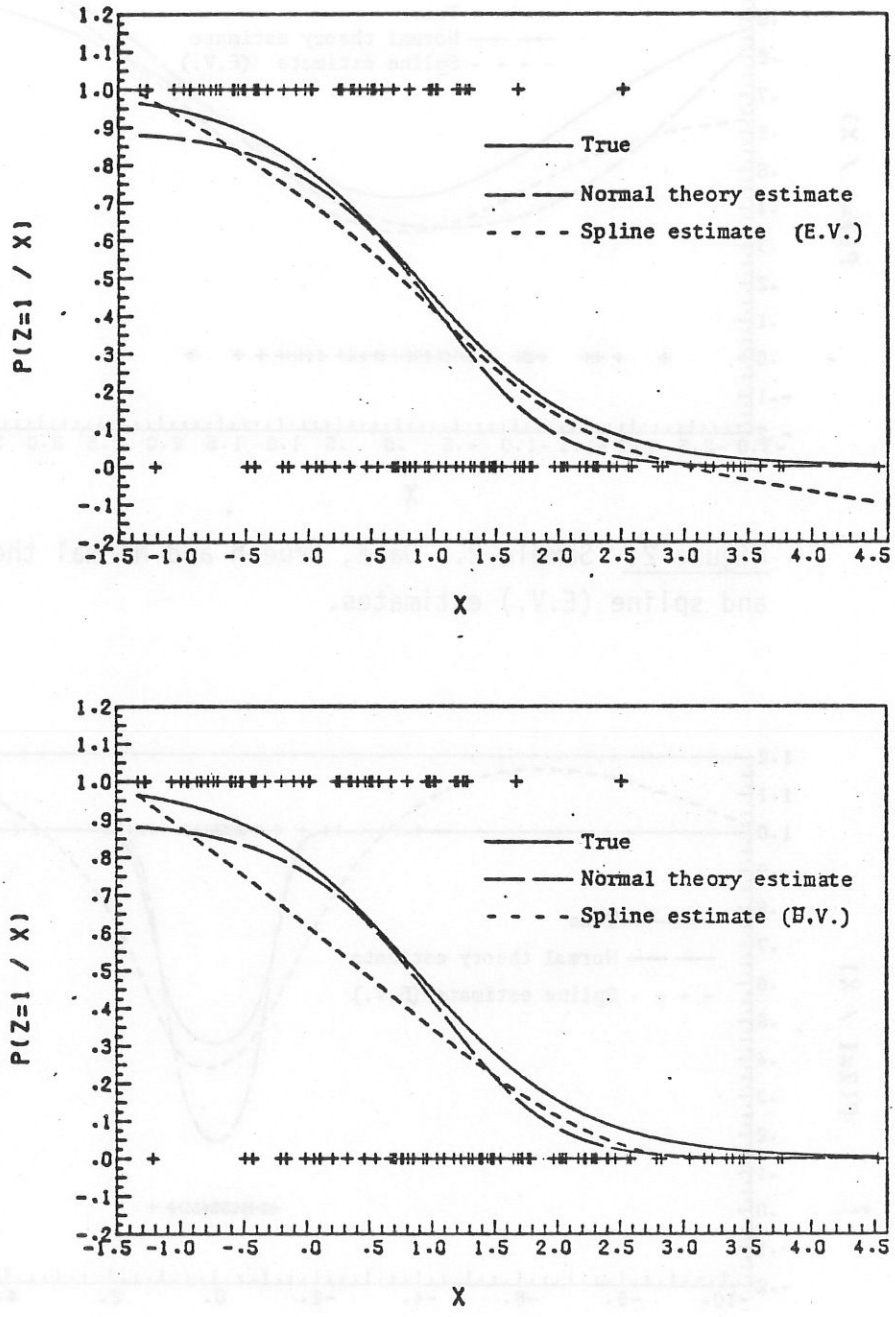

Figure 1: Sample 1. Data, True h and Normal theory, spline (E.V.) and spline (U.V.) estimates.

