Part II*

1. The polynomial smoothing spline.

2. Leaving-out-one, GCV and other smoothing parameter estimates.

3. The thin plate smoothing spline.

4. Generalizations: Different kinds of observations: Non-gaussian, indirect, constrained.

5. Examples: The histospline, convolution equations with positivity constraints. GCV with inequality constraints.

*Part II of 'An Introduction to Model Building With Reproducing Kernel Hilbert Spaces', by Grace Wahba, Univ. of Wisconsin Statistics Department TR 1020, Overheads for Interface 2000 Short Course.

♣♣ The Polynomial Smoothing Spline

- The polynomial smoothing spline is the forerunner of much more general RKHS models.

Let W_m be the collection of functions on $[0, 1]$ with $\int_0^1 (f^{(m)}(u))^2 du \leq \infty$. The polynomial smoothing spline is the solution to the problem: Find $f \in W_m$ to min

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - f(t(i)))^2 + \lambda \int_0^1 (f^{(m)}(u))^2 du.$$

It can be shown that $W_m = W_m^0 \oplus \pi_m$, where π_m is the span of the polynomials of degree m or less. We rearrange things so that

$$W_m = \mathcal{H}_0 \oplus \mathcal{H}_1$$

where $\mathcal{H}_0 = \pi_{m-1}$, the polynomials of degree $m-1$ or less, on which there will be no penalty, and $\mathcal{H}_1 = W_m^0 \oplus \{k_m\}$. It can be shown that the RK for $\mathcal{H}_1$ with square norm $\|f\|^2 = \int_0^1 (f^{(m)}(u))^2 du$ is

$$K(s, t) = k_m(s)k_m(t) + (-1)^m k_{2m}([s - t]).$$

By an argument generalizing the representer theorem, and upon observing that $\{k_\nu\}_{\nu=0}^{m-1}$ span $\mathcal{H}_0$, it follows that the minimizer f_λ has the form

$$f_\lambda(t) = \sum_{\nu=1}^{m} d_\nu k_{\nu-1}(t) + \sum_{i=1}^{n} c_i K(t(i), t), \quad (1)$$

and that

$$\int_0^1 (f^{(m)}(u))^2 du = \sum_{i,j=1}^{n} c_i c_j K(t(i), t(j)). \quad (2)$$

Upon substituting (1) and (2) into the original variational problem, the solution is obtained by minimizing a quadratic form in $d = (d_1, \cdots, d_m)'$ and $c = (c_1, \cdots, c_n)'$. (There are easier ways to get the polynomial spline, but the present way of going about it is the one which we see will generalize in many ways.)

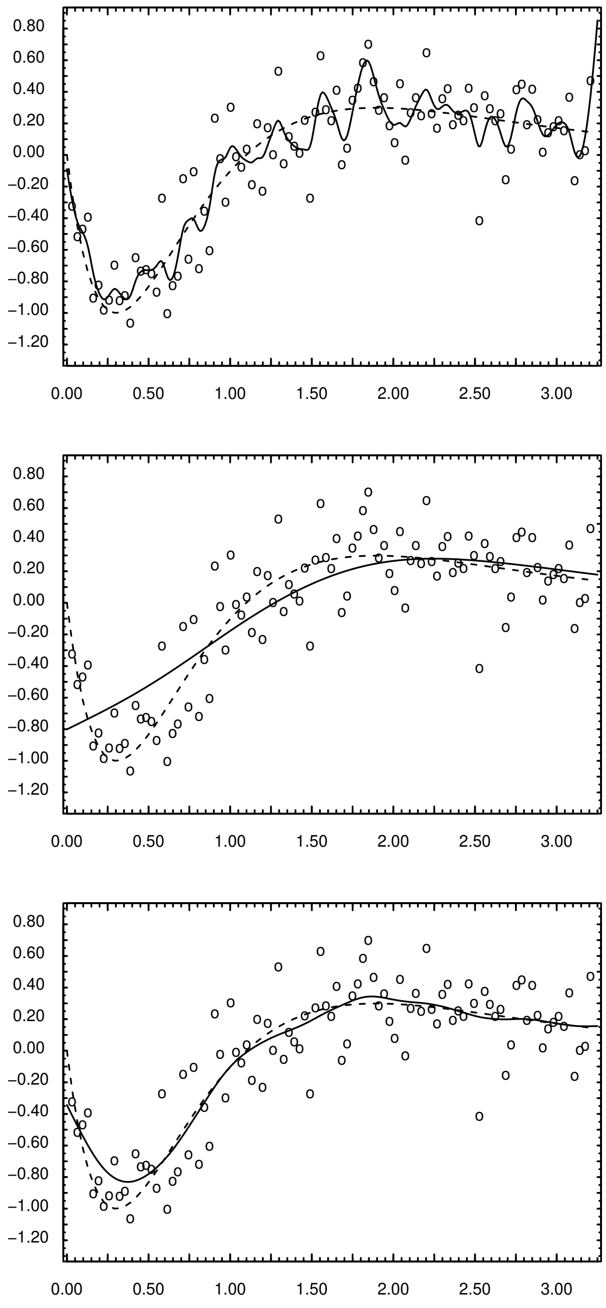


Figure 1: Dashed Lines are Smoothing Splines with λ too small, λ too large, and λ estimated via GCV, from the top. Solid line is[1] ‘truth’.

