JOURNAL OF RESEARCH of the National Bureau of Standards
Vol. 87, No. 1, January–February 1982

An Approach to Peak Area Estimation

John Rice

Department of Mathematics, University of California at San Diego, LaJolla, CA 92093

and

National Bureau of Standards, Washington, DC 20234*

May 27, 1981

We consider the problem, arising in nuclear spectroscopy, of estimating peak areas in the presence of a baseline of unknown shape. We analyze a procedure that chooses the baseline to be as smooth as is consistent with the data and note that the estimates have a certain minimax optimality. Expressions are developed for the systematic and random errors of the estimate, and some large sample approximations are derived. Procedures for choosing a smoothing parameter are developed and illustrated by simulations.

Key words: linear models; minimax; peak area; smoothing; spectroscopy; splines.

1. Introduction

The estimation of peak area in the presence of a baseline of unknown shape is a common problem in nuclear and other spectroscopies. In this paper we analyze some of the properties of a generalization of a procedure proposed by Currie [2][1] and note that the procedure has a certain minimax optimality.

We first introduce the problem and some notation. We suppose that counts are accumulated in n channels over a length of time T, and that the total number of counts has mean $\mu = \nu T$, where ν = mean counting rate per unit time. We let y_j denote the proportional count in the j^{th} channel, i.e. the total count in the j^{th} channel divided by μ, and we assume that

$$y_j = \beta_0 \gamma_j + \beta_j + \varepsilon_j, j = 1, \ldots, n$$

Here, $\Gamma = (\gamma_1, \ldots, \gamma_n)^T$ is a vector representing a peak shape, which is assumed known (Γ might be known from theory or from measurement of pure specimens, for example), β_0 is its unknown amplitude, which we wish to determine, and β_j is the unknown baseline mean in the j^{th} channel. The ε_j's are random counting errors with mean zero and nonsingular covariance matrix $\mu^{-1}W^{-1}$ where W is a matrix which is assumed to be known. (In applications, W is typically estimated rather than known. An application of the δ-method [7] to the perturbation thus introduced shows that the asymptotic means and variances are unchanged.) In vector notation the model can be written

$$\begin{aligned} Y &= [\Gamma : I]\beta + \varepsilon \\ &= A\beta + \varepsilon \end{aligned}$$

where $Y = (y_1, \ldots, y_n)^T$, $\beta = (\beta_0, \beta_1, \ldots, \beta_n)^T$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$. We note that this model is underdetermined, and that even in the limit, with no counting error, there is no unique solution for β_0.

Currie [2] proposed estimating β by forcing the baseline to be as smooth as is consistent with the data (in a sense explained below), taking as measures of smoothness

* Statistical Engineering Division, Center for Applied Mathematics.

[1] Figures in brackets refer to literature references at the end of this paper.

$$S_1 = \sum_{j=1}^{n-1} (\beta_i - \beta_{i+i})^2$$

or

$$S_2 = \sum_{j=1}^{n-2} (\beta_i - 2\beta_{i+1} + \beta_{i+2})^2$$

or generally

$$S_k = \sum_{j=1}^{n-k} (\Delta^k \beta)^2$$

where Δ is a differencing operator. The estimate $\hat{\beta}$ is formed by minimizing S_k subject to the constraint

$$(Y - A\beta)^T W (Y - A\beta) = c$$

where the constraint c is obtained from the χ^2 distribution. Using the technique of Lagrange multipliers, the solution is found to be

$$\beta = (A^T W A + \lambda U^T U)^{-1} A^T W Y$$

when λ is chosen to force β to satisfy the constraint and S_k is expressed as

$$S_k = \|U\beta\|^2.$$

By considering numerical examples, Currie reached some empirical conclusions about the statistical behavior of the method, with special attention to the bias, or systematic error, of the method.

Techniques of this kind have been used in solving ill-posed problems such as integral equations of the first kind [1] and in smoothing data via smoothing splines [8, 11]. Motivated by such problems, Kuks and Olman [5] and Speckman [9] have considered the problem of estimating a linear functional $h^T\beta$ by linear functionals of the data, $\ell^T Y$. Their result is the following: Consider the linear model

$$Y = A\beta + \varepsilon$$

where ε has a nonsingular covariance matrix $\sigma^2 W^{-1}$, and assume that $\|U\beta\|^2 \leq \alpha^2$ for some matrix U such that $N(U) \cap N(A) = \phi$ ($N(A)$ = null space of A). Then the estimate $\ell_0^T Y$ for which

$$E(\ell_0^T Y - h^T\beta)^2 = \min_{\ell} \max_{\|U\beta\|^2 \leq \alpha^2} E\,(\ell_0^T Y - h^T\beta)^2$$

is unique and is given by

$$\ell_0^T Y = h^T (A^T W A + (\sigma^2/\alpha^2) U^T U)^{-1} A^T W Y.$$

Identifying λ with σ^2/α^2 this solution is seen to be formally the same as the estimate proposed by Currie for estimating the peak amplitude $\beta_0 = (1, 0, \ldots, 0)\beta$. An operational difference is that the minimax theorem assumes the smoothness parameter α^2 to be known, whereas Currie implicitly estimates it from the data. It should be noted that the estimate is minimax for estimating any single linear functional but is not generally minimax for estimating several linear functionals simultaneously [10].

In the next section we will consider the more general problem of several peaks of known shape and unknown amplitudes, superposed on an unknown baseline (Currie considered only the single peak case). We will

develop expressions for the bias and variance of the amplitude estimates and limiting approximations as the expected total count $\mu \to \infty$ which give some insight into the properties of the method. In section 4 a procedure for choosing λ from the data is discussed and is illustrated by some simulations.