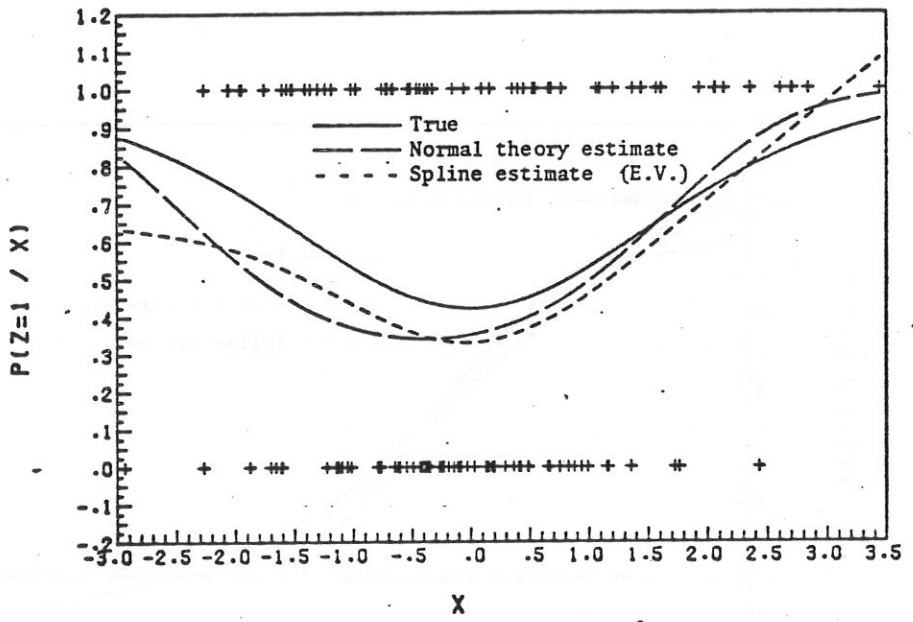

Figure 2: Sample 2. Data, true h and Normal theory and spline (E.V.) estimates.

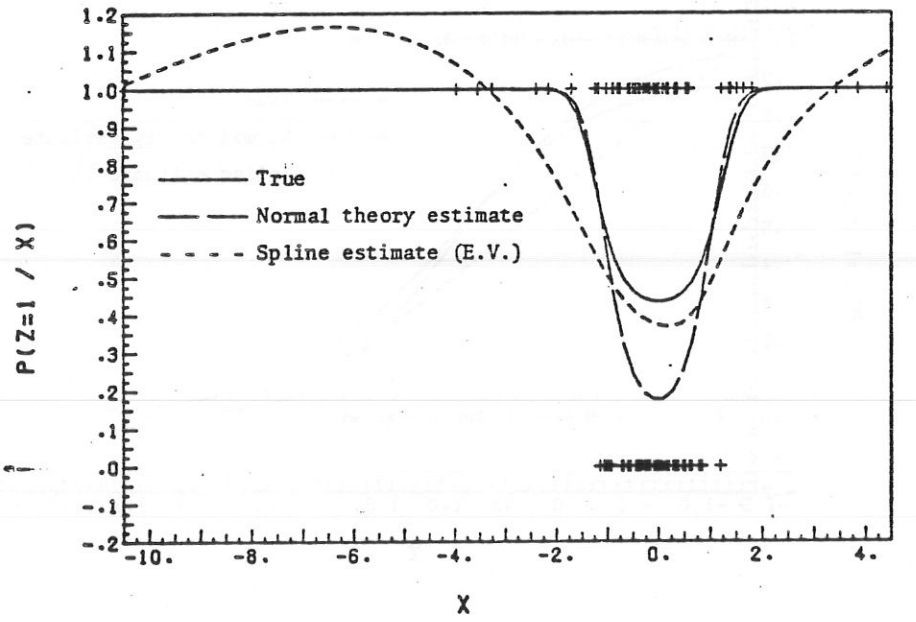

Figure 3: Sample 3. Data, true h and Normal theory and spline (E.V.) estimates.

