Numerical tools for analysis and solution of Fredholm integral equations of the first kind

1992 Inverse Problems 8 849

(http://iopscience.iop.org/0266-5611/8/6/005)

Inverse Problems 8 (1992) 849–872. Printed in the UK

Numerical tools for analysis and solution of Fredholm integral equations of the first kind

Per Christian Hansen†

UNI•C, Danish Computing Center for Research and Education, Building 305, Technical University of Denmark, DK-2800 Lyngby, Denmark

Received 17 September 1991, in final form 7 January 1992

Abstract. We survey several numerical tools that can be used for the analysis and solution of systems of linear algebraic equations derived from Fredholm integral equations of the first kind. These tools are based on the singular value decomposition (SVD) and the generalized SVD, and they allow the user to study many details of the integral equation. The tools also aids the user in choosing a good regularization parameter that balances the influence of regularization and perturbation errors.

1. Introduction

Fredholm integral equations of the first kind play an important role in many problems from science and engineering. Some examples are antenna design [42], astrometry [18], computerized tomography [53], early vision [9], image restoration [4], inverse geo- and helioseismology [57, 54], and mathematical biology [47]. Other examples of applications can be found in [6, 30, 65]. A Fredholm integral equation of the first kind, in one dimension, has the following generic form

$$\int_a^b K(s,t)\, f(t)\, \mathrm{d}t = g(s) \qquad c \leqslant s \leqslant d \tag{1}$$

where the functions K (the *kernel*) and g (the *right-hand side*) are known functions, at least in principle, while f is the unknown, sought function. In many—but not all—practical applications of (1) the kernel K is given exactly by the underlying mathematical model, while the right-hand side g typically consists of measured quantities, i.e. g is only known with a certain accuracy and only in a finite set of points $s_1, \ldots, s_m$.

Fredholm integral equations of the first kind are inherently ill-posed problems [46, section 15.1], i.e. the solution is extremely sensitive to arbitrarily small perturbations of the system. This means that all the classical numerical methods, such as LU and Cholesky factorization, fail to compute a meaningful solution once (1) has been discretized. On the other hand, this does not mean that a solution cannot be computed. The development of stable and reliable numerical methods particularly suited for the solution of (1) has therefore always been a challenge

† E-mail: unipch@wuli.uni-c.dk.

0266-5611/92/060849+24$07.50

[5, 20, 21]. Previously, the rather slow computational speed of many computers put a severe restriction on the computational complexity and sophistication of the practical numerical methods available for solving (1). However, modern computers and workstations have much faster numerical performance, and thus allow the present-day user to apply much more advanced numerical tools in order to analyse and solve the integral equations.

In the light of this new situation, the purpose of this paper is to give an overview of some of the author's recent results related to such advanced numerical tools. In particular, we demonstrate how the singular value decomposition and the generalized singular value decomposition are used to achieve important insight into the ill-posed problem (1), and thus aid the computation of a meaningful solution to the integral equation. Another important tool that we want to advocate is the so-called L-curve which is a plot of the norm or seminorm of the solution versus the residual norm. The L-curve is a simple means of showing the influence of regularization, and it aids the user in choosing a good regularization parameter.

We emphasize that the main topic of this paper is *numerical tools* for treating ill-posed problems, in the sense that we assume that the problem has already been discretized and that we are faced with a matrix problem $\mathbf{A}x = b$ or $\min \|\mathbf{A}x - b\|_2$. We shall not deal with the multitute of discretization methods, but instead refer to the surveys in [5, 20, 21]. Often, at least one of the dimensions of the matrix $\mathbf{A}$ is fixed by the particular problem—for example when the right-hand side b consists of a fixed set of measurements from an experiment which cannot easily be repeated. The main purpose of the numerical process is then to 'squeeze out' as much information as possible from the given problem.

