CHAPTER 6

Partial Spline Models

6.1 Estimation.

As before, let $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_0$ is M-dimensional, and let $L_1, \ldots, L_n$ be n bounded linear functionals on $\mathcal{H}$. Let $\psi_1, \ldots, \psi_q$ be q functions such that the $n \times q$ matrix S with irth entry

$$S_{ir} = L_i \psi_r$$

is well defined and finite. Letting the matrix $T_{n \times M}$ have $i\nu$th entry $L_i \phi_\nu$ as before, where $\phi_1, \ldots, \phi_M$ span $\mathcal{H}_0$, we will need to suppose that the $n \times (M + q)$ matrix

$$X = (S : T) \tag{6.1.1}$$

is of full column rank (otherwise there will be identifiability problems). The original abstract spline model was

$$y_i = L_i f + \epsilon_i, \quad i = 1, \ldots, n$$

$$f \epsilon \mathcal{H}_0 \oplus \mathcal{H}_1.$$

Find $f_\lambda \in \mathcal{H}$ to minimize

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - L_i f)^2 + \lambda \|P_1 f\|_{\mathcal{H}}^2.$$

The partial spline model is

$$y_i = \sum_{r=1}^{q} \beta_r L_i \psi_r + L_i f + \epsilon_i, \quad i = 1, \ldots, n \tag{6.1.2}$$

where

$$f \in \mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1,$$

as before. Now we find $\beta = (\beta_1, \ldots, \beta_q)'$ and $f \epsilon \mathcal{H}$ to minimize

$$\frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{r=1}^{q} \beta_r L_i \psi_r - L_i f \right)^2 + \lambda \|P_1 f\|_{\mathcal{H}}^2. \tag{6.1.3}$$

I originally believed I was the first to think up these wonderful models around 1983 while enjoying the hospitality of the Berkeley Mathematical Sciences Research Center. Their generation was an attempt to extend the applicability of thin-plate splines to semiparametric modeling of functions of several variables with limited data, and the result appears in Wahba (1984b, 1984c). The work was also motivated by the ideas in Huber (1985) on projection pursuit concerning "interesting directions." I soon found that the idea of partial splines, which has a wealth of applications, had occurred to a number of other authors in one form or another—Ansley and Wecker (1981), Anderson and Senthilselvan (1982), Shiller (1984), Laurent and Utreras (1986), Eubank (1986), Engle, Granger, Rice and Weiss (1986), to mention a few. (The work of Laurent and Utreras and Engle et al. appears earlier in unpublished manuscripts in 1980 and 1982, respectively.)

The application of Engle et al. is quite interesting. They had electricity sales y_i billed each month i for four cities, over a period of years. They also had price ψ_1, income ψ_2, and average daily temperatures x, for each month, by city. The idea was to model electricity demand h as the sum of a smooth function f of monthly temperature x, and linear functions of ψ_1 and ψ_2, along with 11 monthly dummy variables $\psi_3, \ldots, \psi_{13}$, that is, the model was

$$h(x, \psi_1, \ldots, \psi_{13}) = \sum_{\nu=1}^{13} \beta_r \psi_r + f(x)$$

where f is "smooth."

Engle et al. did not observe the daily electricity demand directly, but only certain weighted averages $L_i h$ of it resulting from the fact that the total monthly sales billed reflected the staggered monthly billing cycles. Thus, their model was

$$y_i = \sum_{r=1}^{13} \beta_r \ L_i \psi_r + L_i f + \epsilon_i, \quad i = 1, 2, \ldots, n.$$

An additional twist of their model was the assumption that, rather than being independent, the ϵ_i's followed a first-order autoregressive scheme

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

where the δ_i's are independently and identically distributed for some ρ. This assumption appeared reasonable in the light of the staggered data collection scheme. For the right ρ the quasi-differences

$$\tilde{y}_i = y_i - \rho y_{i-1}$$

result in a new model with independent errors

$$\tilde{y}_i = \sum_{r=1}^{13} \beta_r \tilde{L}_i \psi_r + \tilde{L}_i f + \delta_i, \quad i = 1, \ldots, n,$$

where $\tilde{L}_i = L_i - \rho L_{i-1}$. They fit the model (6.1.2) with $\|P_1 f\|^2 = \int (f''(x))^2 dx$, using gridpoint discretization, whereby the function f is approximated by a vector of its values on a (fine) grid. $\tilde{L}_i f$ is a linear combination of the values of f, $\int (f''(x))^2 dx$ is replaced by a sum of squares of second divided differences, and so forth. The influence matrix $A(\lambda)$ can be found and the GCV estimate $\lambda = \lambda_\rho$ found. Little detail was given on their selection of ρ but Engle et al. noted that all of their estimates for ρ appeared to be quite similar. Altman (1987) has reiterated that care must be taken when the errors are correlated and has studied in depth some procedures appropriate in that case, including the selection of ρ.

Returning to the abstract partial spline model, from the geometric point of view of Kimeldorf and Wahba (1971), we have not done anything new, except adjoin span $\{\psi_r\}_{r=1}^q$ to $\mathcal{H}$, giving a new Hilbert space $\tilde{\mathcal{H}}$

$$\tilde{\mathcal{H}} = \mathcal{H}_{00} \oplus \mathcal{H}_0 \oplus \mathcal{H}_1,$$

where $\mathcal{H}_{00} \equiv \text{span}\{\psi_r\}$. Then $\tilde{\mathcal{H}}_0 = \mathcal{H}_{00} \oplus \mathcal{H}_0$ is the (new) null space of the penalty functional. By the same argument as in Kimeldorf and Wahba, one shows that

$$h = \sum_{r=1}^{q} \beta_r \psi_r + \sum_{\nu=1}^{M} d_\nu \phi_\nu + \sum_{i=1}^{n} c_i \xi_i, \tag{6.1.4}$$

and the problem becomes: Find β, c, d to minimize

$$\frac{1}{n}\|y - S\beta - Td - \Sigma c\|^2 + \lambda c' \Sigma c.$$

Letting $\alpha = \begin{pmatrix} \beta \\ d \end{pmatrix}$, we get

$$\frac{1}{n}\|y - X\alpha - \Sigma c\|^2 + \lambda c' \Sigma c,$$

$$(\Sigma + n\lambda I)c + X\alpha = y,$$
$$X'c = 0,$$

and the GCV estimate of λ can be obtained as before.

We did not say anything concerning properties of the functions $\psi_1, \ldots, \psi_q$, other than the fact that the $L_i \psi_r$ must be well defined and the columns of $X = (S : T)$ must be linearly independent with $q + M \leq n$. It does not otherwise matter whether or not the ψ_r are, say, in $\mathcal{H}$ as the following way of looking at the problem from a geometric point of view will show.

Let $\tilde{\mathcal{H}}$ be a Hilbert space with elements

$$h = (h_0, f_0, f_1)$$

where $h_0 \epsilon \mathcal{H}_{00} \equiv$ span $\{\psi_r\}, f_0 \epsilon \mathcal{H}_0$ and $f_1 \epsilon \mathcal{H}_1$, $\mathcal{H}_0$ and $\mathcal{H}_1$ being as before. We define the projection operators P_{00}, P_0, and P_1 in $\tilde{\mathcal{H}}$ as

$$\begin{aligned} P_{00}h &= (h_0, 0, 0), \\ P_0 h &= (0, f_0, 0), \\ P_1 h &= (0, 0, f_1), \end{aligned}$$

and the squared norm

$$\begin{aligned}\|h\|^2_{\tilde{\mathcal{H}}} &= \|h_0\|^2_{\mathcal{H}_{00}} + \|f_0\|^2_{\mathcal{H}_0} + \|f_1\|^2_{\mathcal{H}_1}\\ &= \|P_{00}h\|^2_{\tilde{\mathcal{H}}} + \|P_0h\|^2_{\tilde{\mathcal{H}}} + \|P_1h\|^2_{\tilde{\mathcal{H}}}.\end{aligned}$$

By convention $L_ih = L_ih_0 + L_if_0 + L_if_1$. Now consider the following three problems.