♣♣ Choosing λ.

• Leaving-out-one.

Let $f_\lambda^{[k]}(\cdot)$ be the minimizer of

$$\frac{1}{n}\sum_{\substack{i=1\\ i\neq k}}^{n}(y_i - f(t(i)))^2 + \lambda\int_0^1 (f^{(m)}(u))^2 du.$$

The leaving-out-one estimate of λ is the minimizer of

$$V_0(\lambda) = \frac{1}{n}\sum_{k=1}^{n}(y_k - f_\lambda^{[k]}(t(k)))^2.$$

• GCV (Generalized Cross Validation).

The influence matrix $A(\lambda)$ with kkth entry $a_{kk}(\lambda)$ plays an important role. The influence matrix relates the data to the predicted data:

$$\begin{pmatrix} f_\lambda(t(1)) \\ f_\lambda(t(2)) \\ \vdots \\ f_\lambda(t(n)) \end{pmatrix} \equiv A(\lambda)y.$$

• GCV (continued)

We have the Lemma:

$$V_0(\lambda) \equiv \frac{1}{n} \sum_{k=1}^{n} \left(\frac{(y_k - f_\lambda(t(k))}{(1 - a_{kk}(\lambda))} \right)^2$$

The GCV estimate of λ:

$$\begin{aligned} min\, V(\lambda) &= \frac{\frac{1}{n}\sum_{k=1}^{n}(y_k - f_\lambda(t(k)))^2}{(1 - \frac{1}{n}\sum_{\ell=1}^{n} a_{\ell\ell}(\lambda))^2}. \\ &\equiv \frac{\|(I - A(\lambda)y\|^2}{\frac{1}{n}(I - trA(\lambda)))^2} \end{aligned}$$

• Unbiased Risk (if you know σ^2).

$$min\, U(\lambda) = \|(I - A(\lambda)y\|^2 + 2\sigma^2 trA(\lambda)$$

• Generalized Max Likelihood (GML, aka REML).

$$min\, M(\lambda) = \frac{y'(I - A(\lambda))y}{[\det^+(I - A(\lambda))]^{1/(n-M)}}$$

$\det^+$ = product of the $n - M$ non-zero eigenvalues.

- Cubic smoothing spline with GCV, SOFTWARE.

Codes for cubic (or higher order) smoothing splines with GCV to choose the smoothing parameter. Approximately reverse chronological order.

...

Code- Author- Where Found $(* = freeware)$

—- —– —–

$*$ pspline-Jim Ramsay-`http:www.r-project.org`

smooth.spline()-Trevor Hastie-Splus

$*$ sbart-Finbarr O'Sullivan-`http://www.netlib.org/gcv`

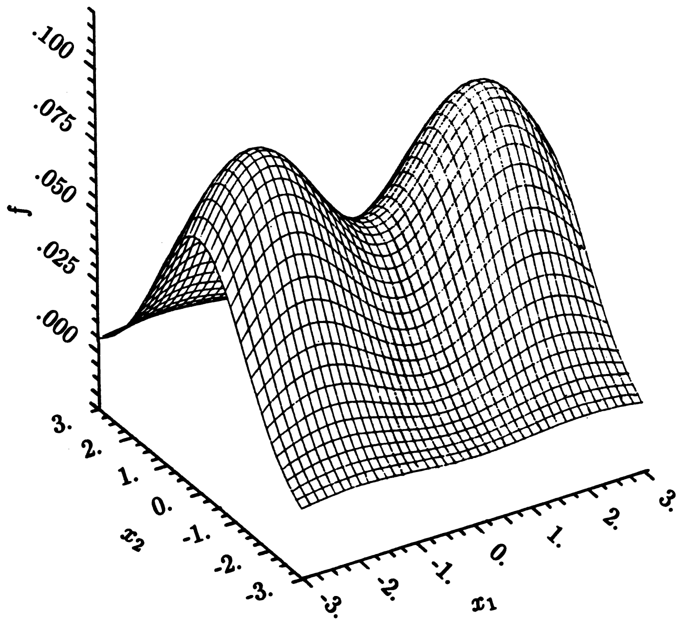
$*$ gcvspl-H. J. Woltring-`http://www.netlib.org/gcv`

