Control and Cybernetics

vol. **23** (1994) No. 4

A cross–validation method for non–parametric estimation

by

B.A. Mair*

Department of Mathematics,
University of Florida
201 Walker Hall
P.O.Box 11 8000
Gainesville, FL 32611–8000
USA

This paper deals with the general problem of estimating a signal from a finite number of independent blurred observations. The method consists of obtaining unbiased estimators for a family of smoothed versions determined by approximations to the identity (or, delta sequences), and using a regularization of the inverse of the blurring operator. This results in the problem of determining two regularizing parameters. Both apriori and data–driven methods are presented for choosing them. Numerical examples are presented which demonstrate satisfactory estimation even for small data sets.

1. Introduction

The problem of estimating an unknown signal from incomplete noisy observations plays a significant role in many applications of mathematics. This paper investigates the general problem of estimating an unknown element f of a Hilbert space from a finite number of independent blurred observations. Thus is applicable to the deconvolution problem, non–parametric density estimation, non–parametric regression, and random design models for general inverse problems. The blurring is represented by a bounded, selfadjoint, linear operator K on the Hilbert space, whose spectral decomposition is known. Here, K is not necessarily compact. For non–Hermitian K, the method applies to the equivalent problem involving K^*K, so the Hermitian assumption is only for convenience. The method of estimation consists of first obtaining an estimate

* Partially supported by NSF Grant DMS-9311513

for a "smoothed" version of the blurred signal Kf. This is achieved by a method which generalizes the method of delta sequences in Walter, Blum (1979), and is realizable in the Hilbert space $L^2(\mathbb{R})$ as convolution with an approximate identity. The second part of the method is to regularize the (unbounded) inverse of K by using some family of spectral functions, such as a smoothing or a spectral cut–off family (cf. Carroll, Rooij, Ruymgaart, 1991). Thus, two parameters need to be determined, one for the delta sequence, and the other for the regularization of K.

This method introduces an additional degree of freedom to the method proposed in Carroll, Rooij, Ruymgaart (1991). This freedom can be used advantageously to determine a suitable family of delta sequences by applying wavelet theory. This introduces the problem of obtaining a "best basis" as in the recently introduced wavelet methods (cf. Donoho, Johnstone, 1993). This generalization of the methods presented in this paper will be the subject of further research.

Both apriori and data–driven methods are presented for the choice of these parameters. The data–driven method is the usual cross–validation one, (cf. Bowman, 1984; Dey, Mair, Ruymgaart, 1994; Rudemo, 1982; Silverman, 1986; Thompson, Tapia, 1990; Wahba, 1977; Wahba, Wold, 1975) which results here in a cross–validation score of two variables, rather than the usual single variable encountered in traditional applications.

The general method developed in Section 2 is illustrated in Section 3 with a deconvolution problem. Numerical examples for estimating normal and log–normal distributions from samples of size 100 corrupted by the double exponential density are presented.

2. A general method

Let $\mathbb{H}$ be a separable Hilbert space and K be an injective, positive, selfadjoint, bounded linear mapping from $\mathbb{H}$ into $\mathbb{H}$. Consider the general problem of determining an element $f \in \mathbb{H}$ based on some statistical knowledge of

$$p = Kf \tag{2.1}$$

To quantify this information, introduce a family of bounded, positive, selfadjoint operators $\{A_\sigma : \sigma > 0\}$ on $\mathbb{H}$ with the property that, for all $g \in \mathbb{H}$

$$A_\sigma g \to g, \text{ as } \sigma \to 0+ \tag{2.2}$$

The data is assumed to consist of unbiased estimators $\hat{q}_{\sigma,1}, \hat{q}_{\sigma,2}, \dots, \hat{q}_{\sigma,n}$ of $A_\sigma p$. That is, if $\mathbb{E}$ is the expectation operator, for all j

$$\mathbb{E}\,[\hat{q}_{\sigma,j}] = A_\sigma p \tag{2.3}$$

To clarify the meaning of $\mathbb{E}$, let $\hat{q}$ be a random element of $\mathbb{H}$. That is, there is an underlying probability space $(\Omega, \mathcal{F}, \mathcal{P})$ and $\omega \mapsto \hat{q}(\omega)$ is a measurable function on Ω with values in $\mathbb{H}$. Then $\mathbb{E}\,[\hat{q}]$ is the unique element of $\mathbb{H}$ which satisfies, for all $g \in \mathbb{H}$

$$< \mathbb{E}\,[\hat{q}], g > \; = \int < \hat{q}(\omega), g > d\mathcal{P}(\omega) = \mathbb{E}\,[< \hat{q}, g >]$$