2. Bias and variance

In this section we will assume the following, multi-peak model:

$$\begin{aligned} Y &= \beta_{11}\Gamma_1 + \cdots + \beta_{1p}\Gamma_p + \beta_2 + \varepsilon \\ &= [\Gamma : I]\,\beta + \varepsilon \\ &= A\beta + \varepsilon \end{aligned}$$

where Y is an n-vector, $\Gamma = [\Gamma_1, \Gamma_2, \ldots, \Gamma_p]$, $\beta_2 = (\beta_{21}, \ldots, \beta_{2n})^T$ is the vector of mean background counts, $\beta^T = (\beta_1^T, \beta_2^T)$, and ε is a vector of random errors with nonsingular covariance matrix $\mu^{-1}W^{-1}$. We will derive expressions for the bias and variance of the estimate

$$\hat{\beta} = (A^T W A + \lambda U^T U)^{-1} A^T W Y$$

when U is of the form

$$\underset{(n+p-k)\ x\ (n+p)}{U} = \begin{bmatrix} \underset{pxp}{0} & \underset{pxn}{0} \\ \underset{(n-k)xp}{0} & \underset{(n-k)xn}{U_1} \end{bmatrix},$$

and thus $U^T U$ is of the form

$$\underset{(n+p)\ x\ (n+p)}{U^T U} = \begin{bmatrix} \underset{pxp}{0} & \underset{nxn}{0} \\ \underset{pxp}{0} & \underset{nxn}{D} \end{bmatrix},$$

where

$$D = U_1^T U_1$$

(D is not diagonal) and $\lambda = 1/\mu\alpha^2$ is given. If λ is estimated from the data these expressions are conditional on λ. The unconditional bias and variance are different.

We will focus attention on the estimate β_1 of the vector of peak amplitudes, which is of primary interest. It is thus useful to partition the matrix $(A^T W A + \lambda U^T U)^{-1}$:

$$\begin{aligned} (A^T W A + \lambda U^T U)^{-1} &= \begin{bmatrix} \Gamma^T W \Gamma & \Gamma^T W^T \\ W\Gamma & W + \lambda D \end{bmatrix}^{-1} \\ &= \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \end{aligned}$$

From an identity for the inverse of a partitioned matrix [7],

$$\begin{aligned} B_{11} &= (\Gamma^T W \Gamma)^{-1} + (\Gamma^T W \Gamma)^{-1}\Gamma^T W^T [W + \lambda D - W\Gamma(\Gamma^T W \Gamma)^{-1}\Gamma^T W^T]^{-1} W\Gamma(\Gamma^T W \Gamma)^{-1} \\ &= G^{-1} + G^{-1}\Theta^T R^{-1}\theta G^{-1} \end{aligned}$$

where $G = (\Gamma^T W \Gamma)$, $\theta = W\Gamma$, and R is the matrix given in square brackets. With this notation,

$$B_{12} = -G^{-1}\theta^T R^{-1}.$$

$B_{12} = B_{21}^T$; we will not need B_{22}. Now,

$$E\hat{\beta} = (A^T WA + \lambda U^T U)^{-1} A^T WA\beta$$

and

$$A^T WA\beta = \begin{bmatrix} \Gamma^T W\Gamma\beta_1 + \Gamma^T W\beta_2 \\ W\Gamma\beta_1 + W\beta_2 \end{bmatrix}$$

$$= \begin{bmatrix} G\beta_1 + \theta^T\beta_2 \\ \theta\beta_1 + W\beta_2 \end{bmatrix},$$

so that

$$E\hat{\beta}_1 = (G^{-1} + G^{-1}\theta^T R^{-1}\theta G^{-1})(G\beta_1 + \theta^T\beta_2) - G^{-1}\theta^T R^{-1}(\theta\beta_1 + W\beta_2).$$

We thus have, after simplification, an expression for the bias of $\hat{\beta}_1$:

$$\beta_1 - E\hat{\beta}_1 = -G^{-1}\theta^T[I - R^{-1}(W - \Theta G^{-1}\Theta^T)]\beta_2. \tag{1}$$

Note that the bias does not involve β_1 and that the derivation of the bias expression has not assumed that $\mu^{-1}W^{-1}$ is the true covariance matrix of the random errors. In the appendix it is shown that the bias is zero if $U_1\beta_2 = 0$.

A simple bound for the bias may be obtained as follows: from the expression above, the squared bias for a particular component β_{1k}, say, may be written in the form

$$|\beta_{1k} - E\hat{\beta}_{1k}|^2 = |r^T\beta_2|^2$$

Let $P = U_1^T(U_1U_1^T)^{-1}U_1$ be the matrix which projects onto $N(U_1)$, let $Q = I - P$ project onto $N(U_1)$, and express $\beta_2 = P\beta_2 + Q\beta_2$. Noting from above that $r^TQ\beta_2 = 0$, we may write

$$\begin{aligned} |r^T\beta_2|^2 &= |r^TU_1^T(U_1U_1^T)^{-1}U_1\beta_2|^2 \\ &\leqslant \sup_{\{\beta_2:\ \|U_1\beta_2\|^2 \leqslant \alpha^2\}} |r^TU_1^T(U_1U_1^T)^{-1}U_1\beta_2|^2 \\ &= \alpha^2|r^TU_1^T(U_1U_1^T)^{-2}U_1r|^2 \end{aligned}$$

We now consider the variance of the estimate. Under the assumption that the covariance matrix of the errors is $\mu^{-1}W^{-1}$, it is immediate that the covariance matrix of $\hat{\beta}$ is

$$\Sigma = \mu^{-1}(A^T WA + \lambda U^T U)^{-1}A^T WA(A^T WA + \lambda U^T U)^{-1}$$

In an appendix it is shown how this matrix may be partitioned and that the covariance matrix of β_1 can be expressed as

$$\Sigma_{11} = \mu^{-1}F^TF \tag{2}$$

where

$$F = W^{-1/2}[I-(W-\theta G^{-1}\theta^T)R^{-1}]\Theta G^{-1} \tag{3}$$

and $W^{1/2}$ is the symmetric square root of W.

