Minimax Estimation of Linear Functionals, Particularly in Nonparametric Regression and Positron Emission Tomography

Julian Stander[1] and Bernard W. Silverman[2]

[1]School of Mathematics and Statistics, University of Plymouth, Plymouth PL4 8AA, United Kingdom
J.Stander@pa.plym.ac.uk
[2]School of Mathematics, University of Bristol, Bristol BS8 1TW, United Kingdom
B.W.Silverman@bristol.ac.uk

Summary

Often it is required to estimate a linear functional of a function f given observations of some kind from f. We consider the linear minimax estimation of a bounded linear functional T with respect to the mean square error loss function over a certain class of functions $\mathcal{F}$ defined by means of a semi-norm, which in practice usually provides some restriction on roughness. We review the remarkable result essentially due to Speckman (1979) that the linear minimax estimate of $T(f)$ is $T(\hat{f})$, where $\hat{f}$ is a type of penalized least squares estimate of f. Hence the amount and method of smoothing carried out in the minimax estimation of $T(f)$ does not depend on the functional T under consideration. The function f^* in $\mathcal{F}$ for which the maximum of the loss function is attained does, however, depend on T.

We discuss the result in the case of nonparametric regression and positron emission tomography (PET). In the former case we see a surprising consequence of the dependence of f^* on T. In the latter case we show how the result can be applied to PET and illustrate it numerically, highlighting certain interesting features of the PET process.

Keywords: linear functionals; minimax estimation; nonparametric regression; positron emission tomography; splines.

1 Introduction

There are many problems where one is interested in estimating a functional of a function f given observations of some kind from f. A simple example is nonparametric regression, where a function f is observed with error at various points $t_1, \ldots, t_n$ and the functional of interest is the value or the derivative of f at some point, or perhaps the integral of f over a particular interval. Another problem arises in various kinds of tomography, where f is an intensity function in two or three dimensional space, and one observes the integral of f along a number of lines in space, or over a number of tubes defined by pairs of detectors. Here the functionals of interest may again be the value of f at a particular point or the integral of f over a region.

In this paper we consider the estimation of bounded linear functionals $T(f)$, which include all the examples given above, by means of a linear minimax approach over a suitable function class $\mathcal{F}$, defined by means of a semi-norm, using the mean square error (or something closely related to it) as the loss function. In practice the class $\mathcal{F}$ usually provides some restriction on roughness. We review a result essentially due to Speckman (1979) which tells us that the amount and method of smoothing carried out in the estimation of $T(f)$ does not depend on the functional T under consideration. More precisely, the linear minimax estimator of $T(f)$ is $T(\hat{f})$, where $\hat{f}$ is a type of penalized least squares estimator of f. Hence, the same $\hat{f}$ yields the linear minimax estimator for all functionals T. We refer to the function f^* in $\mathcal{F}$ that maximizes the loss function as the *least favourable function* and show that f^* may indeed depend on the functional being estimated.

After discussing the general theory, we consider two specific examples: nonparametric regression and positron emission tomography (PET). In the first example we assume that a real valued function f defined on a finite interval $I \subset R$ and belonging to a certain smoothness class is observed, subject to error, at points $t_1, \ldots, t_n$. We show that although a suitable $\hat{f}$ yields the linear minimax estimator of $f(t)$ for all $t \in I$ with respect to the mean square error loss function, $\hat{f}$ does not yield the linear minimax estimator subject to a more global loss function defined by the sum of the mean square error losses over the points $t_1, \ldots, t_n$. We discuss how this paradox can be explained by the dependence of the least favourable function on the functional being estimated. In the second example, we apply the general theory to the PET case using a slightly modified loss function called the surrogate mean square error. We give the form of the linear minimax estimator, of the least favourable function f^* and of the value of the surrogate mean square error evaluated at f^* and the minimax estimator. We present numerical examples that illustrate that the minimaxed surrogate mean square error does not decrease monotonically as the number of detectors increases, and that f^* does depend upon the functional being estimated.

The layout of the paper is as follows. In Section 2 we discuss the main result in detail. We consider its application to nonparametric regression in Section 3 and to PET in Section 4. We present our conclusions in Section 5. Finally, in Section 6, we give a proof of the main result.

2 The Main Result

Before discussing our main result, we need to introduce some notation, assumptions and definitions.

2.1 Preliminaries: notation, assumptions and definitions

Suppose that $\mathcal{H}$ is a real vector space of real valued functions equipped with a semi-norm $\|\cdot\|_s$ and associated semi-inner-product $(\cdot\,,\,\cdot)_s$. Let $Y_1, \ldots, Y_n$ be observations of linear functionals $\eta_1(f), \ldots, \eta_n(f)$ such that $\mathbf{E}(Y_i) = \eta_i(f)$, and the Y_i are uncorrelated and have variance one. Write Y for the vector $(Y_1, \ldots, Y_n)^T$ and $\eta(f)$ for the vector $(\eta_1, \ldots, \eta_n)^T$. Let $T : \mathcal{H} \to R$ be a linear functional. This paper considers the estimation of $T(f)$ for $f \in \mathcal{H}$ by linear estimates of the form $w^T Y$ for vectors w. Let $\mathcal{K}$ be the kernel of $\|\cdot\|_s$, that is

$$\mathcal{K} = \{k \in \mathcal{H} : \|k\|_s = 0\}.$$

Given any f and g in $\mathcal{H}$, define a bilinear form by

$$[f, g] = \sum_{i=1}^{n} \eta_i(f)\, \eta_i(g).$$

Let $\mathcal{H}_0$ be the orthogonal complement of $\mathcal{K}$ with respect to the bilinear form, so that

$$\mathcal{H}_0 = \{f \in \mathcal{H} : [f, g] = 0 \text{ for all } g \in \mathcal{K}\}.$$

We now make two assumptions.

Assumption 1 *If* $\|f\|_s = 0$ *and* $[f, f] = 0$*, then* $f = 0$.

We discuss the validity of this assumption for nonparametric regression in Section 3 where it reduces to a particularly simple form. If Assumption 1 holds, it can easily be shown that, in algebraic terms, $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{K}$, and that $\|\cdot\|_s$ is a norm and $(\cdot\,,\,\cdot)_s$ is an inner product on $\mathcal{H}_0$.

Assumption 2 *The space $\mathcal{H}_0$ equipped with the inner product $(\cdot\,,\,\cdot)_s$ is a Hilbert space, and the functionals η_i and T are bounded linear functionals on $\mathcal{H}_0$.*

We discuss the validity of this assumption for nonparametric regression in Section 3 and for PET in Section 4 where we see that it is not particularly restrictive. Because of Assumption 2 we can apply Riesz's representation lemma to η_i and T on $\mathcal{H}_0$ to conclude that there exist unique representers $\tilde{\eta}_i$ and $\tilde{T}$ in $\mathcal{H}_0$ such that $\eta_i(f_0) = (f_0\,,\,\tilde{\eta}_i)_s$ and $T(f_0) = \left(f_0\,,\,\tilde{T}\right)_s$ for all $f_0 \in \mathcal{H}_0$; see Kreyszig (1978), page 188.

Write $\tilde{\eta}$ for the vector $(\tilde{\eta}_1, \ldots, \tilde{\eta}_n)^T$ and let H be the matrix with entries $H_{ij} = (\tilde{\eta}_i\,,\,\tilde{\eta}_j)_s$ so that we can write H as $(\tilde{\eta}\,,\,\tilde{\eta}^T)_s$. For arbitrary $f_0 \in \mathcal{H}_0$ write $(f_0\,,\,\tilde{\eta})_s$ for the vector

$$\{(f_0\,,\,\tilde{\eta}_1)_s\,, \ldots, (f_0\,,\,\tilde{\eta}_n)_s\}^T\,.$$

We show in Section 6 that, under Assumption 1, $\mathcal{K}$ is a finite dimensional vector space with dimension $r \leq n$ and that the bilinear form $[\cdot,\cdot]$ restricted to $\mathcal{K}$ is an inner product on $\mathcal{K}$. We can accordingly let $\phi_1, \ldots, \phi_r$ be an orthonormal basis for $\mathcal{K}$ relative to the inner product $[\cdot,\cdot]$. Let ϕ be the vector $(\phi_1, \ldots, \phi_r)^T$, Φ be the $n \times r$ matrix with entries $\Phi_{ij} = \eta_i(\phi_j)$, and t be the vector $\{T(\phi_1), \ldots, T(\phi_r)\}^T$.

