Available online at www.sciencedirect.com
ScienceDirect
Journal of Computational and Applied Mathematics 198 (2007) 493–504
JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS
www.elsevier.com/locate/cam

Statistical inverse problems: Discretization, model reduction and inverse crimes

Jari Kaipio[a], Erkki Somersalo[b,*,1]

[a] *Department of Applied Physics, University of Kuopio, P.O. Box 1627, FIN–70211 Kuopio, Finland*
[b] *Institute of Mathematics, Helsinki University of Technology, P.O. Box 1100, FIN–02015 TKK, Finland*

Received 30 November 2004; received in revised form 28 July 2005

Abstract

The article discusses the discretization of linear inverse problems. When an inverse problem is formulated in terms of infinite-dimensional function spaces and then discretized for computational purposes, a discretization error appears. Since inverse problems are typically ill-posed, neglecting this error may have serious consequences to the quality of the reconstruction. The Bayesian paradigm provides tools to estimate the statistics of the discretization error that is made part of the measurement and modelling errors of the estimation problem. This approach also provides tools to reduce the dimensionality of inverse problems in a controlled manner. The ideas are demonstrated with a computed example.

Keywords: Inverse problems; Bayesian statistics; Discretization; Modelling error

1. Introduction

In this article, we consider inverse problems from the point of view of Bayesian statistics. In this paradigm, the traditional deterministic formulation of an inverse problem is replaced by a statistical quest of inference. For the sake of definiteness and simplicity, the discussion here is restricted to linear inverse problems with additive noise,

$$y = Ax + e, \tag{1}$$

where the left side y is the measured data, x is the unknown of interest, A is a presumably known linear operator and, finally, e represents the unknown additive noise. Although the discussion is limited to this simple model, the general ideas carry over to more complicated problems, as will be briefly explained later in this article.

The particular issues of interest in this article are mainly related to discretization. Suppose that x represents an element in an infinite-dimensional space, e.g., a Hilbert space. To treat the problem (1) numerically, we need to represent x by means of finitely many degrees of freedom. Here, the finite-dimensional approximations of the model (1) are called discretizations. We shall study, by means of statistical analysis, the effects of discretizations to the solution of the inverse problem.

* Corresponding author.
E-mail addresses: kaipio@venda.uku.fi (J. Kaipio), erkki.somersalo@hut.fi (E. Somersalo).
[1] The work of E.S. partly supported by the Academy of Finland, project 204753.

0377-0427/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.cam.2005.09.027

A closely related problem is the model reduction. Often, a reliable numerical approximation of the forward problem $x \mapsto Ax$ requires a discretization with a high number of degrees of freedom. On the other hand, solving the inverse problem, i.e., estimating x from the noisy data y plus possible complementary information, with such a high number of variables may be prohibitively computer intensive. Thus, it is desirable to base the inverse solver on a coarser model. Since inverse problems are typically ill-posed and therefore sensitive to errors, whether originating from data collection systems or from modelling, the discrepancy between the forward and inverse models may have a dramatic effect on the quality of the solution. We propose here a model reduction strategy that models the statistics of the discrepancy and hence is able to reduce the misinterpretations due to model reduction.

The third issue to be discussed here is also related to the discrepancy between the coarse inverse model and the fine forward model. A relatively common strategy to test inverse solvers' performance is to generate simulated noisy data with the same model that the inverse solver is based on. In some cases, this strategy that ignores the discretization errors leads to excessively optimistic expectations about the performance of the method. When this happens, the testing strategy is sometimes called an inverse crime. We address this issue in the light of statistical error modelling with a numerical example.

Much of the material in this article is based on the book [3] where further results and ideas can be found.

2. The setting

In this section, we present the basic tools needed for the statistical analysis of discretization effects. These include the fundamentals of Bayesian inversion and some results concerning infinite-dimensional random variables.

2.1. Statistical inversion

Consider Eq. (1). We assume that we have finite amount of data, i.e., $y \in \mathbb{R}^m$. To define the basic setting of the statistical approach, assume first that $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{m\times n}$. We remark that it is possible to develop a satisfactory Bayesian inversion theory directly in infinite-dimensional function space setting. The probability densities are then defined via their finite-dimensional projections; see, e.g., [7,6].

In the Bayesian framework, we model the variables x, y and e as random variables. The interpretation of this modelling is that the information concerning their values is incomplete; rather, our information—or lack of it—is expressed by their distribution as random variables.

Assume that the probability distributions are absolutely continuous with respect to the Lebesgue measure, i.e., the probability distributions can be written in terms of probability densities that are measureable functions. By using the relatively standard notational conventions, the joint probability density of x and y can be written in terms of the conditional densities as

$$\pi(x, y) = \pi(x|y)\pi(y) = \pi(y|x)\pi(x),$$

leading to the well-known Bayes formula,

$$\pi(x|y) = \frac{\pi(y|x)\pi(x)}{\pi(y)}. \tag{2}$$

In (2), $\pi(x)$ is the prior probability density, expressing our information of x prior to measurement of y; the conditional density $\pi(y|x)$ is the likelihood density, expressing the probability density of the observation y if x were known. The left-hand side, $\pi(x|y)$ is the posterior density, i.e., the probability density of x given the prior information and the measured value of y. In the statistical inversion theory, it is usually considered as the solution of the inverse problem. Finally, the denominator $\pi(y)$ is just a norming constant and has often little importance. For further discussion, we refer to the book [3].