PROBLEM 1. Find $h_\lambda^{(1)}$, the minimizer in $\tilde{\mathcal{H}}$ of

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - L_ih)^2 + \lambda\|P_1h\|^2_{\tilde{\mathcal{H}}}.$$

PROBLEM 2. Find $\hat\beta^{(2)}$ and $h_\lambda^{(2)}$, the minimizer in $(E^q, \tilde{\mathcal{H}})$ of

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{r=1}^{q}\beta_r L_i\psi_r - L_ih\right)^2 + \lambda\|(P_{00} + P_1)h\|^2_{\tilde{\mathcal{H}}}.$$

PROBLEM 3. Find $\hat\beta^{(3)}$, $d^{(3)}$, and $h_\lambda^{(3)}$, the minimizer in $(E^q, E^M, \tilde{\mathcal{H}})$ of

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{r=1}^{q}\beta_r L_i\psi_r - \sum_{\nu=1}^{M} d_\nu L_i\phi_\nu - L_ih\right)^2 + \lambda\|h\|^2_{\tilde{\mathcal{H}}}.$$

It is not hard to convince oneself that, if

$$h_\lambda^{(1)} = \left(\sum_{r=1}^{q}\hat\beta_r\psi_r, \sum_{\nu=1}^{M}\hat d_\nu\phi_\nu, \hat f_1\right)$$

then

$$h_\lambda^{(2)} = \left(0, \sum_{\nu=1}^{M}\hat d_\nu\phi_\nu, \hat f_1\right), \ \hat\beta^{(2)} = \hat\beta$$

and

$$h_\lambda^{(3)} = \left(0, 0, \hat f_1\right), \ \hat\beta^{(3)} = \hat\beta, \ \hat d^{(3)} = \hat d,$$

where

$$\hat\beta = (\hat\beta_1, \ldots, \hat\beta_q)', \ \hat d = (\hat d_1, \ldots, \hat d_M)'.$$

What this says is that explicitly representing an element of a particular subspace in the sum of squares term effectively puts it in the null space of the penalty functional, whether or not it is there already.

Another important application of partial spline models is to model a function of one or several variables as a function that is smooth except for a discontinuity in a low-order derivative at a specific location. Here, let $f \in W_m$ and let

$$h(x) = \sum_{r=1}^{q}\beta_r\psi_r(x) + f(x)$$

where $\psi_r(x) = (x - x_r)_+^{q_r}$. Here h will have jumps in its derivatives at x_r,

$$\lim_{x \downarrow x_r} h^{(q_r)}(x) - \lim_{x \uparrow x_r} h^{(q_r)}(x) = \beta_r \cdot q_r!$$

This problem was discussed by Ansley and Wecker (1981), and Laurent and Utreras (1986). Shiau (1985) considered various classes of jump functions in several variables, and Shiau, Wahba, and Johnson (1986) considered a particular type of jump function in two dimensions that is useful in modeling two-dimensional atmosphere temperature (as a function of latitude and height, say) where it is desired to model the sharp minimum that typically occurs at the tropopause.

Figure 6.1 from Shiau, Wahba, and Johnson (1986), gives a plot of atmospheric temperature $h(z, l)$ as a function of height z and latitude l. In keeping with meteorological convention, this figure is tipped on its side. The model was

$$h(z, l) = \beta\psi(z, l) + f(z, l)$$

where f is a thin plate spline and

$$\psi(z, l) = |z - z^*(l)|.$$

$z^*(l)$ is shown in Figure 6.2 and $\psi(z, l)$ is shown in Figure 6.3.

6.2 Convergence of partial spline estimates.

We will only give details for $q = 1$. The results below can be extended to the general case of $q << n - M$. We want to obtain a simple representation for $\hat{\beta} = \beta_\lambda$, to examine the squared bias and variance. The calculations below follow Shiau (1985), Heckman (1986), and Shiau and Wahba (1988). The model is

$$y_i = \beta L_i\psi + L_i f + \epsilon_i, \ i = 1, \ldots, n \tag{6.2.1}$$

and the partial spline estimator of β and $f \in \mathcal{H}$ is the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \beta L_i\psi - L_i f)^2 + \lambda\|P_1 f\|^2. \tag{6.2.2}$$

Let $s = (L_1\psi, \ldots, L_n\psi)'$ and $A_0(\lambda)$ be the influence matrix for the problem (6.2.2) if β is identically zero. It is easy to see that, for any fixed β,

$$\begin{pmatrix} L_1 f_\lambda \\ \vdots \\ L_n f_\lambda \end{pmatrix} = A_0(\lambda)(y - s\beta),$$

by minimizing (6.2.2) with $(y - s\beta)$ as "data."

As in Section 4.5, letting $P_1 f_\lambda = \Sigma c_i \xi_i$, we have $\|P_1 f_\lambda\|^2 = c'\Sigma c$, and, since $n\lambda c = (I - A_0(\lambda))(y - s\beta)$, after some algebra, we obtain that $n\lambda c'\Sigma c = (y - s\beta)' A_0(\lambda)(I - A_0(\lambda))(y - s\beta)$ and (6.2.2) is equal to

$$\frac{1}{n}(y - s\beta)'(I - A_0(\lambda))(y - s\beta).$$

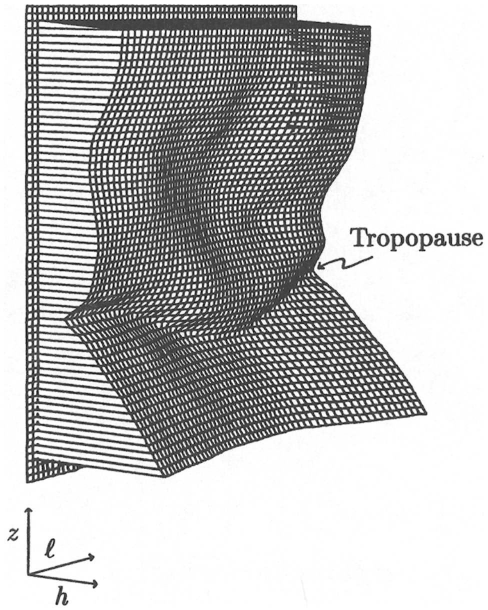

FIG. 6.1. *Estimated temperature $h(z, \ell)$ as a function of height z and latitude ℓ.*

Minimizing over β (and relying on the assumption that $A_0 s \neq s$, which follows from the assumption that s and the columns of T are linearly independent, recall that the the columns of T are the eigenvectors of A_0 with unit eigenvalue), we have

$$\hat{\beta}_\lambda = (s'(I - A_0)s)^{-1}s'(I - A_0)y. \tag{6.2.3}$$

Letting $g = (L_1 f, \ldots, L_n f)'$, we have

$$\hat{\beta}_\lambda - \beta = \frac{s'(I - A_0)(g + \epsilon)}{s'(I - A_0)s} \tag{6.2.4}$$

so that

$$\begin{aligned} \text{bias}\,(\hat{\beta}_\lambda) &= \frac{s'(I - A_0)g}{s'(I - A_0)s} \\ \text{var}\,(\hat{\beta}_\lambda) &= \sigma^2 \frac{s'(I - A_0)^2 s}{[s'(I - A_0)s]^2}. \end{aligned} \tag{6.2.5}$$