♣♣ The Thin Plate Spline

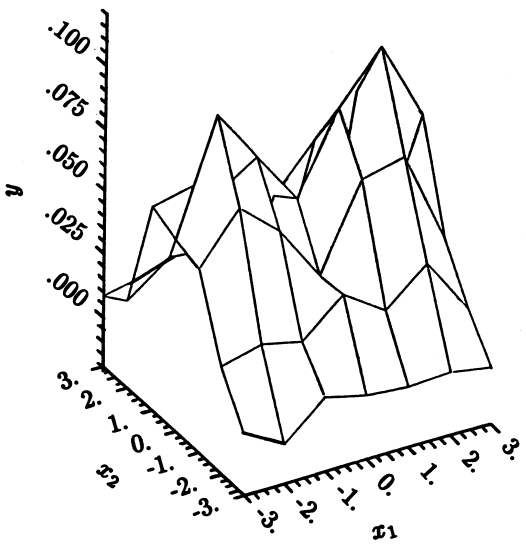
- The thin plate spline is one of the two-dimensional generalizations of the univariate spline.

Letting $t = (t_1, t_2)$, the penalty $\int (f^{(2)})^2$ is replaced by

$$J(f) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [f_{t_1 t_1}^2 + 2 f_{t_1 t_2}^2 + f_{t_2 t_2}^2] dt_1 dt_2.$$

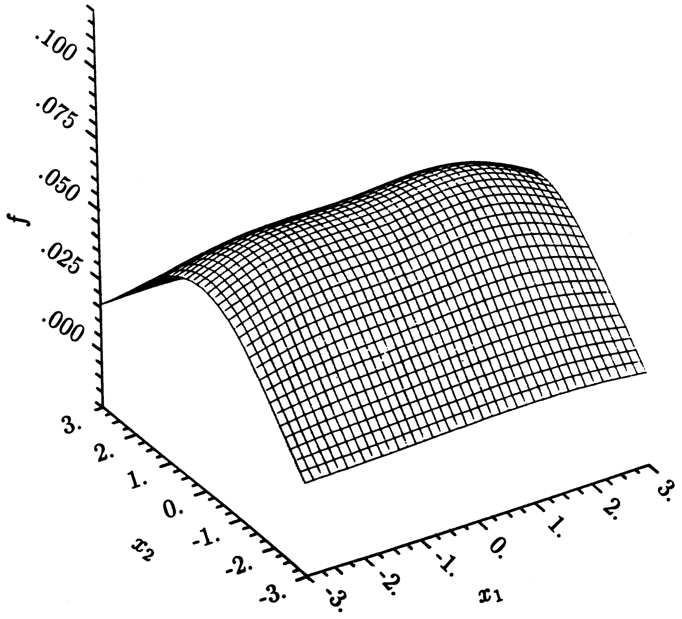


The actual surface.

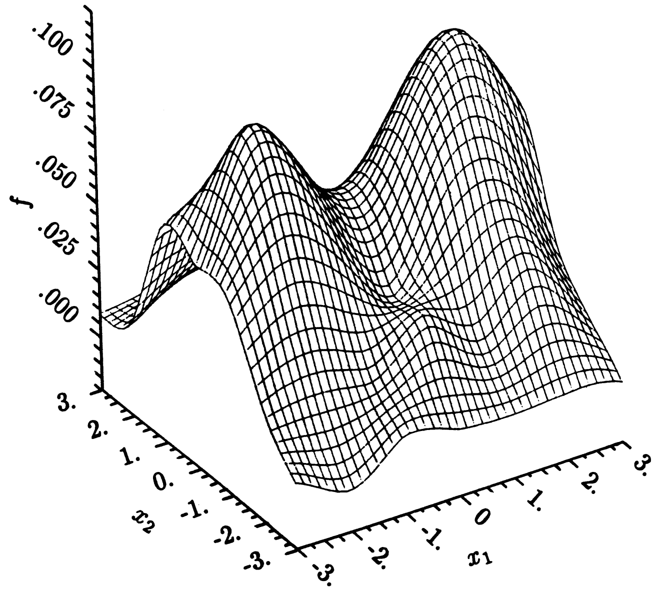


The data.

Figure 2a: Thin plate spline demo. Top: True surface. Bottom: The observations.



f_λ *with* λ *too large,* $\lambda = 100\hat{\lambda}$.



f_λ *with* λ *too small,* $\lambda = .01\hat{\lambda}$.

Figure 2b. Top: f_λ with λ too large. Bottom: f_λ with λ too small.

