A Short Introduction to Inverse Statistical Inference

2001
l´Odyssée de la Statistique
Institut Henri Poincaré
11, rue Pierre et Marie Curie
F-75231 Paris Cedex 05
FRANCE

2002
Department of Statistics
Goettingen University
GERMANY

Frits H. Ruymgaart
Department of Mathematics and Statistics
Texas Tech University
Lubbock, TX 79409
USA

Contents

1 Introduction and Examples

1. General introduction. These notes are concerned with statistical inference about the *input* of a system, when noisy observations on the *output* are given. This justifies the word “ inverse ” in the title. Typically, the output is a transform of the input. To recover information regarding the input from the output of the system, this transformation, which is supposed to be known,

$$\xrightarrow[\text{input}]{\text{unknown}} \boxed{\begin{array}{c}\text{known}\\ \text{transformation}\end{array}} \xrightarrow[\text{output}]{\text{observed}}$$

must be inverted. Such an inversion is usually an unstable procedure that requires regularization. This is the reason that these inverse problems are in general ill-posed. The term “ ill-posed ” was originally coined by Hadamard in the context of partial differential equations. Since in all cases considered here both the input and output can be represented as functions, we may also refer to the current topic as inference for *indirectly observed curves.* Formally this contains the area of inference regarding ordinary, directly observed curves as a special case, if the transformation is the identity. Since usually the transformation above will be an integral transformation we may also refer to this area as the theory of *noisy integral equations.*

2. Warning. Statistical inference of the type sketched above is of modern type, because it involves an *infinite dimensional parameter* rather than a finite dimensional as in classical statistics. Because it also involves an inverse problem, an interesting blend of analytical and up-to-date statistical techniques is required. We will *not* dwell too much on assumptions and regularity conditions. In particular we will satisfy ourselves with heuristic sketches rather than with rigorous proofs. For the most part, moreover, we will restrict ourselves to the *indirect regression model.* Let us now consider some examples. In each example “ local ” notation will be used.

3. Ridge regression: a f.d. paradigm. Consider the linear model

$$Y = X\beta + \epsilon,$$

where Y and ϵ are $n \times 1$ random vectors, X is an $n \times m$ design matrix of full rank, and β the unknown $m \times 1$ vector-parameter. Since we observe Y we may say that β is observed indirectly and with random error. The classical least squares estimator is

$$\hat{\beta} = (X^*X)^{-1}X^*Y.$$

4. There exists an orthonormal transformation O such that

$$X^*X = O^{-1}\Lambda O, \quad \Lambda = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_m \end{pmatrix},$$

with $\lambda_j > 0$ because X^*X is strictly positive symmetric. Note that $\hat{\beta}$ is unbiased. For the risk we find

$$\begin{aligned} \mathbf{E}\,\|\hat{\beta} - \beta\|^2 &= \mathbf{E}\,\|O^{-1}\Lambda^{-1}OX^*(X\beta + \epsilon) - \beta\|^2 \\ &= \mathbf{E}\,\|\Lambda^{-1}OX^*\epsilon\|^2 \\ &= \mathbf{E}\,\epsilon^* XO^{-1}\Lambda^{-2}OX^*\epsilon \\ &= \sigma^2\ \mathbf{tr}\ O^{-1}\Lambda^{-1}O \\ &= \sigma^2 \sum_{j=1}^{m} 1/\lambda_j. \end{aligned}$$

If some columns of X are close (collinearity) there may be small eigenvalues and the risk will be large.

5. This is why modified estimators, the so-called *ridge estimators*

$$\hat{\beta}_\alpha = (\alpha I + X^*X)^{-1}X^*Y, \;\; \alpha \geq 0,$$

are proposed in the literature. We will see that they may have smaller risk in some cases but they are no longer unbiased. Note that

$$\alpha I + X^*X = O^{-1}\Lambda_\alpha O, \; \Lambda_\alpha = \alpha I + \Lambda = \begin{pmatrix} \alpha + \lambda_1 & 0 & \dots & 0 \\ 0 & \alpha + \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \alpha + \lambda_m \end{pmatrix}.$$

We have

$$\begin{aligned}\beta_\alpha &= \mathbf{E}\,\hat{\beta}_\alpha\\ &= (\alpha I + X^*X)^{-1}X^*X\beta\\ &= O^{-1}I_\alpha O\beta,\end{aligned}$$

$$I_\alpha = \begin{pmatrix} \lambda_1/(\alpha+\lambda_1) & 0 & \dots & 0\\ 0 & \lambda_2/(\alpha+\lambda_2) & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & \lambda_m/(\alpha+\lambda_m)\end{pmatrix},$$

and for the risk we now find

$$\begin{aligned}\mathbf{E}\,\|\hat{\beta}_\alpha-\beta\|^2 &= \mathbf{E}\,\|\hat{\beta}_\alpha-\beta_\alpha\|^2 + \|\beta_\alpha-\beta\|^2\\ &= \mathbf{E}\,\|\Lambda_\alpha^{-1}OX^*\epsilon\|^2 + \|\beta_\alpha-\beta\|^2\\ &= \sigma^2\ \mathbf{tr}\ O^{-1}\Lambda\Lambda_\alpha^{-2}O + \|(I_\alpha - I)O\beta\|^2\\ &= \sigma^2\sum_{j=1}^m \lambda_j/(\alpha+\lambda_j)^2 + \alpha^2\sum_{j=1}^m \widetilde{\beta}_j^{\,2}/(\alpha+\lambda_j)^2\\ &= r_1(\alpha)+r_2(\alpha),\ \ \alpha\ge 0,\end{aligned}$$

where $\widetilde{\beta} = O\beta$. Because $r_1(\alpha)\downarrow 0$, as $\alpha\uparrow\infty$, with $r_1(0) = \mathbf{E}\,\|\hat{\beta}-\beta\|^2$, and $r_2(\alpha)\uparrow\|\widetilde{\beta}\|^2 = \|\beta\|^2$, as $\alpha\uparrow\infty$, with $r_2(0)=0$,

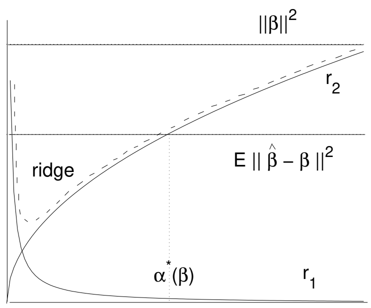

Figure 1:

it is clear from the picture that for each β there exists $\alpha^*(\beta)$ such that

$$\mathbf{E}\,\|\hat{\beta}_\alpha-\beta\|^2 \le \mathbf{E}\,\|\hat{\beta}-\beta\|^2,\ \text{ for all }\ 0\le\alpha\le\alpha^*(\beta).$$

Although in this finite dimensional situation the operator X^*X still has a continuous inverse, so that the regularization of the inverse in §1.5 is not absolutely necessary, the situation becomes different in the infinite dimensional case where regularization will be indispensible.

6. Computer tomography. Consider a fixed plane through a physical body, and let $\rho(x, y)$ denote its density at the point (x, y) of the plane. Let L be any line in the plane. Supposed that a thin beam of X-rays is directed

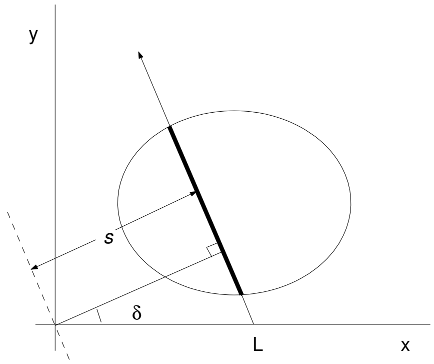

Figure 2:

into the body along this line L and that the attenuation of the intensity by going through the body is measured for several lines. How can the density ρ be recovered from these data?

7. Parametrize L by (s, δ). Using complex numbers the points on the ray $L = L_{s,\delta}$ can be written

$$se^{i\delta} + iue^{i\delta}, \;\; u \in \mathbb{R}.$$

Note that $e^{i\delta} \perp ie^{i\delta}$. The attenuation of the intensity I is approximately described by $\Delta I = -\gamma \cdot \rho(u) \cdot I \cdot \Delta u$, for some constant γ. The total attenuation along the ray then satisfies

$$\log I = -\gamma \int_{-\infty}^{\infty} \rho(se^{i\delta} + iue^{i\delta})\, du.$$

In principle, from these attenuation factors we can compute all the line integrals

$$(R\rho)(s, \delta) = \int_{-\infty}^{\infty} \rho(se^{i\delta} + iue^{i\delta})\, du, \;\; s \in \mathbb{R}, \;\; \delta \in [0, \pi).$$

The l.h.s. is called the Radon transform of ρ.

8. The problem can be significantly simplified when we may assume that ρ is radially symmetric. This entails that it suffices to take rays in one direction only, let us say perpendicular to the x-axis. Then we have $\rho = \rho(r),\ r = \sqrt{x^2+y^2}$, and the ray $L = L_x$ passing through $(x, 0)$ can be parametrized by $(x, u),\ u \in \mathbb{R}$. But then we have

$$\log I = V(x) = -2\gamma \int_0^\infty \rho(\sqrt{x^2+u^2})\, du,$$

the factor 2 being due to symmetry. If ρ has compact support in the disc

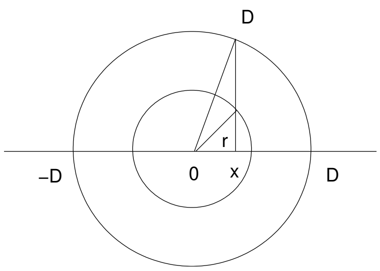

Figure 3:

$\{(x, y) : \sqrt{x^2+y^2} \leq D\}$ this integral equals

$$V(x) = -2\gamma \int_0^{\sqrt{D^2-x^2}} \rho(\sqrt{x^2+u^2})\, du.$$

Making the change of variables $u = \sqrt{r^2-x^2}$ we obtain $x^2+u^2 = r^2,\ du = (r/\sqrt{r^2-x^2})\, dr,\ u = 0$ iff $r = x,\ u = \sqrt{D^2-x^2}$ iff $r = D$, and we arrive at

$$V(x) = -2\gamma \int_x^D \frac{r}{\sqrt{r^2-x^2}} \rho(r)\, dr.$$

Making a further change of variable set $z = D^2 - r^2$, and write $y = D^2 - x^2$. Then we have $dr = (-1/2r)\, dz$, $r = x$ iff $z = D^2 - x^2$, $r = D$ iff $z = 0$, $r^2 - x^2 = y - z$, and the equation reduces to

$$V(\sqrt{D^2-y}) = -\gamma \int_0^y \frac{\rho(\sqrt{D^2-z})}{\sqrt{y-z}}\, dz,\ 0 \leq y \leq D.$$

If we write $\widetilde{\rho}(z) = \rho(\sqrt{D^2 - z})$ and $\widetilde{V}(y) = V(\sqrt{D^2 - y})$ we finally arrive at

$$\widetilde{V}(y) = -\gamma \int_0^y \frac{\widetilde{\rho}(z)}{\sqrt{y-z}}\, dz,\ \ 0 \le y \le D.$$

This is an example of *Abel's integral equation.*

9. Typically the data are random variables

$$\xi_i = (X_i, Y_i),\ \ Y_i = \widetilde{V}(X_i) + \epsilon_i,\ \ i = 1, \ldots, n,$$

where the X_i are given fixed design points or i.i.d. random variables and the ϵ_i are i.i.d. random measurement errors, independent of the X_i. The problem is to recover $\widetilde{\rho}$ from these data.

10. Spectroscopic binary orbits. Suppose that only the projections of vectors onto a plane perpendicular to the line of sight can be measured. Let the random variable L denote the actual length of the vector with density f, and the random variable X the length of the projection with density g. Let the random variable Θ be the angle between the actual vector and the line of sight. We will assume that Θ and L are indepedent and that Θ

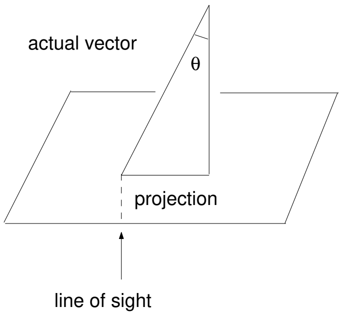

Figure 4:

has the Uniform $(0, \frac{\pi}{2})$ distribution. The data consist of independent copies $X_1, \ldots, X_n$ of X and the problem is to recover f from the sample.

11. Because $X = L\sin\Theta$ it follows that

$$\begin{aligned}
g(x) &= \frac{d}{dx}\mathbf{P}\{X \le x\} \\
&= \frac{2}{\pi}\frac{d}{dx}\int_0^{\pi/2} \mathbf{P}\{L\sin\Theta \le x | \Theta = \theta\}\ d\theta \\
&= \frac{2}{\pi}\frac{d}{dx}\int_0^{\pi/2} F\left(\frac{x}{\sin\theta}\right)\ d\theta \\
&= \frac{2}{\pi}\int_0^{\pi/2} f\left(\frac{x}{\sin\theta}\right)\frac{1}{\sin\theta}\ d\theta.
\end{aligned}$$

Replacing x with $1/\sqrt{x}$ leads to

$$g\left(\frac{1}{\sqrt{x}}\right) = \frac{2}{\pi}\int_0^{\pi/2} f\left(\frac{1}{\sqrt{x}\sin\theta}\right)\frac{1}{\sin\theta}\ d\theta.$$

With the new variables $\sqrt{y} = \sqrt{x}\sin\theta,\ \sqrt{t} = \sqrt{x}\cos\theta$, we have $x = y + t,\ d\theta = dy/2\sqrt{yt},\ \theta = 0$ iff $y = 0, \theta = \frac{\pi}{2}$ iff $y = x$,and we obtain

$$\frac{1}{\sqrt{x}}\, g\left(\frac{1}{\sqrt{x}}\right) = \frac{1}{\pi}\int_0^x \frac{1}{y}\, f\left(\frac{1}{\sqrt{y}}\right)\frac{1}{\sqrt{x-y}}\ dy,$$

which is equivalent with

$$\widetilde{g}(x) = \frac{1}{\pi}\int_0^x \frac{\widetilde{f}(y)}{\sqrt{x-y}}\ dy.$$

This is again an equation of Abel type.

12. Abel's equation also occurs in *Wicksell's unfolding problem in stereology*, where the frequency distribution of the actual unobservable diameter of spherical particles is to be recovered from a sample of planar cuts. The equation was originally derived by Abel in order to solve the *tautochrone* problem. *Geological prospecting*, the problem of determining the location, shape, etc. of geological anomalies in the earth's interior from measurements on its surface, leads to another type of integral equation involving a bounded kernel on a compact interval.

13. Error-in-variables. We are interested in the density f of a random variable U, but we can only observe this random variable with independent additive random error ϵ. This error is unobservable but has known density ψ. Our data are n independent copies $X_1, \ldots, X_n$ of $X = U + \epsilon$. Note that

$$\begin{aligned}
p(x) &= \frac{d}{dx}\mathbf{P}\{X \leq x\} \\
&= \frac{d}{dx}\mathbf{P}\{U + \epsilon \leq x\} \\
&= \frac{d}{dx}\int_{-\infty}^{\infty} \mathbf{P}\{U + \epsilon \leq x | U = u\} f(u)\ du \\
&= \frac{d}{dx}\int_{-\infty}^{\infty} \Psi(x-u)f(u)\ du \\
&= \int_{-\infty}^{\infty} \psi(x-u)f(u)\ du \\
&= (\psi * f)(x).
\end{aligned}$$

This is an integral equation of *convolution* type. The problem is to recover f from the data.

14. Convolutions can be formulated and dealt with for functions on abstract locally compact Abelian groups. In *picture restoration* one has to deal with convolutions for functions on $\mathbb{Z} \times \mathbb{Z}$. *Inverse heat conduction* may be described by means of a convolution for functions on $\mathbb{R}^3$.