Let A_0 be as in (1.3.23) and (4.4.2), that is,

$$I - A_0(\lambda) = n\lambda Q_2(Q_2'(\Sigma + n\lambda I)Q_2)^{-1}Q_2',$$

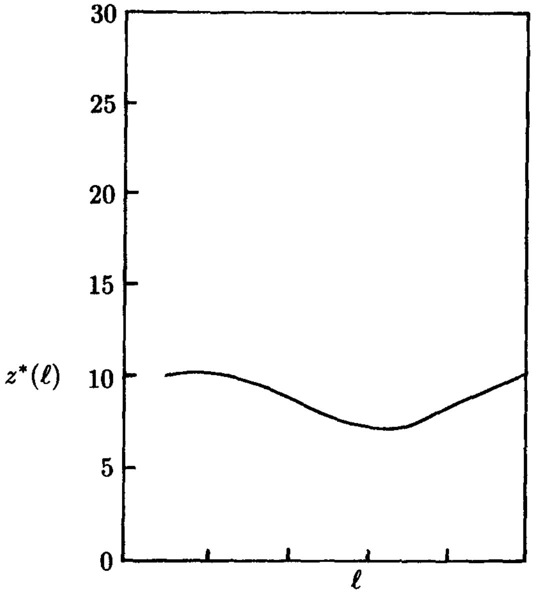

FIG. 6.2. *The tropopause,* $z^*(\ell)$.

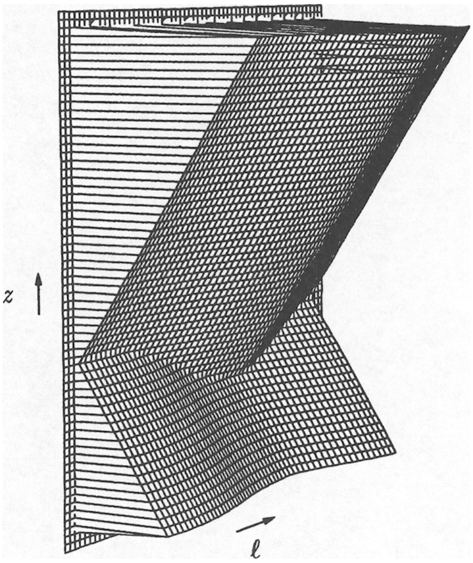

FIG. 6.3. *The tropopause break function* $\psi(z,\ell)$.

$$\begin{aligned} Q_2'\Sigma Q_2 &= UDU', \\ \Gamma &= Q_2U. \end{aligned}$$

Let

$$h = \Gamma' g, \qquad u = \Gamma' s.$$

Then we have

$$\begin{aligned} (\text{bias})\,(\hat{\beta}_\lambda) &= \sum \frac{(n\lambda)u_{\nu n}h_{\nu n}}{\lambda_{\nu n} + n\lambda} \Big/ \left(\sum \frac{n\lambda}{\lambda_{\nu n} + n\lambda} u_{\nu n}^2 \right) \qquad (6.2.6) \\ \text{var}\,(\hat{\beta}_\lambda) &= \sigma^2 \left(\sum \left(\frac{n\lambda}{\lambda_{\nu n} + n\lambda} \right)^2 u_{\nu n}^2 \Big/ \left(\sum \frac{n\lambda}{\lambda_{\nu n} + n\lambda} u_{\nu n}^2 \right)^2 \right). \end{aligned}$$

Asymptotic theory for the squared bias and variance of $\hat{\beta}_\lambda$ has been developed by several authors under various assumptions.

Heckman (1986) considered the case where the components of s (and hence the $u_{\nu n}$) behaved as white noise, and $\|P_1 f\|^2 = \int_0^1 (f^{(m)}(u))^2\,du$. She proved that if $n\lambda^{1/2m} \to \infty$ and either $\lambda \to 0$, or $\|P_1 f\|^2 = 0$, then $\sqrt{n}(\hat{\beta}_\lambda - \beta)$ is asymptotically normal with mean zero and finite variance. See also Chen (1988). In this case we note that a parametric rate for MSE$(\hat{\beta}_\lambda)$ is obtained in the case of an infinite-dimensional "nuisance" parameter g. See Severini and Wong (1987) for a discussion on infinite-dimensional nuisance parameters.

Rice (1986) considered the case where $s_i = s(t_i) + \delta_i$ where $s(\cdot)$ is "smooth" and δ_i behaves like white noise, and found that the parametric rate $1/\sqrt{n}$ cannot hold.

Shiau and Wahba (1988) considered the case where $s_i = s(t_i)$ with s a "smooth" function, such that

$$u_{\nu n} \simeq \sqrt{n}\nu^{-p}, \qquad (6.2.7)$$

and

$$h_{\nu n} \simeq \sqrt{n}\nu^{-q},$$

$$\lambda_{\nu n} \simeq n\nu^{-2m}\ , \ \nu = 1, \ldots, n - M.$$

We also assumed that $q > p > \frac{1}{2}$ so that s is "rougher" than f, and $s \in \mathcal{L}_2$.

To analyze the behavior of bias^2 $(\hat{\beta}_\lambda)$ and var $(\hat{\beta}_\lambda)$, one substitutes (6.2.7) into (6.2.6) and sums, using the lemma

$$\sum_{\nu=1}^{n} \left(\frac{\lambda\nu^{r-\theta}}{(1 + \lambda\nu^r)} \right)^k = \begin{cases} O(\lambda^{(k\theta-1)/r}) & \text{if } \frac{k\theta-1}{k} < r \\ O(\lambda^k) & \text{if } \frac{k\theta-1}{k} > r, \end{cases} \qquad (6.2.8)$$

good for $r > 0, \theta > 0$ and $k \geq 1$. Some of the results are given in Table 6.1. The third column of Table 6.1, λ_{opt}, is the rate of decay for λ that minimizes MSE $(\hat{\beta}_\lambda) = \text{bias}^2(\hat{\beta}_\lambda) + \text{var}(\hat{\beta}_\lambda)$, and the fourth column is the MSE for $\lambda = \lambda_{\text{opt}}$.

TABLE 6.1
Bias, variance, λ_{opt}, *and* MSE (λ_{opt}) *for* $\hat{\beta}_\lambda$.

