

# ITERATIVE INVERSION METHODS FOR STATISTICAL INVERSE PROBLEMS

NICOLAI BISSANTZ

*Institute for Mathematical Stochastics, University of Göttingen, Maschmühlenweg 8-10, 37083 Göttingen, Germany  
E-mail: bissantz@math.uni-goettingen.de*

In this paper we discuss general regularization estimators. This class includes Tikhonov type and spectral cut-off estimators as well as iterative methods, such as  $\nu$ -methods and the Landweber iteration. The latter estimators achieve the same (optimal) convergence rates as spectral cut-off, but do not require explicit spectral information on the operator and are often much faster to compute than Tikhonov regularization. We demonstrate application of a  $\nu$ -method by an example involving the backwards heat equation.

## 1. Introduction

In this paper we are concerned with Inverse Problems. Here we aim to estimate some quantity of interest, which cannot be observed directly. In more detail, suppose we want to estimate a quantity described by an element  $f$  in a Hilbert space  $\mathbb{H}_1$  from indirect noisy measurements

$$Y = (Kf)(X) + \sigma \cdot \xi, \quad (1)$$

where  $K$  is a known operator  $K : \mathbb{H}_1 \rightarrow \mathbb{H}_2$  mapping  $\mathbb{H}_1$  to another Hilbert space  $\mathbb{H}_2$ . The observations  $Y$  and  $\mathbb{H}_2$  are Hilbert-space-valued processes described below in more detail, and  $\sigma$  is the variance of the noise. We assume that  $K$  is linear, bounded and injective, but not necessarily compact.

Inverse problems are prevalent in science. Typical examples include parameter identification problems in partial differential equations, e.g. the backwards heat equation. Here  $K$  is the so-called “parameter-to-solution” operator, which simply means solving the partial differential equation for the parameter  $f$ . Another typical class of problems emerges if  $K$  is an integral operator, whence (1) may be an inverse regression or an inverse density estimation problem, e.g. estimation of the density of globular cluster luminosities in the Antennae galaxies from noisy observations<sup>1, 2</sup>.

The organization of this paper is as follows. In section 2 we show how model (1) relates to inverse regression and inverse density (quasi-)deconvolution problems, and briefly discuss Tikhonov and spectral cut-off estimators for  $f$ , which are the most frequently used spectral regularization estimators in practical applications. Moreover, we introduce iterative spectral regularization methods, which are often computationally more feasible than the afore-

mentioned. In section 3 we apply  $\nu$ -methods to the backwards heat equation.

It is beyond the scope of this paper to discuss in detail the technical assumptions required for the results presented. Instead, we refer to Bissantz, Hohage, Munk & Ruymgaart<sup>1</sup>.

## 2. Methodology

### 2.1. The noise model

In this section we discuss how model (1) is related to practically relevant statistical models. We assume that the noise  $\xi$  is a Hilbert-space valued process, which is centered and has variance 1. Important applications of model (1) are the following.

**Error-in-variables, deconvolution:** Suppose that the following observations are at our disposal

$$X_1, \dots, X_n \sim X = F + W,$$

where  $F, W$  are stochastically independent, with densities  $f, w \in L^2$  and  $w$  known. Our aim is to estimate  $f$ . In this case the density  $g$  of  $X$  is related to  $f$  by the convolution operator

$$g = Kf = w * f.$$

It will be shown below that estimation of  $f$  by spectral regularization methods can be achieved by estimating  $q := K^*g = K^*Kf$  from the observations in the first step. In the density deconvolution or error-in-variables problem, an unbiased,  $\sqrt{n}$ -consistent estimator of  $q$  is

$$\hat{q}_n(\cdot) = \frac{1}{n} \sum_{i=1}^n w(X_j - \cdot), \quad (2)$$

and the noise process  $\xi$  is given by

$$K^*\xi = (\hat{q}_n - q)/\sigma,$$

where  $\sigma = (\|g\|_{L^\infty} + \|g\|_{L^2}^2)^{1/2} / \sqrt{n}$ .