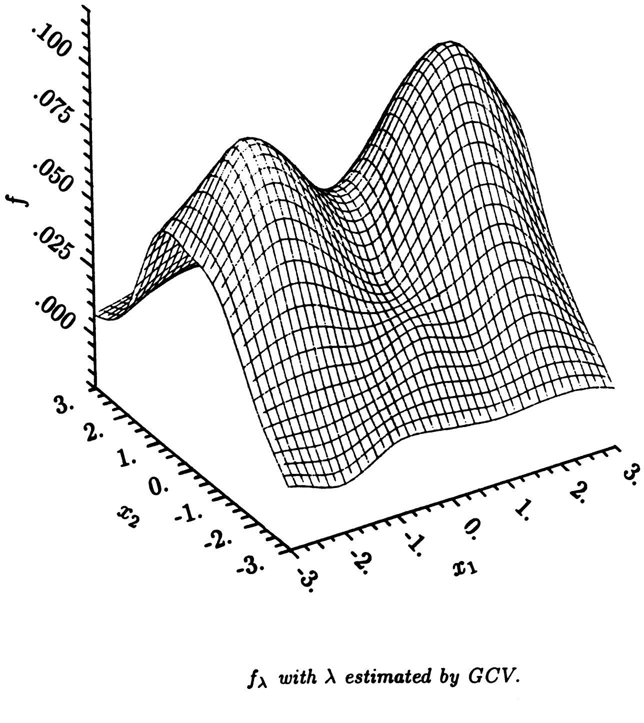


Figure 2c. $f_{\widehat{\lambda}}$ with λ estimated by GCV.

- Thin plate spline with GCV, SOFTWARE.

Codes for thin plate smoothing splines with GCV to choose the smoothing parameter. Approximately reverse chronological order.

...

Code- Author- Where Found $(* = freeware)$

—- —– —–

tpspline-Dong Xiang-SAS

$*$ funfits-Doug Nychka-`http://www.cdg.ucar.edu/` `/stats/software.shtml`

ANUSPLIN-M. Hutchinson-`http://cres20.anu.edu/` `/au/software/anusplin.html`

$*$ GCVPACK-Bates et al-`http://www.netlib.org/gcv`

♣♣ The Representer Theorem (more general case)

Let $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_0$ is a finite dimensional space spanned by $\phi_\nu, \nu = 1, \cdots M$, and $\mathcal{H}_1$ is an RKHS with square norm $\|f\|^2_{\mathcal{H}_K}$. Find $f = f_0 + f_1$ with $f_0 \in \mathcal{H}_0$ and $f_1 \in \mathcal{H}_1$ to min

$$I_\lambda(y, f) = \frac{1}{n} \sum_{i=1}^{n} g_i(y_i, f(t(i))) + \lambda \|f_1\|^2_{\mathcal{H}_K}.$$

Suppose g is convex in f and the minimizer of $\sum_{i=1}^{n} g_i(y_i, f(t(i)))$ in $\mathcal{H}_0$ is unique. Then the minimizer f_λ of I_λ is unique and has a representation

$$f(\cdot) = \sum_{\nu=1}^{M} d_\nu \phi_\nu(\cdot) + \sum_{i=1}^{n} c_i K_{t(i)}(\cdot).$$

where (d, c) minimize

$$\sum_{i=1}^{n} g_i(y_i, (Td + Kc)_i) + \lambda c' K c.$$

Here, $T_{n\times M} = \{\phi_\nu(t(i))\}$, $K_{n\times n} = \{K(t(i), t(j))\}$ and $(x)_i$ means the ith component of x.

♣♣ Quick List of Generalizations

- In the ‘distance’ of f from observations.

1. $g(y, f)$ = log likelihood

2. $g(y, f)$ = robust functional

3. $g(y, f)$ = support vector machine (SVM) functional

4. $g(y, f)$ = indicator functionals, e. g. $g(y, f) = 0$ or ∞ according as $f \in [y + u, y - l]$

- In the kinds of observations.
- In the imposition of constraints.
- In the domain of the model, $\mathcal{T} \rightarrow \mathcal{T}^{(1)} \otimes \cdots \otimes \mathcal{T}^{(d)}$

- In the kinds of observations: Integrals:

Replace $f(t)$ by $L_t f$ where the $L_t f$ are bounded linear fuctionals in $\mathcal{H}$: Example: Tomography.

$$L_t f = \int_{\mathcal{T}} H(t,u) f(u) du.$$

Then K_t is replaced by

$$\xi_t(\cdot) = \int_{\mathcal{U}} H(t,u) K(u,\cdot) du$$

and $< K_s, K_t >$ is replaced by

$$< \xi_s, \xi_t >= \int_{\mathcal{U}} \int_{\mathcal{U}} H(s,u) K(u,v) H(t,u) du du.$$

ξ_t is called the representer of L_t in $\mathcal{H}$,

$$L_t f \equiv < \xi_t, f >= \int_{\mathcal{T}} H(t,u) f(u) du.$$

Where does this come from?