	$\mathrm{bias}^2(\hat{\beta}_\lambda)$	$\mathrm{var}(\hat{\beta}_\lambda)$	$\lambda_{\mathrm{opt}}(\hat{\beta}_\lambda)$	$\mathrm{MSE}_{\lambda_{\mathrm{opt}}}(\hat{\beta}_\lambda)$
$2m > p+q-1$	$\lambda^{\frac{2(q-p)}{2m}}$	$(n\lambda^{\frac{2p-1}{2m}})^{-1}$	$n^{-\frac{2m}{2q-1}}$	$n^{\frac{-2(q-p)}{2q-1}}$
$p+q-1 > 2m > 2p-1$	$\lambda^{\frac{2(2m-2p+1)}{2m}}$	$(n\lambda^{\frac{2p-1}{2m}})^{-1}$	$n^{-\frac{2m}{4m-2p+1}}$	$n^{\frac{-2(2m-2p+1)}{4m-2p+1}}$
$2p-1 > 2m$	$O(1)$			

We now compare the rates of convergence of λ^* and λ_{opt}, where λ^* is the optimal rate for minimizing the predictive mean-square error $T(\lambda)$ in smoothing splines. We have

$$\begin{aligned}
ET(\lambda) &= \frac{1}{n}E\|g+s\beta-\hat{g}_\lambda-s\hat{\beta}_\lambda\|^2 \\
&= \frac{1}{n}E\|(g+s\beta)-A_0(\lambda)(g+s\beta+\epsilon-s\hat{\beta}_\lambda)-s\hat{\beta}_\lambda\|^2 \\
&= \frac{1}{n}E\|(I-A_0(\lambda))(g-s(\hat{\beta}_\lambda-\beta))-A_0(\lambda)\epsilon\|^2 \\
&= \frac{1}{n}E\|(I-A_0(\lambda))(g-s(s'(I-A_0(\lambda))s)^{-1}s'(I-A_0(\lambda))g) \\
&\qquad -(I-A_0(\lambda))s(s'(I-A_0(\lambda))s)^{-1}(s'(I-A_0(\lambda))\epsilon)-A_0(\lambda)\epsilon\|^2 \\
&= \frac{1}{n}\|(I-A_0(\lambda))(g-s\cdot(\mathrm{bias}(\hat{\beta}_\lambda)))\|^2 \text{ (squared bias term)} \\
&+ \frac{\sigma^2}{n}\{\mathrm{tr}A_0^2(\lambda)+2s'(I-A_0(\lambda))A_0(\lambda)(I-A_0(\lambda))s[s'(I-A_0(\lambda))s]^{-1} \\
&+ s'(I-A_0(\lambda))^2 s\ \mathrm{var}(\hat{\beta}_\lambda)/\sigma^2\} \text{ (variance term)}.
\end{aligned}$$

We just consider the case $2m-2q+1>0$. Then under the assumption we have made on the $h_{\nu n}$ and $\lambda_{\nu n}$, we have that the two main terms in the squared bias term are:

$$\frac{1}{n}\|(I-A_0(\lambda))g\|^2 = O(\lambda^{(2q-1)/2m}),$$

$$\begin{aligned}
\frac{1}{n}\|(I-A_0(\lambda))s\|^2\,\mathrm{bias}^2(\hat{\beta}_\lambda) &= O(\lambda^{(2p-1)/2m})\cdot O(\lambda^{2(q-p)/2m}) \\
&= O(\lambda^{(2q-1)/2m})
\end{aligned}$$

and it can be shown that the variance term is dominated by

$$\frac{\sigma^2}{n}\,\mathrm{tr}\,A_0^2(\lambda) = O(n^{-1}\lambda^{-1/2m}).$$

If there is no cancellation in the squared bias term, then we get

$$\text{Squared bias} + \text{variance} = O(\lambda^{(2q-1)/2m}) + O(n^{-1}\lambda^{-1/2m})$$

and

$$\lambda^* = O(n^{-2m/2q}),$$

whereas from Table 6.1, we have

$$\lambda_{\text{opt}}(\hat{\beta}_\lambda) = O(n^{-2m/(2q-1)}) \text{ when } 2m > p+q-1.$$

So that in this case, $\lambda_{\text{opt}}(\hat{\beta}_\lambda)$ goes to zero at a faster rate than λ^*.

Speckman (1988) and Denby (1986) have proposed an estimator $\tilde{\beta}_\lambda$ that can have a faster convergence rate than $\hat{\beta}_\lambda$ of (6.2.3). It was motivated as follows.

Let $\tilde{y} = (I - A_0(\lambda))y$ be the residuals after removing the "smooth" part $A_0(\lambda)y$, and let $\tilde{s} = (I - A_0(\lambda))s$. Then by regressing $\tilde{y}$ on $\tilde{s}$, the Denby–Speckman estimate $\tilde{\beta}_\lambda$ of β is obtained:

$$\tilde{\beta}_\lambda = (\tilde{s}'\tilde{s})^{-1}\tilde{s}'\tilde{y} = [s'(I - A_0(\lambda))^2 s]^{-1} s'(I - A_0(\lambda))^2 y.$$

Formulas analogous to (6.2.6) for $\tilde{\beta}_\lambda$ are obvious. Table 6.2 gives the square bias, variance λ_{opt}, and MSE (λ_{opt}) for $\tilde{\beta}_\lambda$, from Shiau and Wahba (1988).

TABLE 6.2
Bias, variance λ_{opt}, and MSE_{opt} *for* $\tilde{\beta}_\lambda$.

	$\text{bias}^2(\tilde{\beta}_\lambda)$	$\text{var}(\tilde{\beta}_\lambda)$	$\lambda_{\text{opt}}(\tilde{\beta}_\lambda)$	$\text{MSE}_{\lambda\ \text{opt}}(\tilde{\beta}_\lambda)$
$2m > \frac{p+q-1}{2}$	$\lambda^{\frac{2(q-p)}{2m}}$	$(n\lambda^{\frac{2p-1}{2m}})^{-1}$	$n^{-\frac{2m}{2q-1}}$	$n^{-\frac{2(q-p)}{2q-1}}$
$\frac{p+q-1}{2} > 2m > \frac{2p-1}{2}$	$\lambda^{\frac{2(4m-2p+1)}{2m}}$	$(n\lambda^{\frac{2p-1}{2m}})^{-1}$	$n^{-\frac{2m}{8m-2p+1}}$	$n^{-\frac{2(4m-2p+1)}{8m-2p+1}}$
$\frac{2p-1}{2} > 2m$	$O(1)$			

It can be seen by comparison of Tables 6.1 and 6.2 that there are combinations of p, q, and m for which $\text{MSE}_{\lambda_{\text{opt}}}(\tilde{\beta}_\lambda)$ is of smaller order than $\text{MSE}_{\lambda_{\text{opt}}}(\hat{\beta}_\lambda)$, and combinations for which they are the same order (see Shiau and Wahba (1988) for more details).

6.3 Testing.

Returning to our Bayes model of

$$y_i = \sum_{\nu=1}^{M} \theta_\nu L_i \phi_\nu + b^{1/2} L_i X + \epsilon_i,\ i = 1, \ldots, n, \tag{6.3.1}$$

we wish to test the null hypothesis

$$b = 0, \; y_i = \sum_{\nu=1}^{M} \theta_\nu L_i \phi_\nu + \epsilon_i, \; i = 1, \ldots, n \tag{6.3.2}$$

versus the alternative

$$b \neq 0. \tag{6.3.3}$$

Letting $T_{n \times M} = \{L_i \phi_\nu\}$, and $\Sigma = \{L_{i(s)} L_{j(t)} Q(s,t)\}$, where $EX(s)X(t) = Q(s,t)$, we have

$$y \sim \mathcal{N}(T\theta, b\Sigma + \sigma^2 I)$$

under the "fixed effects" model, and

$$y \sim \mathcal{N}(0, \xi TT' + b\Sigma + \sigma^2 I)$$

under the "random effects" model.

The most interesting special case of this is

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n$$

with the null hypothesis f a low-degree polynomial, versus the alternative, f "smooth." Yanagimoto and Yanagimoto (1987) and Barry and Hartigan (1990) considered maximum likelihood tests for this case. Cox and Koh (1989) considered the case $f \in W_m$ and obtained the locally most powerful (LMP) invariant test. Cox, Koh, Wahba, and Yandell (1988) considered the LMP invariant test in the general case, and obtained a relation between the LMP invariant and the GCV test (to be described). In preparation for the CBMS conference, I obtained a similar relation for the GML test (to be described) and did a small Monte Carlo study to compare the LMP invariant, GML, and GCV tests.