**Inverse regression, Fredholm equation:** Next we consider the regression setting, where we want to estimate the input function  $f$  from  $n$  discrete, noisy i.i.d. observations

$$Y_i = Kf(X_i) + \epsilon_i, \quad i = 1, \dots, n,$$

where  $(X_i, \epsilon_i)$  are stochastically independent design variables  $X_i$  and noise terms  $\epsilon_i$ , and

$$\mathbb{E}[Y|X] = Kf(X)$$

for a linear integral operator  $K$ . Similarly as in the deconvolution case the generalized empirical process

$$\hat{q}_n(\cdot) = \frac{1}{n} \sum_{i=1}^n Y_i K(X_i, \cdot)$$

estimates  $q = (K^*K)f$  in an unbiased and  $\sqrt{n}$ -consistent manner. Moreover, the noise process  $\xi$  can again be defined as

$$K^*\xi = (\hat{q}_n - q)/\sigma,$$

where  $\sigma = (\text{Var}\epsilon_1 + \|Kf\|_{L^\infty}^2 + \|Kf\|_{L^2}^2)^{1/2} / \sqrt{n}$ .

**Quasi-deconvolution:** Reconsider the deconvolution case, but now assume that the density  $g$  of  $X$  is given by

$$g = \int_{\mathbb{R}^d} h(\cdot - y|y)f(y)dy =: Kf,$$

where  $h(\cdot|y)$  is the conditional density of  $W$  given  $Y = y$ . Note that  $K$  is a convolution operator if  $Y$  and  $W$  would be stochastically independent. However, in many practical applications the variance of the noise term  $W$  depends on  $F$ . For a typical example consider observations of the brightness of a globular cluster belonging to some remote galaxy. Here the measurement gets increasingly difficult with fainter cluster brightness, and the measurement noise increases. For quasi-deconvolution we replace the estimator (2) of  $q := K^*g$  by

$$\hat{q}_n(z) := \frac{1}{n} \sum_{j=1}^n h(X_j - z|z).$$

## 2.2. Inverse estimators

We assume that the operator  $K : \mathbb{H}_1 \rightarrow \mathbb{H}_2$  is bounded and injective. Therefore its generalized (Moore-Penrose) inverse

$$K^\dagger = (K^*K)^{-1}K^* : R(K) \oplus R(K)^\perp \rightarrow \mathbb{H}_1$$

is in general unbounded, and noise in the measurements  $Y$  is blown up by the inversion  $\hat{f} := (K^*K)^{-1}\hat{q}_n$ , which, in general, yields useless results  $\hat{f}$ . A possible solution to this problem consists in *regularization*, i.e. to replace  $K^\dagger$  by a sequence of bounded operators  $R_\alpha$  with *regularization parameter*  $\alpha$ , such that  $R_\alpha \rightarrow K^\dagger$  for  $\alpha \searrow 0$  (pointwise).

How can we construct such regularization estimators for general inverse problems? The fundamental tool is *Halmos' spectral theorem*<sup>3</sup>: Let  $A : \mathbb{H} \rightarrow \mathbb{H}$  be a bounded, self-adjoint operator defined on a separable Hilbert space  $\mathbb{H}$ . Then there exists a  $\sigma$ -compact space  $\mathbb{S}$ , a Borel measure  $\Sigma$  on  $\mathbb{S}$ , a unitary operator  $U : \mathbb{H} \rightarrow L^2(\Sigma)$ , and a measurable function  $\rho : \mathbb{S} \rightarrow \mathbb{R}$  such that

$$UAf = \rho \cdot Uf, \quad \Sigma - \text{almost everywhere,}$$

for all  $f \in \mathbb{H}$ . The spectral theorem justifies the *functional calculus*, which will be used to define general spectral regularization estimators. Let  $\Phi : \sigma(A) \rightarrow \mathbb{R}$  be a bounded function on the spectrum  $\sigma(A)$  of  $A$ . Then

$$\Phi(A) = U^*M_{\Phi(\rho)}U,$$

where  $M_{\Phi(\rho)}$  is the operator given by multiplication with  $\Phi(\rho)$ , and  $U^*$  the adjoint of  $U$ . For example, if  $K$  is the operator generated by convolution with some (known) density  $w$  on  $\mathbb{R}$ ,  $K^*$  its adjoint and  $A := K^*K$ , then the unitary transform  $U$  which appears in the spectral theorem and in functional calculus are the Fourier transformation  $\mathcal{F}$ , and  $\rho$  is the Fourier transform of  $w$ .