We will now develop approximations to the bias and Σ_{11} for large samples by examining their behavior as T and thus $\mu\to\infty$ and $\lambda\to 0$. The expressions for Σ_{11} and the bias both involve the matrix

$$I - R^{-1}(W-\Theta G^{-1}\Theta T) = I - [W+\lambda D - W\Gamma(\Gamma^T W\Gamma)^{-1}\Gamma^T W]^{-1}\cdot(W - W\Gamma(\Gamma^T W\Gamma)^{-1}\Gamma^T W)$$

As $\lambda\to 0$, $R\to W - W\Gamma(\Gamma^T W\Gamma)^{-1}\Gamma^T W$, but this matrix is singular (the null space is spanned by $\Gamma_1, \ldots, \Gamma_p$). A further complication is that D will typically not be of full rank (for example, D may annihilate constant and linear functions). However, our assumption that $N(U)\cap N(A) = \phi$ guarantees that $D\Gamma_j \neq 0, j=1, \ldots, p$ and thus that the matrix R is invertible. In the appendix we prove the following:

LEMMA. *Suppose that* C *is an* nxn *non-negative definite matrix with* p *dimensional null space spanned by* $v_1, \ldots, v_p$. *Suppose that* D *is another* nxn *non-negative definite matrix and that* $N(C)\cap N(D) = \phi$. *Then as* $\lambda\to 0$

$$I - (C + \lambda D)^{-1}C = V(V^T D V)^{-1}V^T D + O(\lambda)$$

where $V = [v_1, \ldots, v_p]$ *is an* nxp *matrix*.

Applying this lemma to the expressions for Σ_{11} and the bias of β_1, with $W - W\Gamma(\Gamma^T W\Gamma)^{-1}\Gamma^T W$ corresponding to C and Γ corresponding to V we have,

COROLLARY: *Under the assumptions of our linear model, as* $\lambda\to 0$ ($\mu\to\infty$),

$$\beta_1 - E\hat{\beta}_1 = -(\Gamma^T D\Gamma)^{-1}\Gamma^T D\beta_2 + O(\lambda) \tag{1}$$

$$\mu\Sigma_{22} = (\Gamma^T D\Gamma)^{-1}(\Gamma^T D W^{-1} D\Gamma)(\Gamma^T D\Gamma)^{-1} + O(\lambda). \tag{2}$$

The expression for the bias is simpler to understand if we write it as

$$\beta_1 - E\hat{\beta}_1 \to -[(U_1\Gamma)^T(U_1\Gamma)]^{-1}(U_1\Gamma)^T(U_1\beta_2)$$

and keep in mind that U_1, is a differencing operator. The bias is determined by the relationships of the vectors $U_1\Gamma_j$, $j=1, \ldots, p$ and $U_1\beta_2$. If the baseline β_2 is quite smooth $U_1\beta_2$ will be small. If a particular peak shape Γ_j does not overlap any other peaks then the limiting ($\mu\to\infty$) bias of the estimate of its amplitude is simply

$$\beta_{1j} - E\hat{\beta}_{1j} \cong \frac{(U_1\Gamma_j)^T(U_1\beta_2)}{\|U_1\Gamma_j\|^2} < \frac{\alpha}{\|U_1\Gamma_j\|}$$

which follows from the rule for the inverse of a partitioned matrix and the Cauchy-Schwartz inequality. The large components of $U_1\Gamma_j$ will be those near the peak center and if the true background β_2 is smooth in this region, the bias will be small.

When two peaks overlap substantially, however, the bias will typically be worse than the bias if either one of the peaks were absent, since corresponding elements of the matrix $[(U_1\Gamma)^T(U_1\Gamma)]^{-1}$ will be large.

Finally, we note that this limiting bias does not depend on the weighting matrix W and that it depends linearly on the baseline proportion.

The variance of the estimate β_{1j} of a peak amplitude can also be expressed simply in the case that the matrix W is diagonal and the peak does not overlap other peaks:

$$Var(\hat{\beta}_{1j}) \cong \frac{(U_1\Gamma_j)^T U_1 W^{-1} U_1^T(U_1\Gamma_j)}{\mu\|U_1\Gamma_j\|^2}$$

but in the case that there is considerable peak overlap the variance may be inflated considerably.

It is of some interest to consider the relative size of the bias to the standard error and to understand qualitatively how this is affected by varying the baseline amplitude. To this end we consider a single peak model with a peak shape standardized so that $\Sigma\gamma_j = 1$ and a standard baseline profile with $\sum_{j=1}^{n}\beta j = 1$. Any mixture of this peakshape and background profile with peak proportion β_0 and background proportion $1-\beta_0$ can be expressed as $\beta_0\Gamma+(1-\beta_0)\beta$, where $0 \leqslant \beta_0 \leqslant 1$. Denoting $D\Gamma$ by $V = (V_1, \ldots, V_n)^T$ and taking $W^{-1} = \text{diag}\,(\beta_0\gamma_j + (1-\beta_0)\beta_j)$, the appropriate bias ($B$) and standard error ($\sigma$) of β_0 given by the equations above are

$$|B| \simeq (1-\beta_0)\,\Sigma\, V_i\beta_i/\Sigma V_i\gamma_i$$

$$\sigma \simeq \frac{1}{\sqrt{\mu}}\,[(1-\beta_0)\,\Sigma V_i^2\beta_i + \beta_0\,\Sigma V_i^2\gamma_i]^{1/2}/\Sigma F_i\gamma_i$$

From these expressions we may make some observations that agree with observations made by Currie on the basis of empirical experiments: (1) The bias is proportional to the background proportion; (2) For small values of β_0 the standard error is proportional to the square root of the background proportion; (3) Since $\Sigma V_i^2\beta_i$ is typically less than $\Sigma V_i^2\gamma_i$, the standard error increases with increasing peak area proportion.