Let $T = (Q_1 : Q_2) \begin{pmatrix} R \\ 0 \end{pmatrix}$ as in (1.3.18). As before

$$T'Q_2 = 0.$$

Let

$$w_1 = Q_1' y, \qquad w_2 = Q_2' y;$$

then

$$w_2 \sim \mathcal{N}(0, bQ_2' \Sigma Q_2 + \sigma^2 I)$$

for either the fixed or random effects model. Letting $UDU' = Q_2' \Sigma Q_2$ and

$$z = U' w_2$$

we obtain

$$z \sim \mathcal{N}(0, bD + \sigma^2 I), \tag{6.3.4}$$

where the diagonal entries of D are $\lambda_{\nu n}, \nu = 1, \ldots, n-M$, that is

$$z_\nu \sim \mathcal{N}(0, b\lambda_{\nu n} + \sigma^2), \; \nu = 1, \ldots, n-M.$$

Cox and Koh (1989) showed that the LMP test (at $b = 0$), invariant under translation by columns of T, rejects when

$$t_{\text{LMP}} = \sum_{\nu=1}^{n-M} \lambda_{\nu n} z_\nu^2 \tag{6.3.5}$$

is too large.

A test based on the GCV estimate for λ may be obtained by recalling that $\lambda = \infty$ corresponds to $f_\lambda \epsilon \mathcal{H}_0$. In the notation of this section, we have

$$\begin{aligned} n^{-1}V(\lambda) &= \frac{\|(I - A(\lambda))y\|^2}{(\text{tr}(I - A(\lambda)))^2} \\ &= \sum_{\nu=1}^{n-M} \left(\frac{n\lambda}{n\lambda + \lambda_{\nu n}}\right)^2 z_\nu^2 \Bigg/ \left(\sum_{\nu=1}^{n-M} \frac{n\lambda}{n\lambda + \lambda_{\nu n}}\right)^2 . \end{aligned} \tag{6.3.6}$$

We have the following theorem.

THEOREM 6.3.1. *$V(\lambda)$ has a (possibly local) minimum at $\lambda = \infty$, whenever*

$$t_{\text{LMP}} \equiv \sum_{\nu=1}^{n-M} \lambda_{\nu n} z_\nu^2 \le \bar{\lambda} \sum_{\nu=1}^{n-M} z_\nu^2, \tag{6.3.7}$$

where $\bar{\lambda} = 1/(n-M)\sum_{\nu=1}^{n-M} \lambda_{\nu n}$.

Proof. Let $\gamma = 1/n\lambda$, and define $\tilde{V}(\gamma)$ as $V(\lambda)$ with $1/n\lambda$ replaced by γ, that is,

$$n^{-1}\tilde{V}(\gamma) = \sum_{\nu=1}^{n-M} \left(\frac{1}{1 + \gamma\lambda_{\nu n}}\right)^2 z_\nu^2 \Bigg/ \left(\sum_{\nu=1}^{n-M} \frac{1}{1 + \gamma\lambda_{\nu n}}\right)^2 . \tag{6.3.8}$$

Differentiating $\tilde{V}(\gamma)$ with respect to γ, one obtains that

$$\tilde{V}'(\gamma)\Big|_{\gamma=0} \ge 0 \text{ iff } \sum_{\nu=1}^{n-M} \lambda_{\nu n} z_\nu^2 \le \bar{\lambda} \sum_{\nu=1}^{n-M} z_\nu^2 .$$

We note that $\sum_{\nu=1}^{n-M} z_\nu^2$ is the residual sum of squares after regression of y onto the columns of T, that is, the residual sum of squares after fitting the null model.

An approximate LMP invariant test when σ^2 is unknown is to use

$$t_{\text{LMP approx}} = \sum_{\nu=1}^{n-M} \lambda_{\nu n} z_\nu^2 / \sum_{\nu=1}^{n-M} z_\nu^2 .$$

The GCV test is

$$t_{\mathrm{GCV}} = \text{const.}\,\frac{\tilde{V}(\hat{\gamma})}{\tilde{V}(0)} = \frac{\sum\left(z_\nu^2/(1+\hat{\gamma}\lambda_{\nu n})^2\right)}{\left(\sum(1/(1+\hat{\gamma}\lambda_{\nu n}))^2\right)} \times \frac{1}{\sum z_\nu^2}$$

where $\hat{\gamma} = 1/n\hat{\lambda}$.

The (invariant) likelihood ratio test statistic t_{GML} for $\lambda = \sigma^2/nb = \infty$ is

$$t_{\mathrm{GML}} = \text{ const.}\,\frac{M(\tilde{\lambda})}{M(\infty)},$$

where $\tilde{\lambda}$ minimizes $M(\lambda)$ of (4.8.4). Upon letting $\tilde{\gamma} = 1/n\tilde{\lambda}$, we have

$$t_{\mathrm{GML}} = \frac{\sum\left(z_\nu^2/(1+\tilde{\gamma}\lambda_{\nu n})\right)}{\prod(1+\tilde{\gamma}\lambda_{\nu n})^{-1/(n-M)}} \times \frac{1}{\sum z_\nu^2}.$$

It can also be shown that $M(\lambda)$ has a (possibly local) minimum at $\lambda = \infty$, whenever (6.3.7) holds.

An experiment was designed to examine the relative power of $t_{\mathrm{LMP}}, t_{\mathrm{GCV}}$, and t_{GML} by simulation. It was to be expected that t_{LMP} would have the greatest power for nearby alternatives but would not be so good for "far out" alternatives, and it was tentatively conjectured that t_{GML} would be better than GCV for "random" alternatives (to be defined) but GCV would be better for "smooth" alternatives. We considered the one-dimensional cubic smoothing spline with $n = 100$ data points at $x_i = i/n$; there are 98 eigenvalues $\lambda_{\nu n}$, which are plotted in Figure 6.4. This corresponds to the null hypothesis that f is linear. We also considered a two-dimensional thin-plate spline with $m = 2$ on a 12×12 regular grid with $(x_1, x_2) = (i/12, j/12)$, $i = j = 1, \ldots, 12$, thus $n = 144$, $n - M = 141$. The null hypothesis corresponds to $f(x_1, x_2)$, a plane. The 141 eigenvalues are also plotted in Figure 6.4. The eigenvalues $\lambda_{\nu n}$ for these splines decay at the rate $\nu^{-2m/d}$ where $m = 2$ here and d is the dimension of x; here $d = 1$ and $d = 2$. The decay rate of the eigenvalues is readily evident.