We now define the regularized inverse of  $K^\dagger$  as

$$\Phi_\alpha(K^*K)K^*, \quad (3)$$

where  $\Phi_\alpha : \sigma(A) \rightarrow \mathbb{R}$ ,  $\alpha > 0$  are bounded functions which satisfy

$$\lim_{\alpha \searrow 0} \Phi_\alpha(t) = \frac{1}{t}, \quad \text{for all } t \in \sigma(A),$$

in particular. The (generalized) spectral regularization estimator for  $f$  is given by

$$\hat{f}_{\alpha,n} := \Phi_\alpha(K^*K)\hat{q}_n.$$

For a detailed discussion on admissible collections of functions  $\Phi_\alpha$ , and the necessary regularity properties, we refer to Engl, Hanke & Neubauer<sup>4</sup> and Bissantz, Hohage, Munk & Ruymgaart<sup>1</sup>.

In practical applications, Tikhonov regularization type methods and spectral cut-off are the most frequently used methods. Tikhonov regularization

results if  $(K^*K)^{-1}$  is replaced by  $(K^*K + \alpha I)^{-1}$ , and can therefore be computed without referring to the spectral information  $\sigma(A), U$ . However, it can also be defined as the regularization estimator (3) for  $\Phi_\alpha^{\text{Tik}}(t) := 1/(t + \alpha)$ . In the case of spectral cut-off methods we have

$$\Phi_\alpha^{\text{SC}} := \begin{cases} t^{-1}, & t \geq \alpha \\ 0, & t < \alpha \end{cases}$$

To provide a specific example, consider density deconvolution on  $\mathbb{R}$  (cf. Section 2.1). Then the spectral cut-off estimator of  $f$  reads

$$\hat{f}_{\alpha,n} = \mathcal{F}^{-1} \int_{\rho(\omega) \geq \alpha} \frac{\mathcal{F}\hat{q}_n(\omega)}{\rho(\omega)} d\omega, \quad (4)$$

where  $\rho = |\mathcal{F}w|^2$ . Note from eq. (4) that the regularization property of spectral cut-off is achieved by neglecting the high-frequency information in the observations  $\hat{q}_n$ . This is because  $\rho(\omega) \searrow 0$  for  $\omega \rightarrow \infty$ , and division by  $\rho$  would blow up the measurement noise by an arbitrarily large amount for increasing frequency  $|\omega|$  if no regularization is performed.

Both Tikhonov regularization and spectral cut-off methods require setting up a matrix representing the operator  $K$ , and moreover a matrix inversion or eigenvalue decomposition. This can be computationally very costly, e.g. in the case of parameter identification problems in partial differential equations. Another reason can be that estimates  $\hat{f}_{\alpha,n}$  are computed for many different values of the regularization parameter  $\alpha$  in the case of data-driven regularization parameter selection methods such as cross-validation.

On the other hand, iterative methods can be defined for suitable collections of functions  $\Phi_\alpha$ , which require for their computation only to apply the matrix representing the operator  $K$  and its transpose to a solution vector. This is an important advantage since for many problems there exist algorithms to apply the matrix to a given vector at a much smaller computational cost than the cost of setting up the matrix.

Important iterative spectral regularization methods are Landweber iterations and  $\nu$ -methods. For these methods the regularization parameter  $\alpha$  is given by the stopping index  $k$  of the iterations. The more iterations are performed, the less regularization is imposed on the solution.