We conclude this section with a brief consideration of the problem of mis-specification of Γ. Suppose that the true peak profile is $\Gamma_0 = \Gamma+\delta\Gamma$; from calculations similar to those done above for the bias, we find that the additional bias introduced by $\delta\Gamma$ is

$$G^{-1}\Theta^T\,[I-R^{-1}(W-\Theta G^{-1}\Theta^T)]\delta\Gamma\beta_1$$

which, as $\mu\to\infty$, tends to

$$(\Gamma^T D\Gamma)^{-1}\,(\Gamma^T D\delta\Gamma)\beta_1.$$

In the single peak case, the Cauchy-Schwarz inequality shows that this quantity is bounded in absolute value by $\beta_1\|U_1\delta\Gamma\|/\|U_1\Gamma\|$. Thus a variation $\delta\Gamma$ such that $U_1\delta\Gamma$ is highly correlated with $U_1\Gamma$ will give rise to a relatively large bias proportional to the peak amplitude.

3. Choosing λ

If the parameter α^2 is known, the minimax λ is $\lambda = 1/\mu\alpha^2$. In the absence of this knowledge, λ must be chosen from the data. In this section we discuss a class of such procedures and illustrate them with examples.

Given a non-negative definite matrix B, one might attempt to choose λ to minimize

$$E(\hat{Y}(\lambda) - EY)^T B\,(\hat{Y}(\lambda) - EY) = ET_B(\lambda)$$

where

$$\begin{aligned}\hat{Y}(\lambda) &= A(A^TWA + \lambda U^TU)^{-1}A^TWY\\ &= A(\lambda)Y.\end{aligned}$$

$ET_B(\lambda)$ is a weighted mean-square error. This quantity may be estimated from the data by using

$$\begin{aligned}RSS_B(\lambda) &= (Y - \hat{Y}(\lambda))^T B\,(Y-\hat{Y}(\lambda)).\\ &= Y^T\,(I-A(\lambda))^T B\,(I-A(\lambda))Y\\ &= Y^TGY\ .\end{aligned}$$

The expectation of $RSS_B(\lambda)$ can be computed to be

$$ERSS_B(\lambda) = ET_B(\lambda) + \mu^{-1}tr\,(BW^{-1}) - 2\mu^{-1}\,tr\,(BA(\lambda)W^{-1})$$

and thus an unbiased estimate of $ET_B(\lambda)$ is

$$\hat{T}_B(\lambda) = RSS_B(\lambda) - \mu^{-1}tr\,(BW^{-1}) + 2\mu^{-1}\mathrm{tr}\,(\mathrm{BA}(\lambda)\mathrm{W}^{-1})\,.$$

We note that if Y follows a Gaussian distribution, then

$$Var\hat{T}_B(\lambda) = 2\,\mu^{-2}tr\,(CW^{-1})^2 + 4\,\mu^{-1}\,(A\beta)^TGW^{-1}A\beta$$

For a given B we propose choosing λ to minimize $\hat{T}_B(\lambda)$. (Similar procedures with $B=\mathrm{I}$ have been discussed in [3, 6].)

If it were possible, we might choose B so that $ET_B(\lambda) = E\|\beta_1 - \beta_1(\lambda)\|^2$, the total mean square error of the estimates of the peak amplitudes. However, if we write

$$EY = [\Gamma : I]\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix}$$

$ET_B(\lambda)$ may be expressed as

$$\begin{aligned}ET_B(\lambda) =& E(\beta_1 - \hat{\beta}_1(\lambda))^T\Gamma^TB\Gamma\,(\beta_1 - \hat{\beta}_1(\lambda))\\ &+ E(\beta_2 - \hat{\beta}_2(\lambda))^TB\,(\beta_2 - \hat{\beta}_2(\lambda))\\ &+ 2E(\beta_1 - \hat{\beta}_1(\lambda))^T\Gamma^TB\,(\beta_2 - \hat{\beta}_2(\lambda))\end{aligned}$$

from which it is apparent that it is imposible to choose B so that the second two terms vanish and the first does not.

We have experimented with three choices of B: $B_1=I$, $B_2=\Gamma(\Gamma^T\Gamma)^{-1}\Gamma^T$ and $B_3=\Gamma(\Gamma^T\Gamma)^{-2}\Gamma^T$. B_2 is the matrix which projects onto the column space of Γ; the motivation for choosing B_2 is that $\beta_2 - \hat{\beta}_2(\lambda)$ will hopefully not be highly correlated with the columns of Γ and thus the second two terms will be small and the first term will dominate. Choosing B_3 reduces the first term to $E\|\beta_1 - \hat{\beta}_1(\lambda)\|^2$ and hopefully causes the other terms to be small. A disadvantage in using B_2 or B_3 is that if there are two or more peaks with considerable overlap, the variance of $\hat{T}_B(\lambda)$ may be rather large, causing the procedure to be rather unstable.

Currie suggests choosing λ so that $RSS_W(\lambda) = n/\mu$. The motivation for this is that $\mu\cdot RSS_W$ would follow a chi-square distribution with n degrees of freedom if $E\hat{Y}(\lambda) = EY$ and no parameters were estimated from the data. In fact, however, parameters have been estimated from the data, although it is not clear how many "degrees of freedom" remain, and $E\hat{Y}(\lambda) \neq EY$. Thus the application of the x^2 distribution is questionable. The procedure outlined above with $B=W$ would choose λ to minimize

$$\hat{T}_W(\lambda) = RSS_W(\lambda) - n\mu^{-1} + 2\mu^{-1}\,tr\,A(\lambda)$$

which would cause $RSS_W(\lambda)$ to be somewhat smaller than n/μ. (In a vague sense, the "degrees of freedom" of the Chi-square statistic are reduced.)