- In the kinds of observations: The Eta Theorem:

Theorem: Let L be a bounded linear functional in an RKHS $\mathcal{H}_K$ with RK K. By the Riesz representation theorem there exists an η in $\mathcal{H}_K$ such that

$$Lf =< \eta, f >, \quad all\ f \in \mathcal{H}_K.$$

We may find η by observing that

$$\eta(s) =< K_s, \eta > .$$

Since

$$< K_s, \eta > \equiv LK_s,$$

the representer of any bounded linear functional in $\mathcal{H}_K$ may be found by applying the bounded linear functional to K_s, and then looking at the result as a function of s.

This allows us to estimate f based on observations on integrals and even derivatives if $\mathcal{H}_K$ is chosen so that these are bounded linear functionals. Derivatives up to the $m-1$st are bounded linear functionals in W_m.

- In the kinds of observations:The Representer Theorem (even more general case)

Let $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_0$ is a finite dimensional space spanned by $\phi_\nu, \nu = 1, \cdots M$, and $\mathcal{H}_1$ is an RKHS with square norm $\|f\|^2_{\mathcal{H}_K}$. Find $f = f_0 + f_1$ with $f_0 \in \mathcal{H}_0$ and $f_1 \in \mathcal{H}_1$ to min

$$I_\lambda(y, f) = \frac{1}{n}\sum_{i=1}^{n} g_i(y_i, L_i f) + \lambda\|f_1\|^2_{\mathcal{H}_K},$$

where the $\{L_i\}$ are bounded linear functionals on $\mathcal{H}$. Suppose g is convex in f and the minimizer of $\sum_{i=1}^n g_i(y_i, L_i f)$ over f in $\mathcal{H}_0$ is unique. Then the minimizer f_λ of I_λ is unique and has a representation

$$f(\cdot) = \sum_{\nu=1}^{M} d_\nu \phi_\nu(\cdot) + \sum_{i=1}^{n} c_i \xi_i(\cdot). \tag{3}$$

where ξ_i is the representer for L_i in $\mathcal{H}_1, L_i f \equiv < \xi_i, f >$ and (d, c) minimize

$$\frac{1}{n}\sum_{i=1}^{n} g_i(y_i, (Td + Kc)_i) + \lambda c' K c.$$

Here, $T_{n\times M} = \{L_i \phi_\nu\}$, $K_{n\times n} = \{< \xi_i, \xi_j >\}$.

•. The histospline. Given area integrals.

Female lung cancer rates in Wisconsin, by county.

$$y_i = \frac{1}{|\Omega_i|} \int_{\Omega_i} f(A) dA + \epsilon_i.$$

The thin plate penalty is used for the histospline.