For *Landweber iterations* we have  $\Phi_{1/(k+1)}(t) :=$

$\sum_{j=0}^{k-1} (1-t)^j$ , but the method can be implemented by the recursion formula

$$\hat{f}_{0,\sigma} = 0, \quad \hat{f}_{k+1,\sigma} = \hat{f}_{k,\sigma} - A\hat{f}_{k,\sigma} + K^*Y, \quad k = 0, 1, \dots,$$

i.e. we do not require the spectral information  $U, \rho$  of  $A = K^*K$ . Here, the regularization parameter can be identified as  $\alpha = 1/(k+1)$ , and the norms on  $\mathbb{H}_1$  and  $\mathbb{H}_2$  have to be scaled such that  $\|A\| \leq 1$ .

Better numerical convergence than for Landweber iterations can be achieved by  $\nu$ -methods<sup>5</sup>. For a given parameter  $\nu > 0$ , the estimator  $\hat{f}_{k,\sigma}$  can be computed by the three-term recursion

$$\begin{aligned} \hat{f}_{k,\sigma} &= \hat{f}_{k-1,\sigma} + \theta_k \left( \hat{f}_{k-1,\sigma} - \hat{f}_{k-2,\sigma} \right) \\ &\quad + \omega_k K^* \left( Y - K\hat{f}_{k-1,\sigma} \right), \quad k \geq 2, \end{aligned}$$

with starting values  $\hat{f}_{0,\sigma} := 0, \hat{f}_{1,\sigma} = \omega_1 K^*Y$ , coefficients  $\theta_1 = 0, \omega_1 = (4\nu + 2)/(4\nu + 1)$  and

$$\theta_k = \frac{(k-1)(2k-3)(2k+2\nu-1)}{(k+2\nu-1)(2k+4\nu-1)(2k+2\nu-3)},$$

$$\omega_k = 4 \frac{(2k+2\nu-1)(k+\nu-1)}{(k+2\nu-1)(2k+4\nu-1)},$$

for  $k \geq 2$ . Now the regularization parameter can be identified by  $\alpha = (1+k)^{-2}$ , which implies that the number of iterations required for  $\nu$ -methods typically are of order square root the number of required Landweber iterations.

Bissantz, Hohage, Munk & Ruymgaart<sup>1</sup> analyzed the convergence of general spectral regularization methods of the form (3), and determined their rates of convergence, which depend on the smoothness properties of the input function  $f$ . It turns out that all methods defined in (3) converge with the same rates of convergence as spectral cut-off, which are in many cases optimal (cf. Mair & Ruymgaart<sup>6</sup>). However, this only holds true as long as the smoothness of  $f$  is within the *qualification* of the respective method. Spectral cut-off and Landweber iterations have infinite qualification, and  $\nu$ -methods are available for arbitrary qualification, but Tikhonov regularization has small qualification 1. This implies that in many cases Tikhonov methods cannot converge with optimal order. For details we refer to Engl, Hanke & Neubauer<sup>4</sup> and Bissantz, Hohage, Munk & Ruymgaart<sup>1</sup>.

### 3. The backwards heat equation

Finally, we briefly discuss an application of  $\nu$ -methods to the backwards heat equation. To this end consider the inverse problem of reconstructing the temperature distribution at time  $t = 0$  on some compact domain  $H \subset \mathbb{R}^2$  from discrete, noisy observations

$$Y_i = g(X_i) + \varepsilon_i$$

of the temperature distribution at time  $t = T$ , where the design points  $X_i$  form a regular mesh on  $H$ , and  $\varepsilon_i$  is a centered, i.i.d. noise term with standard deviation  $\sigma$ .

For the backwards heat equation, the forward “parameter-to-solution” problem is described by the partial differential equation of parabolic type

$$\begin{aligned} \partial_t u(x, t) &= \Delta u(x, t), & x \in H, t \in (0, T) \\ u(x, t) &= 0, & x \in \partial H, t \in (0, T] \\ u(x, 0) &= f(x), & x \in H, \end{aligned} \quad (5)$$

with an initial temperature distribution  $f \in L^2(H)$  and the final temperature distribution  $g(x) := u(x, T)$ ,  $x \in H$ .