We now briefly discuss the results of some simulations of this technique. The configurations are the following: (1) two slightly overlapping peaks on a linear baseline, (2) the same peaks on a quadratic baseline, (3) two highly overlapped peaks on a quadratic baseline, and (4) a single peak on a quadratic baseline which also contains a small "unsuspected" peak obscured by the dominant peak. All the simulations were done over a width of 20 channels with a total count $\mu = 10^5$. The sum of squared second differences was used as the smoothness measure. Computations were done on the Univac 1100 at the National Bureau of Standards. Subroutines from the IMSL library were used to generate random numbers and for matrix calculations. The most numerically sensitive calculation is the inversion of the matrix $A^TWA + \lambda U^TU$, which in theory is

positive definite; however, the matrix may be for practical purposes numerically singular for very small or very large values of λ, so it is important that a good algorithm be used and that diagnostic messages be printed when instabilities arise. (An alternative to actually forming and inverting this matrix is to simultaneously diagonalize A^TWA and U^TU; having done this once, $(A^TWA + \lambda U^TU)^{-1}$ may be computed quite rapidly for various values of λ.)

1. Two peaks on a linear baseline; the peak shapes were Gaussian with locations at channels 8 and 12 and standard deviations 1.5. Each peak contained 30 percent of the total area. The baseline was $\beta_j = c(1 + j)$ where c was chosen so that the baseline area was 40 percent. For this configuration the optimal (minimum variance unbiased) method of peak area estimation is weighted linear least squares; we are interested in seeing what "price" has to be paid for the additional flexibility of the smoothing method in this null case. Table 1a shows the bias, variance, and total mean square error of the peak area estimates for various values of λ. From the table we see that ETB decreases as λ increases (for λ greater than 10^7 numerial problems develop). For $\lambda = 10^5$ the variance is very close to that for the linear least squares.

TABLE 1a.

λ	Bias β_{11}	Var β_{11}	Bias β_{12}	Var β_{12}	Total MSE	$ETB_1 \times 10^5$	$ETB_2 \times 10^5$	$ETB_3 \times 10^5$
10^0	0	0.559(−4)	0	0.593(−4)	0.115(−3)	0.664	0.180	0.985
10^1	0	.315(−4)	0	.345(−4)	.660(−4)	.404	.178	.975
10^2	0	.108(−4)	0	.126(−4)	.234(−4)	.278	.176	.963
10^3	0	.692(−5)	0	.772(−5)	.146(−4)	.237	.175	.956
10^4	0	.519(−5)	0	.596(−5)	.111(−4)	.217	.174	.952
10^5	0	.487(−5)	0	.575(−5)	.106(−4)	.214	.174	.951
least squares ($\lambda = \infty$)	0	.486(−5)	0	.575(−4)	.106(−4)			

Table 1b shows the results for one realization with random Poisson noise added. As stated above, the total count was 10^5. $\hat{TB}_1$ is minimized at $\lambda = 10^3$ and $\hat{TB}_2$ and $\hat{TB}_3$ are minimized at $\lambda = 10^5$. (In this and in the later simulations in which noise was added, the weighting matrix W was estimated from the data.)

TABLE 1b.

λ	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$TB_1 \times 10^5$	$TB_2 \times 10^5$	$TB_3 \times 10^5$
10^0	0.295	0.298	0.692	0.180	0.984
10^1	.295	.299	.515	.178	.972
10^2	.297	.302	.389	.173	.948
10^3	.299	.302	.340	.171	.942
10^4	.299	.300	.370	.171	.933
10^5	.299	.300	.381	.170	.932

2. Two peaks on a quadratic baseline—the peaks were as above and the background was $\beta_j = c(1 + j + j^2/20)$ above c was chosen so that $\Sigma\beta_j = 0.4$. This shape deviates only slightly from a linear baseline. Table 2a exhibits the biases, variance, and total mean square error for various values of λ; as λ increases the variance decreases and the bias increases. For this discretization the minimum total mean square error occurs for $\lambda = 350$ ($MSE = .17 \times 10^{-4}$). The mean square error for the least squares method is much larger, being dominated by the bias ($MSE = 0.42 \times 10^{-3}$). The minima of ETB_1, ETB_2, and ETB_3 occur at $\lambda = 250$, 450, and 550 respectively, over which range the MSE does not change appreciably.

Table 2b summarizes the results of a single realization with random Poisson noise. $\hat{TB}_1$, $\hat{TB}_2$, and $\hat{TB}_3$ are minimized at $\lambda = 350$, 250 (or 350), and 350, respectively. It is noteworthy that the estimates do not change substantially over the tabulated range of λ. Other realizations gave similar results.

For this example there is little difference in the results for B_1, B_2, or B_3—any choice would give satisfactory results. B_1 is somewhat easier to compute.

TABLE 2a.

λ	Bias β_{11}	Bias β_{12}	Var β_{11}	Var β_{12}	Total $MSE \times 10^4$	$ETB_1 \times 10^5$	$ETB_2 \times 10^5$	$ETB_3 \times 10^5$
50	−0.327(−4)	−0.291(−4)	0.144(−4)	0.169(−4)	0.312	0.304	0.174	0.949
150	−.649(−4)	−.185(−4)	.899(−5)	.111(−4)	.201	.272	.173	.944
250	.245(−3)	.468(−3)	.787(−5)	.976(−5)	.179	.266	.1723	.942
350	.448(−3)	.750(−3)	.740(−5)	.911(−5)	.1727	.267	.17219	.9413
450	.658(−3)	.102(−2)	.714(−5)	.868(−5)	.1728	.270	.17216	.94099
550	.867(−3)	.127(−2)	.696(−5)	.836(−5)	.177	.275	.17218	.94097
650	.107(−2)	.151(−2)	.684(−5)	.811(−5)	.184	.281	.17225	.94117
750	.128(−2)	.173(−2)	.673(−5)	.791(−5)	.193	.288	.1723	.942
850	.148(−2)	.195(−2)	.665(−5)	.774(−5)	.204	.296	.173	.942
950	.167(−2)	.215(−2)	.657(−5)	.759(−5)	.216	.304	.173	.943
least squares ($\lambda = \infty$)	.156(−1)	.128(−1)	.465(−5)	.579(−5)	4.19			

TABLE 2b.