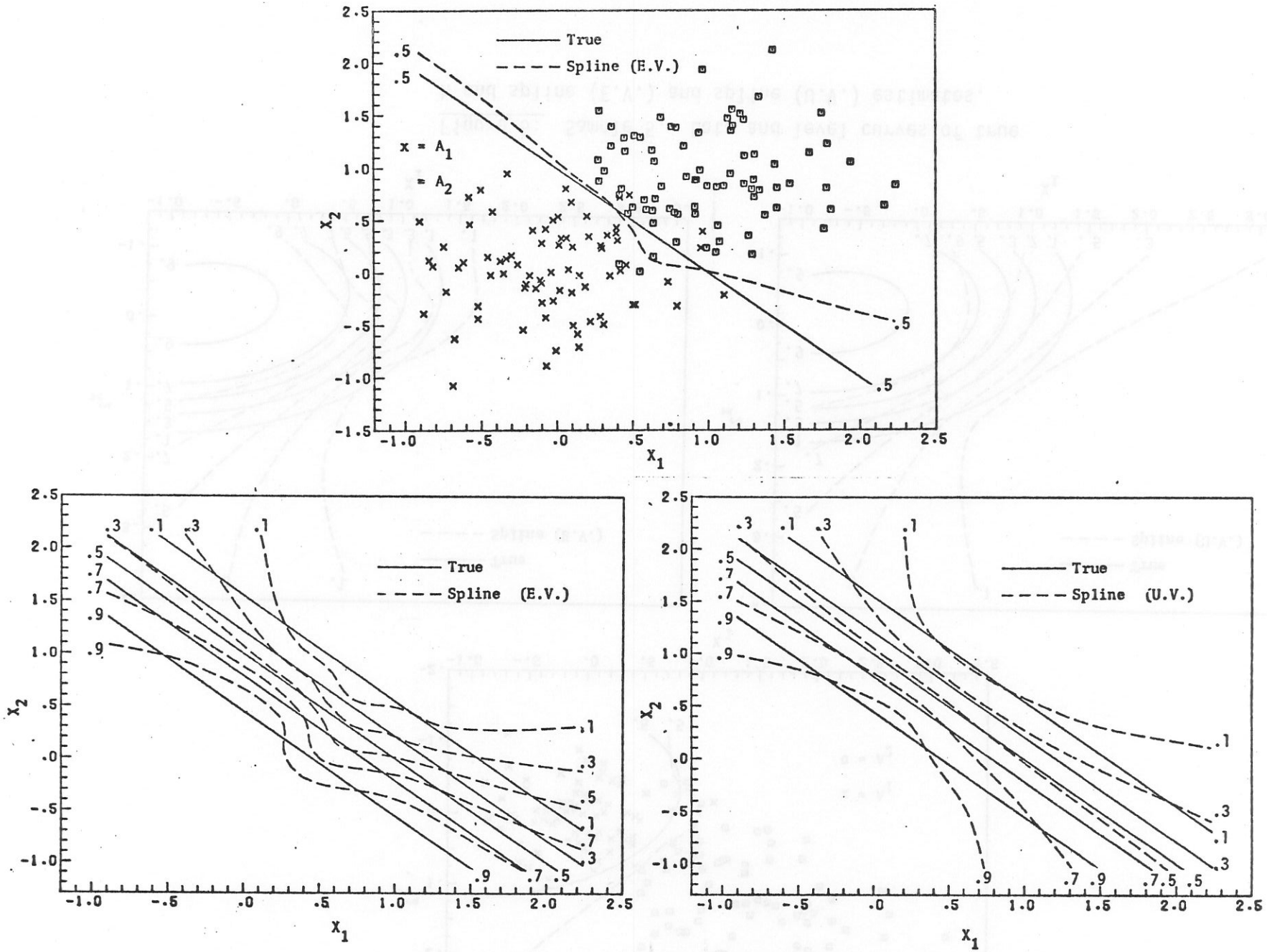

Figure 4: Sample 4. Data and level curves of true h and spline (E.V.) and spline (U.V.) estimates.