We have implemented the backwards heat equation for a two-dimensional, approximately heart-shaped, smooth domain  $H$  and defined the operator  $K$  as the evolution of the heat equation from time  $t = 0$  to  $T = 0.001$ . For the observations the sample size is  $n = 200$  and  $\sigma = 0.001$ . Moreover, the Laplace operator on the domain  $H$  was discretized by a finite difference scheme using 16038 unknowns. The matrix representing the forward solution operator  $K$  is therefore a dense  $16038 \times 16038$  matrix, which would require a huge amount of computation time to be set up. However, the application of the operator  $K$  to a vector  $f$  can be implemented efficiently by time stepping methods. We have used a BDF multistep method.

To apply a  $\nu$ -method to this problem we first have to estimate  $q = K^*g$ . To this end we first estimate  $g := Kf$  with a locally linear estimator  $\hat{g}$  from the observations  $Y_i$ . In the second step we compute  $\hat{q} := K^*\hat{g}$ . This estimator of  $q$  is not unbiased because the local polynomial estimator  $\hat{g}$  used in the first step is not either. However, in numerical simulations this approach turned out to be very stable. For a discussion of local polynomial estimators cf. Wand & Jones<sup>7</sup>.

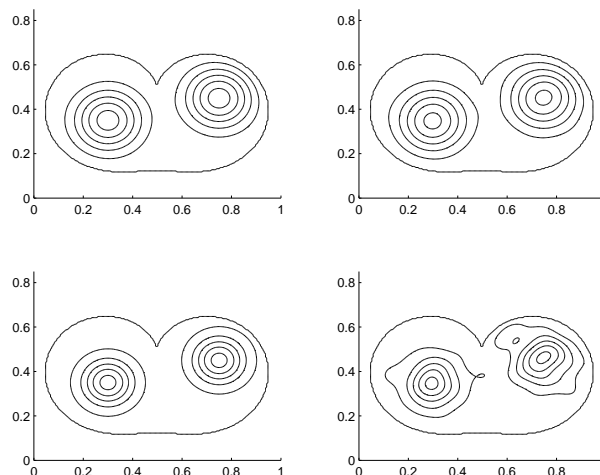


Fig. 1. A typical simulation of the backwards heat equation. Upper row (from left to right): True  $q$  and estimate  $\hat{q}$  from  $n = 200$  observations  $Y_i$ . Contour levels are 0.1, 0.2, 0.3, ... Lower row (from left to right): True  $f$  and estimate  $\hat{f}$ . Here the contour levels are 0.1, 0.3, 0.5, ... The outer, heart-shaped contour indicates the boundary of the domain under consideration.

Fig. 1 shows a typical example of a simulation, where a  $\nu$ -method was used with  $\nu = 1$  and 8 iterations, which amounts to approximately 5 minutes of CPU time on a Pentium IV 1.7 Ghz processor.

### Acknowledgments

The author would like to thank A. Blümel and T. Hohage for help and interesting discussions.

### References

1. N. Bissantz, T. Hohage, A. Munk and F. Ruymgaart, Convergence rates of general regularization methods for statistical inverse problems and applications, submitted.
2. P. Anders, N. Bissantz, L. Boysen, U. F. v. Alvensleben and R. de Grijs, The luminosity distribution of young massive clusters in the Antennae galaxies, in preparation.
3. P. R. Halmos, *Amer. Math. Monthly* **70**, 241 (1963).
4. H. W. Engl, M. Hanke and A. Neubauer, *Regularization of Inverse Problems* (Kluwer Academic Publisher, Dordrecht, Boston, London, 1996).
5. H. Brakhage, On ill-posed problems and the method of conjugate gradients, in *Inverse and Ill-Posed Problems*, eds. H. W. Engl and C. W. Groetsch (Academic Press, Orlando, 1987) pp. 191–205.
6. B. A. Mair and F. Ruymgaart, *SIAM J. Appl. Math.* **56**, 1424 (1996).
7. M. P. Wand and M. C. Jones, *Kernel Smoothing* (Chapman & Hall, London, 1995).