λ	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$TB_1 \times 10^5$	$TB_2 \times 10^5$	$TB_3 \times 10^5$
50	0.300	0.305	0.256	0.173	0.945
150	.299	.303	.206	.1721	.9415
250	.299	.302	.196	.17190	.9405
350	.298	.302	.195	.17190	.9404
450	.298	.301	.199	.17196	.9406
550	.298	.301	.205	.1721	.9409
650	.298	.301	.213	.1722	.9412
750	.297	.300	.222	.1723	.9416
850	.297	.300	.232	.1724	.9421
950	.297	.300	.242	.173	.943

3. Two peaks on a quadratic baseline; the peaks were close enough together (centers 9, 11, $\sigma = 1.5$) so that there was no trough between them when they were superimposed. The peak areas were 0.3 and 0.3 again and the baseline was as in the previous example. On a grid of λ values spaced linearly by 150 the minimum *MSE* occurred at $\lambda = 800$ ($MSE = 0.20610 \times 10^{-4}$); the minimum of ETB_1 was at $\lambda = 350$ ($MSE = 0.213 \times 10^{-4}$); the minimum of ETB_2 was at $\lambda = 650$ ($MSE = 0.20611 \times 10^{-4}$); the minimum of ETB_3 was at $\lambda = 950$ ($MSE = 0.207 \times 10^{-4}$). The *MSE* for a linear least squares fit was 0.241×10^{-3}. Table 3 records the minimizing values of λ for $\hat{TB}_1$, $\hat{TB}_2$, and $\hat{TB}_3$, and the corresponding *MSE*'s for 4 realizations. The results suggest that $\hat{TB}_1$ may be a more stable criterion function in this situation, but we would not wish to make a conclusion on the basis of a sample size of 4!

TABLE 3. *Minimizing values of* λ *and corresponding MSE's for four realizations.*

	TB_1	TB_2	TB_3
1	50(.279 × 10⁻⁴)	1400(.217 × 10⁻⁴)	1300(.217 × 10⁻⁴)
2	500(208 × 10⁻⁴)	3000(288 × 10⁻⁴)	2150(246 × 10⁻⁴)
3	500(208 × 10⁻⁴)	1100(210 × 10⁻⁴)	2600(267 × 10⁻⁴)
4	950(207 × 10⁻⁴)	5000(407 × 10⁻⁴)	6500(501 × 10⁻⁴)

4. A single peak (center = 10, $\sigma = 2$) on a quadratic baseline with a hidden peak centered at 12 with standard deviation 2. The peak area of the dominant peak was 0.8 and the area of the hidden peak was 0.02. In an attempt to mimic a situation in which the hidden peak is unsuspected, a single peak model was fit. The behaviors of ETB_1, ETB_2, and ETB_3 were somewhat different. ETB_1 had a minima at $\lambda = 10$ ($MSE = 0.55 \times 10^{-4}$) whereas ETB_2 and ETB_3 had minimum at $\lambda = 10^4$ ($MSE = 0.96 \times 10^{-4}$). The *MSE* was minimum at $\lambda = 10^7$ ($MSE = 0.18 \times 10^{-4}$). The *MSE* of the linear least squares procedure was 0.21×10^{-4}. The reason that ETB_1 was minimized for a smaller value of λ is that this criterion gives greater weight

to fitting the baseline as well as the peak than do the other two, which concentrate more on the peak. The baseline (which includes the hidden peak) is fit well with small values of λ since it is not very smooth. Since the hidden peak has substantial correlation with the modelled peak, however, B_2 and B_3 fail to choose λ large enough.

On several realizations with random noise $\hat{T}B_1$ achieved a minimum at small values of λ and $\hat{T}B_2$ and $\hat{T}B_3$ at larger values of λ. On some occasions $\hat{T}B_2$ and $\hat{T}B_3$ also had local minima at small values of λ. Figure 1 shows the estimated baseline for $\lambda = 20$, which was the attained minimum for $\hat{T}B_1$ on a particular realization. The unsuspected peak shows quite clearly, giving valuable diagnostic information! The estimated baseline for the larger value of $\lambda = 10^4$ at which $\hat{T}B_2$ and $\hat{T}B_3$ were minimized smooths over the peak (fig. 2). We also plotted residuals on a square root scale to stabilize the variance, $r_i = \sqrt{y_i} - \sqrt{\hat{y}_i(\lambda)}$. Figure 3 shows the residual plot for $\lambda = 10^4$; there is a hint of a discrepancy near channel 12.

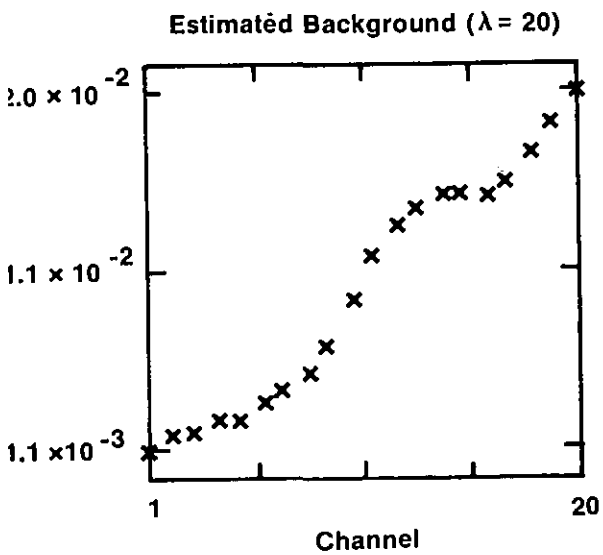

FIGURE 1.