The random alternative was

$$z_\nu \sim \mathcal{N}(0, b\lambda_{\nu n} + \sigma^2)$$

where the size of b controlled the distance of the alternative from the null hypothesis, and, without loss of generality we set $\sigma^2 = 1$. The "smooth" fixed function corresponded to

$$z_\nu \sim \mathcal{N}(\sqrt{b}h_{\nu n}, \sigma^2)$$

with $\sum(h_{\nu n}^2/\lambda_{\nu n}) < \infty$. Here the $h_{\nu n}$ are related to $f(x)$ by

$$\begin{pmatrix} h_{1n} \\ \vdots \\ h_{n-M,n} \end{pmatrix} = U'Q_2' \begin{pmatrix} f(x(1)) \\ \vdots \\ f(x(n)) \end{pmatrix}.$$

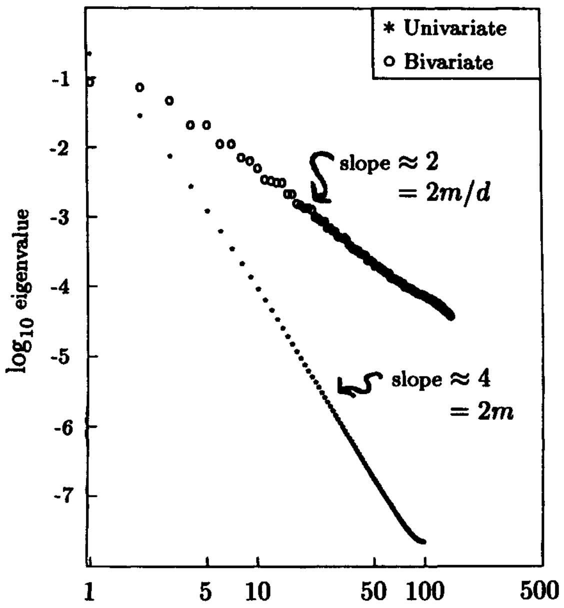

FIG. 6.4. *Univariate* ($n = 98$) *and bivariate* ($n = 141$) *eigenvalues.*

Figure 6.5 gives a plot of the univariate alternative $f(x)$ at $x(i)$, $i = 1, \ldots, n$ with $b = 1$; actually f was chosen so that the first four $h_{\nu n}$'s were 1,1,-1, and 1 and the rest 0. Figure 6.6 gives a plot of the bivariate smooth alternative $f(x_1, x_2)$. This function was chosen so that the first three $h_{\nu n}$'s were 1 and the rest 0.

The distributions of $t_{\mathrm{LMP}}, t_{\mathrm{GCV}}$, and t_{GML} for the univariate example under the null hypothesis were estimated by drawing 1,000 replicates of $n - M = 98$ $\mathcal{N}(0,1)$ z_ν's, and computing 1,000 values of each statistic. Global search in increments of $\log\gamma$ was used for the minimizations of the GCV and GML functions. Care must be taken that the search increment is sufficiently fine and over a sufficiently wide range.

Figures 6.7 and 6.8 give histograms of $-\log t_{\mathrm{GCV}}$ and $-\log t_{\mathrm{GML}}$ under the null hypothesis. If V or M is minimized for $\gamma = 1/\lambda = \infty$, then $-\log t_{\mathrm{GCV}}$ or $-\log t_{\mathrm{GML}}$ is zero. We note that these were 581 samples of $-\log t_{\mathrm{GCV}} = 0$ and 628 samples of $-\log t_{\mathrm{GML}} = 0$. Defining $S/N = (b \sum \lambda_{\nu n}/n\sigma^2)^{1/2}$ and $(b \sum h_{\nu n}^2/n\sigma^2)^{1/2}$ for the random and smooth alternatives, respectively, 1,000 replicates of $t_{\mathrm{LMP}}, t_{\mathrm{GCV}}$, and t_{GML} for a series of values of S/N were generated. The same 98 × 1,000 random numbers were used for the different values of S/N, so the data in Figures 6.7–6.14 are not independent. The histograms

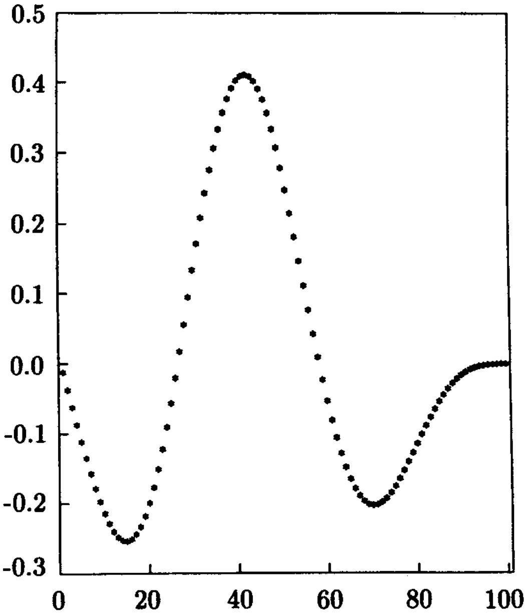

FIG. 6.5. *The univariate alternative.*

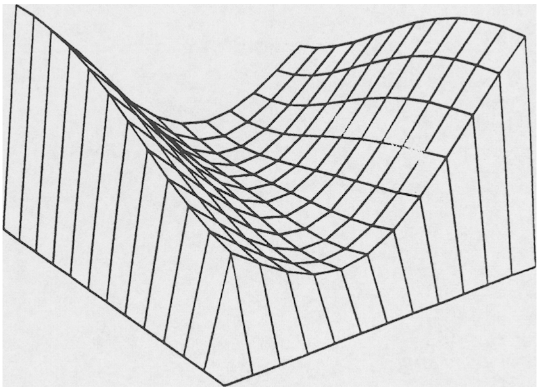

FIG. 6.6. *The bivariate alternative.*

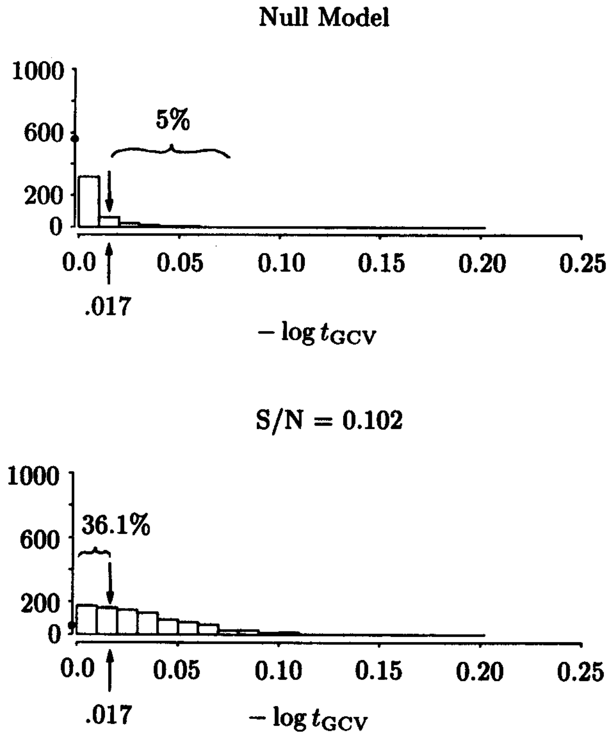

FIG. 6.7. *Histograms for* $-\log t_{\mathrm{GCV}}$*, univariate deterministic example.*