2.2 The main result

We consider linear estimates of $T(f)$ of the form $w^T Y$ for vectors w. The approach that we take is a minimax one over a particular class $\mathcal{F}$ where

$$\mathcal{F} = \{f : \|f\|_s \leq 1\}\,, \tag{1}$$

the unit ball in $\mathcal{H}$ with respect to $\|\cdot\|_s$. We now define what we mean for an estimator to be linear minimax for estimating T with respect to the mean square error loss function.

Definition 1 *The estimator $\hat{w}^T Y$ is said to be* linear minimax for estimating T *with respect to the mean square error loss function if and only if $\hat{w}$ is the minimizer over vectors w of*

$$\sup_{f \in \mathcal{F}} \boldsymbol{E}\left\{T(f) - w^T Y\right\}^2 = \sup_{f \in \mathcal{F}} \left\{T(f) - w^T \eta(f)\right\}^2 + w^T w. \tag{2}$$

The equality in the definition follows by standard properties of the mean square error, making use of **Var**$(Y) = I$. We call the function $f^* \in \mathcal{F}$ that attains the supremum of the mean square error over $\mathcal{F}$ the *least favourable function* and write $\hat{T}$ for the linear minimax estimator of T.

As our main object is to demonstrate a connection between the linear minimax estimators of functionals and the penalized least squares estimator of f itself, we make another definition:

Definition 2 *The function $\hat{f}$ is the* penalized least squares estimator of f *if and only if $\hat{f}$ minimizes*

$$\{Y - \eta(f)\}^T \{Y - \eta(f)\} + \|f\|_s^2 \tag{3}$$

over $f \in \mathcal{H}$.

The first term in (3) is a measure of the fidelity of $\eta(f)$ to the data, while the second term is a measure of the roughness of f. Penalized least squares estimators are discussed in detail by Green and Silverman (1994). They replace expression (3) by one of the form 'fidelity + α roughness', where $\alpha > 0$ is an unknown smoothing parameter that controls the trade-off between fidelity and roughness in such a way that the higher the value of α, the smoother the resulting minimizer. Note, however, that expression (3) does not require any choice of smoothing parameter. See Sections 2.3 and 3 for further discussion of this point. We now state our main theorem:

Theorem 1 *The linear minimax estimator of $T(f)$ is $T(\hat{f})$, where $\hat{f}$ is the penalized least squares estimate of f.*

The theorem is essentially given, in a somewhat different form, in an unpublished manuscript by Speckman (1979). In Section 6 we present a more straightforward proof than the one given by Speckman (1979). Our proof is valid under slightly weaker assumptions than Speckman's. Both Li (1982) and Ibragimov and Khas'minskii (1984) consider much more abstract results in related contexts. See also Donoho and Liu (1991) and Donoho and Low (1992). While we make no claim to the originality of either the result or the idea behind its proof, we do believe that our proof is substantially more accessible than those that have so far appeared.

The explanation of the theorem is that the linear minimax way to proceed is to construct the estimator $\hat{f}$ and then to evaluate $T(\hat{f})$. Thus the amount and method of smoothing carried out in the minimax estimation of $T(f)$ does not depend upon T at all, and the same function estimate $\hat{f}$ yields the linear minimax estimator of all linear functionals simultaneously.

2.3 Further discussion of the main result

Theorem 1 applies in the more general case when **Var**$(Y) = \Sigma$. In that case, to use Theorem 1 just apply the linear transformation $\Sigma^{-1/2}$ to Y and $\eta(f)$. Also more general $\mathcal{F}$ of the form

$$\mathcal{F} = \{f : \|f\|_s \leq C\},$$

for $C > 0$, can be reduced to the case given by (1) by rescaling the semi-norm by a factor of C^{-1}. As far as the penalized least squares estimator of f is concerned, this factor has the same effect as the smoothing parameter α already discussed. We see an example of such a transformation and rescaling in Section 3. Furthermore, general ellipsoidal conditions on f are coped with by working with the function space equipped with the semi-norm generated by the quadratic form defining the ellipsoid. We see an example of such an ellipsoidal condition in Section 4.

In the next theorem we present expressions for the least favourable function f^*. We also give formulae for $\hat{w}$, for the value taken by the mean square error at f^* and $\hat{w}$, and for $\hat{f}$. The derivations of these results also appear in Section 6.

Theorem 2 *The least favourable function f^* takes the form*

$$f^* = \pm \frac{\tilde{T} - w^T \tilde{\eta}}{\left\| \tilde{T} - w^T \tilde{\eta} \right\|_s} + k \textit{ for any } k \in \mathcal{K},$$

in terms of w. The w that minimizes (2) takes the form

$$\hat{w} = (I + H)^{-1} \left(\tilde{T}, \tilde{\eta} \right)_s + \Phi t. \tag{4}$$

The value of the mean square error at f^* and $\hat{w}$ is

$$\left(\tilde{T}, \tilde{T} \right)_s - \left(\tilde{\eta}^T, \tilde{T} \right)_s (I + H)^{-1} \left(\tilde{T}, \tilde{\eta} \right)_s + t^T t. \tag{5}$$

The penalized least squares estimator of f takes the form

$$\hat{f} = Y^T (I + H)^{-1} \tilde{\eta} + Y^T \Phi \phi. \tag{6}$$

Although expressions (4), (5) and (6) are given in terms of ϕ and Φ, their values are independent of the precise choice of orthonormal basis. Note that $\|f^*\|_s = 1$ so that f^* lies on the boundary of $\mathcal{F}$.

In the case when $\|\cdot\|_s$ is a norm $\|\cdot\|$, we have that $\mathcal{H}_0 = \mathcal{H}$ and $\mathcal{K} = \{0\}$, and so Assumption 1 holds automatically. Accordingly, we only need to check Assumption 2. If Assumption 2 holds, then

$$f^* = \pm \frac{\tilde{T} - w^T \tilde{\eta}}{\left\|\tilde{T} - w^T \tilde{\eta}\right\|} \tag{7}$$

in terms of w, where $\tilde{T}$ and $\tilde{\eta}$ are defined with respect to the inner product $(\cdot\,,\,\cdot)$ associated with the norm $\|\cdot\|$. Immediately, we see that if $\|\cdot\|_s$ is a norm the least favourable function is unique up to sign, but otherwise it takes the form given in (7) plus any function in $\mathcal{K}$. Note that f^* depends on the particular T under consideration, unlike $\hat{f}$. Moreover, when $\|\cdot\|_s$ is a norm, the final terms in the other three expressions given in Theorem 2 do not appear, as we see in Section 4.

2.4 An illustrative example

Before discussing our two substantive examples, we sketch a simple illustrative example suggested by a referee. Suppose that $\mathcal{H}$ is the space of square-summable sequences $f = (f_1, f_2, \ldots)$, with the usual l^2 inner product and norm. Let T be a bounded linear functional on $\mathcal{H}$, so that $T(f) = \sum_i t_i f_i$ for some square-summable sequence (t_i). Let p and n be positive integers with $p < n$. Define the squared seminorm $\|f\|_s^2 = (f, f)_s = \sum_{i>p} f_i^2$, so that $\mathcal{K}$ is the space of all f of the form $(f_1, \ldots, f_p, 0, \ldots)$, and $\tilde{T}$ is $(0, \ldots, 0, t_{p+1}, t_{p+2}, \ldots)$.

Suppose that Y consists of the first n observations of a sequence Y_i of independent $\mathcal{N}(f_i, 1)$ random variables. It is then easy to verify the following, either by application of Theorems 1 and 2 or by direct minimization:

- The penalized least squares estimator of f (*i.e.* the minimizer of $\sum_{i=1}^n (Y_i - f_i)^2 + \|f\|_s^2$) is given by $\hat{f}_i = Y_i$ for $i \leq p$; $\hat{f}_i = \frac{1}{2} Y_i$ for $p < i \leq n$; and $\hat{f}_i = 0$ for $i > n$.
- The linear minimax estimator of $T(f)$ under the restriction $\|f\|_s^2 \leq 1$ is $T(\widehat{f})$, so that $w_i = t_i$ for $i \leq p$ and $w_i = \frac{1}{2} t_i$ for $p < i \leq n$. The least favourable f^* satisfies

$$f_i^* = \begin{cases} \text{arbitrary} & \text{for } i = 1, \ldots, p \\ \frac{1}{2}\, C\, t_i & \text{for } i = p+1, \ldots, n \\ C\, t_i & \text{for } i > n, \end{cases}$$

where the constant C is chosen to ensure that $||f^*||_s^2 = 1$. Note that the least favourable function f^* depends on T, while the estimator $\hat{f}$ does not.