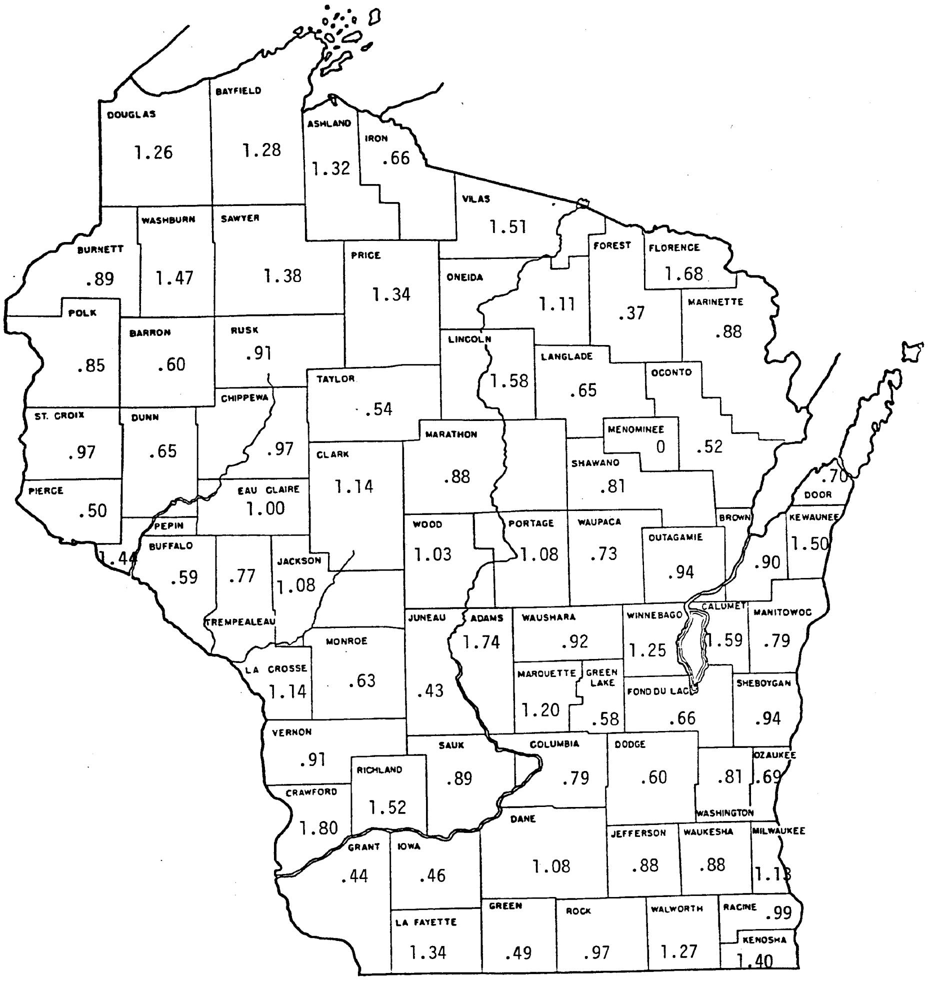


Fig. 3.13. 1970-1975 Female Lung Cancer, Revised SMR's by County

Figure 3a. Female lung cancer rates, by county.

-27-

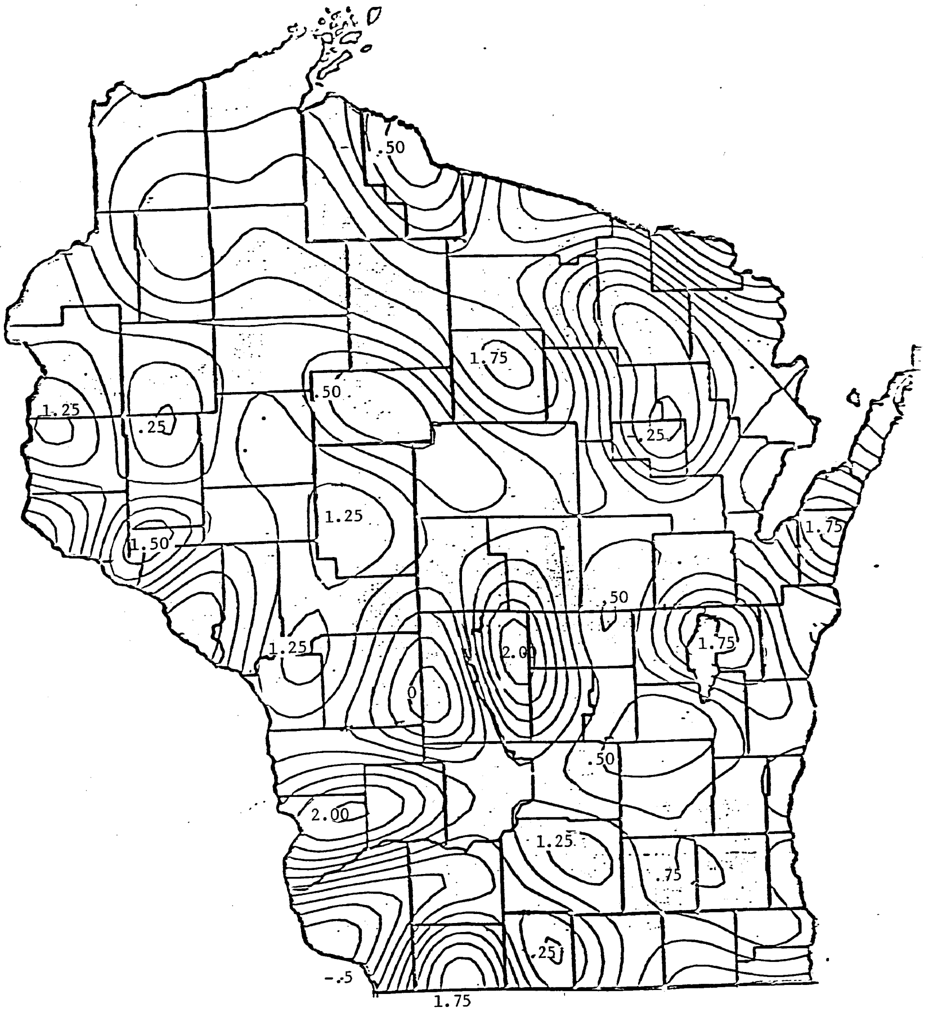


Figure 3.14. Female Lung Cancer Revised SMR's. Volume Matching Histospline. Contour Interval: 0.25.