15. Summarizing, in all the examples the output, i.e. the l.h.s. of the equation, can be estimated from the data. To recover the input one is inclined to apply the inverse of the operator involved to the estimated output. We have seen that even in the finite dimensional situation (§1.3 - §1.5) it is better to regularize this inverse, although it is still continuous. In infinite dimensional cases the inverse is typically no longer continuous and regularization of some sort becomes pertinent. Hence the input will be recovered by applying a regularized inverse to the estimated output. Like in the case of ridge regression the MISE will display a trade-off between variance and bias.

2 The General Indirect Regression Model

1. Our first assumption is that input and output can be represented as elements of possibly different real separable Hilbert spaces $\mathbb{H} = L^2(\mathbb{T}, \mathcal{T}, \tau) =$

$L^2(\tau)$ and $\mathbb{L} = L^2(\mathbb{X}, \mathfrak{X}, \mu) = L^2(\mu)$. Suppose that $K : L^2(\tau) \to L^2(\mu)$ is a *bounded, injective, linear* operator, and that $\mu(X) = 1$. In the *general indirect regression model* the data consists of n independent copies $(X_1, Y_1), \cdots,$ (X_n, Y_n) of (X, Y), where all random elements are defined on an underlying probability space $(\Omega, \mathcal{W}, \mathbf{P})$. The random variable X will represent the *random design* variable and we assume for convenience that

$$X =_d \text{ Uniform } (\mathbb{X}).$$

The real valued random variable Y is the *response* variable and related to X according to

$$Y = (Kf)(X) + \epsilon, \quad f \in L^2(\tau),$$

where ϵ is a real valued continuous type random error variable with density ψ, finite variance, mean $\mathbf{E}\epsilon = 0$, and stochastically independent of X ($X \perp \epsilon$). The problem is to recover f from the data.

2. Because $\mathbf{E}(Kf)^2(X) = \int_{\mathbb{X}} g^2(x)\, d\mu(x) = \|g\|^2 < \infty$, the random variable Y has finite variance. It is easy to see that

$$\mathbf{E}(Y|X) = (Kf)(X).$$

For an arbitrary Borel set B in $\mathbb{R}$ and $A \in \mathfrak{X}$ we have

$$\begin{aligned}
\mathbf{P}\{X \in A, Y \in B\} &= \mathbf{EP}\{X \in A, Y \in B|X\} \\
&= \int_A \mathbf{P}\{(Kf)(x) + \epsilon \in B\}\, d\mu(x) \\
&= \int_A \mathbf{P}\{\epsilon \in B - (Kf)(x)\}\, d\mu(x) \\
&= \int_A \int_B \psi(y - (Kf)(x))\, dy\, d\mu(x).
\end{aligned}$$

This entails that

$$\boxed{p(x, y) = p_{f,\psi}(x, y) = \psi(y - (Kf)(x)),\ (x, y) \in \mathbb{X} \times \mathbb{R}}$$

is the density of (X, Y) with respect to $\mu \times \lambda$ (λ Lebesgue measure on $\mathbb{R}$). Throughout the nuisance parameter ψ will be kept fixed.

3. The case where $L^2(\tau) = L^2(\mu)$ and $K = I$, the identity operator, is formally included as a special case. The structure of the data now simplifies to

$$Y = f(X) + \epsilon, \;\; f \in L^2(\mu),$$

which corresponds to the ordinary *direct regression* model. Although we will not focus on this special case we don't want to exclude it either. Therefore we will not require K to be an integral operator, since I is not an integral operator. It should also be observed that we may write

$$Y = g(X) + \epsilon, \;\; g = Kf, \;\; f \in L^2(\mu),$$

so that we have a direct regression model in terms of g. This means that in general an unbiased, $\sqrt{n}$ - consistent estimator of g will not exist. Moreover, the operator K is still quite arbitrary. In the following we will replace the original operator equation with an integral equation based on an operator in a restricted class and with a l.h.s. that can be better estimated.

4. Let $Q : L^2(\mu) \to L^2(\tau)$ be a *bounded, injective, linear, integral* operator. Its real, measurable kernel will be denoted $Q(t,x), \; (t,x) \in \mathbb{T} \times \mathbb{X}$, so that

$$(Qg)(t) = \int_{\mathbb{X}} Q(t,x)\, g(x) \; d\mu(x), \;\; t \in \mathbb{T}, \;\; g \in L^2(\mu).$$

We are free to choose this operator, the only restriction being that its composition with K,

$$R = QK : L^2(\tau) \to L^2(\tau), \text{is } \textit{strictly positive Hermitian.}$$

Because all operators are injective we see that

$$g = Kf \text{ is equivalent with } q = Qg = QKf = Rf.$$

If K is itself an integral operator as is the case in all the examples we have considered the choice

$$Q = K^* \text{ yields } R = K^*K,$$

and satisfies all the requirements. This process of replacing the original equation with an equivalent one that might be easier to deal with is referred to as *preconditioning*. The *standard* choice that usually gives good results is $Q = K^*$. In some cases, however, the operator K^*K is still not simple enough to deal with and a nonstandard choice has to be made.

5. Recall that $R : \mathbb{H} \to \mathbb{H}$ is *Hermitian* if it is bounded, linear and satisfies

$$\langle Rf, g\rangle = \langle f, Rg\rangle, \quad \text{for all } f, g \in \mathbb{H}.$$

It is *strictly positive* if

$$\langle Rf, f\rangle > 0, \quad \text{for all } f \in \mathbb{H} \text{ with } f \neq 0.$$

For a Hermitian R to be injective it is necessary and sufficient that it is strictly positive or that its range is dense in $\mathbb{H}$. It has a bounded inverse R^{-1} if and only if the range of R equals $\mathbb{H}$. If $K : \mathbb{H} \to \mathbb{L}$ is bounded and linear, the adjoint operator $K^* : \mathbb{L} \to \mathbb{H}$ is uniquely determined by the requirement

$$\langle Kf, g\rangle = \langle f, K^*g\rangle, \quad \text{for all } f \in \mathbb{H},\ g \in \mathbb{L}.$$

Under the present conditions the adjoint is bounded and linear, and K^*K is Hermitian. If, in addition K is injective the operator K^*K will also be injective.

6. We will see now that it is possible to construct an unbiased and $\sqrt{n}$ - consistent estimator of q. This estimator is

$$\boxed{\hat{q}(t) = \frac{1}{n}\sum_{k=1}^{n} Y_k Q(t, X_k),\ t \in \mathbb{T}}$$

This simple form is partly due to the assumption that X is Uniform $(\mathbb{X})$. It is immediate that

$$\begin{aligned}
\mathbf{E}\,\hat{q}(t) &= \mathbf{E}\,YQ(t, X) \\
&= \mathbf{E}\,\{(Kf)(X) + \epsilon\}Q(t, X) \\
&= \mathbf{E}\,Q(t, X)\,g(X) \\
&= \int_{\mathbb{X}} Q(t, x)\,g(x)\ d\mu(x) \\
&= q(t),
\end{aligned}$$

so that $\hat{q}$ is *unbiased* indeed.

7. Let us now assume that $\mathbf{E}\,\|YQ(\cdot,X)\|^2<\infty$. Then we have

$$\begin{aligned}
\mathbf{P}\left\{\|\hat{q}-q\|\geq\epsilon\right\} &\leq \frac{1}{\epsilon^2}\mathbf{E}\,\|\hat{q}-q\|^2\\
&= \mathbf{E}\int_{\mathbb{T}}\{\hat{q}(t)-q(t)\}^2\;d\tau(t)\\
&= \int_{\mathbb{T}}\;\mathbf{Var}\;\;\hat{q}(t)\;d\tau(t)\\
&= \frac{1}{n}\int_{\mathbb{T}}\;\mathbf{Var}\;\;YQ(t,X)\;d\tau(t)\\
&\leq \int_{\mathbb{T}}\;\mathbf{E}\,Y^2Q^2(t,X)\;d\tau(t)\\
&= \frac{1}{n}\mathbf{E}\,\|YQ(\cdot,X)\|^2\\
&= O(\frac{1}{n}),\;\text{ as } n\to\infty.
\end{aligned}$$

This entails that the estimators are $\sqrt{n}$ - *consistent.*

8. Example. We will consider an example where the standard pre-conditioning may not be the easiest to work with. Let us consider the Abel equation (cf. §1.8, §1.9, §1.11, §1.12)

$$g(x)=\int_x^1\frac{f(t)}{\sqrt{t-x}}\;dt=(Kf)(x),\;\;0\leq x\leq 1.$$

It is known that $K:L^2(0,1)\to L^2(0,1)$ is compact and injective, and that its action is taking " half " an antiderivative. Indeed we see that K^2 has kernel

$$\begin{aligned}
K^2(x,t) &= \int_{s=0}^1 1_{[x,1]}(s)\frac{1}{\sqrt{s-x}}1_{[s,1]}(t)\frac{1}{\sqrt{t-s}}\;ds\\
&= \int_x^t\frac{1}{\sqrt{(s-x)(t-s)}}\;ds\\
&= \int_0^1\frac{1}{\sqrt{u(1-u)}}\;du\\
&= \pi,\quad 0<x\leq t<1,
\end{aligned}$$

and 0 otherwise. In other words, we have

$$(K^2)(x,t) = \pi 1_{[0,t]}(x),\ \ (x,t) \in [0,1] \times [0,1],$$

which essentially boils down to taking a " full " antiderivative.

9. This suggests that a suitable preconditioning operator might be $Q = (K^2)^* K$. This leads to the compact, strictly positive Hermitian $R = (K^2)^* K^2$. It is clear that

$$((K^2)^*)(t,x) = \pi 1_{[0,t]}(x),\ \ (x,t) \in [0,1] \times [0,1],$$

so that Q has kernel

$$\begin{aligned}
Q(s,t) &= \pi \int_0^1 1_{[0,s]}(x) 1_{[x,1]}(t) \frac{1}{\sqrt{t-x}}\ dx \\
&= \pi \int_0^{s\wedge t} \frac{1}{\sqrt{t-x}}\ dx \\
&= \pi \int_{t-s\wedge t}^{t} \frac{1}{\sqrt{u}}\ du \\
&= 2\pi(\sqrt{t} - \sqrt{t - s\wedge t}),\ \ (s,t) \in [0,1] \times [0,1].
\end{aligned}$$

Finally, R has the kernel

$$\begin{aligned}
R(s,t) &= \int_0^1 ((K^2)^*)(s,t)(K^2)(x,t)\ dx \\
&= \pi^2 \int_0^1 1_{[0,s]}(x) 1_{[0,t]}(x)\ dx \\
&= \pi^2 (s\wedge t),\ \ (s,t) \in [0,1] \times [0,1].
\end{aligned}$$

This kernel is well-known from Brownian motion. We'll see below that all the relevant properties are known for this kernel.

10. Note that we also have $R(s,t) = \int_0^1 Q(s,u)\, K(u,t)\, du$. This means that we must have

$$\begin{aligned}
\pi^2(s\wedge t) &= 2\pi \int_0^1 (\sqrt{u} - \sqrt{u - s\wedge u})\, 1_{[u,1]}(t) \frac{1}{\sqrt{t-u}}\ du \\
&= 2\pi \int_0^t \frac{\sqrt{u} - \sqrt{u - s\wedge u}}{\sqrt{t-u}}\ du.
\end{aligned}$$

It is, indeed, easier to find the kernel of R by the method of §2.9.

11. The estimator of $q = Rf$ is now given by

$$\hat{q}(s) = \frac{2\pi}{n} \sum_{k=1}^{n} Y_k(\sqrt{X_k} - \sqrt{X_k - s \wedge X_k}), s \in [0,1].$$

By following the general pattern of §2.6 it is immediate that $\hat{q}$ is an unbiased estimator of q. For the $\sqrt{n}$ - consistency we need to show that $\mathbf{E}\|YQ(\cdot, X)\|^2 < \infty$. This follows from

$$\begin{aligned}
\mathbf{E}\,\|YQ(\cdot, X)\|^2 &= \mathbf{E}\,Y^2 \int_0^1 Q^2(s, X)\; ds \\
&= \mathbf{E}\,Y^2 \int_0^1 (X + X - s \wedge X - 2\sqrt{X}\sqrt{X - s \wedge X})\; ds \\
&\leq 5\,\mathbf{E}\{(Kf)(X) + \epsilon\}^2 \\
&= 5\,(\|Kf\|^2 + \mathbf{E}\epsilon^2) < \infty.
\end{aligned}$$

Finally, we have

$$p(x, y) = \psi\left(y - \int_x^1 \frac{f(t)}{\sqrt{t - x}}\; dt\right),\ (x, y) \in [0,1] \times \mathbb{R},$$

for the density of (X, Y).

3 Recovering the Input

1. Recovering the input will require inversion of the operator R in §2.4. In most situations R^{-1} will be unbounded and therefore, although $R^{-1}q = f$, the expression " $R^{-1}\hat{q}$ " may not be defined. If it is, it might not be a good estimator of f, even though $\hat{q}$ is a good estimator of q. A kind of regularization of the inverse will be needed. Here we will focus on *spectral methods*. It is possible to formulate a more general inversion pattern, of which the spectral method is a special case and that is in particular useful if the inverse of the operator R or K is known in the time domain, but we will not pursue this here. For a few remarks, however, see Chapter 8, §6 -§10.

2. Let $\mathbb{H}$ be a separable Hilbert space, and let $R : \mathbb{H} \to \mathbb{H}$ be strictly positive Hermitian. Then R is unitarily equivalent to a multiplication operator: there exists a σ-finite measure space $(\mathbb{S}, \mathfrak{S}, \Sigma)$, a real valued $\rho \in L^\infty(\mathbb{S}, \mathfrak{S}, \Sigma)$ satisfying $\rho > 0$, and unitary $U : \mathbb{H} \to L^2(\Sigma) = L^2(\mathbb{S}, \mathfrak{S}, \Sigma)$, such that

$$R = U^{-1} M_\rho U.$$

The following diagram summarizes the actions of these operators.

$$\begin{array}{ccc} \mathbb{H} & \xrightarrow{R} & \mathbb{H} \\ {\scriptstyle U}\Big\downarrow & & \Big\uparrow{\scriptstyle U^{-1}} \\ L^2(\Sigma) & \xrightarrow{M_\rho} & L^2(\Sigma) \end{array}$$

Recall that U *unitary* means that U preserves inner products

$$\langle f, g \rangle = \langle Uf, Ug \rangle, \text{ for all } f, g \in \mathbb{H},$$

and hence preserves norms. The action of the multiplication operator is defined as

$$M_\rho \varphi = \rho \cdot \varphi, \;\; \varphi \in L^2(\Sigma).$$

We have $M_\rho : L^2(\Sigma) \to L^2(\Sigma)$ for bounded ρ. The theorem in this § is a version of the *spectral theorem* due to Halmos. It generalizes the well-known theorem on the diagonalizatoin of symmetric matrices of finite dimension. Below we will formulate two important special cases.

3. Let $\Psi : (0, \infty) \to \mathbb{R}$ be measurable. Then we define

$$\Psi(R) = U^{-1} M_{\Psi(\rho)} U.$$

This operator may be unbounded and its domain of definition is determined by the domain of the multiplication operator, i.e. by $\{\varphi \in L^2(\Sigma) : \Psi(\rho).\, \varphi \in L^2(\Sigma)\}$. In particular we have

$$\Psi(t) = t^2,\, t > 0 : \Psi(R) = R^2 = U^{-1} M_{\rho^2} U,$$

$$\Psi(t) = \sqrt{t},\, t > 0 : \Psi(R) = \sqrt{R} = U^{-1} M_{\sqrt{\rho}} U,$$

and the one of particular importance here,

$$\Psi(t) = 1/t,\, t > 0 : \Psi(R) = R^{-1} = U^{-1} M_{1/\rho} U.$$

These operators satisfy the usual properties which justifies the shorter expressions on the left. The first two are bounded, but the last one will be in general unbounded because $1/\rho$ will usually be an unbounded function. We will now discuss a regularization method for this operator.