Hence, for any bounded linear $A : \mathbb{H} \to \mathbb{H}$, $\mathbb{E}\,[A\hat{q}] = A\,\mathbb{E}\,[\hat{q}]$. These properties will be used freely throughout this paper.

From (2.3) it is natural to consider the following unbiased estimator of $A_\sigma p$.

$$\hat{p}_\sigma = \frac{1}{n}\sum_{j=1}^{n} \hat{q}_{\sigma,j} \tag{2.4}$$

Since in many applications, the operator K does not have a bounded inverse, an estimator of f will be obtained by using a regularization approach (cf. Carroll, Rooij, Ruymgaart, 1991; Rooij, Ruymgaart, 1992; Ruymgaart, 1994). Although there are many such methods, by virtue of the spectral theory of operators, many of these are simply various forms of approximating spectra. Thus, they can be covered by introducing spectral functions ρ_α satisfying the following properties.

DEFINITION 2.1 *For each $\alpha > 0$, $\rho_\alpha : [0,\infty) \to [0,\infty)$ is a measurable function such that*

(i) $\sup_{t\geq 0} \rho_\alpha(t) < \infty$, *for all* $\alpha > 0$

(ii) $\sup_{\alpha>0,t\geq 0} t\rho_\alpha(t) < \infty$

(iii) $\rho_\alpha(t) \to t^{-1}$, *as* $\alpha \to 0$.

The choice $\rho_\alpha(t) = 1/(\alpha+t)$ corresponds to the classical method of Tikhonov regularization. For applications to infinite dimensional problems however, it seems more appropriate to employ the spectral cut–off function

$$\rho_\alpha(t) = \frac{1}{t} 1_{[\alpha,\infty)}(t) \tag{2.5}$$

which usually renders the regularized approximation more amenable to numerical computation.

To employ this spectral method of regularization it seems reasonable that the approximations to the identity A_σ be chosen in a manner consistent with K. This can be described as follows.

By the spectral theorem there exists a $\sigma-$finite Borel measure μ on a space S, a unitary operator $U : \mathbb{H} \to L^2(S,\mu)$ and a positive function $h \in L^\infty(S,\mu)$ such that

$$K = U^* M_h U \tag{2.6}$$

The needed commutativity assumption on A_σ takes the form of the existence of a positive function $\psi_\sigma \in L^\infty(S,\mu)$ such that

$$A_\sigma = U^* M_{\psi_\sigma} U. \tag{2.7}$$

Now, as is well–known, a family of regularizing operators for K is given by

$$R_\alpha = U^* M_{\rho_\alpha(h)} U. \tag{2.8}$$

That is, $\lim_{\alpha \to 0+} R_\alpha K f = f$.

However, since the assumed estimator $\hat{p}_\sigma$ is for $A_\sigma p$ rather than p, these regularizers are not immediately applicable. From (2.2) and the regularizing property of (2.8) it is easy to see that the family $\{R_\alpha A_\sigma : \alpha, \sigma > 0\}$ is also a regularizing family for K.

THEOREM 2.1 $\lim\limits_{(\alpha,\sigma)\to(0,0)} R_\alpha A_\sigma p = f.$

DEFINITION 2.2 $\hat{f}_{\alpha,\sigma} = R_\alpha \hat{p}_\sigma$

Since $\hat{p}_\sigma$ is an unbiased estimator for $A_\sigma p$, it seems natural to estimate $R_\alpha A_\sigma p$ by $\hat{f}_{\alpha,\sigma}$. The usual integrated mean square error (IMSE) will be used to measure the quality of the estimator $\hat{f}_{\alpha,\sigma}$.