2.2. Discretization

Let H be a separable Hilbert space. We assume that the unknown x is an H-valued vector. To study the discretization, we assume that H has a multiresolution structure. Let $\{V_n | n \in \mathbb{N}\}$ be a family of multiresolution subspaces characterized

by the following properties:

(1) Each $V_n \subset H$ is a finite-dimensional subspace.
(2) The spaces are nested, i.e., $V_n \subset V_{n+1}$.
(3) The spaces are exhaustive, i.e.,

$$\overline{\bigcup_{n=0}^{\infty} V_n} = H.$$

Assume that

$$V_n = \mathrm{sp}\{\phi_j^n | 1 \leqslant j \leqslant d_n = \dim(V_n)\},$$

where the elements $\{\phi_j^n | 1 \leqslant j \leqslant d_n\}$ are orthonormal. We define the family of discretization operators,

$$H \to \mathbb{R}^{d_n}, \quad x \mapsto x^n = \begin{bmatrix} \langle x, \phi_1^n \rangle \\ \langle x, \phi_2^n \rangle \\ \vdots \\ \langle x, \phi_{d_n}^n \rangle \end{bmatrix} = (\Phi^n)^{\mathrm{T}} x.$$

The discrete approximation of x on the discretization level n is

$$\tilde{x}^n = \sum_{j=1}^{d_n} \langle x, \phi_j^n \rangle \phi_j^n = \Phi^n x^n \in V_n. \tag{3}$$

Hence, we have the projection,

$$P^n : H \to V_n, \quad x \mapsto \Phi^n (\Phi^n)^{\mathrm{T}} x.$$

We define also projections between different discretization levels by setting

$$P^{nk} : \mathbb{R}^{d_k} \to \mathbb{R}^{d_n}, \quad x^k \mapsto (\Phi^n)^{\mathrm{T}} \Phi^k x^k.$$

Clearly, $P^{nn} = I$, the identity matrix.

An example of multiresolution spaces will be given later in this article. The multiresolution spaces play a central role in the wavelet analysis, see, e.g., [8].

2.3. Hilbert-space valued random variables

In general, the forward model $x \mapsto Ax$ is not given in a discretized form. We assume here that $x \in H$, H being a separable Hilbert space, and $A : H \to \mathbb{R}^m$ is continuous. For the purposes of the statistical interpretation, let x be an H-valued random variable.

Assume that x has finite first and second moments, i.e., there is an element $x_* \in H$ and a linear trace class mapping $\Gamma : H \to H$ such that for all $\psi, \phi \in H$,

$$\mathbf{E}\{\langle x, \phi \rangle\} = \langle x_*, \phi \rangle,$$
$$\mathbf{E}\{\langle x - x_*, \phi \rangle \langle x - x_*, \psi \rangle\} = \langle \phi, \Gamma \psi \rangle.$$

The vector x_* and the operator Γ are called the mean and covariance of x, respectively. We write $\Gamma = \mathrm{cov}(x)$.

Consider a discretization x^n of x. It is an $\mathbb{R}^{d_n}$-valued random variable. The mean and the covariance of this random variable are

$$\mathbf{E}\{x^n\} = x_*^n \in \mathbb{R}^{d_n}, \quad (x_*^n)_j = \langle x_*, \phi_j^n \rangle,$$
$$\mathbf{E}\{(x^n - x_*^n)(x^n - x_*^n)^{\mathrm{T}}\} = \Gamma^n \in \mathbb{R}^{d_n \times d_n}, \quad \Gamma_{ij}^n = \langle \phi_i^n, \Gamma \phi_j^n \rangle, \tag{4}$$

or concisely,

$$x_*^n = (\Phi^n)^{\mathrm{T}} x_*, \quad \Gamma^n = (\Phi^n)^{\mathrm{T}} \Gamma \Phi^n.$$

The H-valued random variable x is said to be Gaussian, if for any discretization level n, the variable x^n is Gaussian, i.e., we have

$$\mathbf{E}\{\exp(-\mathrm{i}\xi^{\mathrm{T}}(x^n - x_*^n))\} = \exp(-\tfrac{1}{2}\xi^{\mathrm{T}}\Gamma^n\xi), \quad \xi \in \mathbb{R}^{d_n}.$$

In particular, if the matrix Γ^n is invertible, it follows that the probability density of x^n is

$$\pi(x^n) = \left(\frac{1}{(2\pi)^{d_n}\det(\Gamma^n)}\right)^{1/2} \exp\left(-\frac{1}{2}(x^n - x_*^n)^{\mathrm{T}}(\Gamma^n)^{-1}(x^n - x_*^n)\right).$$

The theory of Gaussian random variables in Hilbert spaces is well understood. We refer to the book [10] for discussion of this topic.

3. Discretized Bayesian model

Having introduced the basic concepts in the previous section, we are ready now to put the pieces together and develop a discretized Bayesian model for Eq. (1).

Consider the discrete approximation (3) of x. We have

$$A\tilde{x}^n = A\Phi^n x^n = A^n x^n, \quad A^n = A\Phi^n \in \mathbb{R}^{m\times d_n}.$$

We start by writing a simple identity. If (1) holds, we have

$$y = A^n x^n + (Ax - A^n x^n) + e. \tag{5}$$

The term

$$w^n = w^n(x) = Ax - A^n x^n = A(I - P^n)x \tag{6}$$

is the *discretization error*. When a discrete model for (1) is set up, this term is usually neglected, and the analysis is started directly with a model

$$y = A^n x^n + e. \tag{7}$$

In the classical approach to inverse problems, the effect of the discretization error is difficult to analyze as it depends on the unknown x. The catch in the Bayesian approach is that while the discretization error is still unknown, its *statistics* is not. Hence, rather than neglecting it, we are able to include it as a part of the statistical model.