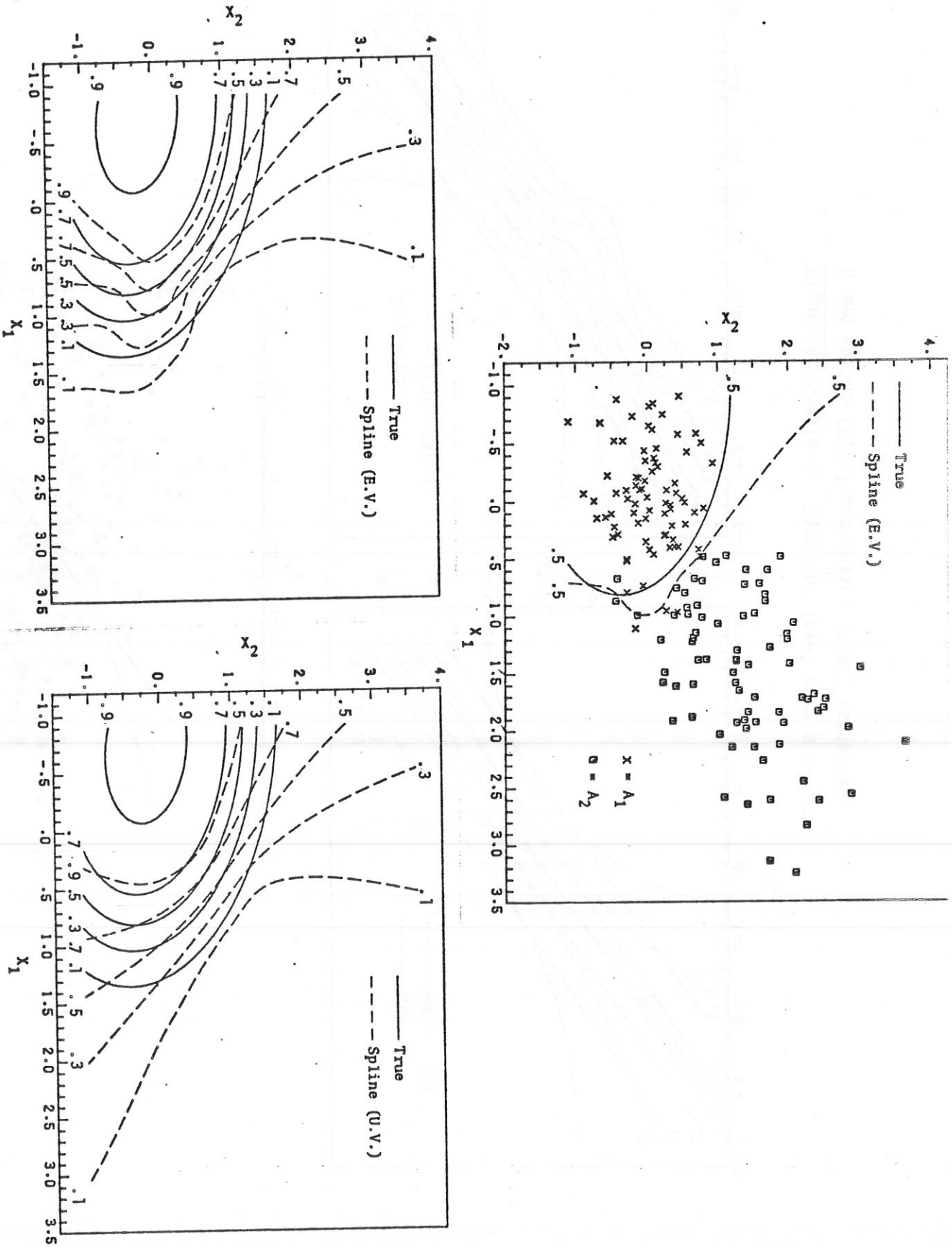

Figure 5: Sample 5. Data and level curves of true h and spline (E.V.) and spline (U.V.) estimates.