THEOREM 2.2 *For all j,*

$$\begin{aligned}\mathbb{E}\,[\|\hat{f}_{\alpha,\sigma} - f\|^2] \;\le\; & \int (\rho_\alpha^2(h)(2h^2|1-\psi_\sigma|^2 + \frac{1}{n}\mathbb{E}\,[|U\hat{q}_{\sigma,j}|^2]) \\ & +2|1-h\rho_\alpha(h)|^2|Uf|^2)d\mu\end{aligned}$$

PROOF. Since $\hat{p}_\sigma$ is an unbiased estimator of $A_\sigma p$,

$$\mathbb{E}\,[\|\hat{f}_{\alpha,\sigma} - f\|^2] = \mathbb{E}\,[\|R_\alpha(\hat{p}_{\dot\sigma} - A_\sigma p)\|^2] + \|R_\alpha A_\sigma p - f\|^2.$$

Now

$$\begin{aligned}\|R_\alpha A_\sigma p - f\|^2 &\le 2(\|R_\alpha A_\sigma p - R_\alpha p\|^2 + \|R_\alpha p - f\|^2) \\ &= 2\int (h^2\rho_\alpha^2(h)|1-\psi_\sigma|^2 + |1-h\rho_\alpha(h)|^2)|Uf|^2 d\mu\end{aligned}$$

Since $\hat{q}_{\sigma,1}, \hat{q}_{\sigma,2}, \ldots, \hat{q}_{\sigma,n}$ are independent, unbiased estimators of $A_\sigma p$, the stochastic term can be estimated as follows

$$\begin{aligned}\mathbb{E}\,[\|R_\alpha(\hat{p}_\sigma - A_\sigma p)\|^2] &= \mathbb{E}\,[\|M_{\rho_\alpha(h)}U(\hat{p}_\sigma - A_\sigma p)\|^2] \\ &= \mathbb{E}\,[\frac{1}{n^2}\|\sum_{j=1}^{n} M_{\rho_\alpha(h)}U(\hat{q}_{\sigma,j} - A_\sigma p)\|^2] \\ &= \mathbb{E}\,[\frac{1}{n^2}\sum_{j=1}^{n} \|M_{\rho_\alpha(h)}U(\hat{q}_{\sigma,j} - A_\sigma p)\|^2] \\ &= \frac{1}{n^2}\sum_{j=1}^{n}\int \rho_\alpha^2(h)\mathbb{E}\,[|U(\hat{q}_{\sigma,j} - A_\sigma p)|^2]d\mu\end{aligned}$$

Now, observe that $[\mathbb{E}\,[|U(\hat{q}_{\sigma,j} - A_\sigma p)|^2] = \mathbb{E}\,[|U\hat{q}_{\sigma,j}|^2] - |UA_\sigma p|^2]$ hence,

$$\mathbb{E}\,[||R_\alpha(\hat{p}_\sigma - A_\sigma p)||^2] \leq \frac{1}{n}\int \rho_\alpha^2(h)\mathbb{E}\,[|U\hat{q}_{\sigma,j}|^2]d\mu.$$

As demonstrated in the next section, this result can be used to obtain asymptotic choices for the parameters α and σ based on regularity assumptions on the true solution f.

However, the main aim of this paper is to obtain a completely data–driven method for choosing α and σ independent of apriori assumptions. As usual the method of cross–validation will be employed.

For each k, let

$$\hat{p}_{\sigma,k} = \frac{1}{n-1}\sum_{j\neq k}\hat{q}_{\sigma,j}. \tag{2.9}$$

THEOREM 2.3

$$\begin{aligned}\mathbb{E}\,[||A_\sigma K(\hat{f}_{\alpha,\sigma} - f)||^2] &= \mathbb{E}\,[||A_\sigma K\hat{f}_{\alpha,\sigma}||^2 - \frac{2}{n}\sum_{k=1}^{n}\langle A_\sigma KR_\alpha\hat{p}_{\sigma,k}, \hat{q}_{\sigma,k}\rangle] \\ &\quad +||A_\sigma Kf||^2.\end{aligned}$$

PROOF.