To understand the effect of the discretization error on the solution of the inverse problem, let us consider a simple but still useful case, where x and e are Gaussian and independent. Without loss of generality, we may assume that both x and e have zero mean. The covariances are denoted as

$$\mathrm{cov}(x) = \Gamma_x : H \to H, \quad \mathrm{cov}(e) = \Gamma_e \in \mathbb{R}^{m\times m}.$$

The objective is to estimate $x^n \in \mathbb{R}^{d_n}$ based on observed $y \in \mathbb{R}^m$.

Let us consider first the simple discrete model (7) ignoring the discretization error. The joint probability distribution of x^n and y is Gaussian and zero mean. Let us denote by $\Gamma_x^n \in \mathbb{R}^{d_n\times d_n}$ the discretized covariance matrix (4) of x^n. The model (7) and the assumed independency of x and e imply that

$$\mathbf{E}\{x^n y^{\mathrm{T}}\} = \mathbf{E}\{x^n (x^n)^{\mathrm{T}}\}(A^n)^{\mathrm{T}} = \Gamma_x^n (A^n)^{\mathrm{T}},$$

and

$$\mathbf{E}\{yy^{\mathrm{T}}\} = A^n \mathbf{E}\{x^n (x^n)^{\mathrm{T}}\}(A^n)^{\mathrm{T}} + \mathbf{E}\{ee^{\mathrm{T}}\} = A^n \Gamma_x^n (A^n)^{\mathrm{T}} + \Gamma_e.$$

Thus, the joint covariance matrix is in this case is

$$\mathbf{E}\left\{\begin{bmatrix} x^n \\ y \end{bmatrix}[(x^n)^{\mathrm{T}}\ y^{\mathrm{T}}]\right\} = \begin{bmatrix} \Gamma_x^n & \Gamma_x^n(A^n)^{\mathrm{T}} \\ A^n\Gamma_x^n & A_n\Gamma_x^n(A^n)^{\mathrm{T}} + \Gamma_e \end{bmatrix}.$$

The maximum a posteriori solution based on the simplified model (7) is then

$$x_{\mathrm{s}}^n = \Gamma_x^n(A^n)^{\mathrm{T}}(A^n\Gamma_x^n(A^n)^{\mathrm{T}} + \Gamma_e)^{-1}y, \tag{8}$$

see [3] for the derivation. We also remind that if Γ_x^n and Γ_e are invertible, the above solution assumes an equivalent form,

$$x_{\mathrm{s}}^n = ((A^n)^{\mathrm{T}}(\Gamma_x^n)^{-1}A^n + \Gamma_e^{-1})^{-1}(A^n)^{\mathrm{T}}y. \tag{9}$$

The previous formula corresponds to the classical *Wiener filtered* solution, while the latter one is the *Tikhonov regularized* solution.

The solution above was based on ignoring the discretization error. A corrected version that we call the *enhanced error model* is obtained by assuming that the discretization error is non-zero but independent of x. This assumption is tantamount to adding the covariance of w^n to the noise covariance. We have

$$\Gamma_w^n = \mathbf{E}\{w^n(w^n)^{\mathrm{T}}\} = A(I - P^n)\Gamma_x(I - P^n)^{\mathrm{T}}A^{\mathrm{T}} \in \mathbb{R}^{m\times m}. \tag{10}$$

By replacing Γ_e by $\Gamma_e + \Gamma_w^n$ in formulas (8) or (9) leads to an estimator

$$\begin{aligned} x_{\mathrm{e}}^n &= \Gamma_x^n(A^n)^{\mathrm{T}}(A^n\Gamma_x^n(A^n)^{\mathrm{T}} + \Gamma_e + \Gamma_w^n)^{-1}y \\ &= ((A^n)^{\mathrm{T}}(\Gamma_x^n)^{-1}A^n + (\Gamma_e + \Gamma_w^n)^{-1})^{-1}(A^n)^{\mathrm{T}}y, \end{aligned} \tag{11}$$

the latter form, of course, assuming that the covariances are invertible.

Finally, let us consider the *complete error model* (5) in which the discretization error is modelled correctly. We have

$$\mathbf{E}\{x^n y^{\mathrm{T}}\} = \mathbf{E}\{x^n x^{\mathrm{T}}\}A^{\mathrm{T}} = (\Phi^n)^{\mathrm{T}}\Gamma_x A^{\mathrm{T}} \in \mathbb{R}^{d_n\times m}$$

and

$$\mathbf{E}\{yy^{\mathrm{T}}\} = A\mathbf{E}\{xx^{\mathrm{T}}\}A^{\mathrm{T}} + \mathbf{E}\{ee^{\mathrm{T}}\} = A\Gamma_x A^{\mathrm{T}} + \Gamma_e \in \mathbb{R}^{m\times m}.$$

Hence, the joint covariance matrix of x^n and y, under this model, is

$$\mathbf{E}\left\{\begin{bmatrix} x^n \\ y \end{bmatrix}[(x^n)^{\mathrm{T}}\ y^{\mathrm{T}}]\right\} = \begin{bmatrix} \Gamma_x^n & (\Phi^n)^{\mathrm{T}}\Gamma_x A^{\mathrm{T}} \\ A\Gamma_x\Phi^n & A\Gamma_x A^{\mathrm{T}} + \Gamma_e \end{bmatrix}.$$

The maximum a posteriori estimator based on this covariance model is found to be

$$x_{\mathrm{c}}^n = (\Phi^n)^{\mathrm{T}}\Gamma_x A^{\mathrm{T}}(A\Gamma_x A^{\mathrm{T}} + \Gamma_e)^{-1}y. \tag{12}$$

In the following sections we shall discuss the performance of the various error models and the computational issues related to them.