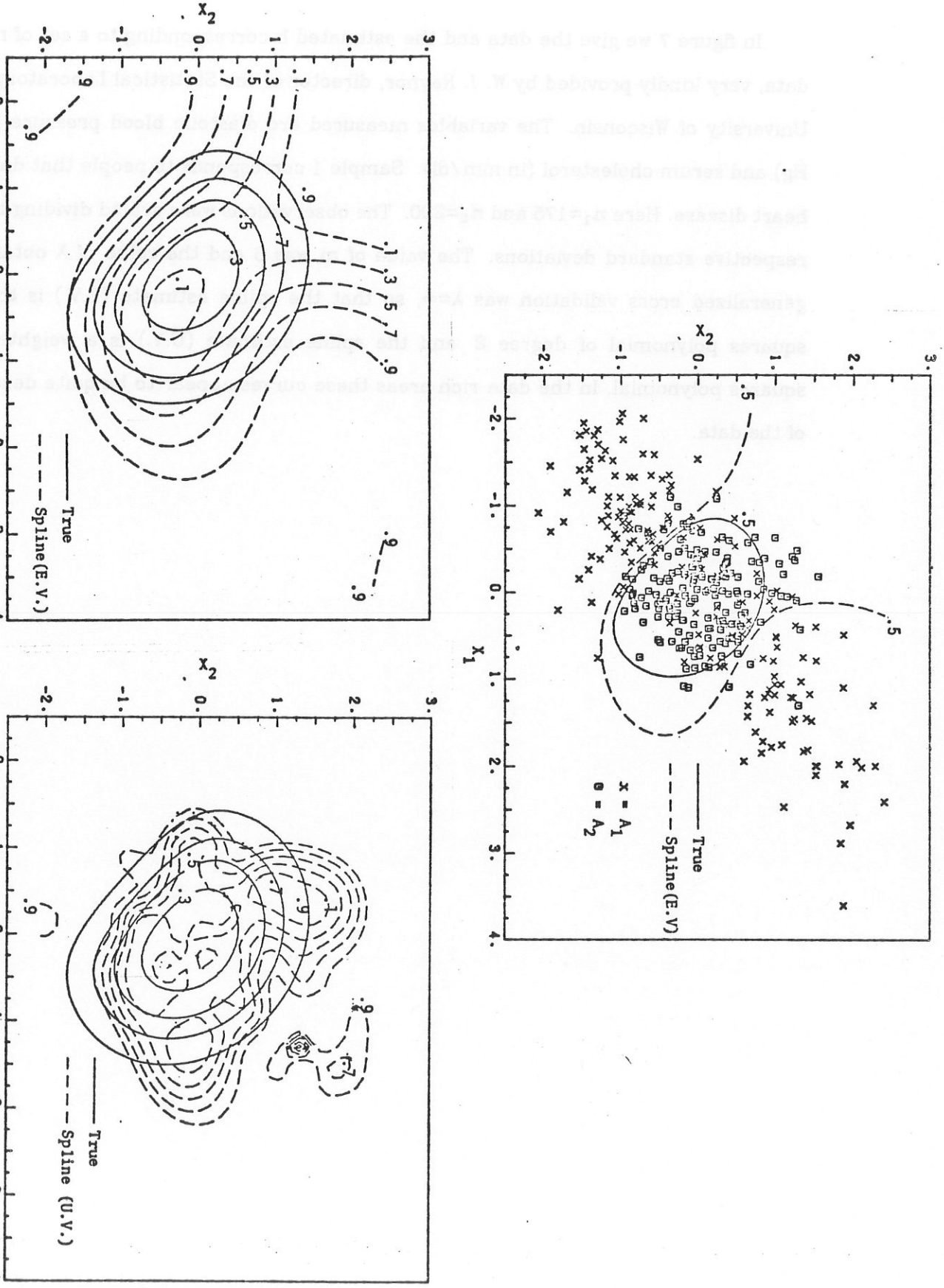

Figure 6: Sample 6. Data and level curves of true h and spline (E.V.) and spline (U.V.) estimates.