$$\begin{aligned}\mathbb{E}\,[\langle A_\sigma KR_\alpha\hat{p}_{\sigma,k}, \hat{q}_{\sigma,k}\rangle] &= \frac{1}{n-1}\sum_{j\neq k}\mathbb{E}\,[\langle A_\sigma KR_\alpha\hat{q}_{\sigma,j}, \hat{q}_{\sigma,k}\rangle] \\ &= \frac{1}{n-1}\sum_{j\neq k}\int \psi_\sigma h\rho_\alpha(h)\mathbb{E}\,[U\hat{q}_{\sigma,j}\cdot U\hat{q}_{\sigma,k}]\,d\mu \\ &= \int \psi_\sigma^2 h\rho_\alpha(h)|Up|^2 d\mu \\ &= \mathbb{E}\,[\langle A_\sigma^2 KR_\alpha\hat{p}_\sigma, p\rangle] \\ &= \mathbb{E}\,[\langle A_\sigma K\hat{f}_{\alpha,\sigma}, A_\sigma p\rangle] \\ &= \mathbb{E}\,[\langle A_\sigma K\hat{f}_{\alpha,\sigma}, A_\sigma Kf\rangle]\end{aligned}$$

The result follows easily.

By the law of large numbers and Theorem 2.3, a purely data–driven method consists of determining α and σ to minimize the cross–validation score

$$M(\alpha,\sigma) = ||A_\sigma K\hat{f}_{\alpha,\sigma}||^2 - \frac{2}{n}\sum_{k=1}^{n}\langle A_\sigma KR_\alpha\hat{p}_{\sigma,k}, \hat{q}_{\sigma,k}\rangle \tag{2.10}$$

which can be rewritten in terms of the spectral information as

$$\begin{aligned} M(\alpha,\sigma) &= \int \psi_\sigma^2 h^2 (\rho_\alpha \circ h)^2 |\frac{1}{n}\sum_{j=1}^{n} U\hat{q}_{\sigma,j}|^2 d\mu \\ &- \frac{2}{n(n-1)}\sum_{k=1}^{n}\sum_{j\neq k}\int \psi_\sigma h(\rho_\alpha \circ h)(U\hat{q}_{\sigma,j})(U\hat{q}_{\sigma,k})d\mu. \end{aligned} \tag{2.11}$$

3. Application to deconvolution

This section demonstrates the applicability of the method developed in Section 2 to obtain both apriori asymptotic, and data–driven choices of the regularizing parameters. Numerical examples demonstrate good estimates even for samples of size 100.

Consider the problem of determining a density f based on independent samples $X_1, X_2, \ldots, X_n$ from a random variable

$$X = F + E \quad . \tag{3.1}$$

where F and E are stochastically independent variables with densities f and w respectively. Hence X has density p where $p = w * f$.

Assume that $w \in L^1(\mathbb{R})$ and that the characteristic function

$$\tilde{w}(s) = \int_{-\infty}^{\infty} e^{ist} w(t)dt$$

is strictly positive and even on $\mathbb{R}$.

Then, the operator K, defined on the Hilbert space $\mathbb{H} = L^2(\mathbb{R})$ to be convolution with w, is bounded, injective and does not have a bounded inverse.

Let $\mathcal{F}$ denote the unitary Fourier transform on $L^2(\mathbb{R})$. Then $K = \mathcal{F}^*\mathcal{M}_{\tilde{w}}\mathcal{F}$.

To illustrate the results in Section 2, consider the approximate identity determined by the normal distribution. That is

$$A_\sigma g = \varphi_\sigma * g \tag{3.2}$$

where

$$\varphi_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-x^2/2\sigma^2) \tag{3.3}$$

Other possibilities for A_σ include those generated by sinc functions (cf. Ikebe, Kowalski, Stenger, preprint). Other wavelet–based methods could be obtained by choosing a family of approximate identities based on cardinal spline or Daubechies scaling functions (cf. Daubechies, 1992). Thus, treating these possibilities as a template, this data–driven method could be expanded to include searching for the "best approximate identity" corresponding to a given signal. This is similar to the "best basis approach" in Donoho, Johnstone (1993) and will be developed in forthcoming articles.