4. Setting up the prior covariance

The Bayes formula (2) contains two factors in the numerator, the prior and the likelihood. The first question that we consider here is how the discretization and in particular, a model reduction, affects the prior density and how the prior should be set up. We consider here two examples. In the first one, the prior density is set up directly in the underlying Hilbert space and then discretized, in the second one we define a hierarchy of discrete prior densities in the multiresolution spaces.

Let $H = L^2([0, 1])$. To define the multiresolution spaces, we define

$$\phi(t) = \begin{cases} 1 & \text{if } 0 \leqslant t < 1, \\ 0 & \text{if } t < 0 \text{ or } t \geqslant 1. \end{cases}$$

Define V^n, $0 \leqslant n < \infty$, as a subspace of piecewise constant functions,

$$V^n = \text{span}\{\phi_k^n \,|\, 1 \leqslant k \leqslant 2^n\},$$

where

$$\phi_k^n(t) = 2^{n/2}\phi(2^n t - k - 1).$$

Hence, $d_n = \dim(V_n) = 2^n$.

Assume that our prior knowledge of the unknown $x \in H$ is that (i) $t \mapsto x(t)$ is smooth, and (ii) x contains details of a given size $\sim \ell < 1$. The parameter ℓ is called the *correlation length*. To set up a prior density with these properties, we write a stochastic model for x,

$$x(t) = K_\ell w(t), \quad K_\ell : \phi \mapsto \int_{-\infty}^{\infty} k_\ell(t - s)\phi(s)\,\mathrm{d}s, \quad 0 \leqslant t \leqslant 1, \tag{13}$$

where k_ℓ is a smooth kernel of width $\sim \ell$ and w is Gaussian white-noise process on $\mathbb{R}$. Since the white-noise process is distribution valued (see [2]), the convolution product above needs to be interpreted in the distributional sense.

The covariance operator of the white noise is, by definition, the identity operator, so we can solve the covariance of x,

$$\begin{aligned} \mathbf{E}\{\langle x, \varphi\rangle\langle x, \psi\rangle\} &= \mathbf{E}\{\langle w, K_\ell^{\mathrm{T}}\varphi\rangle\langle w, K_\ell^{\mathrm{T}}\psi\rangle\} \\ &= \langle \varphi, K_\ell K_\ell^{\mathrm{T}}\psi\rangle. \end{aligned} \tag{14}$$

As an example, assume that k_ℓ is a Gaussian kernel with variance λ^2. The parameter λ will be adjusted to match with the correlation length ℓ: the integral kernel $\gamma(t - t')$ of the operator $K_\ell K_\ell^{\mathrm{T}}$ is

$$\begin{aligned} \gamma(t - t') &= \int_{-\infty}^{\infty} k_\ell(t - s)k_\ell(t' - s)\,\mathrm{d}s \\ &= \frac{1}{2\pi\lambda^2}\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\lambda^2}((t - s)^2 + (t' - s)^2)\right)\mathrm{d}s \\ &= \left(\frac{1}{4\pi\lambda^2}\right)^{1/2} \exp\left(-\frac{1}{4\lambda^2}(t - t')^2\right). \end{aligned}$$

We can calculate the discretized covariance matrix Γ_x^n on any discretization level by the formula

$$\Gamma_{ij}^n = \int_0^1 \int_0^1 \phi_i^n(t)\gamma(t - t')\phi_j^n(t')\,\mathrm{d}t\,\mathrm{d}t',$$

and the resulting prior densities at each discretization level are consistent with each other. To fix the parameter λ, consider the correlation of two point values of x. By choosing the test functions φ and ψ as approximations of the Dirac distributions at t and t', respectively, and passing to the limit, we find that

$$\mathbf{E}\{x(t)x(t')\} = \gamma(t - t').$$

We require that when $|t - t'| = \ell$, then $\gamma(t - t') = \gamma(0)/10$, leading to the relation $\lambda = \ell/2\sqrt{\log 10} \approx \ell/3$. The choice of the factor 10 here is somewhat arbitrary.

To understand the meaning of this prior, we calculate random draws from this prior at different discretization levels. To this end, let us calculate the projection matrices for passing to one level to another. It is not hard to see that in this

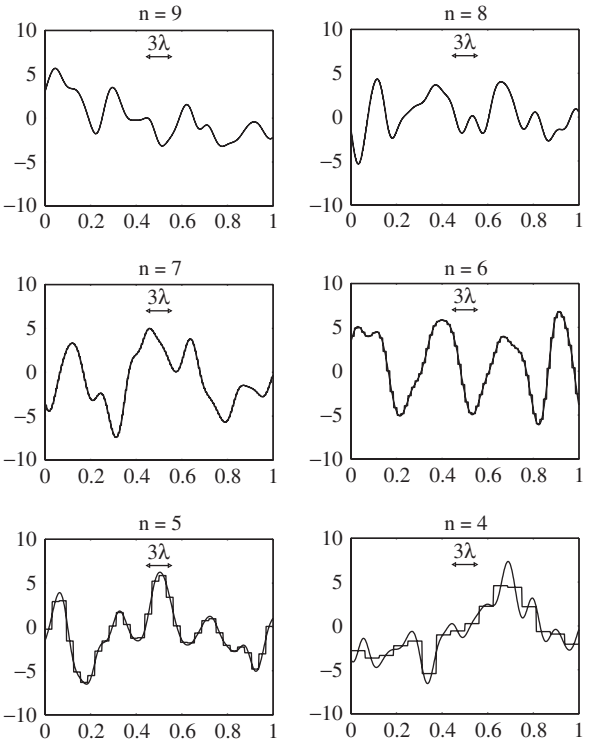

Fig. 1. Six random draws from the density corresponding to the stochastic model (13). The images show also the projections $\widetilde{x}^n = \Phi^n x^n$ with decreasing n.