In figure 7 we give the data and the estimated h corresponding to a set of medical data, very kindly provided by W. J. Raynor, director of the Statistical Laboratory at the University of Wisconsin. The variables measured are diastolic blood pressure (in mm Hg) and serum cholesterol (in mm/dl). Sample 1 corresponds to people that died from heart disease. Here n_1=175 and n_2=200. The observations were scaled dividing by their respective standard deviations. The value of m was 3 and the value of λ obtained by generalized cross validation was $\lambda=\infty$, so that the spline estimate (E.V.) is the least squares polynomial of degree 2 and the spline estimate (U.V.) is a weighted least squares polynomial. In the data rich areas these curves appear to be quite descriptive of the data.

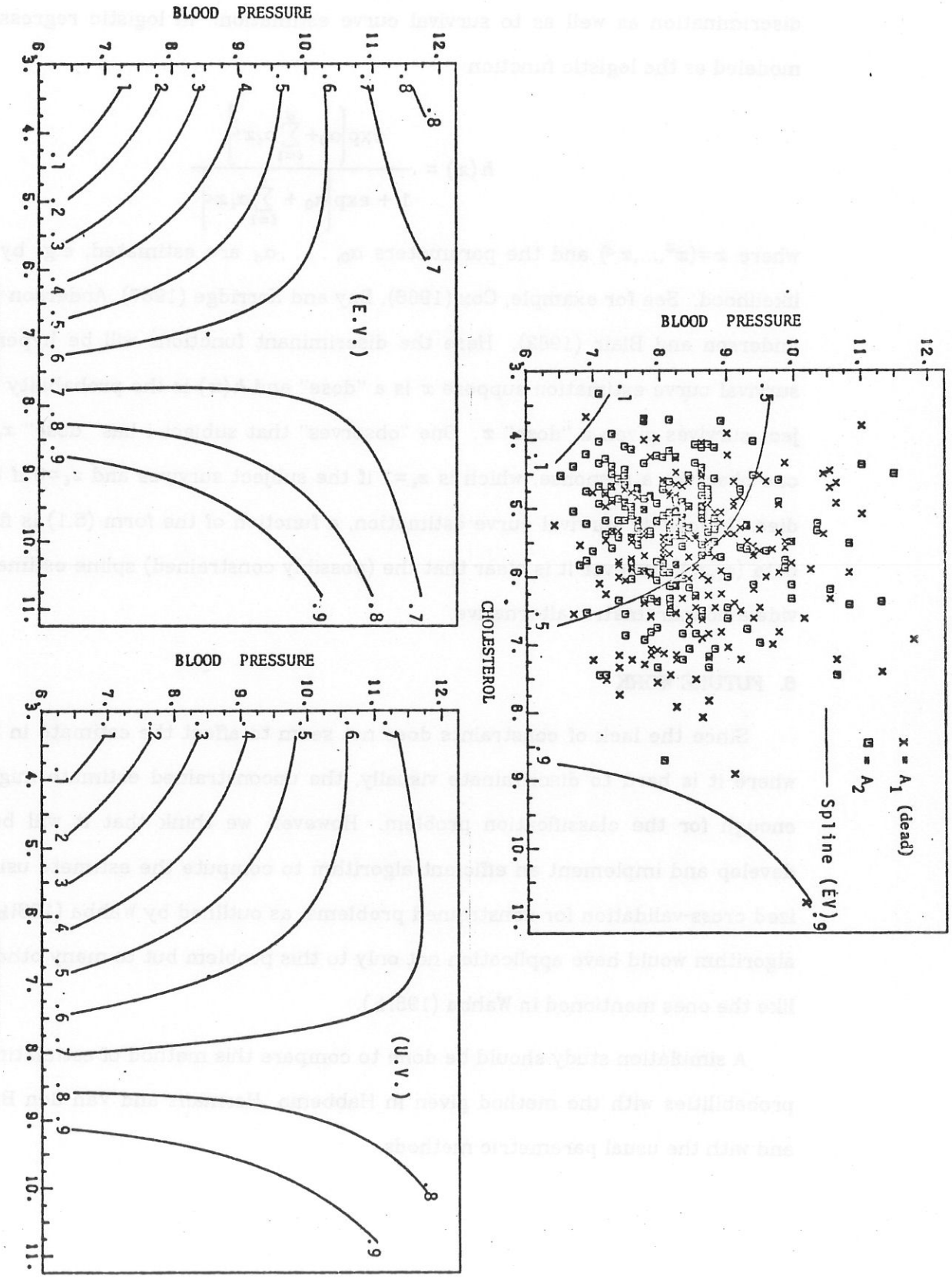

Figure 7: Sample 7. Data and level curves of spline (E.V.) and spline (U.V.) estimates.