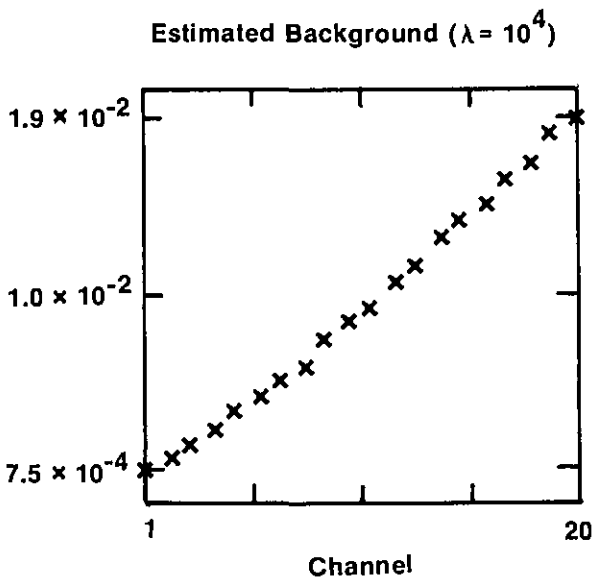

FIGURE 2.

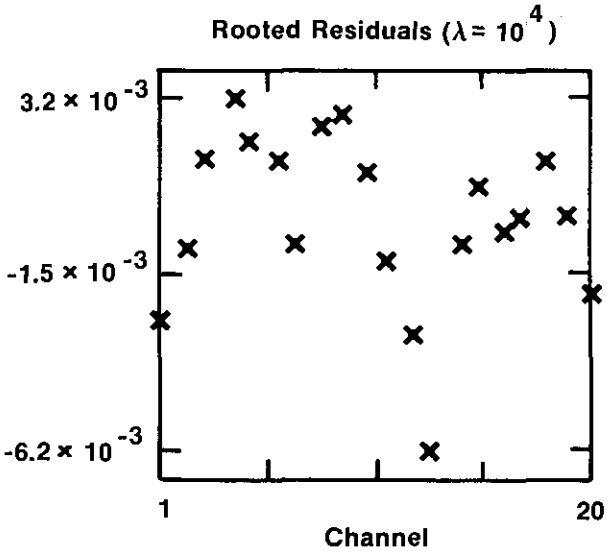

FIGURE 3.

If the hidden peak is incorporated into the model, the total *MSE*, ETB_1, ETB_2 and ETB_3 are all minimized for $\lambda \approx 10^3$. The total *MSE* is 0.36×10^{-4} and the individual *MSE*'s are 0.20×10^{-4} and 0.16×10^{-4} for the large and small peaks respectively. The bias and variance for the small peak are 0.99×10^{-3} and 0.15×10^{-4} so that the relative error in estimating this peak area is quite large. For the linear least squares method the total *MSE* is 0.13×10^{-3}; the bias and variance for the small peak are $.36 \times 10^{-2}$ and $.12 \times 10^{-4}$.

On the basis of these computations there is no clear evidence that would favor B_2 or B_3 over B_1, despite the fact that they were designed to focus more on the peak. The last example shows that focusing on the peak may hide unsuspected features of the baseline. The computations suggest that choosing λ to minimize $\hat{T}B(\lambda)$ is reasonable, but they are not nearly extensive enough to give insight into the stochastic behavior of the minimizing λ.

There are many possibilities we have not investigated. Other choices of B are possible; for example $B = \Gamma_j((\Gamma_j^T\Gamma_j)^{-1}\Gamma_j^T$ would focus on the jth peak if there were more than one peak, $B = W^{-1}$ would weight the deviations according to the variances of the observed counts; a possible advantage of this choice is that the statistics $RSS_W(\lambda)$ might be compared with the percentiles of a χ^2 distribution (above, however, we have noted some difficulties with this procedure). Another possibility is to attempt to choose between several smoothness criteria by computing $\hat{T}B^{(k)}(\lambda)$ for $k = 1, 2, 3, \ldots, K$ and choosing the solution corresponding to

$$\min_k \inf_\lambda \hat{T}B^{(k)}(\lambda) .$$

4. Final Comments

The results above leave several questions unanswered and suggest problems for further research. The following is perhaps the most immediate: in many applications the peak vector is not known exactly, but is assumed to have a parametric form such as $\gamma_j = \gamma_j(\mu,\sigma) = \frac{1}{\sigma}\gamma\frac{j-\mu}{\sigma}$, where γ is a given function μ and σ are location and shape parameters and must be estimated from the data. If the peak profile Γ is estimated from other experiments, for example from pure sources, the variability of the estimate will affect subsequent analyses in which it is used. We plan to pursue the analysis of these problems in the future.

An alternative approach to the problem is to use the method of maximum likelihood with the assumption of Poisson statistics; which might be more appropriate for small counts. The likelihood function of β could be maximized subject to the constraint $\|U\beta\|^2 = \alpha^2$. Although we conjecture that the large sample properties of the estimates would be equivalent to the results above, the small sample properties would be different.

Finally, we note again that in the multi-peak situation the estimates we have considered are minimax for any single peak amplitude but are probably not jointly minimax. One might attempt to solve the simultaneous minimax problem by numerical optimization; we conjecture that the results would not be substantially different.

5. References

[1] Cullum, J. Ill-posed problems, regularization, and the singular value decomposition. IBM Tech. Rep. RC 6465; 1977.

[2] Currie, L. Model uncertainty and bias in the evaluation of nuclear spectra. J. Radioan. Chem. **39**(1–2): 223–237; 1977.

[3] Golub, G.; Heath, M.; Wahba, G. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics **21**(2): 215–223; 1979 May.