model case, the projector $P^{n-1,n} : x^n \mapsto x^{n-1}$ is given by

$$P^{n-1,n} = I_{n-1} \otimes \mathbf{1}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & 1 & \dots & 0 & 0 \\ \vdots & & & & & & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & 1 \end{bmatrix} \in \mathbb{R}^{d_{n-1} \times d_n},$$

where $I_{n-1} \in \mathbb{R}^{d_{n-1} \times d_{n-1}}$ is the unit matrix, $\mathbf{1}_2 = [1\ 1]$ and $\otimes$ is the Kronecker product. Recursively, for $k > 1$,

$$P^{n-k,n} = P^{n-k,n-k+1} \cdots P^{n-1,n} = \frac{1}{2^{k/2}} I_{n-k} \otimes \mathbf{1}_{2^k},$$

with obvious notations.

The consistency of the prior model means that the passage from finer to coarser level ($N \to n$, $n < N$), is done by projections,

$$\Gamma^n = P^{n,N} \Gamma^N (P^{n,N})^{\mathrm{T}}.$$

In Fig. 1, we have plotted six random draws of the functions $\widetilde{x}^n = \Phi^n x^n$ from the prior density constructed above. In this example, we have $\lambda = 0.05$, i.e., the correlation length is $\ell \approx 0.15$. The discretization levels correspond to indices $n = 4, 5, \dots, 9$. In practice, we treat the finest model $n = 9$ as the continuous case and generate the white-noise process in this grid.

The consistency of the discrete priors is seen in the fact that on each discretization level, the random draws consist of details of the size $\sim \ell$. Naturally, if we pass further into coarser models in which the length of the discretization interval $1/2^n$ becomes larger than the correlation length, the model loses its capacity to represent the finest details. Evidently, such discretizations should be avoided. Also, we expect that when the discretization interval approaches the correlation length, the modelling error takes over and becomes the predominant source of error in the inverse solutions.

In the previous example, the prior was written directly on the Hilbert space level. In some cases, a continuous model is not available, but rather a hierarchy of discrete densities is sought for. As a first observation, note that the expectation of the norm of the discretization error (6) must go to zero as the discretization level increases. Indeed, assuming that the discretized prior models arise from an underlying Gaussian Hilbert space distribution, we have

$$\begin{aligned}\mathbf{E}\{\|w^n\|^2\} &\leqslant \|A\|^2\mathbf{E}\{\|(I-P^n)x\|^2\}\\ &= \|A\|^2\mathrm{Tr}\,(\mathbf{E}\{(I-P^n)xx^{\mathrm{T}}(I-P^n)^{\mathrm{T}}\})\\ &= \|A\|^2\mathrm{Tr}\,((I-P^n)\Gamma_x(1-P^n)^{\mathrm{T}}) \longrightarrow 0,\end{aligned}$$

as $n \to \infty$, since the operator Γ_x is in the trace class and the family $\{V_n\}$ is exhaustive. Since we are here interested in numerical applications rather than the theoretical construction, we may argue that the discretization error can be neglected in practice if the discretization is fine enough. We shall not discuss the question whether the discrete models converge to a Hilbert space model as the discretization becomes finer. Rather, the existence of the underlying Hilbert space process is taken here as granted. Thus, we assume that when $n \geqslant N$ for some $N > 0$, $w^n = 0$ up to computing precision. In practice, we write $H = V^N$ and treat this discrete model as a reference model.

After this remark, we shall consider a concrete and widely used example in the light of model reduction.

Let $H = L^2([0, 1])$ with the same multiresolution structure as before. Let N be large. In this example, let us define a second order smoothness prior,

$$\pi(x^N) \propto \exp(-\tfrac{1}{2}\alpha\|L^N x^N\|^2) = \exp(-\tfrac{1}{2}(x^N)^{\mathrm{T}}[\alpha(L^N)^{\mathrm{T}}L^N]x^N),$$

where $\alpha > 0$ and $L^N \in \mathbb{R}^{d_N \times d_N}$ is the finite difference matrix

$$L^N = 2^{2N}\begin{bmatrix} -2 & 1 & 0 & \dots & 0\\ 1 & -2 & 1 & & \\ 0 & 1 & -2 & & \vdots\\ \vdots & & & \ddots & 1\\ 0 & \dots & & 1 & -2\end{bmatrix} \in \mathbb{R}^{d_N \times d_N}.$$

The matrix L^N is invertible and the prior covariance in this case is

$$\Gamma^N = [\alpha(L^N)^{\mathrm{T}}L^N]^{-1}. \tag{15}$$

Having set up the reference prior density, assume that, e.g., for computational reasons, we need to reduce the model and pass to a coarser representation of the unknown.