4. We will consider a family of functions $\{\Psi_\alpha,\ \alpha > 0\}$ such that $\Psi_\alpha(t) \approx 1/t$ for α small. We will almost exclusively deal here with the family

$$\Psi_\alpha(t) = \frac{1}{t}\,1_{[\alpha,\infty)}(t),\ \ t > 0,\ \alpha > 0.$$

This yields the family of *regularized inverses of spectral-cut-off* type

$$R_\alpha^{-1} = \Psi_\alpha(R) = U^{-1} M_{(1/\rho)1_{\{\rho\le\alpha\}}} U,\ \ \alpha > 0.$$

It should be noted that

$$R_\alpha^{-1} : \mathbb{H} \to \mathbb{H} \text{ is bounded for each } \alpha > 0,$$

and that, writing $R_\alpha^{-1} R = I_\alpha$

$$\|R_\alpha^{-1} Rf - f\| = \|I_\alpha f - f\| \to 0,\ \text{ as } \alpha \downarrow 0,\ \text{ for each } f \in \mathbb{H},$$

by the dominated convergence theorem. A second choice is based on

$$\Psi_\alpha(t) = \frac{1}{\alpha + t},\ \ t > 0,\ \alpha > 0,$$

and leads to the *Moore - Penrose* or *penalized least squares* type of regularized inverses

$$R_\alpha^{-1} = U^{-1} M_{\rho/(\alpha+\rho)} U = (\alpha I + R)^{-1},\ \ \alpha > 0.$$

The relation with penalized least squares will be discussed in §13.

5. Special case : compact operators. Let us now assume that $\mathbb{H} = L^2(\tau)$ and that R is compact in addition to being strictly positive Hermitian, meaning that the image of the closed unit ball $\{Rf : f \in \mathbb{H},\ \|f\| \le 1\}$, is contained in a compact subset of $\mathbb{H}$. It can be shown that in this case R has a pure point spectrum with all eigenvalues strictly positive and that the eigenvalues can be arranged in a sequence decreasing to 0 :

$$\rho_1 \ge \rho_2 \ge \cdots \downarrow 0.$$

All multiplicities are finite and the corresponding sequence of normalized eigenfunctions

$$e_1, e_2, \cdots,$$

forms an orthonormal basis of $\mathbb{H}$. Here we may choose $\mathbb{S} = \mathbb{N}$, $\mathfrak{S} =$ the family of all subsets of $\mathbb{N}$, and $\Sigma =$ counting measure. This means that $L^2(\Sigma) = l^2$. The unitary operator U is the operator that assigns to each $f \in \mathbb{H}$ the sequence of its Fourier coefficients in the basis of eigenfunctions. The multiplication operator is coordinatewise multiplication in l^2 with the bounded vector of eigenvalues. We see that we have

$$R = U^{-1} M_\rho U = \sum_{m \geq 1} \rho_m \, e_m \otimes e_m,$$

by combining the above facts. Clearly

$$R_\alpha^{-1} = \sum_{m:\, \rho_m \geq \alpha} \frac{1}{\rho_m} e_m \otimes e_m, \ \alpha > 0,$$

is the spectral-cut-off family of inverses.

6. Special case: convolution operators. Let $r \in L^1(\mathbb{R})$ be symmetric about 0 with

$$\begin{aligned} \widetilde{r}(t) &= \int_{-\infty}^{\infty} e^{itx} r(x) \ dx \\ &= \int_{-\infty}^{\infty} r(x) \cos tx \ dx \ > 0, \ \text{for all } t \in \mathbb{R}. \end{aligned}$$

Under these conditions

$$(Kf)(x) = \int_{\infty}^{\infty} r(x-t) f(t) \ dt = (r * f)(x), \ x \in \mathbb{R},$$

defines a strictly positive Hermitian operator mapping $L^2(\mathbb{R})$ into itself. Define the Fourier transform as

$$(Ff)(s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{isx} f(x) \ dx, \ s \in \mathbb{R}, \ f \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}),$$

and by continuous extension to all of $L^2(\mathbb{R})$. Then F is a unitary operator mapping $L^2(\mathbb{R})$ onto $L^2(\mathbb{R})$. It is not hard to show that

$$F(r * f) = \widetilde{r} \cdot Ff, \ \text{in other words } K = F^{-1} M_{\tilde{r}} F.$$

Hence here we have another instance of the spectral theorem in §3.2.

7. Now let $K : \mathbb{H} \to \mathbb{L}$ be bounded and injective and hence $R = K^*K$ strictly positive Hermitian (§2.5). We define

$$|K| = (K^*K)^{1/2},\ \mathfrak{R}_K = \text{ range of } K,\ \mathfrak{R}_{|K|} = \text{ range of } |K|,$$

with closures $\overline{\mathfrak{R}_K}$ and $\overline{\mathfrak{R}_{|K|}}$ respectively. A bounded linear operator $V : \mathbb{H} \to \mathbb{L}$ is called a *partial isometry* from $\overline{\mathfrak{R}_{|K|}} \subset \mathbb{H}$ to $\overline{\mathfrak{R}_K} \subset \mathbb{L}$ if

$$\begin{aligned}
&\{Vf : f \in \overline{\mathfrak{R}_{|K|}}\} = \overline{\mathfrak{R}_K}\\
&\langle Vf, Vg\rangle = \langle f, g\rangle \text{ for all } f, g \in \overline{\mathfrak{R}_{|K|}}\\
&Vf = 0 \text{ if } f \perp \overline{\mathfrak{R}_{|K|}}.
\end{aligned}$$

Then $V^* : \mathbb{L} \to \mathbb{H}$ is a partial isometry from $\overline{\mathfrak{R}_K}$ to $\overline{\mathfrak{R}_{|K|}}$. According to the *polar decomposition* there exists a partial isometry such that

$$K = V|K|.$$

It can be shown that V^*V is the orthogornal projection onto $\overline{\mathfrak{R}_{|K|}}$ and VV^* the orthogonal projection onto $\overline{\mathfrak{R}_K}$.

8. We have assumed that Q is an integral operator (§2.4) and will assume that U is an integral operator (on a dense subset of $\mathbb{H}$). This implies that $U^{-1} = U^*$ has kernel

$$U^*(t, s) = \overline{U(s, t)},\ t \in \mathbb{T},\ s \in \mathbb{S}.$$

Hence we have

$$\begin{aligned}
U(Q(\cdot, x)) &= \int_{\mathbb{T}} U(s, t)\, Q(t, x)\ d\tau(t)\\
&= (UQ)(s, x),\ \ s \in \mathbb{S}, x \in \mathbb{X}.
\end{aligned}$$

We will assume that

$$\sup_{s\in\mathbb{S}, x\in\mathbb{X}} |\,(UQ)(s, x)\,| < \infty.$$

If K is an integral operator we may use the standard preconditioning $Q = K^*$. In this case we have by §3.7

$$(UK^*)(s, x) = \sqrt{\rho(s)}\,(UV^*)(s, x),\ \ s \in \mathbb{S}, x \in \mathbb{X},$$

where the kernel UV^* must be bounded.

9. In order to recover f we will now use the estimator

$$\boxed{\hat{f}_\alpha = R_\alpha^{-1}\,\hat{q}}$$

for suitable $\alpha > 0$, where $\hat{q}$ is defined in §2.6, and R_α^{-1} is the spectral-cut-off inverse in §3.4. Note that we have

$$\begin{aligned}\hat{f}_\alpha &= \frac{1}{n}\sum_{k=1}^{n} Y_k\, R_\alpha^{-1} Q(\bullet, X_k)\\ &= \frac{1}{n}\sum_{k=1}^{n} Y_k\, U^{-1}\,\frac{1}{\rho}\,1_{\{\rho\geq\ \alpha\}}\,(UQ)(\bullet, X_k).\end{aligned}$$

If we precondition with $Q = K^*$ we have

$$\hat{f}_\alpha = \frac{1}{n}\sum_{k=1}^{n} Y_k\, U^{-1}\,\frac{1}{\sqrt{\rho}}\,1_{\{\rho\geq\alpha\}}\,(UV^*)(\bullet, X_k).$$

If we further specialize to the case where K is itself Hermitian and hence $K = K^*$, we have $V = I$ and hence

$$\hat{f}_\alpha = \frac{1}{n}\sum_{k=1}^{n} Y_k\, U^{-1}\,\frac{1}{\sqrt{\rho}}\,1_{\{\rho\geq\alpha\}}\,U(\bullet, X_k).$$

10. Example. Let us continue the example of §2.7. The operator R in §2.8 is compact and strictly Hermitian with eigenvalues

$$\rho_m = \frac{1}{(m-\frac{1}{2})^2},\ \ m = 1, 2, \cdots,$$

and eigenfunctions

$$e_m(s) = \sqrt{2}\sin(m - \frac{1}{2})\pi s,\ \ 0 \leq s \leq 1, m = 1, 2, \cdots.$$

Hence we have from §3.5

$$\begin{aligned}
\hat{f}_\alpha &= R_\alpha^{-1}\hat{q} \\
&= \frac{1}{n}\sum_{k=1}^{n} Y_k\, R_\alpha^{-1}(2\pi(\sqrt{X_k} - \sqrt{X_k - \bullet \wedge X_k})) \\
&= \frac{2\pi}{n}\sum_{k=1}^{n} Y_k \sum_{m:\rho_m \geq \alpha} \frac{1}{\rho_m}\langle \sqrt{X_k} - \sqrt{X_k - \bullet \wedge X_k}, e_m\rangle\ e_m \\
&= 2\pi \sum_{m:\rho_m \geq \alpha} \frac{1}{\rho_m}\{\frac{1}{n}\sum_{k=1}^{n}\langle \sqrt{X_k} - \sqrt{X_k - \bullet \wedge X_k}, e_m\rangle\}\, e_m
\end{aligned}$$

where ρ_m and e_m are as above.

11. Example. Let us now consider an example of a convolution tailored to the specifics of the current regression model. Let

$$\varphi(t) = (1_{[-\frac{1}{6},\frac{1}{6}]} * 1_{[-\frac{1}{6},\frac{1}{6}]}) \cdot \cos t,\ \ t \in \mathbb{R},$$

and restrict the input f to the class

$$\mathcal{F} = \{\text{ all } f \in L^2(\mathbb{R}) \text{ with support in } [-\frac{1}{6}, \frac{1}{6}\,]\,\}.$$

It is then obvious that

$$Kf = \varphi * f,\ \ f \in \mathcal{F},$$

has support in $[-\frac{1}{2}, \frac{1}{2}]$. Suppose we observe independent copies of

$$Y = (\varphi * f)(X) + \epsilon,$$

where X has the Uniform$(-\frac{1}{2}, \frac{1}{2})$ distribution. Note that $\varphi \in L^1(\mathbb{R})$ and that, with $\Delta = 1_{[-\frac{1}{6},\frac{1}{6}]} * 1_{[-\frac{1}{6},\frac{1}{6}]}$,

$$\begin{aligned}
\widetilde{\varphi}(s) &= \int_{-\infty}^{\infty} e^{ist}\,\Delta(t)\ \cos t\ dt \\
&= \frac{1}{2}\int_{-\infty}^{\infty} e^{ist}\,\Delta(t)\,(e^{-it} + e^{it})\ dt \\
&= \frac{1}{2}\{\int_{-\infty}^{\infty} e^{i(s-1)t}\,\Delta(t)\ dt + \int_{-\infty}^{\infty} e^{i(s+1)t}\ dt\} \\
&= \frac{1}{2}\{\widetilde{\Delta}(s-1) + \widetilde{\Delta}(s+1)\} \\
&= \frac{1}{18}\{\text{sinc}^2\frac{1}{6}(s-1) + \text{sinc}^2\frac{1}{6}(s+1)\} > 0,\ \ s \in \mathbb{R},
\end{aligned}$$

where $\operatorname{sinc} x = (\sin x)/x$. Hence the operator K is strictly positive Hermitian. We'll precondition with K itself so that

$$\hat{q}(t) = \frac{1}{n}\sum_{k=1}^{n} Y_k\, \varphi(t - X_k),\ \ t \in \mathbb{R}.$$

Note that the unbiasedness is due to the fact that the support of $\varphi * f$ is contained in $[-\frac{1}{2}, \frac{1}{2}]$:

$$\begin{aligned}
\mathbf{E}\hat{q}(t) &= \mathbf{E}(q * f)(X)\, \varphi(t - X) \\
&= \int_{-1/2}^{1/2} \varphi(t-x)(\varphi * f)(x)\ dx \\
&= \int_{-\infty}^{\infty} \varphi(t-x)(\varphi * f)(x)\ dx \\
&= ((\varphi * \varphi) * f)(t) \\
&= q(t),\ \ t \in \mathbb{R}.
\end{aligned}$$

12. In order to calculate the input estimator let us write the empirical characteristic function of the X_k as

$$\hat{\chi}(s) = \frac{1}{n}\sum_{k=1}^{n} e^{isX_k},\ \ s \in \mathbb{R}.$$

Next note that

$$\begin{aligned}
(F\varphi(\bullet - X))(s) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{ist}\, \varphi(t - X)\, dt \\
&= \frac{1}{\sqrt{2\pi}}\, e^{isX}\, \widetilde{\varphi}(s),\ \ s \in \mathbb{R}.
\end{aligned}$$

Now we obtain (§3.6)

$$\begin{aligned}
\hat{f}_\alpha &= \frac{1}{n}\sum_{k=1}^{n} Y_k\, R_\alpha^{-1}\, Q(\bullet, X_k) \\
&= \frac{1}{n}\sum_{k=1}^{n} Y_k\, F^{-1}\, \frac{1}{\widetilde{\varphi}^2}\, 1_{\{\widetilde{\varphi}^2 \geq \alpha\}} F\varphi(\bullet - X_k) \\
&= \frac{1}{\sqrt{2\pi}}\sum_{k=1}^{n} Y_k\, F^{-1}\, \frac{1}{\widetilde{\varphi}}\, 1_{\{\widetilde{\varphi}^2 \geq \alpha\}}\, \hat{\chi}.
\end{aligned}$$

There is another way to compute $\hat{f}_\alpha$ by realizing that

$$F^{-1}(g \cdot h) = \frac{1}{\sqrt{2\pi}}\,(F^{-1}g) * (F^{-1}h),$$

for $g, h \in L^2(\mathbb{R})$. This yields

$$\hat{f}_\alpha = \frac{1}{\sqrt{2\pi}}\,(F^{-1}\,\frac{1}{\widetilde{\varphi}^2}\,1_{\{\widetilde{\varphi}^2 \geq \alpha\}}) * \hat{q}.$$

Because $\widetilde{\varphi}$ is not monotonically decreasing the spectral-cut-off inverse may be hard to deal with and even fail to give optimal results.