With the choice of φ_σ ·in (3.3), the function

$$\psi(s) = \tilde{\varphi}_\sigma(s) = \exp(-\sigma^2 s^2/s). \tag{3.4}$$

For each j, choose the following unbiased estimator for $\varphi_\sigma * p$.

$$\hat{q}_{\sigma,j} = \varphi_\sigma(-X_j) \tag{3.5}$$

Regularization will be achieved by using the spectral cut-off function in (2.5) which eliminates the need for evaluation of infinite integrals. The range of integration will be determined by

$$I_\alpha = \{s : \tilde{w}(s) \geq \alpha\} \tag{3.6}$$

The estimator in Definition 2.1 becomes

$$\hat{f}_{\alpha,\sigma}(t) = \frac{1}{2\pi n} \int_{I_\alpha} \frac{\psi_\sigma(s)}{\tilde{w}(s)} \left(\sum_{j=1}^{n} e^{-is(t-X_j)} \right) ds \tag{3.7}$$

Now, consider the theoretical problem of asymptotic choices based on apriori assumptions on the regularity of f.

As indicated by Theorem 2.2, it is necessary to estimate $\mathbb{E}\ [|U\hat{q}_{\sigma,j}|^2]$, which, in this case becomes

$$\mathbb{E}\ [|\mathcal{F}\hat{q}_{\sigma,j}|^2] = \frac{1}{2\pi}\psi_\sigma^2(s). \tag{3.8}$$

Using the classical Sobolev space characterization for regularity, consider the condition

$$||f||_\nu \leq E \tag{3.9}$$

for some $E, \nu > 0$, where $||f||_\nu^2 = \int (1+s^2)^\nu |\mathcal{F}f(s)|^2 ds$.

To perform the error estimates, information on the degree of ill-posedness of the operator K is necessary. As in many cases, K is finitely smoothing, the following condition is considered, where $a, m, M > 0$, $a \geq 0$ are constants.

$$ms^{-sa} \leq \tilde{w}(s) \leq Ms^{-2a} \tag{3.10}$$

If the errors E are Gaussian, then (3.10) is not satisfied. However this case can be handled by a more general treatment which would be similar to that in Mair, Ruymgaart (1994).

THEOREM 3.1 *Under conditions (3.9) and (3.10), if* $\alpha = n^{-2a/(4a+2\nu+1)}$ *and* $\sigma = n^{-(\nu+2)/2(4a+2\nu+1)}$, *then the estimator* $\hat{f}_{\alpha,\sigma}$ *in (3.7) satisfies*

$$\mathbb{E}\ [||\hat{f}_{\alpha,\sigma} - f||^2] = O(n^{-2\nu/(2\nu+4a+1)}).$$

PROOF. From Theorem 2.2 and (3.8)

$$\mathbb{E}\,[\|\hat{f}_{\alpha,\sigma}-f\|^2] \le 2\int_{I_\alpha}|1-\psi_\sigma(s)|^2|\mathcal{F}f(s)|^2ds+2\int_{I_\alpha^c}|\mathcal{F}f(s)|^2ds$$
$$+\frac{1}{2\pi n}\int_{I_\alpha}\frac{\psi_\sigma^2(s)}{\tilde{w}(s)^2}ds$$

By using (3.6) and (3.10), $s\in I_\alpha \Rightarrow |s|\le (M/\alpha)^{1/2a}$ and $s\in I_\alpha^c \Rightarrow |s| > (m/\alpha)^{1/2a}$.

Hence, by using the inequality $1-e^{-x}\le x$ for $x\ge 0$, and the assumption (3.9), it follows that

$$\mathbb{E}\,[\|\hat{f}_{\alpha,\sigma}-f\|^2]\le C(n^{-1}\alpha^{-(4a+1)/2a}+\sigma^4\alpha^{-2/a}+\alpha^{\nu/a})$$

for some constant C. The result follows by balancing the orders of convergence determined by these three terms.

Easy calculations and (2.11) show that the data–driven method in this case reduces to the problem of minimizing the score

$$M_0(\alpha,\sigma)=\int_{I_\alpha}\psi_\sigma^3(s)((\psi_\sigma(s)-\frac{2n}{n-1})|\frac{1}{n}\sum_{j=1}^{n}e^{isX_j}|^2+\frac{2}{n-1})ds \qquad (3.11)$$

Although there is, as yet no mathematical proof that this has a unique minimum, numerical simulations indicate this to be true. In the numerical examples presented below, it is assumed that the errors have a double exponential distribution, so that $\tilde{w}(s)=1/(1+s^2)$ and the score can be written as an integral over the interval $[0,L]$, where $L=\sqrt{\alpha^{-1}-1}$.