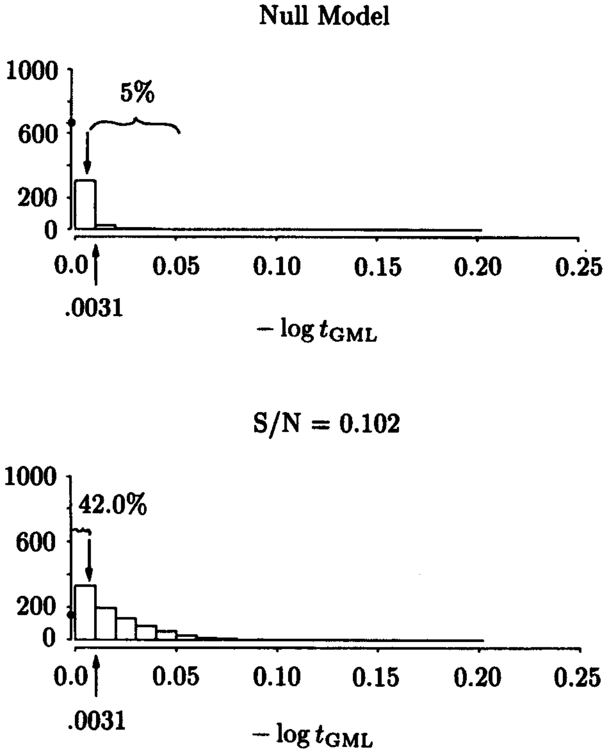

FIG. 6.8. *Histograms for* $-\log t_{\mathrm{GML}}$*, univariate deterministic example.*

for $-\log t_{\mathrm{GCV}}$ and $-\log t_{\mathrm{GML}}$ for the univariate smooth function case for $S/N = .102$ appear in Figures 6.7 and 6.8. It can be seen that a nonnegligible mass point appears at zero. Using the 95 percent points of the simulated null distributions as cutoff points, the probability of accepting the null hypothesis for various values of S/N was estimated by counting how many of the 1,000 simulated values of t fell below the cutoff. These estimated type-two errors are plotted for the three statistics for the univariate smooth alternative in Figure 6.9, for the univariate random alternative in Figure 6.10, and for the bivariate smooth alternative in Figure 6.11. It appears that t_{GCV} is slightly better in the deterministic example and t_{GML} in the random example, but we do not believe these experiments are definitive. The sampling error is fairly large (and its magnitude is not evident in the plots because the same random numbers were used for different S/N), and in other experiments the reverse was found. The results can also be surprisingly sensitive to the search procedure used.

Figures 6.12–6.14 show the histograms for the three test statistics for six values of S/N. We remark that in practice Monte Carlo estimation of distributions such as these can be important.

The distributions of the test statistics $-\log t_{\mathrm{GCV}}$ and $-\log t_{\mathrm{GML}}$ have a mass point at zero, which shrinks as S/N becomes larger. This mass point is not plotted separately in the histograms of Figures 6.12 and 6.13, but is displayed in Figures 6.7 and 6.8. The continuous part of some of the distributions appears to have a bimodal density, although to see the exact behavior near zero probably would require a more delicate search procedure than we have used. Simple asymptotic approximations to the distributions are, in our opinion, not necessarily trustworthy for practical use due to the complex structure of these distributions.