It can be seen that this method presents a nonparametric alternative to logistic discrimination as well as to survival curve estimation. In logistic regression h(x) is modeled as the logistic function

$$h(x) = \frac{\exp\left[\alpha_0 + \sum_{i=1}^{d} \alpha_i x^i\right]}{1 + \exp\left[\alpha_0 + \sum_{i=1}^{d} \alpha_i x^i\right]} \tag{5.1}$$

where $x=(x^1,\ldots,x^d)$ and the parameters $\alpha_0, \ldots, \alpha_d$ are estimated, e.g. by maximum likelihood. See for example, Cox (1966), Day and Kerridge (1967), Anderson (1972), and Anderson and Blair (1982). Here the discriminant functions will be hyperplanes. In survival curve estimation suppose x is a "dose" and $h(x)$ is the probability that a subject survives given a "dose" x. One "observes" that subject i has "dose" x_i, and then one observes a response, which is $z_i=1$ if the subject survives and $z_i=0$ if the subject dies. In logistic survival curve estimation, a function of the form (5.1) is fitted to the data (z_i, x_i), however it is clear that the (possibly constrained) spline estimate will provide a nonparametric alternative.

6. FUTURE WORK

Since the lack of constraints does not seem to affect the estimate in the regions where it is hard to discriminate visually, the unconstrained estimate might be good enough for the classification problem. However, we think that it will be useful to develop and implement an efficient algorithm to compute the estimate using generalized cross-validation for constrained problems, as outlined by Wahba (1981b). Such an algorithm would have application not only to this problem but to many other problems like the ones mentioned in Wahba (1981b).

A simulation study should be done to compare this method of estimating posterior probabilities with the method given in Habbema, Hermans and Van den Broek (1974) and with the usual parametric methods.

Finally, it is important to try to establish some properties of the estimate. We believe that convergence rates may be established using the results in Wahba (1979).

7. ACKNOWLEDGMENTS

We gratefully acknowledge useful discussions with D. Bates, B. Silverman and J. Wendelberger.

REFERENCES.

Anderson, J. A. (1972). "Separate Sample Logistic Discrimination", *Biometrika 59*, 1, pp. 19-35.

Anderson, J. A. and Blair, V. (1982). "Penalized maximum likelihood estimation in logistic regression and discrimination", *Biometrika 69*, 1, pp. 123-136.

Chi, P. Y. and Ryzin, J. Van(1977). "A Simple Histogram Method for Nonparametric Classification ," pp. 395-421. In *Classification and Clustering*, ed. J. Van Ryzin,

Cox, D. R.(1966). "Some Procedures Connected with the Logistic Qualitative Response Curve," pp. 55-71. In *Research Papers in Statistics*, ed. F. N. David,John Wiley & Sons, New York .

Day, N. E. and Kerridge, D. F. (1967). "A general maximum likelihood discriminant.", *Biometrics 23*, pp. 313-323.

Fix, E. and Hodges, J. L.(1951). "Discriminatory analysis, nonparametric discrimination: consistency properties." Report No. 4, Project No. 21-49-004, USAF School of Aviation Medicine, Brooks Air Force, Texas .

Glick, N. (1972). "Sample-based Classification Procedures derived from Density Estimators. ", *Journal of the American Statistical Association 67*, pp. 116-122.

Habbema, J. D. F., Hermans, J., and Van den Broek, K.(1974). "A Stepwise Discriminant Analysis Program using Density Estimation," In *COMPSTAT 1974, Proceedings in Computational Statistics*, ed. G. Bruckmann,Physica Verlag, Wien. .