The numerical results presented below demonstrate this method for recovering normal and log–normal densities from a random sample of size 100.

The first set of results is for the normal distribution with mean 0 and variance 4. The surface in Figure 1 represents the score M_1, which achieves its minimum when $\sigma = 0.355$ and $L = 1.0$. Figure 2 compares the true density and the estimate determined by using these values in (3.7).

The next experiment is to recover the log–normal density

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma x}\exp(-\frac{1}{2\sigma^2}(\log\frac{x}{\mu})^2), \quad x>0$$

with $\mu = 50$ and $\sigma = 0.3$, as considered in Rudemo (1982). The score M_1 (see Figure 3), achieves its minimum when $\sigma = 3.4$ and $L = 0.2$. Figure 4 compares the true density and the estimate.

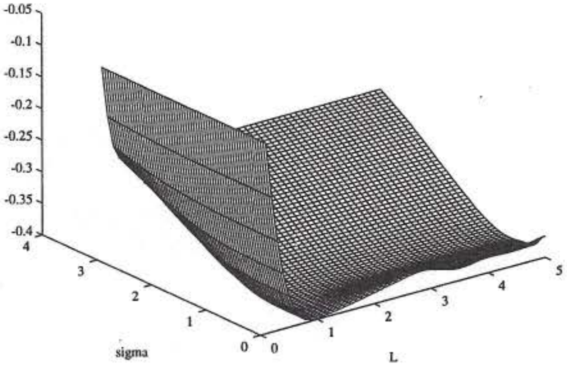

Figure 1

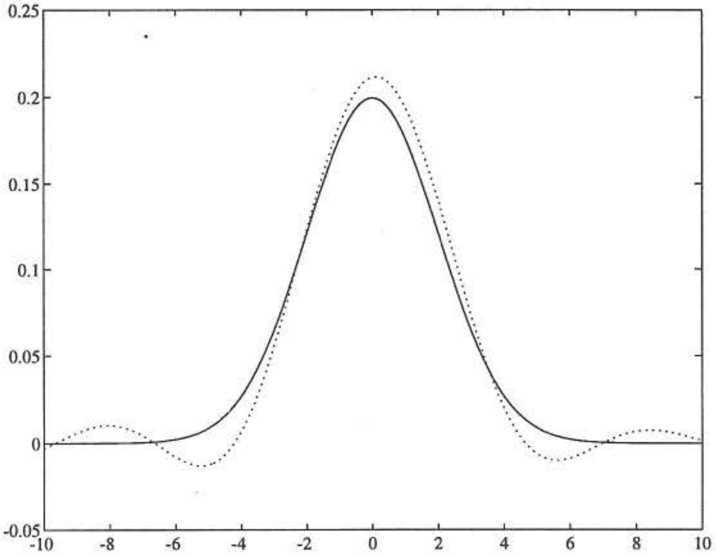

Figure 2

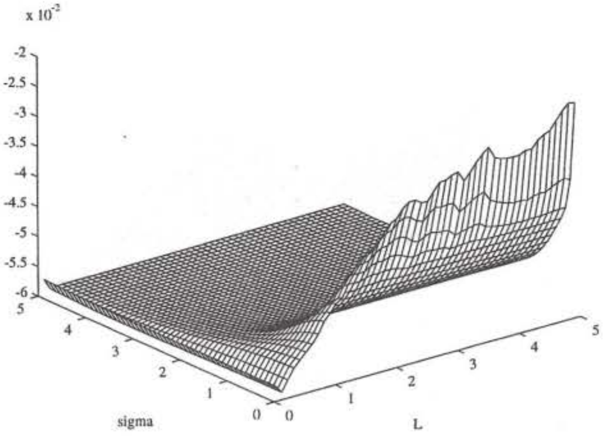

Figure 3

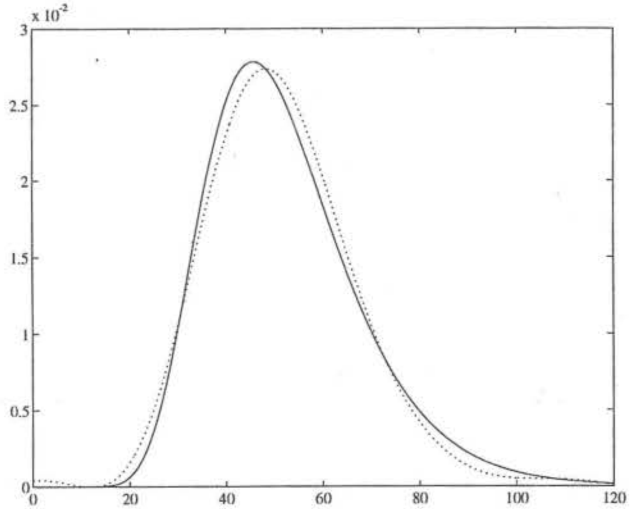

Figure 4

References

BOWMAN A.W. (1984) An alternative method of cross-validation for the smoothing of density estimates, *Biometrika* **71**, 353–360.

CARROLL R.J., ROOIJ A.C.M. VAN AND RUYMGAART F.H. (1991) Theoretical aspects of ill-posed problems in statistics, *Acta Appl. Math.* **24**, 113–140.

DEY A.K., MAIR B.A. AND RUYMGAART F.H. (1994) Cross-validation for parameter selection in inverse estimation problems, submitted.

DAUBECHIES I. (1992) Ten Lectures on Wavelets, *SIAM*, Philadelphia.

DONOHO D.L., JOHNSTONE I.M. (1993) Adapting to unknown smoothness via wavelet shrinkage, Tech. Report, Dept. Statistics, Stanford Univ.

DONOHO D.L., JOHNSTONE I.M., KERKYACHARIAN G. AND PICARD D. (1993) Density estimation by wavelet thresholding, Tech. Report, Dept. Statistics, Stanford Univ.