Making use of a trivial extension of Lemma 3.1 of Speckman (1985) it can be shown that the linear minimax estimator of f itself, under the restriction $||f||_s^2 \leq 1$ and with risk $\mathbf{E}[\sum_{i=1}^n (f_i - \tilde{f}_i)^2]$, is given by

$$\tilde{f}_i = \begin{cases} Y_i & \text{for } i = 1, \ldots, p \\ (1 + n - p)^{-1} Y_i & \text{for } i = p + 1, \ldots, n \\ \text{arbitrary} & \text{for } i > n. \end{cases}$$

Note that this estimate is not the same as the penalized least squares estimate $\hat{f}$; we shall briefly discuss this paradox in Section 3 below. This paradox is somewhat reminiscent of Stein's paradox. When estimating a linear combination of the components of the mean of a multivariate normal, the corresponding linear combination of the components of the sample mean is an admissible estimator; however the sample mean is not, in general, admissible as an estimator of the mean under squared Euclidean distance loss. Of course, the present paper does not deal with estimation of an *unconstrained* mean, and so the comparison is a superficial one.

3 Nonparametric Regression: An Example and Paradox

Let $I \subset R$ be a finite interval and let $\mathcal{H}$ be the set of real valued functions f on I that belong to the space W^2 of functions f such that f has an absolutely continuous first derivative and $\int_I f''(s)^2 ds < \infty$ Assume that f is observed, subject to independent errors with the same variance σ^2, at points $t_1, \ldots, t_n$. Let $t \in I$ be arbitrary and define $T_t(f) = f(t)$. We now derive the minimax estimator $\hat{T}_t$ of T_t where $\mathcal{F}$ is defined as

$$\mathcal{F} = \left\{ f : \int_I \{f''(s)\}^2 ds \leq C^2 \right\},$$

for $C > 0$, which clearly provides some restriction on the roughness of f. Since $[\int_I \{f''(s)\}^2 ds]^{1/2}$ is a semi-norm on $\mathcal{H}$, we may define a new semi-norm $||\cdot||_s$ on $\mathcal{H}$ by rescaling as

$$||f||_s = \frac{1}{C} \left[\int_I \{f''(s)\}^2 ds \right]^{1/2},$$

with associated semi-inner-product

$$(f\,,\,g)_s = \frac{1}{C^2}\int_I f''(s)g''(s)ds.$$

The set $\mathcal{F}$ is the unit ball in $\mathcal{H}$ with respect to $\|\cdot\|_s$. The kernel $\mathcal{K}$ of $\|\cdot\|_s$ is the set of all linear functions $a + bs$, for constants a and b.

Let the observations be $y_1, \ldots, y_n$, and note that $y_i = f(t_i) + \epsilon_i$. If we write ϵ for the vector $(\epsilon_1, \ldots, \epsilon_n)^T$ then $\mathbf{E}(\epsilon) = 0$ and $\mathbf{Var}(\epsilon) = \sigma^2 I$. Now define $Y_i = y_i\,/\,\sigma$ and $\eta_i(f) = f(t_i)\,/\,\sigma$, and note that $\mathbf{E}(Y_i) = \eta_i(f)$ and $\mathbf{Var}(Y) = I$. Since $\mathcal{K}$ is the set of all linear functions and $[f, f] = \sum_{i=1}^n \{f(t_i)\,/\,\sigma\}^2$, we immediately have that Assumption 1 holds provided there are at least two distinct points t_i. If this is true, Assumption 2 holds because of the Sobolev embedding theorem and standard Sobolev properties; see Adams (1975). We may now apply Theorem 1 to obtain that for arbitrary $t \in I$ the linear minimax estimate of $T_t(f)$ is given by $T_t(\hat{f}) = \hat{f}(t)$, where $\hat{f}$ minimizes

$$\sum_{i=1}^n \{y_i - f(t_i)\}^2 + \frac{\sigma^2}{C^2}\int_I \{f''(s)\}^2\,ds, \tag{8}$$

over $f \in \mathcal{H}$. Note that the smoothness of $\hat{f}$ is controlled by C: large values of C give little weight to the integral which measures roughness and so a rough $\hat{f}$ results, while small values of C give much weight to the integral and a smooth $\hat{f}$ results. The minimizing function $\hat{f}$ is the cubic smoothing spline estimator of f with smoothing parameter $\alpha = \sigma^2\,/\,C^2$; see Green and Silverman (1994).

Speckman (1985) considers a different, but related minimax set up. For the n-dimensional space of natural polynomial splines of degree three with simple knots at $t_1, \ldots, t_n$ he introduces the basis of eigenfunctions $\phi_1, \ldots, \phi_n$ and associated eigenvalues $\lambda_1, \ldots, \lambda_n$ determined (essentially uniquely) by the following conditions:

$$\begin{aligned} \sum_{l=1}^n \phi_i(t_l)\phi_j(t_l) &= \delta_{ij}, \\ \int_I \phi_i''(t)\phi_j''(t)dt &= \delta_{ij}\lambda_j, \end{aligned}$$

for $i, j = 1, \ldots, n$, and

$$0 = \lambda_1 = \lambda_2 < \lambda_3 \leq \cdots \leq \lambda_n.$$

He lets $\mathcal{L}$ be the class of all estimators $\tilde{f}$ of functions $f \in \mathcal{F}$ which are linear in the observations $y_1, \ldots, y_n$, and for $\gamma \geq 0$ he defines $f_\gamma \in \mathcal{L}$ as

$$f_\gamma(t) = \sum_{i=1}^{n} \left(1 - \sqrt{\gamma\, \lambda_k}\right)_+ \widehat{\beta_k}\, \phi_k(t),$$

where $\widehat{\beta_k} = \sum_{i=1}^{n} y_i \phi_k(t_i)$ and $(\delta)_+ = \max(\delta, 0)$. He then shows that for any $C > 0$ and $\sigma > 0$ there is a $\gamma_0 = \gamma_0(C, \sigma)$ such that f_{γ_0} is the minimum over functions $\tilde{f} \in \mathcal{L}$ of $\sup_{f \in \mathcal{F}} L(\tilde{f}, f)$ where the loss function $L(\tilde{f}, f)$ that he considers is given by the mean summed square error

$$L(\tilde{f}, f) = \mathbf{E}\left[\sum_{i=1}^{n} \left\{\tilde{f}(t_i) - f(t_i)\right\}^2\right].$$

We remark that the minimizing $\tilde{f} \in \mathcal{L}$ given by the spline f_{γ_0} is different from the minimizer of (8).

This last remark provides us with a paradox as suggested previously in Section 2.4. We might suppose that if $\hat{f}(t)$ is the linear minimax estimator of $f(t)$ simultaneously for all $t \in I$ with respect to mean square error loss, then $\hat{f}$ should be the linear minimax estimator with respect to mean summed square error loss. Indeed, if there were to exist only one least favourable function $f_{\mathrm{lf}}^* \in \mathcal{F}$ such that for the estimation of T_t the supremum over $\mathcal{F}$ was attained by f_{lf}^* for all $t \in I$, then a straightforward argument would establish our supposition. However, this is not the case as we have seen that the least favourable function *does* depend upon the functional being estimated.

4 Minimax Estimation of Linear Functionals of Positron Emission Tomography Images

4.1 Introduction and mathematical setting

Positron emission tomography (PET) is a non-invasive technique for reconstructing the internal metabolic activity of parts of the body, usually the brain. It has recently received much interest in the statisical literature; see, for example, Geman and McClure (1987), Green (1990), Jones and Silverman (1989), Silverman, Jones, Wilson and Nychka (1990) and Vardi, Shepp and Kaufman (1985). In our treatment, we follow the approach and notation of Johnstone and Silverman (1990) and Johnstone and Silverman (1991).

A radioactively labelled metabolite is introduced into the body which causes independent emissions of positrons at positions that cannot be observed directly. Instead, each positron emission gives rise to a pair of photons that fly off in opposite directions along a line with uniformly distributed orientation. These photon pairs are observed by a ring of d detectors. We refer to the space enclosed by the ring as $\mathcal{B}$ (brain space), which we parametrize by either polar coordinates (r, θ) or Cartesian coordinates (x, y) with the ring being the unit circle. We consider the emissions as i.i.d. observations from a density f relative to the uniform density μ on $\mathcal{B}$.