Figure 3b. Volume matching. Min $J(f)$ subject to $\frac{1}{|\Omega_i|}\int_{\Omega_i}\widehat{f}(A)dA = y_i$

-28-

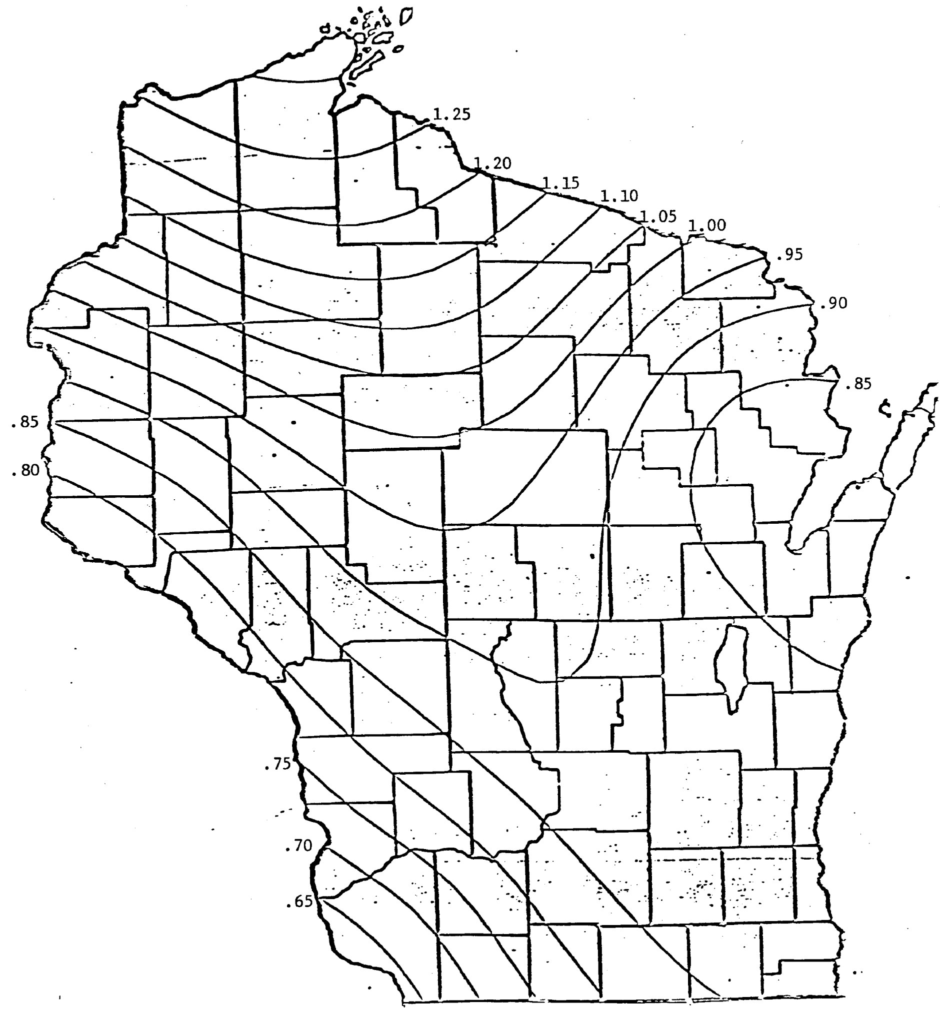


Figure 3.15. 1979-1975 Female Lung Cancer, Histospline Smoothed by GCV. Contour Interval: 0.05.

Figure 3c. Volume smoothing. Find $f_\lambda \in \mathcal{H}$ to min $\sum_i (y_i - \frac{1}{|\Omega_i|} \int_{\Omega_i} f(A) dA)^2 + \lambda J(f)$

- In the kinds of observations: Constraints

Let $\mathcal{H}$ as before. Find $f = f_0 + f_1$ with $f_0 \in \mathcal{H}_0$ and $f_1 \in \mathcal{H}_1$ to min

$$I_\lambda(y, f) = \frac{1}{n} \sum_{i=1}^{n} g_i(y_i, L_i f) + \lambda \|f_1\|^2_{\mathcal{H}_K}$$

subject to

1. Positivity: $f(t) \geq 0$
2. Linear inequality constraints: $N_t f \geq a_t$
3. Constraints via solutions to PDE's:
   $t = (time, space), H(\mathcal{L}f) \leq C$

To compute, the constraints are discretized. The representers of the constraints are incorporated in the representation of the solution. For inequality constraints, the coefficients are obtained by solving a mathematical programming problem. (MINOS) In typical cases where the family of constraints is 'smooth' the addition of a few constraints will lead to the constraints actually being satisfied everywhere.