What is the prior density at the level n, $n < N$? It is not hard to verify that

$$L^n = 2^{2n}\begin{bmatrix} -2 & 1 & 0 & \dots & 0\\ 1 & -2 & 1 & & \\ 0 & 1 & -2 & & \vdots\\ \vdots & & & \ddots & 1\\ 0 & \dots & & 1 & -2\end{bmatrix} = P^{n,N}L^N(P^{n,N})^{\mathrm{T}} \in \mathbb{R}^{d_n \times d_n}.$$

Therefore, one might be tempted to write a smoothness prior

$$\tilde{\pi}_{\mathrm{pr}}(x^n) \propto \exp(-\tfrac{1}{2}\alpha\|L^n x^n\|^2) = \exp(-\tfrac{1}{2}(x^n)^{\mathrm{T}}[\alpha(L^n)^{\mathrm{T}}L^n]x^n). \tag{16}$$

This may or may not work in practice. However, it is definitely *inconsistent* with the choice of $\pi(x^N)$, which can be seen by comparing the covariance matrices. If we require that $x^n = P^{n,N}x^N$, we should have

$$\Gamma^n = \mathbf{E}\{x^n(x^n)^{\mathrm{T}}\} = P^{n,N}\Gamma^N(P^{n,N})^{\mathrm{T}}.$$

But

$$\widehat{\Gamma}^n = [\alpha(L^n)^{\mathrm{T}}L^n]^{-1} \neq P^{n,N}[\alpha(L^N)^{\mathrm{T}}L^N]^{-1}(P^{n,N})^{\mathrm{T}},$$

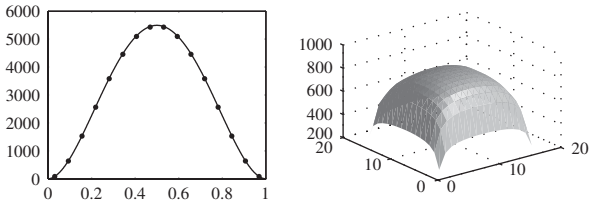

Fig. 2. Left: autocovariances of $\tilde{x}^N(t_j^N) = 2^{N/2}x_j^N$, $N = 9$, $1 \leqslant j \leqslant d^N = 512$ (solid line) and $\tilde{x}_j^n(t_j^n) = 2^{n/2}x_j^n$, $n = 4$ at t_j^n, $1 \leqslant j \leqslant d^n = 16$ (dots). Right: the matrix ratio (18) of the correct and incorrect covariance matrices.

as one can easily verify. One may ask, what practical consequences such inconsistency brings. The answer is immediate: the correlation structure is altered. To understand that, consider the pointwise correlations

$$\mathbf{E}\{\widetilde{x}^n(t_j^n)\widetilde{x}^n(t_k^n)\} = 2^n\Gamma_{jk}^n, \quad 2^n(i-1) < t_i^n < 2^n i. \tag{17}$$

In Fig. 2 we have plotted the autocovariances (17), i.e., $j = k$, with $n = N = 9$ and $n = 4$. The discretization points t_j^n for both models are the midpoints of the discretization intervals. The plot shows clearly the consistency of the models. We also see that the choice of the prior assumes that the functions go to zero at the interval endpoints. This is a consequence of the implicit boundary condition in the definition of the matrix L^N. How to avoid this boundary effect, see [1].

For comparison, we calculate the componentwise matrix ratio

$$r_{jk}^n = \frac{\Gamma_{jk}^n}{\widehat{\Gamma}_{jk}^n}, \quad 1 \leqslant j,\ k \leqslant d_n, \tag{18}$$

Fig. 2 shows this matrix as a surface plot. The plot shows that the variances are few decades off, and the cross correlations and the autocorrelations scale with different constants. In practice, this means that if we insist on using $\widehat{\Gamma}^n$ instead of Γ^n as the covariance, on each reduction level we have to readjust the parameter α in order that the variances be comparable. Such adjustments may be computationally costly, and the outcome does not respect fully the originally defined correlation structure, due to the fact that the componentwise ratio (18) is not constant.

5. A computed example of error models

In this section, we demonstrate with a computed example the effect of discretization errors. We compare the maximum a posteriori estimates (8), (11) and (12) based on the simple discretization, enhanced error model and the complete error model, respectively.

To specify the inverse problem, consider a linear model (1), where A is a convolution operator

$$A : H \to \mathbb{R}^m, \quad x \mapsto \int_0^1 a(t_j - s)x(s)\,\mathrm{d}s,$$

where the convolution kernel is an exponential,

$$a(t) = \exp(-\kappa|t|), \quad \kappa = 20.$$

and the data points are

$$t_j = \frac{1}{2m} + \frac{j-1}{m}, \quad 1 \leqslant j \leqslant m = 64,$$

i.e., the centerpoints of the discretization intervals of the level $n = 5$. The noise is Gaussian white noise with standard deviation σ equal to 1% of the maximum of the noiseless signal.

The prior density is chosen as in the first example of the previous section, i.e., the stochastic model is (13) with the Gaussian kernel k_ℓ. As the true signal, we choose the first random draw shown in Fig. 1.