[4] Greub, W. *Linear Algebra*, 2nd Ed. New York: Springer-Verlag; 1963. 338 p.

[5] Kuks, J.; Olman, V. *A minimax estimator of regression coefficients. (Russian).* Izv. Akad. Nauk. Eston. SSR **21**: 66–72; 1972.

[6] Mallows, C. L. Some comments on Cp. Technometrics **15**(4): 661–675; 1973 November.

[7] Rao, C. R. *Linear statistical inference and its applications*. New York: Wiley; 1965. 624 p.

[8] Reinsch, C. Smoothing by spline functions. Numer. Math. **10**(3): 177–183; 1967 October, 4.

[9] Speckman, P. Minimax estimates of linear functionals in a Hilbert space. Ann. Stat., to appear. 1981.

[10] Speckman, P. On minimax estimation of linear operators in Hilbert spaces from noisy data. Manuscript. 1980.

[11] Whittaker, E. and Robinson, G. The calculus of observations [sec. 151–155]. London: Blackie and Sons; 1944.

6. Appendix

Here we derive an expression for the covariance matrix of $\hat{\beta}_1$ and prove the lemma in section 2 of the text. The covariance matrix of $\hat{\beta}$ is, with the notation of section 2,

$$\begin{aligned}\mu\Sigma &= (A^TWA+\lambda U^TU)^{-1}A^TWA(A^TWA+\lambda U^TU)^{-1} \\ &= \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22}\end{bmatrix}\begin{bmatrix} G & \theta^T \\ \Theta & W\end{bmatrix}\begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22}\end{bmatrix}\end{aligned}$$

We are interested in Σ_{11}. Multiplying through and noting that $B_{21} = B_{12}^T$

$$\begin{aligned}\mu\Sigma_{11} &= B_{11}GB_{11} + B_{12}\Theta B_{11} + B_{11}\Theta^TB_{12}^T + B_{12}WB_{12}^T \\ &= B_{11}\Gamma^TW\Gamma B_{11} + B_{12}W\Gamma B_{11} + B_{11}\Gamma^TWB_{12} + B_{12}^TWB_{12} \\ &= (W^{1/2}\Gamma B_{11} + W^{1/2}B_{12}^T)^T(W^{1/2}\Gamma B_{11} + W^{1/2}B_{12}^T) \\ &= F^TF\ .\end{aligned}$$

Now, using the expressions for B_{11} and B_{12}, and $\Gamma = W^{-1}\Theta$

$$\begin{aligned}F &= W^{1/2}(\Gamma G^{-1} + \Gamma G^{-1}\Theta^TR^{-1}\Theta G^{-1} - R^{-1}\Theta G^{-1}) \\ &= W^{1/2}(W^{-1} + W^{-1}\Theta G^{-1}\Theta^TR^{-1} - R^{-1})\,\Theta G^{-1} \\ &= W^{-1/2}[I - (W-\Theta G^{-1}\Theta^T)R^{-1}]\,\Theta G^{-1}\ ,\end{aligned}$$

which is the expression to be derived.

We now prove the lemma. The key to the proof is the fact that under the assumptions of the lemma C and D may be simultaneously diagonalized [4]; there exists a nonsingular matrix X such that

$$\begin{aligned}X^TCX &= \Omega \\ X^TDX &= M\end{aligned}$$

where Ω and M are diagonal matrices with elements ω_i and μ_i. From this representation we note that the null space of C (resp. D) is spanned by those columns of X corresponding to zero diagonal elements of Ω (resp. M). The assumption of the lemma guarantees that the two null spaces contain no vectors in common. Now expressing C and D in terms of X, Ω, and M, and writing $I = XX^{-1}$

$$I - (C+\lambda D)^{-1}C = X\,[I-(\Omega+\lambda M)^{-1}\Omega]\,X^{-1}$$
$$= XR_\lambda X^{-1}$$

where $R_\lambda = \text{diag}\,[\lambda\mu_i/(\omega_i+\lambda\mu_i)]$.

We note that if $\beta_2 \varepsilon N(U_1) = N(D)$ this representation makes it clear that β_1 is unbiased, for if x_j is a column of X corresponding to $\mu_j = 0$, then

$$XR_\lambda X^{-1}x_j = X\,\frac{\lambda\mu_j}{\omega_j+\lambda\mu_j}\,e_j = 0$$

where e_j is the j^{th} unit vector.

The diagonal elements of R_λ corresponding to $\omega_j = 0$ are 1's, so that

$$XR_\lambda X^{-1} = X\begin{pmatrix}0 & 0\\ 0 & I\end{pmatrix}X^{-1} + \lambda X\begin{pmatrix}N_\lambda & 0\\ 0 & 0\end{pmatrix}X^{-1}$$

where $N_\lambda = \text{diag}\,[\mu_i/(\omega_i+\lambda\mu_i)]$. It is easily verified that the first matrix, call it P, on the right hand side of the expression above has the following properties: (1) it is idempotent with range $N(C)$; (2) $Pv = 0$ if $v \varepsilon N(D)$; (3) for any vector v, $(Pv)^T D(I-P)v = 0$. P is therefore a projection matrix which projects orthojonally with respect to the pseudo inner-product $\langle u,v\rangle = u^T Dv$, and may be written

$$P = V(V^T DV)^{-1}V^T D$$

where $V = (v_1, \ldots, v_p)$ spans the null space of C. Finally noting that N_λ is bounded, we have

$$XR_\lambda X^{-1} = P + 0(\lambda)$$

Finally, we note that expansions for small values of λ (corresponding to large samples) or small values of λ^{-1} (corresponding to a nearly linear background and moderate sample size) may be carried using identities of the form

$$\frac{1}{1+\varepsilon} = 1 - \varepsilon + \frac{\varepsilon^2}{(1+\varepsilon)}.$$