13. Least squares penalization. Although we will for the most part employ spectral-cut-off regularization, it might be interesting to see the relation between Moore-Penrose regularization and least squares penalization hinted at in §4. In the case of standard preconditioning we have $R = K^*K$, and

$$\boxed{(\alpha I + K^*K)^{-1}K^*g = \mathrm{argmin}_{f \in \mathbb{H}}\{\|Kf - g\|_{\mathbb{L}}^2 + \alpha\|f\|_{\mathbb{H}}^2\},\ g \in \mathfrak{R}_K}$$

for all $\alpha > 0$. We'll sketch a proof. Applying the polar decomposition and writing $\kappa = \sqrt{\rho}$ we have $K = VU^{-1}M_\kappa U$, where $V^*V = I$ on $\mathfrak{R}_{|K|}$ and $\|V^*h\| = \|h\|$ for $h \in \mathfrak{R}_K$. Since $Kf - g \in \mathfrak{R}_K$ this entails

$$\|Kf - g\|_{\mathbb{L}}^2 + \alpha\|f\|_{\mathbb{H}}^2 = \|\kappa Uf - UV^*g\|_{\Sigma}^2 + \alpha\|Uf\|_{\Sigma}^2\,.$$

For brevity let us write $Uf = \tilde{f},\ UV^* = \tilde{g}$, and take an arbitrary $\varphi \in L^2(\Sigma)$. For convenience let us assume that all functions are real. If the minimum exists and were attained at g we must have

$$\begin{aligned} 0 &= \frac{d}{dt}\left[\int \{\kappa(\tilde{f} + t\varphi) - \tilde{g}\}^2\,d\Sigma + \alpha\int(\tilde{f} + t\varphi)^2\,d\Sigma\right]_{t=0} \\ &= \int 2\kappa^2\tilde{f}\varphi\,d\Sigma - \int 2\kappa\varphi\tilde{g}\,d\Sigma + \alpha\int 2\tilde{f}\varphi\,d\Sigma\,, \end{aligned}$$

and consequently

$$\int \varphi(\kappa^2\tilde{f} - \kappa\tilde{g} + \alpha\tilde{f})\,d\Sigma = 0,\ \text{for all } \varphi \in L^2(\Sigma).$$

But this can only hold true when

$$\kappa^2 \tilde{f} - \kappa \tilde{g} + \alpha \tilde{f} = 0, \ \Sigma \text{ - a.e.},$$

so that

$$\tilde{f} = \frac{\kappa}{\alpha + \kappa^2} \tilde{g} \, .$$

On the other hand it follows that

$$\begin{aligned} U(\alpha I + K^* K)^{-1} K^* g \\ &= U U^{-1} M_{1/(\alpha+\rho)} U U^{-1} M_\kappa U V^* g \\ &= \frac{\kappa}{\alpha + \rho} \tilde{g} \\ &= \frac{\kappa}{\alpha + \kappa^2} \tilde{g} \, , \end{aligned}$$

which establishes the desired connection.

4 Rate-Optimality for MISE

1. Throughout this chapter it will be assumed that the operator K is strictly positive Hermitian, so that in particular $K^* = K$, and that we will apply standard preconditioning, i.e.

$$Q = K^* = K.$$

Recall from §3.9 that in this case

$$\hat{f}_\alpha = \frac{1}{n} \sum_{k=1}^{n} Y_k U^{-1} \frac{1}{\sqrt{\rho}} 1_{\{\rho \geq \alpha\}} U(\bullet, X_k).$$

2. Unless the model is very restricted nonparametirc curve estimators cannot be unbiased. This is true for the present estimator for which

$$\begin{aligned} f_\alpha &= \mathbf{E} \hat{f}_\alpha \\ &= R_\alpha^{-1} q \\ &= U^{-1} \frac{1}{\rho} 1_{\{\rho \geq \alpha\}} U U^{-1} \rho U f \\ &= U^{-1} 1_{\{\rho \geq \alpha\}} U f. \end{aligned}$$

It follows that

$$\begin{aligned}\|f_\alpha - f\|^2 &= \|Uf_\alpha - Uf\|^2 \\ &= \|1_{\{\rho\ge\alpha\}}Uf\|^2 \\ &= \int_{\{\rho<\alpha\}} |Uf|^2 \, d\Sigma,\end{aligned}$$

exploiting that U is unitary. It is clear that (cf.§3.4)

$$\|f_\alpha - f\|^2 \downarrow 0, \text{ as } \alpha \downarrow 0,$$

by the dominated convergence theorem.

3. Let us now compute the mean integrated squared error (MISE). One often considers the worst MISE over a certain submodel of input functions. Let us introduce

$$\mathcal{F} = \{f \in L^2(\tau) : |(Uf)(s)| \le \lambda(s), s \in \mathbb{S}\},\ \lambda \in L^2(\Sigma).$$

Usually such a submodel is a smoothness class. For instance this would be the case if F were the Fourier transform and if $\lambda(s)$ would decay at a certain rate when $|s| \to \infty$. Exploiting the well-known fact that mean squared error = variance + (bias)^2 we arrive at

$$\begin{aligned}\text{MISE} &= \mathbf{E}\,\|\hat{f}_\alpha - f\|^2 \\ &= \mathbf{E}\,\|\hat{f}_\alpha - f_\alpha\|^2 + \|f_\alpha - f\|^2 \\ &= \mathbf{E}\,\|U\hat{f}_\alpha - Uf_\alpha\|^2 + \|Uf_\alpha - Uf\|^2 \\ &= \int_{\mathbb{S}} \mathbf{Var}(U\hat{f}_\alpha)(s)\, d\Sigma(s) + \|Uf_\alpha - Uf\|^2 \\ &= \frac{1}{n}\int_{\mathbb{S}} \mathbf{Var}\left(\frac{1}{\sqrt{\rho(s)}} 1_{\{\rho\ge\alpha\}}(s) Y U(s,X)\right)\, d\Sigma(s) + \|Uf_\alpha - Uf\|^2 \\ &\le \frac{C}{n}\int_{\{\rho\ge\alpha\}} \frac{1}{\rho(s)}\, d\Sigma(s) + \|Uf_\alpha - Uf\|^2,\end{aligned}$$

for some generic constant C, since the kernel U is bounded. This entails

$$\boxed{\sup_{f\in\mathcal{F}} \mathbf{E}\|\hat{f}_\alpha - f\|^2 \le \frac{C}{n}\int_{\{\rho\ge\alpha\}} \frac{1}{\rho}\, d\Sigma + \int_{\{\rho<\alpha\}} \lambda^2 \, d\Sigma}$$

by the definition of the class $\mathcal{F}$ and the results in §4.2.

4. In many situations it can be shown that

$$\sup_{f\in\mathcal{F}}\mathbf{E}\|\hat{f}_\alpha - f\|^2 \leq C\{\frac{1}{n}(\frac{1}{\alpha})^a + \alpha^b\},$$

for some $\alpha > 0$, $b > 0$. Choosing $\alpha = \alpha(n) = n^{-1/(a+b)}$ makes the two terms on the right of the same order and we then have

$$\sup_{f\in\mathcal{F}}\mathbf{E}\|\hat{f}_\alpha - f\|^2 = O(n^{-b/(a+b)}), \text{ as } n \to \infty.$$

5. In order to derive a lower bound to the MISE let $e_1, e_2, \cdots \in L^2(\tau)$ be *any* orthonormal system, *not* necessarily a basis. Let us define a subclass

$$\mathcal{F} = \{f = \sum_m t_m e_m : |t_m| \leq \lambda\}, \ \sum_m \lambda_m^2 < \infty,$$

of input functions. Furthermore we will consider the class

$$\mathcal{T} = \{T : (\mathbb{X}\times\mathbb{R})^n \to L^2(\tau) : \\ \mathbf{E}\,\|T(X_1, Y_1, \cdots, X_n, Y_n)\|^2 < \infty, \text{ for all } f \in \mathcal{F}\},$$

of all input estimators having finite expected squared norm under the submodel. We are interested in a lower bound for the " minimax risk " $\inf_{T\in\mathcal{T}}\sup_{f\in\mathcal{F}}\mathbf{E}\,\|T - f\|^2$. It turns out that a Baysian version of the information inequality is extremly useful in this context.

6. Suppose that p_θ, $\theta \in \Theta \subset \mathbb{R}$ is a family of probability densities on some measurable space with respect to some dominating measure, where Θ is a compact interval. Let $\theta \mapsto p_\theta$ be continuously differentiable, a.e., and let π be a continuously differentiable density on the real line with support in the interior of Θ. This entails that π equals 0 at the endpoints of Θ. Assume finite Fisher informations

$$\mathfrak{I}_p(\theta) = 4\,\|\frac{\partial\sqrt{p_\theta}}{\partial\theta}\|^2 < \infty, \ \mathfrak{I}_\pi = 4\,\|\frac{\partial\sqrt{\pi(\bullet - t)}}{\partial t}\|^2 < \infty,$$

and let $\theta \mapsto \mathfrak{I}_p(\theta)$ be continuous for $\theta \in \Theta$. Observe that this implies that $\int(\partial p_\theta/\partial\theta) = 0$. Finally, let $\tau : \Theta \to \mathbb{R}$ be continuously differentiable. Then for any real estimator T of τ, based on an i.i.d. sample of size n from p_θ, we have

$$\begin{aligned}\sup_{\theta\in\Theta} E_\theta(T-\tau(\theta))^2 &\geq \int_\Theta E_\theta(T-\tau(\theta))^2\pi(\theta)\ d\theta \\ &\geq \frac{\int_\Theta \{\tau'(\theta)\}^2\pi(\theta)\, d\theta}{\mathfrak{I}_\pi + n\int_\Theta \mathfrak{I}_p(\theta)\pi(\theta)\, d\theta}\ .\end{aligned}$$

The second inequality is called the *van Trees inequality.*

7. Let us write $I_m = [-\lambda_m, \lambda_m]$, identify $\mathcal{F}$ with $I = I_1 \times I_2 \times \cdots$, $f \in \mathcal{F} \subset L^2(\tau)$ with $t = (t_1, t_2, \cdots) \in I \subset l^2$, and the underlying density p_f in §2.2 with p_t, $t \in I$. Let π be a prior on $[-1,1]$ as described in §4.6. Then $\pi_m(\bullet) = (1/\lambda_m)\pi(\bullet/\lambda_m)$ is a prior on I_m with Fisher information

$$\mathfrak{I}(\pi_m) = \left(\frac{1}{\lambda_m}\right)^2 \mathfrak{I}(\pi).$$

Let $\{p_{t_{(m)},\theta},\ \theta \in \Theta = I_m\}$ be the one-parameter family of densities with $t_{(m)} = (t_1, \cdots, t_{m-1}, t_{m+1}, \cdots)$ fixed and $t_m = \theta$. Let us write $\overset{\bullet}{p}_{t_{(m)},\theta} = \partial\sqrt{p_{t_{(m)},\theta}}/\partial\theta$, and assume that the Fisher information in the m-th direction is uniformly bounded:

$$4\ \sup_{t_{(m)},\theta} \|\overset{\bullet}{p}_{t_{(m)},\theta}\|^2 \leq \rho_m^* < \infty.$$

Furthermore, let us write $dt = dt_1 dt_2 \cdots$, $dt_{(m)} = dt_1 \cdots dt_{m-1} dt_{m+1} \cdots$, and $I_{(m)} = I_1 \times \cdots \times I_{m-1} \times I_{m+1} \times \cdots$. We will occasionally write E_t, etc. rather than $\mathbf{E}$ to display the dependence on parameters.

8. Note that for $f \in \mathcal{F}$ we have $\langle f, e_m\rangle = t_m$; let us write $T_m = \langle T, e_m\rangle$ for $T \in \mathcal{T}$. By applying respectively the Bessel and the van Trees inequality with $\tau(\theta) = \theta$ we obtain

$$\begin{aligned}&\sup_{f\in\mathcal{F}} \mathbf{E}\|T-f\|^2 \\ &\geq \sup_{t\in I} \sum_m E_t(T_m - t_m)^2 \\ &\geq \sum_m \int_I \left\{ E_t(T_m-t_m)^2 \prod_j \pi_j(t_j) \right\}\ dt\end{aligned}$$

$$
\begin{aligned}
&= \sum_m \int_{I_{(m)}} \left\{ \int_\Theta E_{t_{(m)},\theta} (T_m - \theta)^2 \pi(\theta)\, d\theta \right\} \left\{ \prod_{j \neq m} \pi_j(t_j) \right\} dt_{(m)} \\
&\geq \sum_m \int_{I_{(m)}} \left\{ \frac{\lambda_m^2}{\mathfrak{I}(\pi) + n\lambda_m^2 \rho_m^*} \right\} \left\{ \prod_{j \neq m} \pi_j(t_j) \right\} dt_{(m)} \\
&= \sum_m \frac{\lambda_m^2}{\mathfrak{I}(\pi) + n\lambda_m^2 \rho_m^*} \\
&\geq C \sum_m \frac{\lambda_m^2}{1 + n\lambda_m^2 \rho_m^*}.
\end{aligned}
$$

It is immediate that

$$
\boxed{\inf_{T \in \mathcal{T}} \sup_{f \in \mathcal{F}} \mathbf{E}\, \|T - f\|^2 \geq C \sum_m \frac{\lambda_m^2}{1 + n\lambda_m^2 \rho_m^*}}
$$

because the lower bound derived above does not depend on T.

9. Under additional regularity conditions one might obtain

$$
\inf_{T \in \mathcal{T}} \sup_{f \in \mathcal{F}} \mathbf{E}\, \|T - f\|^2 \geq C \int_{\mathbb{S}} \frac{\lambda^2}{1 + n\lambda^2 \rho}\, d\Sigma,
$$

an integral over the spectral set $\mathbb{S}$, just like the upper bound. It may then also be shown that there exists a sequence $\alpha = \alpha(n) \to 0$, as $n \to \infty$ such that the upper bound in §4.3 is of the same order as the lower bound in §4.9. In other words, under these assumptions *the spectral-cut-off type estimators* $\hat{f}_{\alpha(n)}$ *are rate-optimal for the* MISE. Rather than proving this in general we will consider an example.

10. Example. Can you see the weight of a cable? To answer this question, a paraphrase of the title of Kac's famous paper, let us first observe that the shape of a cable suspended at its endpoints with coordinates (0,0) and (1,0) is given by the differential equation

$$
-\frac{d^2 g(t)}{dt^2} = f(t),\ 0 \leq t \leq 1,\ g(0) = g(1) = 0.
$$

Apart from the sign the source term f represents the load per horizontal

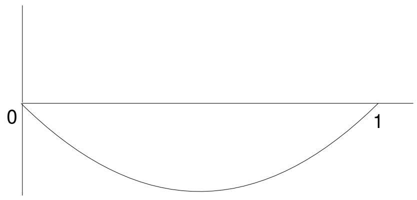

Figure 5:

distance. The problem is to recover the weight distribution or source term from the data $(X_1, Y_1), \cdots, (X_n, Y_n)$ that are independent copies of (X, Y), where

$$Y = g(X) + \epsilon.$$

Let us assume that X has the Uniform $(0, 1)$ distribution.

11. Using the Green's function we can rewrite the differential equation in the form

$$g(x) = \int_0^1 K(x,t)\, f(t)\, dt = (Kf)(t),\ 0 \le x \le 1,$$

where K turns out to be strictly positive Hermitian with kernel

$$K(x,t) = \begin{cases} x(1-t), & 0 \le x \le t \\ t(1-x), & t \le x \le 1 \end{cases}.$$

Preconditioning with $K^* = K$ yields

$$q = K^* g = K^2 f = Rf,\ f \in L^2(0,1).$$

The operator R is compact and strictly positive Hermitian with eigenvalues

$$\rho_m = \left(\frac{1}{\pi m}\right)^4,\ m = 1, 2, \cdots,$$

and corresponding orthonormal basis of eigenfunctions

$$e_m(t) = \sqrt{2} \sin m\pi t,\ 0 \le t \le 1,\ m = 1, 2, \cdots.$$

Following the same pattern as in §3.10 we see that

$$\hat{f}_\alpha(t) = \sum_{m=1}^{M(\alpha)} (\pi m)^2 \left\{ \frac{1}{n} \sum_{k=1}^{n} Y_k \sin m\pi X_k \right\} \sin m\pi t, \ \ 0 \leq t \leq 1,$$

is the spectral-cut-off estimator for suitable $M(\alpha) \in \mathbb{N}$.