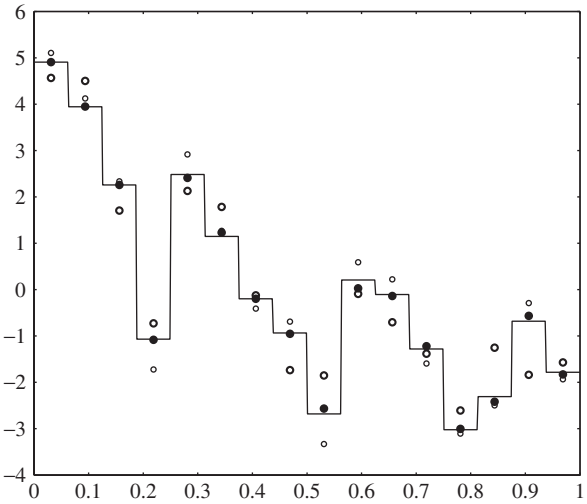

Fig. 3. The projected function $\widehat{x}^n$ (solid line) and the estimated function values with various error models at the midpoints of the discretization intervals. The solid dots are computed with the complete error model, the bold circles correspond to the enhanced error model, the thin circles the simple model.

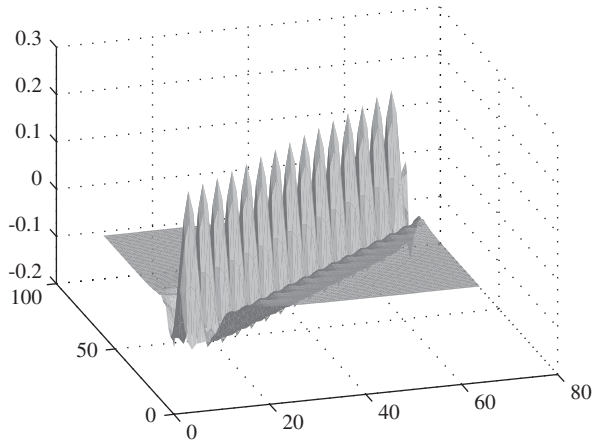

Fig. 4. The covariance matrix Γ_w^n of the discretization error, $n = 4$. The diagonal has $2^n = 16$ minima, corresponding to the midpoints of the 16 discretization intervals.

To get an idea of the performance of the estimators, we compare them to the theoretically best possible solution, that is, the projection of the true solution to the subspace V_n. In Fig. 3 we have plotted the projected piecewise constant function $\widehat{x}^n$, $n=4$ and the reconstructed values $\Phi^n x_{\mathrm{s}}^n$, $\Phi^n x_{\mathrm{e}}^n$ and $\Phi^n x_{\mathrm{c}}^n$, respectively, at the midpoints of the discretization intervals. Clearly, the complete error model outperforms the other two.

To get a more complete picture, we have also plotted the covariance matrix Γ_w^n (10) on this discretization level, see Fig. 4. For comparison, the 1% additive error model has standard deviation $\sigma = 0.0036$. The plot shows that not only is the discretization error level significantly higher than the additive noise, but it has a correlation structure that is not easy to guess by looking at its definition.

To further investigate the performances of the various estimators, consider any linear zero mean estimator on the discretization level n,

$$\widehat{x}^n = T^n y = T^n(Ax + e).$$

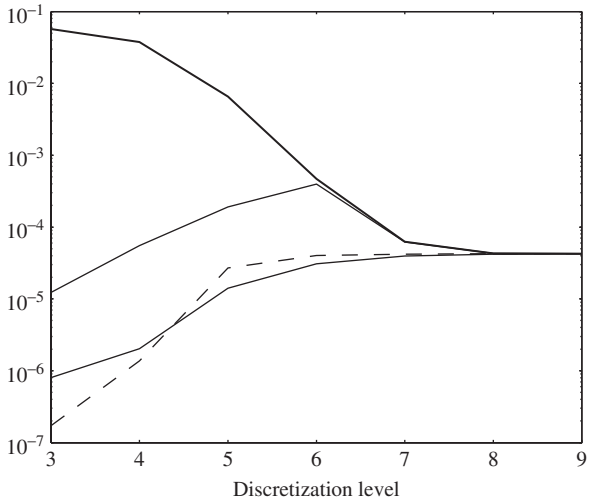

Fig. 5. Relative mean square errors. The topmost solid curve is the simple model, the middle solid curve the enchanced error model and the lowest solid curve is the complete model. The dashed curve corresponds to the unrealistic inverse crime procedure.

The operator T^n stands for any linear estimator. The relative mean square error of this estimator is defined here as

$$\mathrm{RMSE}(\widehat{x}^n) = \frac{\mathbf{E}\{\|\widehat{x}^n - x^n\|^2\}}{\mathbf{E}\{\|x^n\|^2\}}.$$

In terms of the covariance matrices, the numerator gives

$$\begin{aligned}\mathbf{E}\{\|\widehat{x}^n - x^n\|^2\} &= \mathbf{E}\{\|(P^n - T^nA)x\|^2\} + \mathbf{E}\{\|T^ne\|^2\}\\ &= \mathrm{Tr}\{(P^n - T^nA)\Gamma_x(P^n - T^nA)^{\mathrm{T}} + T^n\Gamma_e(T^n)^{\mathrm{T}}\}.\end{aligned}$$

Hence, we get

$$\mathrm{RMSE}(\widehat{x}^n) = \frac{\mathrm{Tr}\{(P^n - T^nA)\Gamma_x(P^n - T^nA)^{\mathrm{T}} + T^n\Gamma_e(T^n)^{\mathrm{T}}\}}{\mathrm{Tr}(\Gamma_x)}.$$

In Fig. 5 we have plotted the estimation errors for the estimators $\widehat{x}^n = x^n_{\mathrm{s}}$, $\widehat{x}^n = x^n_{\mathrm{e}}$ and $\widehat{x}^n = x^n_{\mathrm{c}}$, respectively, with different discretization levels. The corresponding operators corresponding to these models are

$$T^n = \Gamma^n_x(A^n)^{\mathrm{T}}(A^n\Gamma^n_x(A^n)^{\mathrm{T}} + \Gamma_e)^{-1} \tag{19}$$

for the simple model,

$$T^n = \Gamma^n_x(A^n)^{\mathrm{T}}(A^n\Gamma^n_x(A^n)^{\mathrm{T}} + \Gamma_e + \Gamma^n_w)^{-1}$$

for the enhanced model, and

$$T^n = (\Phi^n)^{\mathrm{T}}\Gamma_x A^{\mathrm{T}}(A\Gamma_x A^{\mathrm{T}} + \Gamma_e)^{-1}$$

for the complete model.