Let $\{N_j\}$ be a finite set of discretized constraints, and let $\{\eta_j\}$ be their representers, $< \eta_j, f >= N_j f$. The problem then becomes Find $f = f_0 + f_1$ to min

$$I_\lambda(y, f) = \frac{1}{n} \sum_{i=1}^{n} g_i(y_i, L_i f) + \lambda \|f_1\|^2_{\mathcal{H}_K}$$

subject to

$$< \eta_j, f > \geq a_j, \quad j = 1, \cdots J,$$

and the solution has a representation

$$f(\cdot) = \sum_{\nu=1}^{M} d_\nu \phi_\nu(\cdot) + \sum_{i=1}^{n} c_i \xi_i(\cdot) + \sum_j \tilde{c_j} \eta_j(\cdot).$$

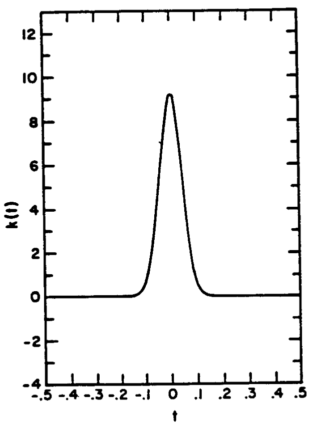


Figure 4a. The convolution kernel

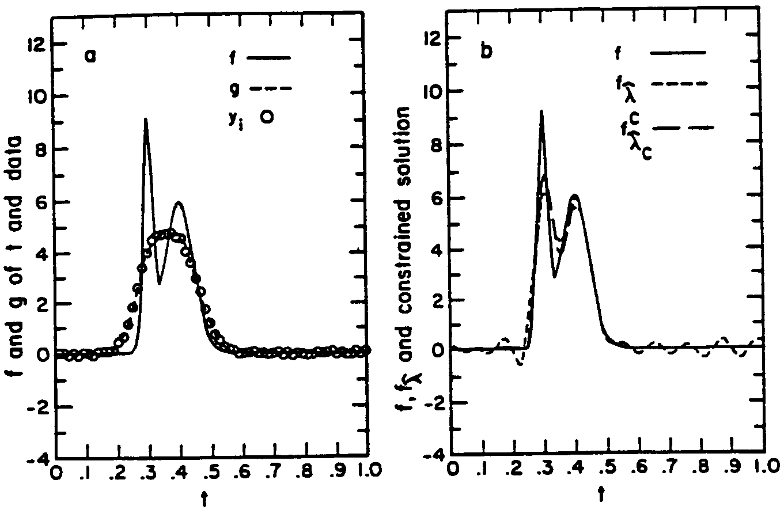
- Positivity constraints in a convolution equation.

$$y_i = \int k(v_i, u) f(u) du + \epsilon_i.$$

Find f_λ to min

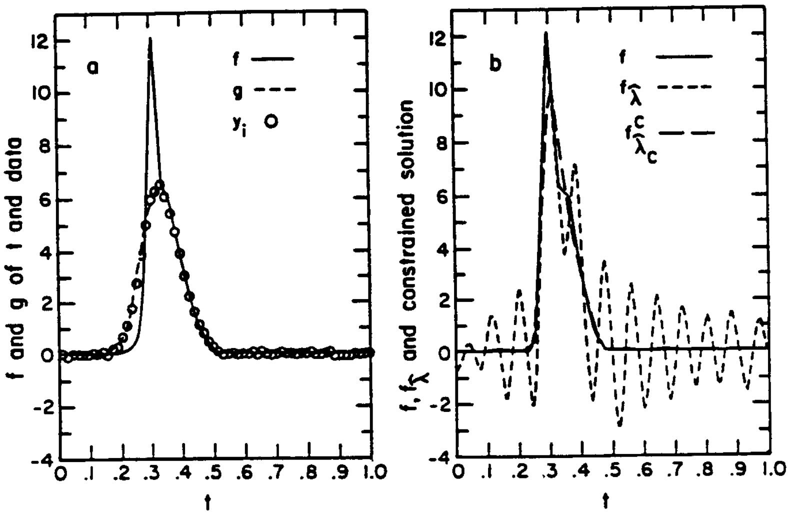
$$\frac{1}{n} \sum_i (y_i - \int k(v_i, u) f(u) du)^2 + \int (f^{(2)})^2,$$

subject to $f_\lambda(u_j) \geq 0$. The GCV for constrained problems: For fixed λ, solve the quadratic programming problem, and find the active constraints. At the solution, the same answer will be obtained by throwing away the inactive constraints and putting in the active inequality constraints as equality constraints. This problem is linear - compute the GCV for it.



f, g, data, $f_{\hat{\lambda}}$ and $f^{C}_{\hat{\lambda}_C}$

peak separation = .10.



f, g, $f_{\hat{\lambda}}$ and $f^{C}_{\hat{\lambda}_C}$

peak separation = .05.

Figure 4b. Two examples: Left panels-true f and observations y_i. Right panels-true f, unconstrained solution (wiggly) and constrained solution $f_{\hat{\lambda}}$.