FAN J. (1991) Global behavior of deconvolution kernel estimates, *Statist. Sin.* **1**, 541–551.

HOERL A. AND KENNARD R.W. (1970A) Ridge regression: biased estimation for nonorthogonal problems, *Technometrics* **12**, 55–67.

HOERL A. AND KENNARD R.W. (1970B) Ridge regression: applications to nonorthogonal problems, *Technometrics* **12**, 69–82.

IKEBE Y., KOWALSKI M. AND STENGER F. Rational approximation of the step, filter, and impulse functions, preprint.

KERYACHARIAN G. AND PICARD D. (1992) Density estimation in Besov spaces, *Statistics and Probability Letters* **13**, 15–24.

MAIR B.A. AND RUYMGAART F.H. (1994) Statistical inverse estimation in Hilbert scales, submitted.

ROOIJ A.C.M. VAN AND RUYMGAART F.H. (1992) Asymptotic optimality for inverse estimation problems, Tech. Report, Dept. Math., Texas Tech Univ.

RUDEMO M. (1982) Empirical choice of histograms and kernel density estimators, *Scand. J. Statist.* **9**, 65–78.

RUYMGAART F.H. (1994) A unified approach to inversion problems in statistics, *Math. Methods Statist.*, to appear.

SILVERMAN B.W. (1986) *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York.

THOMPSON J.R. AND TAPIA R.A. (1990) Nonparametric Function Estimation, Modeling, and Simulation, *SIAM*, Philadelphia.

WAHBA G. (1977) Optimal smoothing of density estimates. In : *Classification and Clustering* (ed. J. Van Ryzin), Academic Press, New York, 423–458.

WAHBA G. AND WOLD S. (1975) A completely automatic French curve: Fitting spline functions by cross validation, *Comm. Statist.* **4**, 1–17.

WALTER G.G. AND BLUM J.R. (1979) Probability estimation using delta sequences, *Ann. Statistic.* **7**, 328–340.

ZHANG C.-H. (1990) Fourier methods for estimating mixing densities and distributions, *Ann Statist.* **18**, 806–831.