12. A natural choice for the submodel seems to be

$$\mathcal{F} = \{ f = \sum_{m=1}^{\infty} t_m e_m, \ |t_m| \leq C m^{-\nu} \}, \ \nu > \frac{1}{2} \ .$$

In the present situation this is a smoothness class. Using $0 < C < \infty$ again as a generic constant we see from §4.3 that

$$\begin{aligned}
\sup\nolimits_{f\in\mathcal{F}} \mathbf{E} \, \|\hat{f}_\alpha - f\|^2 &\leq \frac{C}{n} \sum_{m=1}^{M} m^2 + \sum_{m=M}^{\infty} m^{-2\nu} \\
&\leq \frac{C}{n} \int_0^M x^4 \ dx + \int_M^\infty x^{-2\nu} \ dx \\
&\leq \frac{C}{n} M^5 + M^{1-2\nu} \ .
\end{aligned}$$

Trying $M = n^\delta$ for some $\delta > 0$ we see that the terms are balanced for $\delta = 1/(2\nu + 4)$ so that for this choice

$$\sup\nolimits_{f\in\mathcal{F}} \mathbf{E} \, \|\hat{f}_\alpha - f\,\|^2 \leq C n^{-\frac{2\nu-1}{2\nu+4}}.$$

13. For the bound note that

$$p_t(x,y) = \psi(y - (\sum_{m=1}^{\infty} t_m (K e_m)(x))), \ \ (x,y) \in [0,1] \times \mathbb{R}.$$

Since $K = R^{1/2}$ we have $K e_m = \sqrt{\rho_m} e_m$. If we take $\nu > 1$ the series is even uniformly convergent and

$$\overset{\bullet}{p}_{t_{(m)},\theta}(x,y) = \frac{\psi'(y - (Kf)(x))}{2\sqrt{\psi(y - (Kf)(x))}} \sqrt{\rho_m} e_m(x),$$

so that the Fisher information in the k-th direction equals

$$\begin{aligned}
\|\dot{p}_{t_{(m)},\theta}\| &= \int_0^1 \int_{-\infty}^{\infty} \left\{ \frac{\psi'(y-(Kf)(x))}{2\sqrt{\psi(y-(Kf)(x))}} \right\}^2 \rho_m {e_m}^2(x)\; dy\, dx \\
&= \frac{1}{4} \int_{-\infty}^{\infty} \left\{ \frac{\psi'(y)}{\sqrt{\psi(y)}} \right\}^2 dy \,\cdot \rho_m \int_0^1 {e_m}^2(x)\, dx \\
&= C\,\rho_m\;.
\end{aligned}$$

We see that this number is independent of $t_{(m)}$ and θ so that we may take $\rho_m^* = C\rho_m$ (§4.7). The lower bound in §4.8 reduces in this case to

$$\begin{aligned}
& C \sum_{m=1}^{\infty} m^{-2\nu}/(1+nm^{-2\nu}m^{-4}) \\
& \geq C \int_0^{\infty} \frac{x^{-2\nu}}{1+nx^{-2\nu-4}}\; dx \\
& = \frac{C}{n} \int_0^{\infty} \frac{x^4}{1+(x/n^{1/(2\nu+4)})^2}\; dx \\
& = \frac{C}{n} n^{5/(2\nu+4)} \int_0^{\infty} \frac{y^4}{1+y^{2\nu+4}}\; dy \\
& = C\,n^{-\frac{2\nu-1}{2\nu+4}}.
\end{aligned}$$

Since this lower bound is of the same order as the upper bound in §4.12, the spectral- cut-off estimators are rate-optimal indeed.

5 Estimating Linear Functionals

1. Also throughout this chapter we will assume that K is strictly positive Hermitian and that standard preconditioning with $Q = K^* = K$ is applied. Just like in ordinary, direct curve estimation we cannot expect the input estimator to converge weakly in the Hilbert space $L^2(\tau)$. However certain finite dimensional distributions might converge weakly, and in particular certain linear functionals. Let us introduce the linear submanifold

$$\mathcal{H} = \{h \in L^2(\tau) : \frac{Uh}{\sqrt{\rho}} = \gamma \in L^1(\Sigma) \cap L^2(\Sigma)\}.$$

The condition on the functions in $\mathcal{H}$ will in specific cases turn out to be a smoothness condition. Note that $UQ = \sqrt{\rho}U$ and that this operator (and hence U) has a bounded kernel by assumption(§3.8). Therefore the covariance function

$$\begin{aligned}\Gamma_f(s,t) &= \mathbf{Cov}(YU(s,X), YU(t,X)) \\ &= \mathbf{E}Y^2 U(s,X)\overline{U(t,X)} - (Ug)(s)\overline{(Ug)(t)},\end{aligned}$$

which plays a role below, is then well-defined. We will be interested in using $\langle \hat{f}_\alpha, h\rangle$ as an estimator of $\langle f, h\rangle$ and in particular focus on the question of optimality in the Hájek-LeCam sense.

2. First we will deal with the asymptotic normality and start with centering the random variable at its expectation $\langle f, h\rangle$. Then we have (§4.1,§4.2)

$$\sqrt{n}\,\langle \hat{f}_\alpha - f_\alpha, h\rangle = \frac{1}{\sqrt{n}}\int_{\{\rho\geq\alpha\}} \frac{1}{\sqrt{\rho}}\,\{Y_k U(\bullet, X_k) - Ug\}\,\overline{Uh}\;d\,\Sigma.$$

For each n the terms in the sum are i.i.d. and centered at 0. It follows easily that

$$\begin{aligned}&\mathbf{Var}\ \sqrt{n}\langle \hat{f}_\alpha - f_\alpha, h\rangle \\ &= \mathbf{E}\Big[\int_{\{\rho\geq\alpha\}} \frac{1}{\sqrt{\rho}}\,\{YU(\bullet, X) - Ug\}\,\overline{Uh}\;d\,\Sigma \\ &\quad\times \int_{\{\rho\geq\alpha\}} \frac{1}{\sqrt{\rho}}\,\{Y\overline{U(\bullet, X)} - \overline{Ug}\}\,Uh\;d\,\Sigma\,\Big] \\ &= \int_{\{\rho\geq\alpha\}}\int_{\{\rho\geq\alpha\}} \frac{\overline{(Uh)(s)}}{\sqrt{\rho(s)}}\,\Gamma_f(s,t)\,\frac{(Uh)(t)}{\sqrt{\rho(t)}}\;d\,\Sigma(s)d\,\Sigma(t) \\ &= \int_{\{\rho\geq\alpha\}}\int_{\{\rho\geq\alpha\}} \overline{\gamma(s)}\,\Gamma_f(s,t)\,\gamma(t)\;d\,\Sigma(s)d\,\Sigma(t) \\ &\to \int\int \overline{\gamma(s)}\,\Gamma_f(s,t)\,\gamma(t)\;d\,\Sigma(s)d\,\Sigma(t) = {\sigma_f}^2(h),\ \text{ as } \alpha\downarrow 0.\end{aligned}$$

If we assume that $\mathbf{E}|Y|^{2+\delta} < \infty$, for some $\delta > 0$, due to this convergence the Lyapunov condition for the central limit theorem is fulfilled and it follows that

$$\sqrt{n}\langle \hat{f}_{\alpha(n)} - f_{\alpha(n)}, h\rangle \to_d \ \text{Normal}\ (0, {\sigma_f}^2(h)),\ \text{ as } n\to\infty,$$

for any sequence $\alpha(n) \downarrow 0$, as $n \to \infty$. But then we also have

$$\boxed{\sqrt{n}\langle \hat{f}_{\alpha(n)} - f, h\rangle \to_d \text{ Normal } (0, {\sigma_f}^2(h)), \text{ as } n \to \infty}$$

provided only that $\alpha(n) \downarrow 0$ at such a rate that $\sqrt{n}\,\langle\, f_{\alpha(n)} - f, h\,\rangle \to 0$, as $n \to \infty$.

3. The next question to be answered is whether there exist sequences of estimators of $\langle f, h\rangle$ that might have a limiting distribution which is more concentrated than the normal above in §5.2. If not, the present sequence of estimators might be called asymptotically efficient. We will see that it is not in general. To specify a lower bound to the dispersion let us fix an underlying density by choosing $f_0 \in L^2(\tau)$ and keeping ψ fixed throughout. We then have the density

$$p_0(x, y) = \psi(y - (Kf_0)(x)), \;\; (x, y) \in \mathbb{X} \times \mathbb{R}.$$

Let us write ${\sigma_0}^2(h)$ for the variance of the limiting normal under p_0.

4. The statistical model will be identified with

$$\begin{aligned}\mathcal{P} = \{&\sqrt{p_f(x, y)} = \sqrt{\psi(y - (Kf)(x))}, \, (x, y) \in \mathbb{X} \times \mathbb{R}, \; f \in L^2(\tau)\} \\ &\subset L^2(\mu \times \lambda), \;\; \lambda \text{ Lebesgue measure on } \mathbb{R}.\end{aligned}$$

To find the tangent space to $\mathcal{P}$ at $\sqrt{p_0}$ choose $\dot{f}_0 \in L^2(\tau)$ and consider the curve $t \mapsto \sqrt{p_t(x, y)} = \sqrt{\psi(y - (K(f_0 + t\dot{f}_0))(x)}$ in $L^2(\mu \times \lambda)$. Under reasonable conditions this curve has tangent

$$\boxed{\dot{p}_0(x, y) = \frac{\partial\sqrt{p_t(x, y)}}{\partial t}\Big|_{t=0} = -\frac{\psi'(y - (Kf_0)(x))}{2\sqrt{p_0(x, y)}}(K\dot{f}_0)(x)}$$

Now let

$$\dot{\mathcal{P}}_0 = \text{ closed linear hull of all } \dot{p}_0 : \dot{f}_0 \in L^2(\tau).$$

5. Next we need to find the gradient of the functional $T : \mathcal{P} \to \mathbb{R}$ at $\sqrt{p_0}$. If we set $f_t = f_0 + t\overset{\bullet}{f}_0$ we have

$$\begin{aligned} T(\sqrt{p_t}) &= \langle f_t, h\rangle \\ &= \langle K^{-1}\, K\, f_t, h\rangle \\ &= \langle K\, f_t, (K^{-1})^*\, h\rangle \\ &= \langle K\, f_t, K^{-1}\, h\rangle \\ &= \langle \int_{-\infty}^{\infty} y p_t(\bullet, y)\, dy, U^{-1}\gamma\rangle. \end{aligned}$$

Under reasonable conditions we have on the one hand that

$$\begin{aligned} &\lim_{t\to 0} \frac{1}{t}\, \{T(\sqrt{p_t}) - T(\sqrt{p_0})\} \\ &\quad = \frac{\partial}{\partial t}\, T(\sqrt{p_t})\big|_{t=0} \\ &\quad = \int_{\mathbb{X}} \int_{-\infty}^{\infty} \overset{\bullet}{p}_0(x,y)\, 2y\, \sqrt{p_0(x,y)}\, (U^{-1}\gamma)(x)\ dy\, d\mu(x), \end{aligned}$$

and on the other hand

$$\lim_{t\to 0} \frac{1}{t}\, \{T(\sqrt{p_t}) - T(\sqrt{p_0})\} = \langle \overset{\bullet}{T}_0, \overset{\bullet}{p}_0\rangle.$$

It follows that

$$\boxed{\overset{\bullet}{T}_0(x,y) = 2y\, \sqrt{p_0(x,y)}\, (U^{-1}\gamma)(x)}$$

6. At this point a result due to Hájek and later extended by van der Vaart should be formulated. Tailored to the present situation it says that sufficiently regular sequences $(\hat{T}_n)$ of estimators of T satisfy

$$\sqrt{n}(\hat{T}_n - T) \to_d \ \text{Normal}\ (0, {\Delta_0}^2(h)) * \mathcal{D},\ \text{ as } n \to \infty,$$

under p_0, where

$${\Delta_0}^2(h) = \frac{1}{4}\, \|\ \text{projection of } \overset{\bullet}{T}_0 \text{ onto } \overset{\bullet}{\mathcal{P}}_0\|^2,$$

and where $\mathcal{D}$ is a distribution on the real line. In this context a sequence $(\hat{T}_n^*)$ is called *asymptotically efficient* if for this sequence the distribution $\mathcal{D}$ is degenerate at 0.

7. We have $\sqrt{p_0} \perp \dot{\mathcal{P}}_0$, so that $\dot{\mathcal{P}}_0 \subset [\sqrt{p_0}]^\perp$. Now choose $u \in L^2(\mathbb{R})$, $v \in L^2(\mu)$, such that

$$\langle u, \psi' \rangle = 0.$$

It should be noted that

$$\begin{aligned}
&\int_{\mathbb{X}} \int_{-\infty}^{\infty} \sqrt{p_0(x,y)}\, u(y - (Kf_0)(x))\, v(x)\, \dot{p}_0(x,y)\ dy\, d\mu(x) \\
&= -\frac{1}{2} \int_{\mathbb{X}} \int_{-\infty}^{\infty} \psi'(y - (Kf_0)(x))\, (K\dot{f}_0)(x)\, u(y - (Kf_0)(x))\, v(x)\ dy\, d\mu(x) \\
&= -\frac{1}{2} \langle u, \psi' \rangle \langle K\dot{f}_0, v \rangle\ = 0.
\end{aligned}$$

But this entails that there is a strict inclusion

$$\dot{\mathcal{P}}_0 \subsetneq [\sqrt{p_0}]^\perp.$$

8. Let us now return to ${\sigma_0}^2(h)$ in §5.2. By splitting Γ_0 into its two components we can decompose

$${\sigma_0}^2(h) = {\sigma_{0,1}}^2(h) - {\sigma_{0,2}}^2(h),$$

where

$$\begin{aligned}
{\sigma_{0,1}}^2(h) &= \iint \overline{\gamma(s)}\, \mathbf{E} Y^2 U(s,X)\, \overline{U(t,X)} \gamma(t)\ d\Sigma(s)\, d\Sigma(t) \\
&= \iint \overline{\gamma(s)} \left\{ \iint y^2\, U(s,x) \overline{U(t,x)} p_0(y,x)\ dy\, d\mu(x) \right\} \gamma(t)\ d\Sigma(s)\, d\Sigma(t) \\
&= \iint y^2\, |(U^{-1}\gamma)(x)|^2\, p_0(x,y)\ dy\, d\mu(x), \\
{\sigma_{0,2}}^2(h) &= \iint \overline{\gamma(s)}\, (Ug_0)(s)\, \overline{(Ug_0)(t)} \gamma(t)\ d\Sigma(s)\, d\Sigma(t) \\
&= |\langle \gamma, Ug_0 \rangle|^2 \\
&= |\langle U^{-1}\gamma, g_0 \rangle|^2 \\
&= |\iint (U^{-1}\gamma)(x)\, y\, p_0(x,y)\ dy\, d\mu(x)|^2.
\end{aligned}$$

9. Next we observe that

$$\begin{aligned}
&\frac{1}{4}\|\text{ projection of } \dot{T}_0 \text{ onto } [\sqrt{p_0}\,]^{\perp}\|^2 \\
&= \frac{1}{4}(\|\dot{T}_0\|^2 - |\langle \dot{T}_0, \sqrt{p_0}\,\rangle|^2) \\
&= \frac{1}{4}(\sigma_{0,1}^2(h) - \sigma_{0,2}^2(h)) \\
&= \sigma_0^2(h).
\end{aligned}$$

Because $\dot{\mathcal{P}}_0$ is a strict subset of the orthogonal complement of the line $[\sqrt{p_0}\,]$ this entails

$$\begin{aligned}
{\Delta_0}^2(h) &= \frac{1}{4}\|\text{ projection of } \dot{T}_0 \text{ onto } \dot{\mathcal{P}}_0\|^2 \\
&\leq \frac{1}{4}\|\text{ projection of } \dot{T}_0 \text{ onto } [\sqrt{p_0}\,]^{\perp}\|^2 \\
&= {\sigma_0}^2(h).
\end{aligned}$$

This inequality is strict if we take for ψ the Student(3) distribution, for instance.