Let $\mathcal{D}$ consist of all possible lines of flight of photon pairs, parametrized by pairs (s, ϕ) the length and orientation of the perpendicular from the origin to the detected line, equipped with the measure $d\lambda(s, \phi) = 2\,\pi^{-2}\,(1 - s^2)^{1/2}\,ds\,d\phi$. The density g (relative to λ) of observed lines on $\mathcal{D}$ is related to the density f on $\mathcal{B}$ by the normalized Radon transform P, defined by

$$\begin{aligned} g(s, \phi) &= Pf(s, \phi) \\ &= \tfrac{1}{2}(1 - s^2)^{-\frac{1}{2}} \int_{-\sqrt{1-s^2}}^{\sqrt{1-s^2}} f(s\cos\phi - t\sin\phi, s\sin\phi + t\cos\phi)dt. \end{aligned}$$

In practice, detections are recorded as falling in one of $r_d = \frac{1}{2}d(d-1)$ tubes $\mathcal{D}_j$. Each tube is the collection of all photon lines detected by a particular unordered pair of distinct detectors, and is a data bin in $\mathcal{D}$. The set of these tubes is called *detector space* $\mathcal{D}^d$.

We write $Q_d\,g$ for the probability mass function of observed detections on $\mathcal{D}^d$:

$$(Q_d\,g)_j = \frac{1}{\lambda(\mathcal{D}_j)} \int_{\mathcal{D}_j} g\,d\lambda, \text{ for } j = 1, \ldots, r_d.$$

We let $\mathcal{G}^d$ be the finite dimensional space of vectors indexed by the elements of $\mathcal{D}^d$.

We shall assume that we record a Poisson number $\mathcal{N}$, with mean n, of observations on $\mathcal{D}^d$ drawn from the probability mass function $Q_d\,g$. Let n_j be the number of these observations falling into bin $\mathcal{D}_j$. Let

$$Z_j = n_j/(n\lambda(\mathcal{D}_j)) \tag{9}$$

be the normalized bin count for bin $\mathcal{D}_j$ and note that the Z_j are independent Poisson counts with intensities $(Q_d g)_j$. Much attention has been given in the literature to the estimation of the density function f itself from the Z_j. In this paper we consider the estimation of linear functionals of density functions f.

Bickel and Ritov (1995) have also recently considered this problem, although not from a minimax point of view. From their point of view the appropriate amount of smoothing does depend on the functional being estimated.

Let N_0 be the set

$$N_0 = \{\nu = (j,k) \mid j \geq 0, k \geq 0, j+k > 0\}.$$

For $\nu \in N_0$ we let

$$\phi_\nu(r,\theta) = (j+k+1)^{1/2}\; Z_{j+k}^{|j-k|}(r)\, e^{i(j-k)\theta},$$

where $(r,\theta) \in \mathcal{B}$ and where Z_m^l denotes the Zernike polynomial of degree m and order l (see Born and Wolf (1975), for example). Let

$$\psi_\nu(s,\phi) = U_{j+k}(s)\, e^{i(j-k)\phi},$$

where $(s,\phi) \in \mathcal{D}$ and where $U_m(\cos\theta) = \sin(m+1)\theta \,/\, \sin\theta$ are Chebyshev polynomials of the second kind. Then $\{1, \phi_\nu, \nu \in N_0\}$ and $\{1, \psi_\nu, \nu \in N_0\}$ form orthonormal bases on $\mathcal{B}$ and $\mathcal{D}$ respectively. The probability densities f and g can be written as $f = 1 + \sum_{\nu \in N_0} f_\nu\, \phi_\nu$ and $g = 1 + \sum_{\nu \in N_0} g_\nu\, \psi_\nu$. These functions will automatically integrate to one, but of course conditions will need to be imposed on the coefficients f_ν, to ensure that f is non-negative on $\mathcal{B}$. We adopt where necessary the treatment of Johnstone and Silverman (1990), identifying the complex orthonormal bases $\{1, \phi_\nu\}$ and $\{1, \psi_\nu\}$ with equivalent real orthonormal bases in a standard fashion; see their paper for details.

The mapping P admits a singular value decomposition with respect to the bases $\{1, \phi_\nu\}$ and $\{1, \psi_\nu\}$ such that $P1 = 1$ and

$$P\phi_\nu = b_\nu\, \psi_\nu,$$

for $\nu \in N_0$, where the singular values b_ν are specified by

$$b_\nu = (j+k+1)^{-1/2}.$$

Moreover, we can construct an orthonormal basis $\{\chi_\rho\}$ for $\mathcal{G}^d$ such that, for each $\nu \in N_0$, there exists a subscript $[\nu]$ for which, for certain constants γ_ν,

$$Q_d \psi_\nu \approx \gamma_\nu \chi_{[\nu]};$$

the constants γ_ν are given in Johnstone and Silverman (1991), as is a discussion about the accuracy of the approximation. From now on we assume that the elements of N_0 have been arranged in ordered groups with all the elements in each group mapping on to the same $[\nu]$. We now express the r_d-vector of observations Z as given by (9) in terms of the basis vectors χ_ρ as $Z = \sum_\rho \tilde{Z}_\rho \chi_\rho$. From now on we think of the r_d-vector $Y = \sqrt{n}\,\tilde{Z}$ as being our data.

Let Γ be the matrix representation of Q_d relative to the orthonormal bases we have defined. $\Gamma_{\rho\nu} = \gamma_j$ if $[\nu] = \rho$ and zero otherwise. Let B be a diagonal matrix with diagonal elements b_ν, $\nu \in N_0$. From equation (5) of Johnstone and Silverman (1991) it follows that $\mathbf{E}_f(Y) = \sqrt{n}\,\Gamma B f$, where the expectation is taken with respect to the underlying density function f for the original observations. Finally, let A be the diagonal matrix with elements $a_\nu^2 = (j+1)^{2a}(k+1)^{2a}$ for $\nu = (j,k) \in N_0$. Throughout we shall assume that $a > \frac{1}{2}$.

4.2 The minimax estimation of a linear functional of a PET image

We now have enough notation to begin our discussion of the minimax estimation of linear functionals of density functions f in the PET case. For any fixed $a > \frac{1}{2}$ and $C > 0$ let $\mathcal{H}$ be the Hilbert space of vectors $f = (f_\nu)_{\nu \in N_0}$ satisfying the condition $\sum_{\nu \in N_0} a_\nu^2 f_\nu^2 < \infty$, equipped with the inner product defined by

$$(f\,,\,g) = \frac{1}{C^2}\, f^T A\, g, \tag{10}$$

and associated norm $\|\cdot\|$. With each vector f of $\mathcal{H}$, we may associate the function $1 + \sum_{\nu \in N_0} f_\nu\, \phi_\nu$ on $\mathcal{B}$ that integrates to unity. Let $\mathcal{F}$ be the unit ball in $\mathcal{H}$. Vectors in $\mathcal{F}$ correspond to functions on $\mathcal{B}$ whose $2a^{\text{th}}$ derivatives exist and satisfy a weighted square-integrability condition; see Johnstone and Silverman (1990). If $a > \frac{1}{2}$ and $C \le 2^{a-\frac{1}{2}}$, vectors in $\mathcal{F}$ will correspond to functions on $\mathcal{B}$ that are nonnegative and hence density functions.

Let T be a bounded linear functional. The loss function that we shall consider is the surrogate mean square error

$$M^*\{w, T(f)\} = \{T(f) - w^T \eta(f)\}^2 + w^T w, \tag{11}$$

where $\eta(f) = \mathbf{E}_f(Y) = \sqrt{n}\,\Gamma B f$. For a discussion of the relationship between the surrogate and the exact mean square error

$$M\{w, T(f)\} = \mathbf{E}_f\{T(f) - w^T Y\}^2$$

see Johnstone and Silverman (1991).

A linear functional T on $\mathcal{H}$ is bounded if and only if

$$\sum_{\nu \in N_0} \frac{|\xi_\nu|^2}{a_\nu^2} < \infty, \tag{12}$$

where ξ_ν is the value of T applied to the vector with a one in the ν^{th} position and a zero elsewhere, since $\|T\|^2 = C^2 \sum_{\nu \in N_0} |\xi_\nu|^2 / a_\nu^2$ by elementary functional analysis. This result can also be applied to the linear functionals $\eta_\rho(f)$ to establish that $\|\eta_\rho\|^2 = n\, C^2 \beta_\rho$ where $\beta_\rho = \sum_{[\nu]=\rho} b_\nu^2\, \gamma_\nu^2 / a_\nu^2$. It is easy to show that β_ρ is finite under the assumption that $a > \frac{1}{2}$ with the consequence that the linear functionals η_ρ are bounded. Accordingly, provided (12) holds for T, Assumption 2 is valid. Theorem 1 now tells us that the linear minimax estimator of $T(f)$ with respect to the surrogate mean square error (11) is $T(\hat{f})$, where the vector $\hat{f}$ is the penalized least squares estimate of f given by