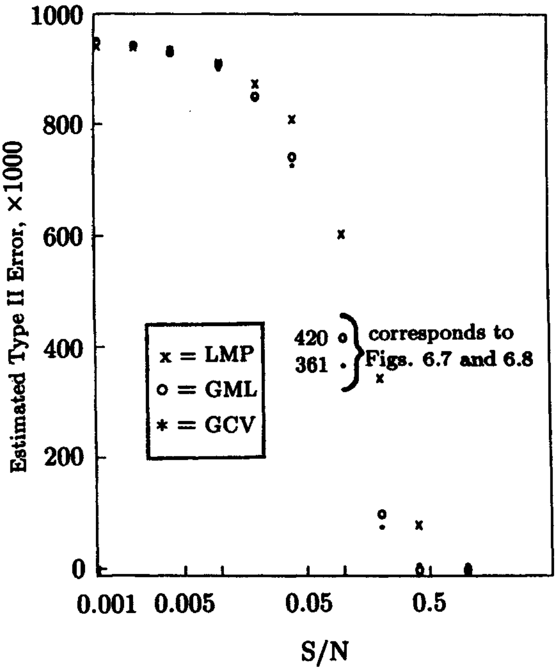

FIG. 6.9. *Univariate example, smooth deterministic alternative.*

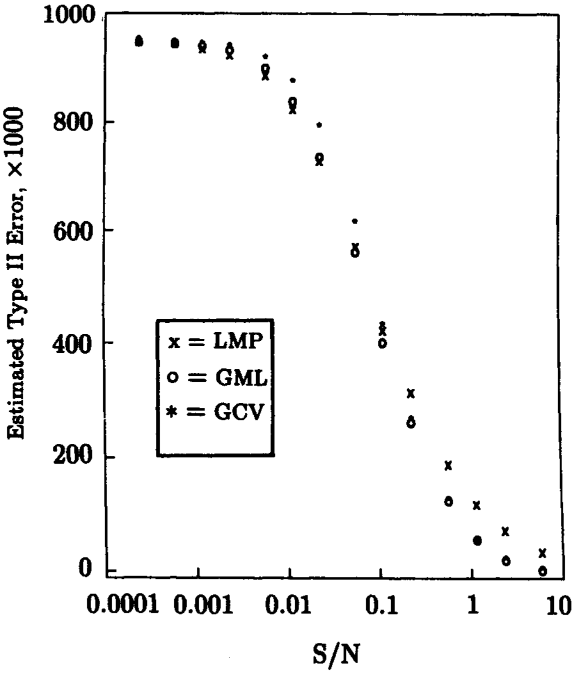

FIG. 6.10. *Univariate example, random alternative.*

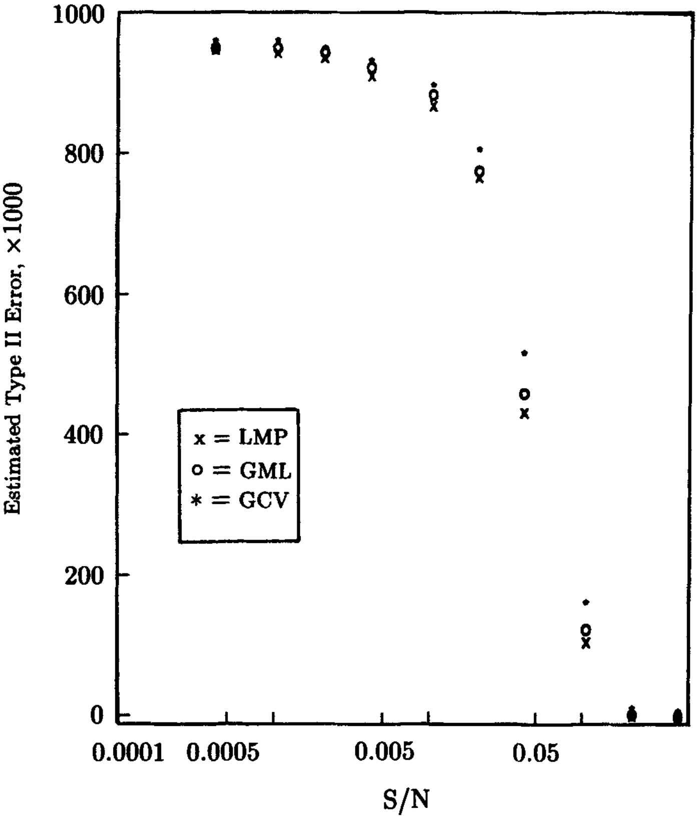

FIG. 6.11. *Bivariate example, deterministic alternative.*

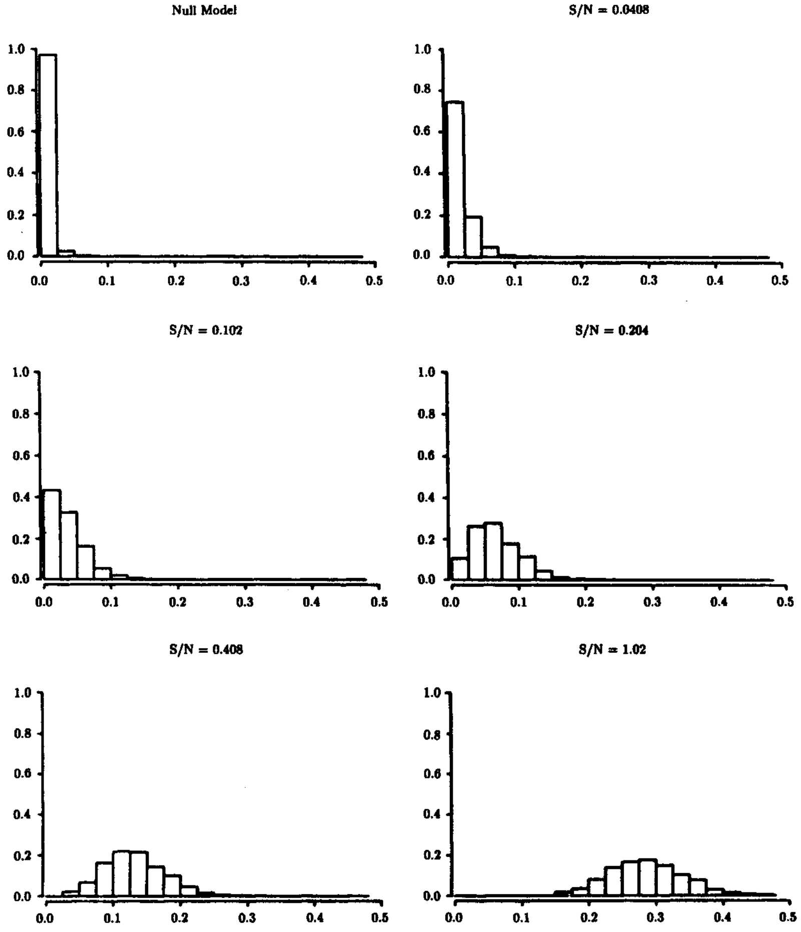

FIG. 6.12. *Histograms for* $-\log t_{\mathrm{GCV}}$, *univariate deterministic example.*

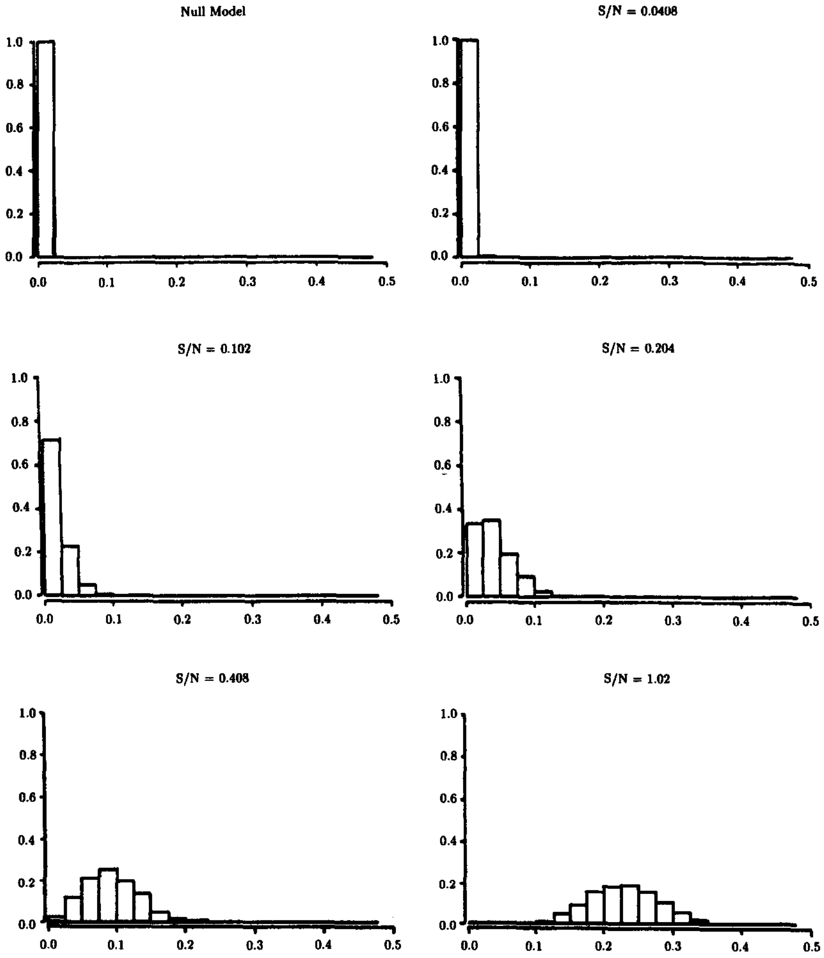

FIG. 6.13. *Histograms for* $-\log t_{\mathrm{GML}}$, *univariate deterministic example.*

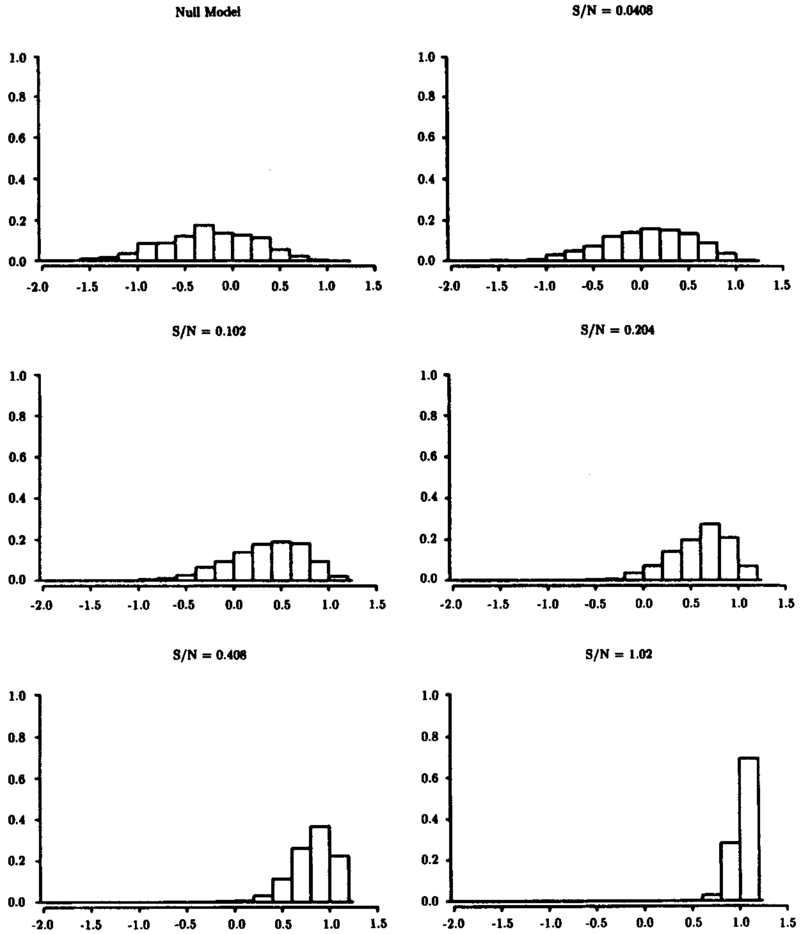

FIG. 6.14. *Histograms for* $-t_{\mathrm{LMP}}$, *univariate deterministic example.*