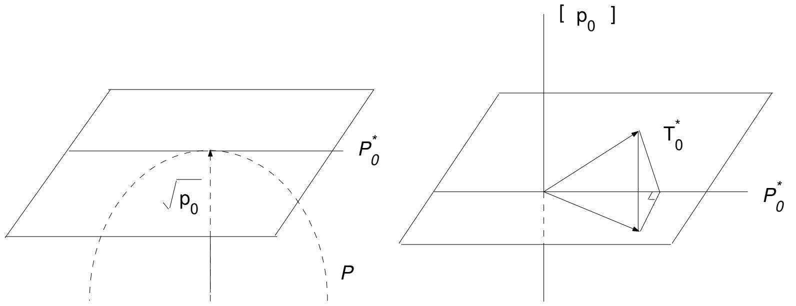

Figure 6:

10. Under an extra technical condition it can be shown that ${\sigma_0}^2(h) = {\Delta_0}^2(h)$ if ψ is *standard normal* . In this case the spectral-cut-off estimator generates an *asymptotically efficient* estimator for the linear functional. There are densities, however, for which they are *not* asymptotically efficient.

Here we have estimated $\langle f, h\rangle$ via an estimator $\hat{f}_\alpha$ of f, because we wanted to investigate how well then $\hat{f}_\alpha$ performs. There is also a simple direct estimator of the linear functional, viz,

$$\langle \hat{\chi}, \gamma\rangle, \text{ where } \hat{\chi}(s) = \frac{1}{n}\sum_{i=1}^{n} Y_i U(s, X_i),\ s \in \mathbb{S}.$$

Note that

$$\begin{aligned}
\mathbf{E}\hat{\chi}(s) &= \mathbf{E}(Kf)(X)U(s, X) \\
&= \int_{\mathbb{X}} U(s,x)(Kf)(x)\ d\mu(x) \\
&= (UKf)(s) = \sqrt{\rho}(Uf)(s),\ s \in \mathbb{S},
\end{aligned}$$

so that

$$\begin{aligned}
\mathbf{E}\langle \hat{\chi}, \gamma\rangle &= \langle \sqrt{\rho}Uf, \frac{1}{\sqrt{\rho}}Uh\rangle \\
&= \langle Uf, Uh\rangle = \langle f, h\rangle.
\end{aligned}$$

Hence this estimator is unbiased. It is easy to see, however, that it has the same limitig normal distribution as $\langle \hat{f}_\alpha, h\rangle$ and consequently is not asymptotically efficient either.

11. On the other hand, it is possible to improve any $\sqrt{n}$ - consistent estimator, like for instance the present estimator $\langle \hat{f}_{\alpha(n)}, h\rangle$, so that it becomes asymptotically efficient. A general version of this procedure can be found in the monograph by Bickel *et al.*; see also Pfanzagl. To formulate this improvement procedure we need to define local quantities at an arbitrary point. Let $e_1, \cdots, e_m \in L^2(\mu)$, and write

$$p_t(x, y) = \psi(y - (Kf_0)(x) - \sum_{i=1}^{m} t_j (Ke_j)(x)),$$

where $t = (t_1, \cdots, t_m) \in \mathbb{R}^m$. Note that

$$\frac{\partial\sqrt{p_t(x,y)}}{\partial t_j} = -\frac{\psi'(y - (Kf_0)(x) - \sum_{j=1}^{m} t_j (Ke_j)(x))(Ke_j)(x)}{2\sqrt{p_t(x,y)}}\ .$$

Let us briefly write $\partial_j\sqrt{p_t}$ for this partial derivative. Denote the tangent space at t to the model by $\dot{\mathcal{P}}_t \subset L^2(\mu\times\lambda)$, and let $\dot{\mathcal{P}}_{t,m}$ be its subspace spanned by $\partial_1\sqrt{p_t},\cdots,\partial_m\sqrt{p_t}$.

12. Suppose we are interested in estimating a sufficiently smooth functional $T(\sqrt{p_t})=\phi(t), t\in\mathbb{R}^m$. Letting $\dot{T}_t$ be the gradient at t we see that

$$\partial_j\phi(t)=\langle\dot{T}_t,\partial_j\sqrt{p_t}\rangle.$$

If we denote by $\Pi_{t,m}$ the orthogonal projection from $L^2(\mu\times\lambda)$ onto $\dot{\mathcal{P}}_{t,m}$, an important quantity is the projected gradient

$$\Pi_{t,m}\dot{T}_t=\widetilde{T}_{t,m}.$$

13. In order to calculate $\widetilde{T}_{t,m}$ we minimize the function $\gamma\mapsto\iint\{\dot{T}_t-\sum_{j=1}^m\gamma_j(\partial_j\sqrt{p_t})\}^2\,d\mu d\lambda$. Exploiting that

$$\langle\partial_j\sqrt{p_t},\partial_k\sqrt{p_t}\rangle=\frac{1}{4}\mathfrak{I}_{m,j,k}(t),$$

where $\mathfrak{I}_m(t)$ is the Fisher information matrix at t, it follows easily that

$$\widetilde{\gamma}=4\,\mathfrak{I}_m^{-1}(t)\bigtriangledown\phi(t),$$

where $\bigtriangledown=(\partial_1,\cdots,\partial_m)^*$ denotes the gradient as a column. By definition the *efficient influence function* (see the monograph by van der Vaart) equals

$$\begin{aligned}\widetilde{T}_{t,m}\cdot\frac{1}{2\sqrt{p_t}}&=\sum_{j=1}^m\{\sum_{k=1}^m\mathfrak{I}^{-1}_{m,j,k}(t)\cdot\partial_k\phi(t)\}\,2\partial_j\sqrt{p_t}/\sqrt{p_t}\\&=\sum_{j=1}^m\{\sum_{k=1}^m\mathfrak{I}^{-1}_{m,j,k}(t)\cdot\partial_k\phi(t)\}\cdot\partial_j\,l_t,\end{aligned}$$

writing, as usual, $\partial_j p_t/p_t=\partial_j l_t$. In matrix notation we now have for the resulting random variable

$$\begin{aligned}\Lambda_{t,m}&=\frac{1}{n}\sum_{i=1}^n\widetilde{T}_{t,m}(X_i,Y_i)/\{2\sqrt{p_t(X_i,Y_i)}\}\\&=\frac{1}{n}\sum_{i=1}^n\{\bigtriangledown\phi(t)\}^*\mathfrak{I}_m^{-1}(t)(\bigtriangledown l_t)(X_i,Y_i).\end{aligned}$$

14. Let us recall that

$$\nabla l_t = \begin{pmatrix} \partial_1 l_t \\ \vdots \\ \partial_m l_t \end{pmatrix}.$$

Also, let us introduce the Hessian matrix

$$Hl_t = \begin{pmatrix} \partial_1\partial_1 l_t & \dots & \partial_1\partial_m l_t \\ \vdots & & \\ \partial_m\partial_1 l_t & \dots & \partial_m\partial_m l_t \end{pmatrix}.$$

It is well-known that if (X, Y) has density p_t we have

$$E_t(\nabla l_t)(X,Y)(\nabla l_t)^*(X,Y) = -E_t(Hl_t)(X,Y) = \mathfrak{I}_m(t).$$

15. Now let $\hat{t}$ be any $\sqrt{n}$ - consistent estimator of t so that $\sqrt{n}\|\hat{t}-t\| = O_p(1)$, as $n \to \infty$. Consequently $\phi(\hat{t})$ will be a $\sqrt{n}$ - consistent estimator of $\phi(t)$, but not in general asymptotically efficient. To improve this estimator, replace it with $\phi(\hat{t}) + \Lambda_{\hat{t},m}$. Indeed, if t is the true parameter we have

$$\sqrt{n}\{\phi(\hat{t}) + \Lambda_{\hat{t},m} - \phi(t)\} \to_d \mathrm{Normal}(0, \Delta^2_{t,m}(\phi)), \text{ as } n \to \infty,$$

where

$$\Delta^2_{t,m}(\phi) = \{\nabla\phi(t)\}^* \mathfrak{I}_m^{-1}(t) \nabla \phi(t) = \frac{1}{4}\|\widetilde{T}_{t,m}\|^2.$$

To sketch a proof, note that we see from §13 and §14 that the standardized estimator on the left equals (J_m is the $m \times m$ identity matrix)

$$\sqrt{n}[\{\nabla\phi(t)\}^*(\hat{t}-t) + \{\nabla\phi(\hat{t})\}^*\mathfrak{I}_m^{-1}(\hat{t})\frac{1}{n}\sum_{i=1}^{n}(\nabla l_t)(X_i,Y_i)$$

$$+ \{\nabla\phi(\hat{t})\}^*\mathfrak{I}_m^{-1}(\hat{t})\frac{1}{n}\sum_{i=1}^{n}(Hl_t)(X_i,Y_i)(\hat{t}-t)] + o_p(1) =$$

$$= \sqrt{n}\{\nabla\phi(t)\}^*\{J_m + \mathfrak{I}_m^{-1}(t)\frac{1}{n}\sum_{i=1}^{n}(Hl_t)(X_i,Y_i)\}(\hat{t}-t)$$

$$+ \frac{1}{\sqrt{n}}\{\nabla\phi(t)\}^*\mathfrak{I}_m^{-1}(t)\sum_{i=1}^{n}(\nabla l_t)(X_i,Y_i) + o_p(1) =$$

$$= \frac{1}{\sqrt{n}}\{\nabla\phi(t)\}^*\mathfrak{I}_m^{-1}(t)\sum_{i=1}^{n}(\nabla l_t)(X_i,Y_i) + o_p(1).$$

Apart from implicit smoothness assumptions in the last transition we use that

$$\frac{1}{n}\sum_{i=1}^{n}(Hl_t)(X_i,Y_i) \to_p E_t(Hl_t)(X,Y) = -\mathfrak{I}_m(t)\ ;$$

see also §14. The asymptotic normality now follows from the central limit theorem. The variance is indeed the one known from the Cramér-Rao lower bound, and its equality to one fourth of the squared length of $\widetilde{T}_{t,m}$ follows from straightforward integration.

16. In order to prepare for improving the estimator $\langle\hat{f}_\alpha, h\rangle$ of $\langle f, h\rangle$ let us consider the tangent space $\dot{\mathcal{P}}_f$ at the parameter $f \in L^2(\mu)$, write p_f for the density (as in §4), $\dot{T}_f$ for the gradient of T at f, and Π_f for the projection of $L^2(\mu \times \lambda)$ onto $\dot{\mathcal{P}}_f$. Let us compute $\widetilde{T}_f = \Pi_f \dot{T}_f$, and first observe that $U^{-1}\gamma = U^{-1}(1/\sqrt{\rho})Uh = K^{-1}h$, which we have also used in §5. We will now have to assume that

$$K^{-1}h \in \mathfrak{R}_K,$$

so that $K^{-2}h = R^{-1}h$ is well-defined. Since $\widetilde{T}_f \in \dot{\mathcal{P}}_f$ there is a function g^* such that (cf §4)

$$\widetilde{T}_f(x,y) = -\frac{\psi'(y-(Kf)(x))}{2\sqrt{p_f(x,y)}}g^*(x),$$

where g^* is a function in the range of K, or a limit of such functions. Since $\overset{\bullet}{T}_f - \widetilde{T}_f \perp \overset{\bullet}{\mathcal{P}}_f$ we must have

$$\iint \{2y\sqrt{p_f(x,y)}(K^{-1}h)(x) + \frac{\psi'(y-(Kf)(x))}{\sqrt{p_f(x,y)}} g^*(x)\} \\ \times \{\frac{\psi'(y-(Kf)(x))}{\sqrt{p_f(x,y)}} g(x)\}\ d\mu(x)\, dy = 0,$$

for all g in the range of K. Since K is Hermitian and injective this range is dense and hence the equality holds for all $g \in L^2(\mu)$. Straightforward calculation now shows that

$$-2\langle K^{-1}h, g\rangle + \mathfrak{I}(\psi)\langle g^*, g\rangle = 0, \text{ for all } g \in L^2(\mu),$$

where $\mathfrak{I}(\psi)$ is the Fisher information of the error density ψ. This entails that

$$g^* = \frac{2}{I(\psi)} K^{-1}h \in \mathfrak{R}_K,$$

and $\widetilde{T}_f$ has been determined.

17. The efficient influence function is now defined as $\widetilde{T}_f/(2\sqrt{p_f})$. From this we may construct the random variable

$$\Lambda_f = -\frac{1}{nI(\psi)} \sum_{i=1}^n \frac{\psi'(Y_i - (Kf)(X_i))}{\sqrt{\psi(Y_i - (Kf)(X_i))}} (K^{-1}h)(X_i).$$

We claim that

$$\boxed{\begin{array}{c} \sqrt{n}(\langle \hat{f}_\alpha, h\rangle + \Lambda_{\hat{f}_\alpha} - \langle f_0, h\rangle) \to_d \text{Normal } (0, \Delta_0^2(h)), \text{ as } n \to \infty \\ \text{when } f_0 \in L^2(\mu) \text{ is the true regression function} \\ \text{and } \Delta_0^2(h) \text{ is the optimal variance in §6} \end{array}}$$

18. As usual the proof will be sketchy and we will forego regularity conditions. Let $e_3, e_4, \cdots$ be an orthonormal basis of $L^2(\mu)$ and let

$$L^2_{0,m} = \text{ linear space spanned by all } f_0 + \sum_{j=1}^m t_j e_j,$$

$t = (t_1, \cdots, t_m) \in \mathbb{R}^m$, where we choose

$$e_1 = f_0,\ e_2 = K^{-1}g^* = \frac{2}{\mathfrak{I}(\psi)} K^{-2}h = \frac{2}{\mathfrak{I}(\psi)} R^{-1}h.$$

Hence $L^2_{0,m} \subset L^2(\mu)$ always contains f_0 and $K^{-1}g^*$. It is easy to see that

$$\frac{\partial}{\partial t_2} \sqrt{p_{f_0 + \sum_{j=1}^m t_j e_j}} \mid_{t=0} = \widetilde{T}_{f_0} = \widetilde{T}_0 \in \dot{\mathcal{P}}_{0,m},$$

see §16 , so that

$$\widetilde{T}_{0,m} = \Pi_{0,m} \dot{T}_0 = \widetilde{T}_0, \text{ for all } m \geq 2.$$

Write

$$\langle\ f_0 + \sum_{j=1}^m t_j e_j, h\ \rangle = \phi(t),$$

and let $\hat{f}_{\alpha,m}$ denote the projection of $\hat{f}_\alpha$ onto $L^2_{0,m}$. There exists a uniquely determined $\hat{t}$ such that

$$\hat{f}_{\alpha,m} = f_0 + \sum_{j=1}^m \hat{t}_j e_j,$$

assuming that $f_0, K^{-1}g^*, e_2, \cdots, e_m$ are linearly independent. Since $\hat{f}_\alpha$ is a $\sqrt{n}$ - consistent estimator of f, $\phi(\hat{t})$ will be a $\sqrt{n}$ - consistent estimator of $\phi(t)$.