$$\{Y - \eta(f)\}^T \{Y - \eta(f)\} + \frac{1}{C^2} f^T A\, f.$$

Routine calculations establish that $\tilde{T} = C^2 A^{-1} \xi$, and that $\tilde{\eta} = \sqrt{n}\, C^2\, \Gamma B A^{-1}$, where $\tilde{\eta}$ is a matrix with r_d rows, the ρ^{th} row of which is the representer of $\eta_\rho(f)$ with respect to the inner product (10). Immediately, $H = n\, C^2\, \Gamma B A^{-1} B \Gamma^T$. Standard, but tedious manipulation yields that

$$\hat{w}_\rho = \frac{\sqrt{n}\, C^2\, \gamma_\rho}{1 + n\, C^2\, \beta_\rho}$$

that the vector f^* takes the form

$$f^* = \pm \frac{A^{-1}\,(\xi - \sqrt{n} B \Gamma^T \hat{w})}{\|A^{-1}\,(\xi - \sqrt{n} B \Gamma^T \hat{w})\|} \tag{13}$$

in terms of $\hat{w}$, where the vector $\xi = (\xi_\nu)_{\nu \in N_0}$, and that the value of M^* at f^* and $\hat{w}$ is

$$C^2 \sum_\rho \left(\alpha_\rho - C^2 \frac{n \gamma_\rho^2}{1 + n\, C^2\, \beta_\rho} \right), \tag{14}$$

where $\alpha_\rho = \sum_{[\nu]=\rho} \xi_\nu^2 / a_\nu^2$ and $\gamma_\rho = \sum_{[\nu]=\rho} b_\nu\, \gamma_\nu\, \xi_\nu / a_\nu^2$. It can be established that these quantities are finite if (12) holds.

It is easy to see from (14) that the minimaxed M^* tends to a strictly positive quantity as $n \to \infty$. Accordingly, it is impossible to reduce the minimax error for estimating a linear functional to zero, even if an infinite

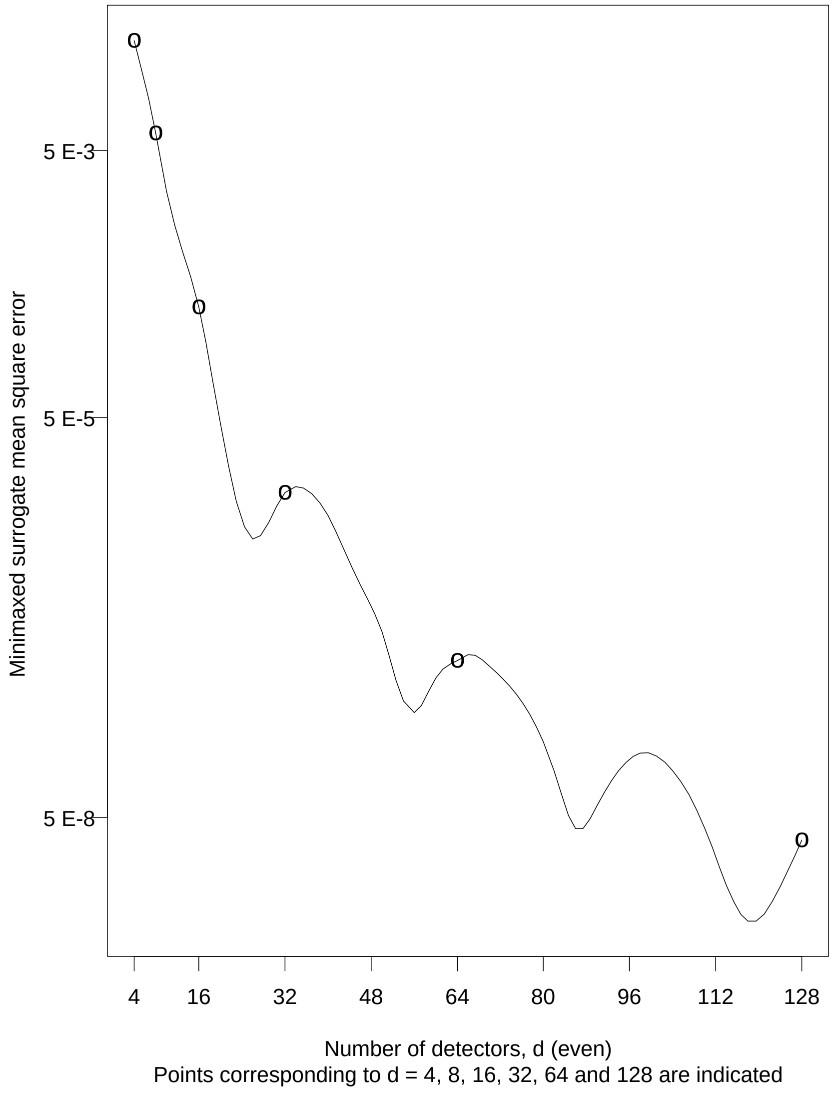

Figure 1: *The minimaxed surrogate mean square error (14) plotted on a logarithmic scale as a function of the number of detectors* d.

number of observations are taken. This is due to the finite nature of the detector ring with the consequence that the exact position of the lines of flight of the photon pairs is not known. Since the information available when there are d detectors can be obtained from the information available when there are $2d$ detectors, the minimaxed M^* decrease when the number of detectors is doubled.

In order to illustrate this phenomenon we define the functional T_1 on $\mathcal{H}$ as

$$T_1(f) = \frac{\displaystyle\int_{D\{(0.7,0.0),0.1\}} f \, d\mu}{\displaystyle\int_{D\{(0.7,0.0),0.1\}} d\mu},$$

where $D\{(x,y),r\}$ is the intersection of the disc of radius r centred at (x,y) with $\mathcal{B}$ and the function f on the right side is defined as $f = \sum_{\nu \in N_0} f_\nu \, \phi_\nu$; note that we do not include the constant function 1 in the definition of the function f. The functional T_1 represents the average of the function f over the disc $D\{(0.7, 0.0), 0.1\}$ and T_1 is a bounded linear functional on $\mathcal{H}$ if $a > 1$. The proof of this result is straightforward and uses the fact that $|Z_m^l(r)| \leq 1$ for $r \in [0,1]$; see Stander (1992).

Figure 1 shows a graph of the minimaxed M^* given by (14) for the bounded linear functional T_1 as a function of the number of detectors d. We have marked on the graph the minimaxed M^* for $d = 2$, 4, 8, 16, 32, 64 and 128. We can see that (14) does not decrease monotonically as d increases, but that doubling d does decrease the minimaxed M^*. The important message here is that merely increasing the number of detectors does not necessarily lead to an improvement in the quality of estimation.

In order to illustrate that the least favourable function f^* does depend on the functional being estimated we introduce a new functional T_2 on $\mathcal{H}$ as

$$T_2(f) = f\{(0.7, 0.0)\},$$

where the function f on the right side is defined as $f = \sum_{\nu \in N_0} f_\nu \, \phi_\nu$ as above. The functional T_2 represents the evaluation of the function f at the point $(0.7, 0.0)$ and T_2 is a bounded linear functional on $\mathcal{H}$ if $a > 1$. The proof of this result is virtually identical to that for T_1; again see Stander (1992).

From now on we shall associate with the least favourable function $f^* \in \mathcal{F}$ the function $f^*(x,y) = \sum_{\nu \in N_0} f_\nu^* \, \phi_\nu(x,y)$; again we do not include the constant function 1 in this definition. In Figure 2 we present for the functionals T_1 and T_2 contour plots of part of the function $f^*(x,y)$ for (x,y) in the square $[0.5, 0.9] \times [-0.2, 0.2] \subset \mathcal{B}$. The plots are produced using

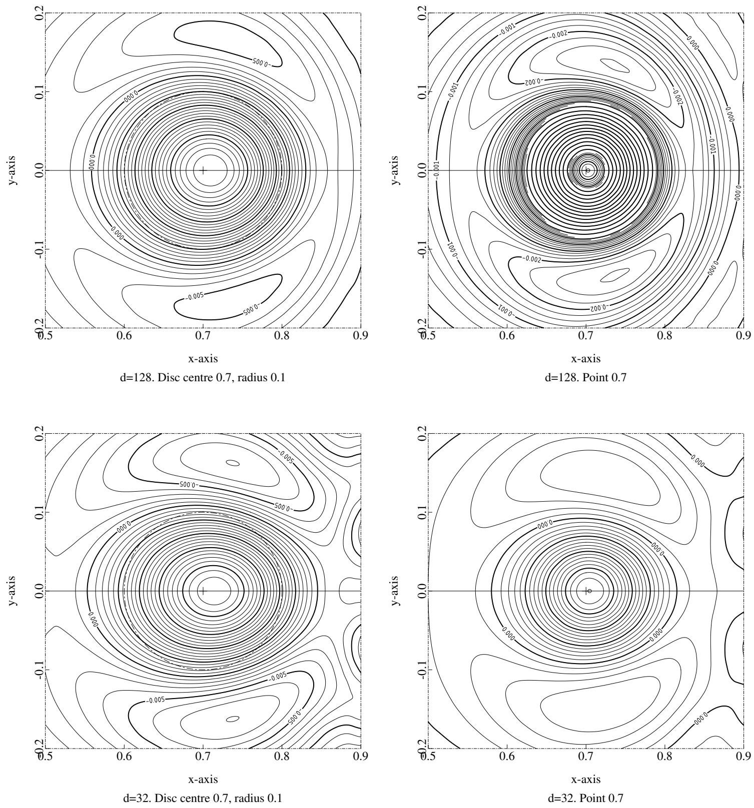

Figure 2: *Least favourable functions* $f^*(x, y)$. The left column corresponds to the functional T_1, while the right column corresponds to the functional T_2; in the top row there are $d = 128$ detectors, while in the bottom row $d = 32$. The point $(x, y) = (0.7, 0.0)$ is indicated on all the plots. In the left column the boundary of the disc $D\{(0.7, 0.0), 0.1\}$ is shown by the broken circle.