Before commenting the results, we include also a fourth modelling error curve that corresponds to the so called *inverse crime*. By inverse crime, it is usually meant the procedure of first simplifying the model, developing an estimator based on this model and then testing it against data produced with the same simplified model. In the present case, a version of such a procedure would be to write the simple model

$$y = A^nx^n + e, \quad x^n \sim \mathcal{N}(0, \Gamma^n), \quad e \sim \mathcal{N}(0, \Gamma_e),$$

and use (19) as a linear estimator. Believing that the above model is the whole truth of y, the mean square error in this case would be

$$\mathbf{E}\{\|T^n y - x^n\|^2\} = \mathrm{Tr}\{(I - T^n A^n)\Gamma^n (I - T^n A^n)^{\mathrm{T}} + T^n \Gamma_e (T^n)^{\mathrm{T}}\}.$$

The relative mean square error is plotted also in Fig. 5.

The figure shows that at all levels of discretization,

$$\mathrm{RMSE}(x_{\mathrm{c}}^n) \leqslant \mathrm{RMSE}(x_{\mathrm{e}}^n) \leqslant \mathrm{RMSE}(x_{\mathrm{s}}^n),$$

the complete model being the only one that is consistently keeping the discretization error under control. The inverse crime curve shows how badly the expectations concerning the performance can be wrong compared to the reality. Finally, notice the rather peculiar feature that the relative errors start to decrease at $n = 5$. The reason for this is that when the length of the discretization intervals grow approaching the correlation length, the reduced models start to loose their capacity to capture the fine details of the signal (see Fig. 1), and the projected vector x^n starts to average out the details. Hence, as n decreases, we are estimating a vector with very small variations, and the relative mean square errors of the enchanced error model and the complete model start to decrease.

6. Discussion

This article discusses the effect of discretization errors in inverse problems by using the statistical approach. With a simple one-dimensional example, it is demonstrated that the statistical method gives tools to control the errors even when the discretization of the problem is rather coarse. The discretization error control is crucial when solving large-scale inverse problems in which the computational cost without a significant model reduction is too high.

There are several extensions of this work that deserve a comment. First, we discussed only linear models with additive noise, and moreover, the Gaussian prior and likelihood densities. When these assumptions are not valid, the discretization error is no longer Gaussian and consequently, it is in general not enough to characterize only the mean and the covariance. However, a Gaussian approximation of the modelling error is often useful. In the case of non-linear inverse problems, the estimation of the mean and the covariance require Monte Carlo techniques. Preliminary results concerning the discretization of the non-linear inverse problem of electrical impedance tomography have been published in the book [3], and further results will be presented elsewhere. For a related work concerning discretization of non-Gaussian priors, see the article [5], where the authors develop the theory for total variation priors. The article discusses also the non-trivial question that was not addressed in the present article: If we have a hierarchy of finite-dimensional prior densities, can we assume that they converge in some sense towards an infinite-dimensional density? The existence of underlying infinite-dimensional densities have been discussed in the dissertations [4,9]. Furthermore, the problem of designing efficient solvers for the various error models for large problems will be discussed elsewhere.

References

[1] D. Calvetti, J. Kaipio, E. Somersalo, Aristotelian prior boundary conditions, Int. J. Math. Comp. Sci. 1 (2006) in press.
[2] I.M. Gelfand, N.Y. Vilenkin, Generalized Functions, vol. 4, Academic Press, New York, 1964.
[3] J. Kaipio, E. Somersalo, Statistical and Computational Inverse Problems, Applied Mathematical Series, vol. 160, Springer, Berlin, 2004.
[4] S. Lasanen, Discretizations of generalized random variables with applications to inverse problems, Ann. Acad. Sci. Fenn. Dissertationes 2002.
[5] M. Lassas, S. Siltanen, Can one use total variation prior for edge-preserving Bayesian inversion?, Inverse Problems 20 (2004) 1537–1563.
[6] M. Lehtinen, L. Päivärinta, E. Somersalo, Linear inverse problems for generalized random variables, Inverse Problems 5 (1989) 599–612.
[7] A. Mandelbaum, Linear estimators and measurable linear transformations on a Hilbert space, Z. Wahrsch. Verw. Gebiete 65 (1884) 385–397.
[8] T. Nguen, G. Strang, Wavelets and Filter Banks, Wellesley, Cambridge, 1996.
[9] H. Pikkarainen, A mathematical model for electrical impedance process tomography, Doctoral dissertation, Helsinki University of Technology, Espoo, Finland, 2005. ISBN 951-22-7651-8.
[10] Ju.A. Rozanov, Infinite-dimensional Gaussian Distributions, Proceedings of the Steklov Institute of Math 108 (1968) (English translation: AMS 1971).