19. This estimator can be improved by the method of §15. It should be noted that

$$\Lambda_{\hat{t},m} \text{ as defined in §13 } = \Lambda_{\hat{f}_{\alpha,m}} \text{ as defined in §17 }.$$

If $t = 0$ is the true parameter it follows that

$$\begin{aligned}
&\sqrt{n}\{\phi(\hat{t}) + \Lambda_{\hat{t},m} - \phi(0)\} = \\
&\quad = \sqrt{n}(\langle \hat{f}_{\alpha,m}, h\rangle + \Lambda_{\hat{f}_{\alpha,m}} - \langle f_0, h\rangle) \to_d \\
&\quad \to_d \text{ Normal } (0, \Delta_0^2(h)), \text{ as } n \to \infty, \text{ for each } m \geq 2,
\end{aligned}$$

because $\frac{1}{4}\|\widetilde{T}_{0,m}\|^2 = \frac{1}{4}\|\widetilde{T}_0\|^2 = \Delta_0^2(h)$ by §18 and §6.

20. The proof can be concluded by a continuity argument. Note that

$$\begin{aligned} S_n &= \sqrt{n}(\langle \hat{f}_\alpha, h\rangle + \Lambda_{\hat{f}_\alpha} - \langle f_0, h\rangle) = \\ &= \sqrt{n}(\langle \hat{f}_{\alpha,m}, h\rangle + \Lambda_{\hat{f}_{\alpha,m}} - \langle f_0, h\rangle) \\ &\quad + \sqrt{n}(\langle \hat{f}_\alpha - \hat{f}_{\alpha,m}, h\rangle + \Lambda_{\hat{f}_\alpha} - \Lambda_{\hat{f}_{\alpha,m}}) = \\ &= S_{n,m} + \bar{S}_{n,m}. \end{aligned}$$

For arbitrary ϵ we can find $m(\epsilon)$, $n(\epsilon)$ sufficiently large so as to ensure that $\mathbf{P}\{|\bar{S}_{n,m(\epsilon)}| \le \epsilon\} \ge 1-\epsilon$ for all $n \ge n(\epsilon)$. For this $m(\epsilon)$ we have that $S_{n,m(\epsilon)}$ tends in distribution to the limiting normal law of the claim, as $n \to \infty$, according to the result in §19. Since this distribution doesn't depend on $m(\epsilon)$ and since ϵ is arbitrary, it follows that S_n must have the desired limit law.

6 Testing Hypotheses

1. Let $\mathcal{L} \subset L^2(\tau)$ be a linear subspace of finite dimension m, and suppose we want to test the null hypothesis

$$H_0 : f \in \mathcal{L}.$$

If we precondition with Q this hypothesis is equivalent with $q = Rf \in R\mathcal{L} = \mathcal{N}$, where $R = QK$ and $\mathcal{N}$ is a linear subspace of $L^2(\mu)$ of the same dimension m. Note that $\mathcal{L}$ and $\mathcal{N}$ are closed. Let Π be the orthogonal projection onto $\mathcal{N}$ and $\Pi^\perp$ the orthogonal projection onto $\mathcal{N}^\perp$. We have

$$H_0 \text{ iff } q \in \mathcal{N} \text{ iff } D^2 = \|q - \mathcal{N}\|^2 = \|\Pi^\perp q\|^2 = 0.$$

Recall that

$$\hat{q}(t) = \frac{1}{n}\sum_{k=1}^{n} Y_k Q(t, X_k), \ t \in \mathbb{T}.$$

It seems to make sense to employ the statistic

$$\hat{D}^2 = \|\hat{q} - \mathcal{N}\|^2 = \|\Pi^\perp \hat{q}\|^2,$$

for testing $H_0 : q \in \mathcal{N}$, i.e. $D^2 = 0$, and to reject for large values of $\hat{D}^2$. Although this procedure does not require inversion of the operator, the problem now is the complexity of the limiting distribution.

2. Let $e_1, e_2, \cdots$ be an orthonormal basis of $L^2(\tau)$ such that $e_1, \cdots, e_m$ span $\mathcal{N}$ and consequently $e_{m+1}, e_{m+2}, \cdots$ span $\mathcal{N}^\perp$. Let us write

$$S_n = (S_{n1}, S_{n2}, \cdots), \text{ with } S_{nk} = \sqrt{n}\langle \hat{q} - q,\, e_k \rangle.$$

The central limit theorem entails at once that

$$S_{nk} \to_d \text{ Normal } (0, {\sigma_k}^2), \text{ as } n \to \infty,$$

where the variance is given by

$$\begin{aligned} {\sigma_k}^2 &= \mathbf{Var}\ \langle YQ(\cdot, X), e_k \rangle \\ &= \mathbf{E}\,\{Y \int Q(t,X) e_k(t)\, d\tau(t)\}^2 - \{\mathbf{E}\, Y \int Q(t,X) e_k(t)\, d\tau(t)\}^2 \\ &= \mathbf{E}\, Y^2 (Q^* e_k)^2(X) - \langle q, e_k \rangle^2. \end{aligned}$$

In pretty much the same way it can be shown that the finite dimensional distributions of S_n as a random element in l^2 converge weakly, and since tightness can be also established we have the basic convergence

$$\boxed{S_n \to_d G, \text{ as } n \to \infty, \text{ in } l^2}$$

where G is a 0 mean Gaussian random variable in l^2. This random variable has covariance matrix (σ_{kl}), the variances ${\sigma_k}^2$ of which are explicitly given above. The covariance matrix depends on the unknown parameter f or q.

3. Under H_0 we have $D^2 = ||\Pi^\perp q||^2 = \sum_{k>m} \langle q, e_k \rangle^2 = 0$ and hence $||\Pi^\perp S_n||^2 = \sum_{k>m} \langle \hat{q}, e_k \rangle^2 = n\hat{D}^2$. It follows from the weak convergence in §6.2 that

$$\boxed{n\hat{D}^2 \to_d \sum_{k>d} {G_k}^2, \text{ as } n \to \infty, \text{ for } q \in \mathcal{N}}$$

Unfortunately, the limiting distribution on the right will be hard to deal with and will depend on the unknown parameter f or q, even under H_0. To circumvent the occurence of such a complicated chi-squared type limiting distribution a modification of this testing problem will be treated. This type of modification was first proposed by Dette & Munk.

4. Let $\delta > 0$ be arbitrary but fixed, and let us enlarge the null hypothesis a little bit and consider

$$H_\delta : 0 \leq D^2 = \|q - \mathcal{N}\|^2 \leq \delta^2.$$

Rather than testing that q is in a linear subspace we test that q is in a " linear slice ". We now propose to base the test procedure on a suitably standard

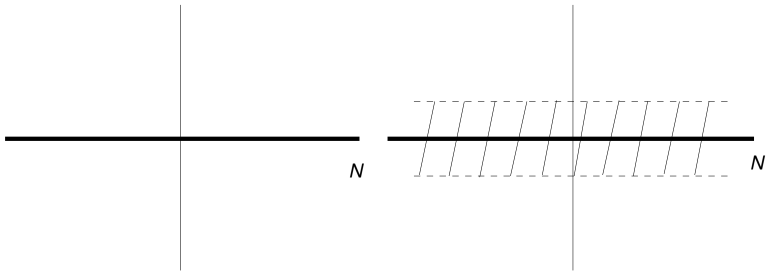

Figure 7:

version of $\hat{D}^2 - \delta^2$. The asymptotics become much simpler now.

5. If $D^2 = 0$ it is obvious from the result in §6.3 that $\sqrt{n}(\hat{D}^2 - D^2) = \sqrt{n}\hat{D}^2$ has a degenerate or Normal(0,0) limiting distribution. For $D^2 > 0$ we have

$$\begin{aligned}
\sqrt{n}(\hat{D}^2 - D^2) &= \sqrt{n} \sum_{k>m} (\langle \hat{q}, e_k \rangle^2 - \langle q, e_k \rangle^2) \\
&= \sum_{k>m} S_{nk}(\langle \hat{q}, e_k \rangle + \langle q, e_k \rangle) \\
&\to_d 2 \sum_{k>m} \langle q, e_k \rangle G_k \\
&=_d \text{Normal}(0, 4\,\mathbf{Var}\,(\sum_{k>m} \langle q, e_k \rangle G_k)), \text{ as } n \to \infty.
\end{aligned}$$

This variance can be further specified :

$$
\begin{aligned}
&\mathbf{Var}\ (\sum_{k>m} \langle q, e_k\rangle G_k) \\
&= \sum_{k>m} \sum_{l>m} \langle q, e_k\rangle \sigma_{kl} \langle q, e_l\rangle \\
&= \sum_{k>m} \sum_{l>m} \{\langle q, e_k\rangle \, \langle q, e_l\rangle \, \mathbf{E}\, Y^2 (Q^* e_k)(X)(Q^* e_l)(X) - \langle q, e_k\rangle^2 \langle q, e_l\rangle^2\} \\
&= \mathbf{E}\, \{\sum_{k>m} \langle q, e_k\rangle Y (Q^* e_k)(X)\}^2 - \{\sum_{k>m} \langle q, e_k\rangle^2\} \\
&= \mathbf{E}\, Y^2 (Q^* \Pi^{\perp} q)^2 (X) - \|\Pi^{\perp} q\|^4.
\end{aligned}
$$

Note that this expression reduces to 0 if $D^2 = 0$. Hence we have shown that

$$
\boxed{\begin{array}{c} \sqrt{n}(\hat{D}^2 - D^2) \to_d \text{Normal}(0, \sigma^2(D)),\ \text{as } n \to \infty \\ \sigma^2(D) = 0 \text{ if } D = 0 \end{array}}
$$

The variance can be consistently estimated by substituting $\hat{q}$ for q above. Denote this estimator by $\hat{\sigma}^2$.

6. suppose that $D^2 = \delta^2$. In this case we see that

$$
\begin{aligned}
\sqrt{n}\, \frac{\hat{D}^2 - \delta^2}{\hat{\sigma}} &= \sqrt{n}\, \frac{\hat{D}^2 - D^2}{\hat{\sigma}} + \sqrt{n}\, \frac{D^2 - \delta^2}{\hat{\sigma}} \\
&\to_d \text{Normal}(0, 1),\ \text{as } n \to \infty.
\end{aligned}
$$

For $0 < D^2 < \delta^2$ it follows similarly that

$$
\sqrt{n}\, \frac{\hat{D}^2 - \delta^2}{\hat{\sigma}} \to_p -\infty,\ \text{as } n \to \infty,
$$

and if $D^2 = 0$ this last claim remains trivially true. Summarizing we have found that (Φ standard normal cdf)

$$
\boxed{\begin{array}{c} \text{the test that rejects } H_\delta : 0 \le D^2 \le \delta^2 \text{ for} \\ \sqrt{n}(\hat{D}^2 - \delta^2)/\hat{\sigma} \ge \Phi^{-1}(1 - \alpha), \\ 0 < \alpha < 1, \text{ has asymptotic level } \alpha \end{array}}
$$

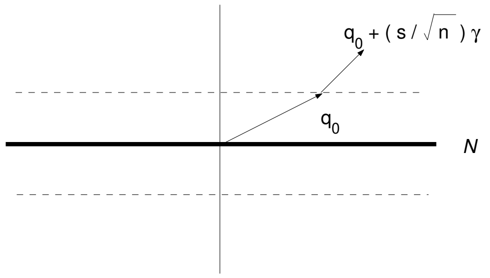

Figure 8:

7. To find the asymptotic power choose f_0, $\varphi \in L^2(\tau)$, such that $\|Rf_0\|^2 = \|q_0\|^2 = \delta^2$ and $\gamma = R\varphi$ has $\langle q_0, \Pi^{\perp}\gamma\rangle > 0$. Let us introduce the local alternatives (including q_0)

$$q_s = q_0 + \frac{s}{\sqrt{n}}\gamma, s \geq 0.$$

It can be shown in a similar manner that

$$\begin{aligned} &\mathbf{P}\,\{\sqrt{n}\,\frac{\hat{D}^2 - \delta^2}{\hat{\sigma}} \geq \Phi^{-1}(1-\alpha)\} \\ &\to 1 - \Phi\left(\Phi^{-1}(1-\alpha) - \frac{2s\langle q_0, \Pi^{\perp}\gamma\rangle}{\sigma}\right), \text{ as } n \to \infty, \end{aligned}$$

which settles the *asymptotic power* of the test.

8. Another advantage of this modification is that it is as easy to test the null hypothesis

$$H'_\delta : D^2 \geq \delta^2.$$

From a practical perspective in many situations this might be the more realistic null hypothesis to be tested. It should be noted, however, that the tests presented here may still have similar drawbacks as the classical goodness of fit tests.

7 Cross-Validation

1. A practical difficulty with the input estimator $\hat{f}_\alpha$ as defined in §3.9 is that it is not clear how one should choose α for a given sample size n.

Therefore in this chapter a data-driven selection method will be presented. Discussion of theoretical properties of the estimator thus obtained is beyond our present scope. In this chapter we return to the *general* preconditioning with an operator Q as in §2.4.

2. For such a geneal Q the MISE equals

$$\begin{aligned}
\mathbf{E}\,\|\hat{f}_\alpha - f\|^2 &= \mathbf{E}\,\|\hat{f}_\alpha - f_\alpha\|^2 + \|f_\alpha - f\|^2 \\
&= \frac{1}{n}\int_{\{\rho\geq\alpha\}} \frac{1}{\rho^2}\ \mathbf{Var}\ Y(UQ)(\bullet, X)\ d\Sigma + \int_{\{\rho<\alpha\}} |Uf|^2\ d\Sigma \\
&= \frac{1}{n}\int_{\{\rho\geq\alpha\}} \frac{1}{\rho^2}\,\{\,\mathbf{E}\,|Y(UQ)(\bullet, X)|^2 - |Uq|^2\}\ d\Sigma + \int_{\{\rho<\alpha\}} |Uf|^2\ d\Sigma \\
&= \frac{1}{n}\int_{\{\rho\geq\alpha\}} \frac{1}{\rho^2}\,\{\,\mathbf{E}\,|Y(UQ)(\bullet, X)|^2 - |Uq|^2\}\ d\Sigma \\
&\qquad\qquad - \int_{\{\rho\geq\alpha\}} |Uf|^2\ d\Sigma + \int_{\mathbb{S}} |Uf|^2\ d\Sigma
\end{aligned}$$

$$= \int_{\{\rho\geq\alpha\}} \frac{1}{\rho^2}\,\{\frac{1}{n}\,\mathbf{E}\,|Y(UQ)(\bullet, X)|^2 - \frac{n+1}{n}\,|Uq|^2\}\ d\Sigma + \int_{\mathbb{S}} |Uf|^2\ d\Sigma.$$

Our aim is to minimize the MISE, for given n, as a function of α. Because in this last expression the term $\int |Uf|^2\, d\Sigma$ does not depend on α, one might as well minimize the function $\alpha \mapsto \mathcal{M}_n(\alpha)$, $\alpha > 0$, where

$$\mathcal{M}_n(\alpha) = \int_{\{\rho\geq\alpha\}} \frac{1}{\rho^2}\,\{\frac{1}{n}\,\mathbf{E}\,|Y(UQ)(\bullet, X)|^2 - \frac{n+1}{n}\,|Uq|^2\}\ d\Sigma,\ \ \alpha > 0.$$

Note, however, that this function still contains the unknown parameter. Therefore it is our purpose to minimize a function $\hat{\mathcal{M}}_n$, say, that only depends on the data and that should be close to $\mathcal{M}_n$.