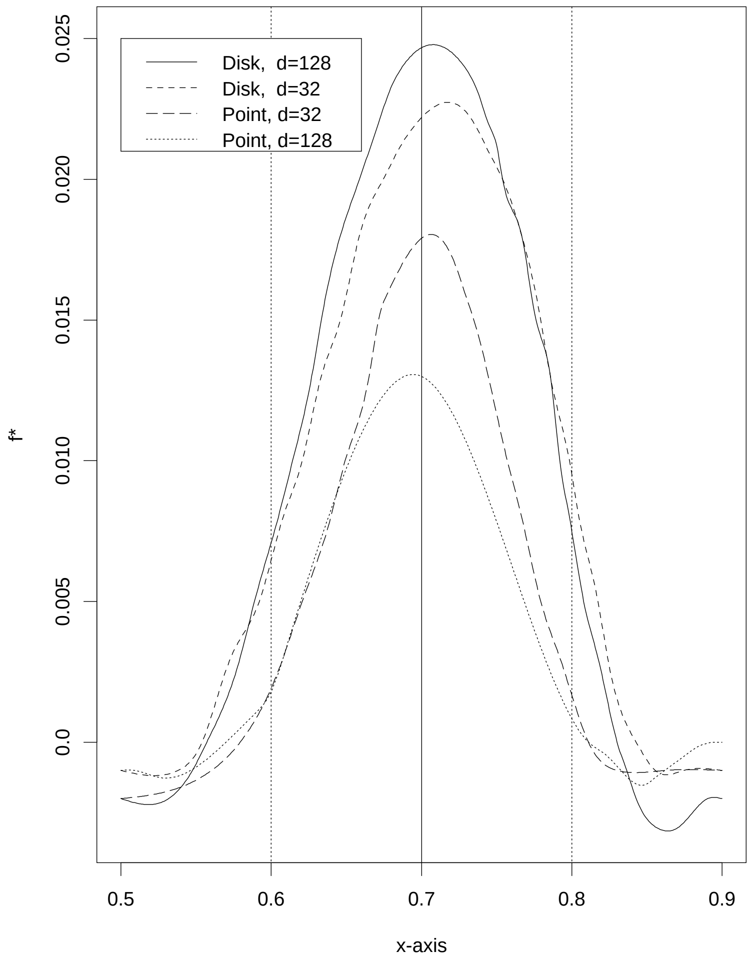

Figure 3: *Cross-section along the x-axis of the least favourable functions $f^*(x, y)$ shown in Figure 2.* The point $(x, y) = (0.7, 0.0)$ is indicated by the unbroken horizontal line, while the boundary of the disc $D\{(0.7, 0.0), 0.0\}$ is indicated by the broken horizontal lines.

the CONICON programs of Sibson (1987). The centre of this square is the point $(0.7, 0.0)$ which is indicated on the x-axis by a short vertical line. We choose to display all the $f^*(x, y)$ only in this square because elsewhere they show much less variation, being virtually flat. Since the remaining part of the expansion is defined up to sign, we choose the sign so that $f^*\{(0.7, 0.0)\}$ is positive. The heights printed on the contour plots follow the convention of having the top of the number pointing in the direction of increasing height; the global maximum of all the $f^*(x, y)$ is near the point $(0.7, 0.0)$. The actual values that the least favourable functionals take are arbitrary in the sense that multiplying C by a particular factor causes the values to be multiplied by that factor. For our figures we take $a = 1.5$ and $C = 2$ so ensuring that all vectors $(f_\nu)_{\nu \in N_0}$ in $\mathcal{F}$ correspond to *densities* $1 + \sum_{\nu \in N_0} f_\nu \phi_\mu$ on $\mathcal{B}$. The left column of Figure 2 corresponds to the functional T_1, while the right column corresponds to T_2. The least favourable function $f^*(x, y)$ shown in the top row correspond to the case when there are are $d = 128$ detectors, while in the bottom row $d = 32$. The boundary of the disc $D\{(0.7, 0.0), 0.1\}$ is indicated on the plots corresponding to T_1 by the broken circle.

In Figure 3 we show the cross-section along the x-axis of the least favourable functions $f^*(x, y)$ presented in Figure 2. The point $(x, y) = (0.7, 0.0)$ is indicated by the unbroken horizontal line, while the boundary of the disc $D\{(0.7, 0.0), 0.0\}$ is indicated by the broken horizontal lines. The $f^*(x, y)$ corresponding to T_1 changes rapidly over the boundary of this disc to reach its global maximum near $(0.7, 0.0)$. Similarly, the $f^*(x, y)$ corresponding to T_2 changes rapidly about $(0.7, 0.0)$ to reach its global maximum near that point. The f^* corresponding to T_1 is more spread out than those corresponding to T_2. This is not surprising since $T_1(f)$ represents the average of the function f over the disc $D\{(0.7, 0.0), 0.1\}$, whereas $T_2(f)$ represents the function f evaluated at the point $(0.7, 0.0)$. Although the plots for T_1 are quite similar to those for T_2, it can easily be seen that the least favourable function *does* depend on the functional being estimated. Moreover, different values of d yield different least favourable functions.

5 Conclusions

In this paper we consider the linear minimax estimation of bounded linear functionals T with respect to the mean square error loss function over a certain class of functions $\mathcal{F}$ defined by means of a semi-norm. We review the result, essentially due to Speckman (1979), that the linear minimax estimate of $T(f)$ is $T(\hat{f})$, where $\hat{f}$ is a type of penalized least squares estimate of f.

We provide a general proof of this result that we believe to be more accessible than those that have so far appeared.

The above result is remarkable in that it says that the amount and method of smoothing carried out in the minimax estimation of $T(f)$ does not depend upon T. Instead, the minimax way to proceed is to construct the estimator $\hat{f}$ and then to evaluate $T(\hat{f})$. The least favourable function $f^* \in \mathcal{F}$ for which the maximum of the loss function is attained does, however, depend upon T.

We apply this result to the case of nonparametric regression where a function f defined on a finite interval $I \subset R$ and belonging to a certain smoothness class is observed, subject to error, at points $t_1, \ldots, t_n$. We consider the minimax estimation of the linear functional T_t, where $T_t(f) = f(t)$, and establish that the minimax estimate of $T_t(f)$ is given by $\hat{f}(t)$, where $\hat{f}$ is a certain cubic smoothing spline, simultaneously for all $t \in I$. However, if a minimax estimator of f were to be sought using a more global loss function, namely the sum of the mean square error over the points $t_1, \ldots, t_n$, the estimator that results would be different from $\hat{f}$. We explain how this provides us with a paradox as we might suppose that $\hat{f}$ should be the linear minimax estimator with respect to mean summed square error loss. We resolved the paradox by noting that f^* depends on T_t.

We also apply the result to the case of Positron Emission Tomography. We give the form of the linear minimax estimator, of the least favourable function f^*, and of the value of the surrogate mean square error evaluated at f^* and the minimax estimator. We present a numerical example that shows that the minimaxed surrogate mean square error does not decrease monotonically as the number of detectors increases; merely increasing the number of detectors does not necessarily lead to an improvement in the quality of estimation. Finally, we display least favourable functions for two linear functionals that clearly show the dependence of the least favourable function on the linear functional under consideration.

6 Some Proofs

In this section we give a proof of Theorem 1 in which we also derive the results of Theorem 2.