Hermans, J. and Habbema, J. D. F. (1975). "Comparison of Five Methods to Estimate Posterior Probabilities", *EDV in Medizin and Biologie 6*, pp. 14-19.

Hermans, J. and Habbema, J. D. F. (1976). *Manual for the ALLOC Discriminant Analysis Programs*, Department of Medical Statistics University of Leiden,

Netherlands

Kimeldorf, G. S. and Wahba, G. (1971). "Some results on Tchebycheffian Spline Functions", *Mathematical Analysis and Applications 33*, pp. 82-95.

Lachenbruch, P. A. and Goldstein, M. (1979). "Discriminant Analysis", *Biometrics 35*, pp. 69-85.

Madison Academic Computing Center, (1981). *Multi-Dimensional Spline Smoothing Routines*, University of Wisconsin, Madison, Wisconsin

Remme, J., Habbema, J. D. F., and Hermans, J. (1980). "A Simulative comparison of linear, quadratic and kernel discrimination", *Journal of Statistical Comp. and Simulation 11*, pp. 87-105.

Silverman, B. W. (1978). "Density Ratios, Empirical Likelihood and Cot Death", *Applied Statistics 27*, pp. 26-33.

Tapia, R. A. and Thompson, J. R. (1978). *Nonparametric Probability Density Estimation*, John Hopkins University Press

Villalobos, M. A. (1982). *Thesis to appear*, Department of Statistics, University of Wisconsin, Madison, Wisconsin

Wahba, G.(1977). "Optimal Smoothing of Density Estimates," pp. 423-458. In *Classification and Clustering*, ed. J. Van Ryzin,Academic Press .

Wahba, G. (1978). "Improper priors, spline smoothing and the problem of guarding against model errors in regression.", *J. Roy. Stat. Soc. B. 40*, 3, pp. 364-372.

Wahba, G.(1979). "Convergence rates of thin plate smoothing splines," pp. 233-245. In *Smoothing Techniques for Curve Estimation. Lecture Notes in Mathematics No. 757*, ed. Th. Gasser and M. Rosenblatt,Springer-Verlag, New York .

Wahba, G. and Wendelberger, J. (1980). "Some New Mathematical Methods for Variational Objective Analysis using Spline and Cross Validation", *Monthly Weather Review 108*, 8, pp. 1122-1143.

Wahba, G.(1980). "Ill Posed Problems: Numerical and Statistical Methods for Mildly, Moderately and Severely Ill Posed Problems with Noisy Data", UW-Madison, TR# 595, to appear in the Proceedings of the International Conference on Ill Posed Problems, M. Z. Nashed, ed.

Wahba, G. (1981a). "Data-Based Optimal Smoothing of Orthogonal Series Density Estimates", *Annals of Statistics 9*, pp. 146-156.

Wahba, G.(1981b). "Constrained Regularization for Ill-Posed Linear Operator Equations,with Applications in Meteorology and Medicine." Statistics Dept.,UW-Madison, TR#646, to appear Third Purdue Symposium on Statistical Decision Theory, S. S. Gupta and J. O. Berger, eds. .

Wendelberger, J.(1981). "The computation of Laplacian smoothing splines with examples." Statistics Dept., UW-Madison, TR#648 .

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Report No. 686	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER /
4. TITLE (and Subtitle) Multivariate Thin Plate Spline Estimates for the Posterior Probabilities in the Classification Problem		5. TYPE OF REPORT & PERIOD COVERED Scientific Interim 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Miguel A. Villalobos and Grace Wahba		8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-K0042
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics University of Wisconsin Madison, WI 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS U.S. Army Research Office P.O. Box 12211 Research Triangle Park, N.C. 27709		12. REPORT DATE July 1982 13. NUMBER OF PAGES 24
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Estimation of posterior probabilities; Discrimination; Maximum penalized log-likelihood; Thin plate spline; Generalized cross validation.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A nonparametric estimate for the posterior probabilities in the classification problem using multivariate thin plate splines is proposed. This method presents a nonparametric alternative to logistic discrimination as well as to survival curve estimation. The degree of smoothness of the estimate is determined from the data using generalized crossvalidation.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)