3. Let us introduce

$$\hat{q}_k = Y_k Q(\bullet, X_k), \quad \hat{q}_{(k)} = \frac{1}{n-1}\sum_{j\neq k} \hat{q}_j\ ,$$

and note that

$$\begin{aligned}\mathbf{E}\, U\hat q_{(k)}\, \overline{U\hat q_k} &= \frac{1}{n-1}\sum_{j\neq k} \mathbf{E}\,(U\hat q_j)\,\overline{(U\hat q_k)}\\ &= \frac{1}{n-1}\sum_{j\neq k} (\mathbf{E}\, U\hat q_j)\,(\mathbf{E}\,\overline{U\hat q_k})\\ &= |Uq|^2,\end{aligned}$$

because the $\hat q_j$ are unbiased and independent. These observations together with the law of large numbers entail that

$$\frac{1}{n}\sum_{k=1}^{n} |U\hat q_k|^2, \quad \frac{1}{n}\sum_{k=1}^{n} (U\hat q_{(k)})(\overline{U\hat q_k}),$$

are unbiased and consistent estimators of

$$\mathbf{E}\,|Y(UQ)(\bullet, X)|^2, \quad |Uq|^2,$$

respectively. This suggests to use the empirical analogue

$$\boxed{\hat{\mathcal{M}}_n(\alpha) = \frac{1}{n}\sum_{k=1}^{n}\int_{\{\rho\geq\alpha\}} \frac{1}{\rho^2}\,\{\frac{1}{n}\,|U\hat q_k|^2 - \frac{n+1}{n}\,(U\hat q_{(k)})(\overline{U\hat q_k})\}\; d\Sigma,\ \alpha>0}$$

This function is an unbiased and consistent estimator of the original function:

$$\mathbf{E}\,\hat{\mathcal{M}}_n(\alpha) = \mathcal{M}_n(\alpha), \quad \hat{\mathcal{M}}_n(\alpha) \to_{\text{a.s.}} \mathcal{M}_n(\alpha),\ \text{as } n\to\infty,\ \text{for each } \alpha>0.$$

4. Minimizing of $\hat{\mathcal{M}}_n$ on $(0,\infty)$ yields an estimator $\hat\alpha = \hat\alpha(n)$ of α and subsequently

$$\boxed{\text{an entirely data-driven input estimator } \hat f_{\hat\alpha}}$$

A simulation study turned out to give quite satisfactory results.

5. In the special case where $K = K^*$ is strictly positive Hermitian and we precondition with $Q = K^*$ we have

$$(U\hat q_k)(s) = \sqrt{\rho(s)}\, Y_k\, U(s, X_k), \quad s\in\mathbb{S},$$

and the expression for $\hat{\mathcal{M}}_n$ simplifies accordingly. If, in addition, K is compact so that $K = \sqrt{R} = \sum_{m=1}^{\infty} \sqrt{\rho_m}\, e_m \otimes e_m$, the integral with respect to Σ reduces to summation over m with

$$\langle \hat{q}_k, e_m \rangle = \sqrt{\rho_m}\, Y_k\, e_m(X_k), \quad m \in \mathbb{S} = \mathbb{N}.$$

6. Example. Let us consider the situation of the example in §2.7-§2.10 and §3.10, where R is strictly positive Hermitian but after nonstandard preconditioning with $Q = (K^2)^* K$. Let ρ_m and e_m be as in §3.10. Then in this case we have

$$\langle \hat{q}_k, e_m \rangle = 2\sqrt{2}\, \pi\, Y_k \int_0^1 (\sqrt{X_k} - \sqrt{X_k - s \wedge X_k}) \sin(m - \frac{1}{2})\pi\ ds.$$

Hence the expression for $\hat{\mathcal{M}}_n$ becomes

$$\begin{aligned} \hat{\mathcal{M}}_n(\alpha) = \frac{1}{n} \sum_{k=1}^{n} \sum_{m=1}^{M(\alpha)} (m - \frac{1}{2})^4 \{ &\frac{1}{n} \langle \hat{q}_k, e_m \rangle^2 \\ &- \frac{n+1}{n} \frac{1}{n-1} \sum_{j \neq k} \langle \hat{q}_j, e_m \rangle \langle \hat{q}_k, e_m \rangle \}, \end{aligned}$$

where $M(\alpha)$ is the largest integer such that $(m - \frac{1}{2})^{-2} \geq \alpha$.

7. Because we don't apply the standard preconditioning here, there is no visible compensation of the factor $1/\rho^2$. In fact, however, this compensation does exist because the inner products $\langle \hat{q}_k, e_m \rangle$ contain the highly oscillating sine functions (when m is large). As a refinement of the Riemann-Lebesgue lemma we have, more precisely,

$$\begin{aligned} &2\sqrt{2}\, \pi \int_0^1 (\sqrt{X} - \sqrt{X - s \wedge X})\ \sin(m - \frac{1}{2})\pi s\ ds \\ &= \frac{\sqrt{2}}{m - \frac{1}{2}} \int_0^X \frac{\cos(m - \frac{1}{2})\pi s}{\sqrt{X - s}}\ ds \\ &= \frac{\sqrt{2X}}{(m - \frac{1}{2})^{3/2}} \int_0^{m - \frac{1}{2}} \frac{\cos \pi s X}{\sqrt{m - \frac{1}{2} - s}}\ ds \\ &= \rho_m^{3/4} \sqrt{2X} \int_0^{m - \frac{1}{2}} \frac{\cos \pi s X}{\sqrt{m - \frac{1}{2} - s}}\ ds. \end{aligned}$$

Since this last integral can be written as a finite alternating series with terms decreasing in absolute value it follows that

$$\int_0^{m-\frac{1}{2}} \frac{\cos \pi s X}{\sqrt{m-\frac{1}{2}-s}}\, ds \le C \frac{1}{\sqrt{2X}},$$

for some generic $0 < C < \infty$. Consequently we see that

$$|\langle \hat{q}_k, e_m \rangle| \le C \rho_m^{3/4},$$

and this entails that the factor $(m-\frac{1}{2})^4$ in the expression for $\hat{\mathcal{M}}_n$ is reduced to $(m-\frac{1}{2}) = \rho m^{-1/2}$. These resulting orders seem to be in line with those that might be expected if standard preconditioning would have been applied.

8 Concluding Remarks

1. Ordinary, direct regression estimation. For developing the theory, we have not required that K be an integral operator (§2.3). As has been observed this enables us to formally include ordinary direct regression estimation as a special case. Now we have the simpler model $K = I$ with the data being independent copies of

$$Y = f(X) + \epsilon, \quad f \in L^2(\tau).$$

We will now precondition with a strictly positive Hermitian integral operator $Q = R : L^2(\tau) \to L^2(\tau)$.

2. The estimator of $q = Rf \in L^2(\tau)$ is now given by

$$\hat{q}(t) = \frac{1}{n} \sum_{k=1}^{n} Y_k\, R(t, X_k), \quad t \in \mathbb{T}.$$

For general Ψ_α as decribed in §3.4 let us write $t\Psi_\alpha(t) = 1_\alpha(t)$, a function which is ≈ 1 for $t > 0$. The estimator of f can then be written as

$$\hat{f}_\alpha = R_\alpha^{-1} \hat{q} = \frac{1}{n} \sum_{k=1}^{n} Y_k\, U^{-1}(1_\alpha(\rho) \cdot U(\bullet, X_k)),$$

with expectation (cf.§3.4)

$$f_\alpha = U^{-1} 1_\alpha(\rho)\, Uf = I_\alpha f.$$

For the spectral-cut-off and Moore-Penrose regularization we have respectively

$$1_\alpha(t) = 1_{[\alpha,\infty)}(t), \quad 1_\alpha(t) = \frac{t}{\alpha + t}.$$

3. For the MISE we now find

$$\begin{aligned}\mathbf{E}\,\|\hat{f}_\alpha - f\|^2 &= \mathbf{E}\,\|\hat{f}_\alpha - f_\alpha\|^2 + \|f_\alpha - f\|^2 \\ &= \frac{1}{n}\int_{\mathbb{S}} 1_\alpha(\rho(s))\ \mathbf{Var}\ YU(s,X)\ d\Sigma(s) \\ &\quad + \int_{\mathbb{S}} (1_\alpha(\rho(s)) - 1)^2 |(Uf)(s)|^2\ d\Sigma(s).\end{aligned}$$

Usually, the lower bound over a subclass like in §4.5 will be

$$\inf\nolimits_{T\in\mathcal{T}} \sup\nolimits_{f\in\mathcal{F}} \mathbf{E}\,\|T - f\|^2 \geq C \sum_k \frac{\lambda_k^2}{1 + n\lambda_k^2},$$

and rate optimality can be established in a similar way.

4. If for R we take a compact operator so that it can be written as $R = \sum_{m=1}^\infty \rho_m e_m \otimes e_m$, if we assume as usual that $\rho_m \downarrow 0$, as $m \to \infty$, and if we apply spectral-cut-off regularization the estimator is of the form

$$\hat{f}_\alpha(t) = \sum_{m=1}^{M(\alpha)} \{\frac{1}{n}\sum_{k=1}^{n} Y_k e_m(X_k)\} e_m(t), \quad t \in \mathbb{T}.$$

This means that it is a truncated series type estimator.

5. If we abandon the condition that the design variable is uniform we may take $L^2(\tau) = L^2(\mathbb{R})$ and for R convolution with a symmetric $r \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$, with $\tilde{r} > 0$ on $\mathbb{R}$ and $\tilde{r}(t)$ strictly decreasing as $|t| \uparrow \infty$. This means that $\{\tilde{r} > \alpha\} = [-A, A]$ for some $A = A(\alpha)$. Using spectral-cut-off regularization we will now arrive at a kernel type estimator with kernel

$$\begin{aligned}
K_\alpha(t) &= \frac{1}{\sqrt{2\pi}} F^{-1} 1_{\{\tilde{r} \geq \alpha\}} \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-its} 1_{\{\tilde{r} \geq \alpha\}}(s)\ ds \\
&= \frac{1}{2\pi} \int_{-A}^{A} e^{-its}\ ds \\
&= \frac{A}{\pi} \operatorname{sinc} At, \quad t \in \mathbb{R},
\end{aligned}$$

where sinc $x = (\sin x)/x$. This means that we obtain the sinc kernel, no matter what reasonable convolution operator we choose.

6. Generalization of inversion pattern. First of all we will no longer require that R be strictly positive Hermitian, just bounded. This means that we may precondition with any bounded operator Q, provided that $R = QK : L^2(\tau) \to L^2(\tau)$ is a *bounded, linear, integral operator.* For $\mathcal{F} \subset L^2(\tau)$, let $\{e_s,\ s \in \mathbb{S}\}$ be a family of measurable functions on $\mathbb{T}$, indexed by $\mathbb{S}$, such that $\int_{\mathbb{T}} |fe_s|\, d\tau < \infty$ for all $s \in \mathbb{S}$ and $f \in \mathcal{F}$. Suppose that it is possible to recover each $f \in \mathcal{F}$ from the numbers

$$[f, e_s] = \int_{\mathbb{T}} f e_s\ d\tau, \quad s \in \mathbb{S}.$$

If K is an integral operator we might for instance precondition with the identity operator I.

7. In many cases an operator $\mathcal{D}$ (not necessarily an integral operator) will exist that is like the adjoint of R^{-1}, such that

$$[f, e_s] = [R^{-1}q, e_s] = [q, \mathcal{D}e_s], \quad s \in \mathbb{S}.$$

If we assume $\mathcal{D}$ to be known we can usually estimate the unknown parameters $f_s = [f, e_s]$ thanks to this identity by means of random variables $\hat{f}_s$, $s \in \mathbb{S}$, say. This would finally enable us to recover f approximately. In this last step a kind of regularization will again be needed.

8. Let us first sketch how the procedure we have followed thus far fits into this general framework. In this case we know that $R = U^{-1} M_\rho U$ and we should choose

$$\boxed{e_s = U(s, \bullet), \quad \mathcal{D}e_s = \frac{1}{\rho(s)} U(s, \bullet)}$$

Indeed, we have

$$\begin{aligned}
[f, e_s] &= \int_{\mathbb{T}} f(t)\, U(s,t)\ d\tau(t) \\
&= (Uf)(s) \\
&= (UR^{-1}q)(s) \\
&= \frac{1}{\rho(s)} \int_{\mathbb{T}} U(s,t) q(t)\ d\tau(t) \\
&= [q, \mathcal{D}e_s].
\end{aligned}$$

9. In principle we now have the freedom to choose a determining system $\{e_s,\ s \in \mathbb{S}\}$ that is suitable to express certain properties, like smoothness, of the input signal; this means that we don't need any longer to describe f in terms of a system that naturally emerges from the " diagonalization " of the operator. Let K be the operator from the Abel equation in §2.8. Let us precondition with the identity operator I, i.e., let us not precondition at all. We don't know the eigenfunctions of K, but an orthonormal basis of trigonometric functions is suitable to describe the input functions in a smoothness class $\mathcal{F}$. So let $\{e_m,\ m \in \mathbb{N}\}$ be such a real valued trigonometric system. It is known that for

$$g(x) = \frac{1}{\sqrt{\pi}} \int_0^x \frac{f(y)}{\sqrt{x-y}}\ dy = (Kf)(x), \quad 0 \le x \le 1,$$

we have $f = -K^* Dg$, where D is differentiation and

$$(K^* g)(x) = \frac{1}{\sqrt{\pi}} \int_x^1 \frac{g(y)}{\sqrt{y-x}}\ dy, \quad 0 \le x \le 1.$$

Now we have $q = g$, and we choose

$$\boxed{e_s = e_m, \quad s = m \in \mathbb{N}, \quad \mathcal{D}e_m = -K^* e_m'}$$

It follows that, indeed,

$$
\begin{aligned}
[f, e_m] &= \langle f, e_m \rangle \\
&= \langle K^{-1} g, e_m \rangle \\
&= \langle g, (K^{-1})^* e_m \rangle \\
&= \langle g, -K^* e'_m \rangle \\
&= [g, \mathcal{D} e_m].
\end{aligned}
$$

10. In the situation of §8.9 we may have to recover irregular inputs, for instance input functions with discontinuous of the first kind. In order to capture the local irregularities one may replace the trigonometric system with an *orthonormal basis* of *localized wavelets.*

The topic of noisy integral equations has its roots in the theory of integral equations with ramifications in functional analysis, numerical analysis, approximation theory, and last but not least, statistics. Only a very few selected references will be given.

References

[1] Bertero, M. (1989). Linear inverse and ill-posed problems. *Advances in Electrononics and Electron Physics* **75,** 1-120.

[2] Bickel, P.J., Klaassen, Ch.A.J., Ritov, Y., & Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* The Johns Hopkins Univ. Press, Baltimore.

[3] Cavalier, L., Golubev, G.K., Picard, D. & Tsybakov, A.B. (2000). Oracle inequalities for inverse problems. Prépublication #602, Universités de Paris 6 & Paris 7.

[4] Cavalier, L., Tsybakov, B. (2000). Sharp adaptation for inverse problems with random noise. Statitics Research Report #2000.001. The Australian National University.

[5] Donoho, D.L. (1995). Nonlinear solutions of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comp. Harmon. Anal.* **2,** 101-126.

[6] Fan, J. (1991). Global behavior of deconvolution kernel estimates. *Statist.Sin.* **1,** 541-551.

[7] Golubev, G.K., & Khasminskii, R.Z. (1999). Statistical approach to some inverse boundary problems for partial differential equations. *Problems Inf. Transmission* **35,** 51-66.

[8] Hackbusch, W. (1995). *Integral Equations.* Birkhäuser, Basel.

[9] Halmos, P.R. (1963). What does the spectral theorem say? *Amer. Math. Monthly* **70,** 241-247.

[10] Ibragimov, I.A. & Has'minskii, R.Z. (1981). *Statistical Estimation.* Springer, New York.

[11] Johnstone, I.M. & Silverman, B.W. (1990). Speed of estimation in positron emission tomography and related inverse problems. *Ann. Statist.* **18,** 251-280.

[12] Kac, M. (1966). Can you hear the shape of a drum? *Amer. Math. Monthly* **73,** 1-23.

[13] Nychka, D. & Cox, D.D. (1989). Convergence rates for regularized solutions of integral equations from discrete noisy data. *Ann. Statist.* **17,** 556-572.

[14] O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statist. Sin.* **4,** 502-527.

[15] van der Vaart, A.W. (1998). *Asymptotic Statistics.* Cambridge Univ. Press.

[16] Wahba, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.