6.1 Notation and some preliminary lemmas

Recall that, under Assumption 1, $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{K}$, and consider for the moment the linear functionals η_i as being restricted to $\mathcal{K}$. The mapping η defined on $\mathcal{K}$ as $\eta(k) = \{\eta_1(k), \ldots, \eta_n(k)\}^T$ for $k \in \mathcal{K}$ is a linear transformation form $\mathcal{K}$ to R^n. By Assumption 1 the null space of η is $\{0\}$. Hence $\mathcal{K}$ is finite dimensional with dimension $r \leq n$, say. Assumption 1 also shows that the bilinear form $[\cdot,\cdot]$ is an inner product on $\mathcal{K}$. Let $\phi_1, \ldots, \phi_r$ be an orthonormal basis for $\mathcal{K}$ relative to the inner-product $[\cdot,\cdot]$, and recall that ϕ is the vector $(\phi_1, \ldots, \phi_r)^T$ and Φ is the $n \times r$ matrix with entries $\Phi_{ij} = \eta_i(\phi_j)$ such that $\Phi^T\Phi = I$.

Given any $f \in \mathcal{H}$, write

$$f = f_0 + \sum_{j=1}^{r} [f, \phi_j]\, \phi_j. \tag{15}$$

By the orthonormality properties of the ϕ_j, it follows at once that $[f_0, \phi_j] = 0$ for all j, and hence $f_0 \in \mathcal{H}_0$. Accordingly, equation (15) represents the decomposition of a function $f \in \mathcal{H}$ according to the relationship $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{K}$. If we define f_1 to be the vector $([f,\phi_1], \ldots, [f,\phi_r])^T$, then the decomposition becomes

$$f = f_0 + f_1^T \phi. \tag{16}$$

By applying Cauchy-Schwarz to $(\phi_j\,,\,g)_s^{\,2}$ and using $\|\phi_j\|_s = 0$, we obtain that $(\phi_j\,,\,g)_s = 0$ for all $g \in \mathcal{H}$. From this we immediately have the following lemma:

Lemma 1 *For any $f \in \mathcal{H}$, $\|f\|_s^2 = \|f_0\|_s^2 = (f_0\,,\,f_0)_s$, where $f_0 \in \mathcal{H}_0$ is defined by (15).*

We can now state another lemma, the proof of which follows directly by applying the linear transformation $\eta : \mathcal{H} \to R^n$ to the decomposition of $f \in \mathcal{H}$ as given by (15):

Lemma 2 *For arbitrary $f \in \mathcal{H}$, we have*

$$\eta(f) = (f_0\,,\,\tilde{\eta})_s + \Phi f_1,$$

where $f_0 \in \mathcal{H}_0$ is given by (15).

Similarly, by substituting the decomposition (15) into the bilinear form $[\cdot,\cdot]$ and using the fact that $\phi_1, \ldots, \phi_r$ form an orthonormal basis for $\mathcal{K}$ relative to the inner product defined by the bilinear form, we obtain the following lemma:

Lemma 3 *For arbitrary $f \in \mathcal{H}$, we have*

$$[f, f] = [f_0, f_0] + f_1^T f_1.$$

We also need a lemma that gives the minimum of a certain expression over $f_0 \in \mathcal{H}_0$.

Lemma 4 *The function $\hat{f}_0 = Y^T(I+H)^{-1}\tilde{\eta}$ mininizes*

$$Y^T Y - 2Y^T\left(f_0\,,\,\tilde{\eta}\right)_s + \left(\tilde{\eta}^T\,,\,f_0\right)_s\left(f_0\,,\,\tilde{\eta}\right)_s + \left(f_0\,,\,f_0\right)_s$$

over $f_0 \in \mathcal{H}_0$.

The lemma follows by noting that for any function $g_0 \in \mathcal{H}_0$

$$\begin{aligned}
&Y^T\left(g_0\,,\,\tilde{\eta}\right)_s - \left(\tilde{\eta}^T\,,\,g_0\right)_s\left(\hat{f}_0\,,\,\tilde{\eta}\right)_s - \left(g_0\,,\,\hat{f}_0\right)_s \\
&\quad= Y^T\left(g_0\,,\,\tilde{\eta}\right)_s - \left(\tilde{\eta}^T\,,\,g_0\right)_s\left(Y^T(I+H)^{-1}\tilde{\eta}\,,\,\tilde{\eta}\right)_s - \left(g_0\,,\,Y^T(I+H)^{-1}\tilde{\eta}\right)_s \\
&\quad= \left(g_0\,,\,Y^T\left\{I - (I+H)^{-1}\left(\tilde{\eta}\,,\,\tilde{\eta}^T\right)_s - (I+H)^{-1}\right\}\tilde{\eta}\right)_s \\
&\quad= \left(g_0\,,\,Y^T\left\{I - (I+H)^{-1}H - (I+H)^{-1}\right\}\tilde{\eta}\right)_s \\
&\quad= 0.
\end{aligned}$$

Writing $f_0 = \hat{f}_0 + g_0$ and using this calculation, we obtain

$$\begin{aligned}
&Y^T Y - 2Y^T\left(f_0\,,\,\tilde{\eta}\right)_s + \left(\tilde{\eta}^T\,,\,f_0\right)_s\left(f_0\,,\,\tilde{\eta}\right)_s + \left(f_0\,,\,f_0\right)_s \\
&\quad= Y^T Y - 2Y^T\left(\hat{f}_0\,,\,\tilde{\eta}\right)_s + \left(\tilde{\eta}^T\,,\,\hat{f}_0\right)_s\left(\hat{f}_0\,,\,\tilde{\eta}\right)_s + \left(\hat{f}_0\,,\,\hat{f}_0\right)_s \\
&\qquad+ \left(\tilde{\eta}^T\,,\,g_0\right)_s\left(g_0\,,\,\tilde{\eta}\right)_s + \left(g_0\,,\,g_0\right)_s. \qquad (17)
\end{aligned}$$

It follows at once that (17) is minimized when $g_0 = 0$. This proves the lemma.

We now produce an expression for $T(f)$ in terms of f_0 and f_1.

Lemma 5 *For arbitrary $f \in \mathcal{H}$, we have*

$$T(f) = \left(f_0\,,\,\tilde{T}\right)_s + f_1^T t.$$

The lemma follows immediately by applying T to the expansion of f given in (15).

We finish this section with a simple lemma:

Lemma 6 *For arbitrary $g_0 \in \mathcal{H}_0$, we have*

$$(\tilde{\eta}^T\,,\,g_0)_s\,\Phi = 0.$$

To prove the lemma consider the jth component of $(\tilde{\eta}^T\,,\,g_0)_s\,\Phi$. We have

$$\{(\tilde{\eta}^T\,,\,g_0)_s\,\Phi\}_j = (\tilde{\eta}^T\,,\,g_0)_s\,\eta(\phi_j) = \eta(g_0)^T\eta(\phi_j) = [g_0, \phi_j] = 0,$$

since $\phi_j \in \mathcal{K}$.

6.2 Some important lemmas

We now give two important lemmas from which the proof of Theorem 1 follows immediately.

6.2.1 The form of the penalized least squares estimator

First we derive the form of the penalized least squares estimator $\hat{f}$.

Lemma 7 *The penalized least squares estimator is given by*

$$\hat{f} = Y^T(I+H)^{-1}\tilde{\eta} + Y^T\Phi\phi.$$

To prove the result, use (16) to expand the penalized least squares in terms of the function f_0 and the vector f_1 to give

$$\begin{aligned}
&\{Y - \eta(f)\}^T\{Y - \eta(f)\} + \|f\|_s^2 \\
&\quad = \; Y^TY - 2Y^T\eta(f) + [f, f] + (f_0\,,\,f_0)_s \\
&\quad = \; Y^TY - 2Y^T\{(f_0\,,\,\eta)_s + \Phi f_1\} + [f_0, f_0] + f_1^Tf_1 + (f_0\,,\,f_0)_s \\
&\quad = \; Y^TY - 2Y^T(f_0\,,\,\eta)_s + (\tilde{\eta}^T\,,\,f_0)_s\,(f_0\,,\,\tilde{\eta})_s \\
&\qquad\quad + (f_0\,,\,f_0)_s + f_1^Tf_1 - 2Y^T\Phi f_1,
\end{aligned}$$

where the first equality follows from the definition of the bilinear form and Lemma 1 and the second equality follows from Lemma 2 and Lemma 3. This expression can now be minimized separately over f_0 and f_1. Lemma 4 yields

$$\hat{f}_0 = Y^T(I+H)^{-1}\tilde{\eta}$$

for the minimization over f_0. Standard matrix manipulations yield

$$\hat{f}_1 = \Phi^TY.$$

Hence, using (16) we obtain the expression for $\hat{f}$ given in the lemma.

6.2.2 The minimax linear estimator of a linear functional

We now turn to the question of finding the linear minimax estimator of $T(f)$.

Lemma 8 *The linear minimax estimator of* $T(f)$ *is given by* $\hat{w}^T Y$ *where*

$$\hat{w} = (I+H)^{-1}\left(\tilde{T}\,,\,\tilde{\eta}\right)_s + \Phi t.$$

To prove the lemma note that we have

$$\begin{aligned}\sup_{f\in\mathcal{F}} \mathbf{E}\left\{T(f) - w^T Y\right\}^2 &= \sup_{f\in\mathcal{F}} \left\{T(f) - w^T\eta(f)\right\}^2 + w^T w \\ &= \sup_{f\in\mathcal{F}} \left\{\left(f_0\,,\,\tilde{T}\right)_s - w^T\left(f_0\,,\,\tilde{\eta}\right)_s + t^T f_1 - w^T\Phi f_1\right\}^2 + w^T w,\end{aligned}$$

where the first equality follows by standard properties of the mean square error and the assumption that $\mathbf{Var}(Y) = I$, and the second equality follows from Lemma 2 and Lemma 5. Finding the maximum of this expression is equivalent to taking an unconstrained maximum over $f_1 \in \mathcal{K}$ and a maximum over f_0 in the unit ball in $\mathcal{H}_0$. The maximum over f_1 will only be finite if

$$\Phi^T w = t, \tag{18}$$

and subject to this constraint we have

$$\begin{aligned}\mathbf{E}\left\{w^T Y - T(f)\right\}^2 &= \left\{\left(f_0\,,\,\tilde{T}\right)_s - w^T\left(f_0\,,\,\tilde{\eta}\right)_s\right\}^2 + w^T w \\ &= \left(f_0\,,\,\tilde{T} - w^T\tilde{\eta}\right)_s^2 + w^T w.\end{aligned}$$

Subject to the constraint (18), the mean square error will not depend upon f_1 at all. The least favourable f_0 in the unit ball in $\mathcal{H}_0$ can be found by standard functional analysis to be

$$f_0 = \pm\frac{\tilde{T} - w^T\tilde{\eta}}{\left\|\tilde{T} - w^T\tilde{\eta}\right\|_s}. \tag{19}$$

Hence, the maximum mean square error will be

$$\left\|\tilde{T} - w^T\tilde{\eta}\right\|_s^2 + w^T w = w^T(I+H)w - 2w^T\left(\tilde{T}\,,\,\tilde{\eta}\right)_s + \left(\tilde{T}\,,\,\tilde{T}\right)_s. \tag{20}$$

A standard Lagrange multiplier argument shows that the constrained minimum of (20) will be given by

$$\hat{w} = (I + H)^{-1}\left\{\left(\tilde{T}\,,\,\tilde{\eta}\right)_s + \Phi\lambda\right\} \tag{21}$$

where λ is a vector of Lagrange multipliers chosen to ensure that $\Phi^T\hat{w} = t$. By Lemma 6 it follows that $\Phi^T\left(\tilde{T}\,,\,\tilde{\eta}\right)_s = 0$. Also $H\Phi = (\tilde{\eta}^T\,,\,\tilde{\eta})_s\,\Phi = 0$ with the consequence that $(I + H)\Phi = \Phi$. Using the expression (21) for $\hat{w}$ we obtain

$$\Phi^T\hat{w} = \Phi^T(I + H)\hat{w} = \Phi^T\left\{\left(\tilde{T}\,,\,\tilde{\eta}\right)_s + \Phi\lambda\right\} = \Phi^T\left(\tilde{\eta}\,,\,\tilde{T}\right)_s + \lambda = \lambda.$$

Therefore $\lambda = t$ and, substituting back into (21), we have

$$\hat{w} = (I + H)^{-1}\left(\tilde{T}\,,\,\tilde{\eta}\right)_s + \Phi t. \tag{22}$$

6.3 Completion of the proof

Armed with the above lemmas, we can easily complete the proof of the theorem.

From the expression for the penalized least squares estimator $\hat{f}$ given in Lemma 7 and the definition of the vector t, we have that

$$\begin{aligned} T(\hat{f}) &= Y^T(I + H)^{-1}T(\tilde{\eta}) + Y^T\Phi\, T(\phi) \\ &= Y^T(I + H)^{-1}\left(\tilde{T}\,,\,\tilde{\eta}\right)_s + Y^T\Phi t \\ &= Y^T\left\{(I + H)^{-1}\left(\tilde{T}\,,\,\tilde{\eta}\right)_s + \Phi t\right\} \\ &= Y^T\hat{w}, \end{aligned}$$

which is the linear minimax estimator of $T(f)$ as required.

We finish by deriving expressions for the least favourable function f^* and the mean square error evaluated at f^* and $\hat{w}$. For the former we easily obtain that f^* takes the form

$$\pm\,\frac{\tilde{T} - w^T\tilde{\eta}}{\left\|\tilde{T} - w^T\tilde{\eta}\right\|_s} + k \text{ for any } k \in \mathcal{K}, \tag{23}$$

by considering the least favourable $f_0 \in \mathcal{H}_0$ given by (19). For the latter we substitute (22) into (20) to obtain

$$\left(\tilde{T}\,,\,\tilde{T}\right)_s - \left(\tilde{\eta}^T\,,\,\tilde{T}\right)_s(I + H)^{-1}\left(\tilde{T}\,,\,\tilde{\eta}^T\right)_s + t^Tt.$$

Acknowledgements

We are grateful to two anonymous referees for helpful comments and suggestions that have clarified the presentation of this article.

References

Adams, R.A. (1975), Sobolev Spaces, Academic Press–New York.

Bickel, P.J. and Ritov, Y. (1995), Estimating Linear Functionals of a PET Image. IEEE Trans. Med. Imgng., 14, 1, 81–87.

Born, M. and Wolf, E. (1975), Principles of Optics, 5th edn., Pergamon–New York.

Donoho, D.L. and Liu, R.C. (1991), Geometrizing Rates of Convergence, III. Ann. Statis., 19, 2, 668–701.

Donoho, D.L. and Low, M.G. (1992), Renormalization Exponents and Optimal Pointwise Rates of Convergence. Ann. Statis., 20, 2, 944–970.

Geman, S. and McClure, D.E. (1987), Statistical Methods for Tomographic Image Reconstruction. Bull. Int. Statist. Inst., 52, 4, 5–23.

Green, P.J. (1990), Bayesian Reconstructions from Emission Tomography Data Using a Modified EM Algorithm. IEEE Trans. Med. Imgng., MI-9, 1, 84–93.

Green, P.J. and Silverman, B.W. (1994), Nonparametric Regression and Generalized Linear Models, Chapman and Hall–London.

Ibragimov, I.A. and Khas'minskii, R.Z. (1984), On Nonparametric Estimation of the Value of a Linear Functional in Gaussiam White Noise. Th. Prob. Appl., 29, 1, 18–32.

Johnstone, I.M. and Silverman, B.W. (1990), Speed of Estimation in Positron Emission Tomography and Related Inverse Problems. Ann. Statist., 18, 1, 251–280.

Johnstone, I.M. and Silverman, B.W. (1991), Discretization Effects in Statistical Inverse Problems. Journal of Complexity, 7, 1, 1–34.

Jones, M.C. and Silverman, B.W. (1989), An Orthogonal Series Density Estimation Approach to Reconstructing Positron Emission Tomography Images. J. Appl. Stat., 16, 2, 177–191.

Kreyszig, E. (1978), Introductory Functional Analysis with Applications, Wiley–New York.

Li K.-C. (1982), Minimaxity of the Method of Regularization on Stochastic Processes, Ann. Statist., 10, 3, 937–942.

Sibson, R. (1987), `CONICON3` Handbook, School of Mathematical Sciences, University of Bath.

Silverman, B.W., Jones, M.C., Wilson, J.D. and Nychka, D.W. (1990), A Smoothed EM Approach to Indirect Estimation Problems, with Particular Reference to Stereology and Emission Tomography (with discussion). J. R. Statist. Soc. B, 52, 2, 271–324.

Speckman, P. (1979), Minimax Estimates of Linear Functionals in a Hilbert Space. Unpublished manuscript.

Speckman, P. (1985), Spline Smoothing and Optimal Rates of Convergence in Nonparametric Regression Models. Ann. Statist., 13, 3, 970–983.

Stander, J. (1992), Some Topics in Statistical Image Analysis, PhD thesis–University of Bath.

Vardi, Y., Shepp, L.A. and Kaufman, L. (1985), A Statistical Model for Positron Emission Tomography. J. Amer. Statist. Assoc., 80, 389, 8–37.