We begin our discussion in section 2 with a few theoretical results on Fredholm integral equations on the first kind, because we feel that this approach gives the best understanding of the difficulties associated with the ill-posed problem. However, in section 3 we switch to the numerical approach which we maintain throughout the rest of the paper. Section 3 shows how the singular value decomposition is used to analyse the matrix derived from discretization of (1). In section 4 we summarize various regularization techniques for solving (1) numerically, and in section 5 we demonstrate how the generalized singular value decomposition provides a means for analysing these regularized problems. Section 6 introduces the L-curve—a plot of the solution norm versus the residual norm—for regularized problems, and we show how this curve yields important information about the problem. In section 7 we point to available software related to the above-mentioned tools and also point out some shortcuts that can reduce the computational demands. Finally, we illustrate the theory with numerical examples in section 8.

2. Inherent difficulties of ill-posed problems

As we have already mentioned above, the integral equation (1) may be extremely difficult to solve because it is an *ill-posed problem*. Such problems are characterized by the fact that arbitrarily small perturbations of the right-hand side g may lead to arbitrarily large perturbations of the solution f. In other words, the solution is extremely sensitive to perturbations.

These difficulties are inseparably connected with the compactness of the operator which is associated with the kernel K [46, theorem 15.4]. In physical terms, the

integration with K in (1) has a 'smoothing' effect on f in the sense that high-frequency components, cusps, and edges in f are 'smoothed out' by the integration. We can therefore expect that the reverse process, i.e. that of computing f from g, will tend to amplify any high-frequency components in g. As we shall illustrate below, this is indeed the case.

2.1. The singular value expansion

The superior analytical tool for analysis of Fredholm integral equations of the first kind is the *singular value expansion* (SVE) of K. By means of the SVE, any square integrable kernel K can be written as the following infinite sum†

$$K(s,t) = \sum_{i=1}^{\infty} \mu_i \, u_i(s) \, v_i(t) \tag{2}$$

(for degenerate kernels, the ∞ should be replaced by the rank of the kernel). The functions u_i and v_i are termed the *singular functions* of K. They are orthonormal with respect to the usual inner product, i.e. $(u_i, u_j) = (v_i, v_j) = \delta_{ij}$, where $(\ ,\ )$ is defined by

$$(\phi, \psi) \equiv \int \phi(t) \, \psi(t) \, \mathrm{d}t. \tag{3}$$

The quantities μ_i are the *singular values* of K; they are non-negative and they can always be ordered in non-increasing order such that

$$\mu_1 \geqslant \mu_2 \geqslant \mu_3 \ldots \geqslant 0.$$

The triplets $\{\mu_i, u_i, v_i\}$ are related to two eigenvalue problems associated with K as follows: $\{\mu_i^2, u_i\}$ are the eigensolutions of the symmetric kernel $\int_a^b K(s,x) \, K(t,x) \, \mathrm{d}x$, while $\{\mu_i^2, v_i\}$ are the eigensolutions of $\int_c^d K(x,s) \, K(x,t) \, \mathrm{d}x$. This illustrates that the triplets $\{\mu_i, u_i, v_i\}$ are characteristic and essentially unique for the given kernel K. For more details, see [7], [46, section 15.4] and [63, section 8].

Perhaps the most important relation between the singular values and functions is the following relation

$$\int_a^b K(s,t) \, v_i(t) \, \mathrm{d}t = \mu_i \, u_i(s) \qquad i = 1, 2, \ldots \tag{4}$$

which shows that any singular vector v_i is mapped onto the corresponding u_i, and that the singular value μ_i is the amplification of this particular mapping. If this relation, together with the SVE (2), is inserted into the integral equation (1), then one obtains the equation‡

$$\sum_{i=1}^{\infty} \mu_i \, (v_i, f) \, u_i(s) = \sum_{i=1}^{\infty} (u_i, g) \, u_i(s) \tag{5}$$

† The equality signs in (2) and (4) hold 'almost everywhere'.

‡ Equations (5) and (6) hold with relatively uniformly absolute convergence [63, section 2.4 and theorem 8.3.2] which implies mean convergence [63, p 55]. Uniform convergence holds if K is continuous [63, p 147].

which, in turn, leads to the following expression for the solution to (1)

$$f(t) = \sum_{i=1}^{\infty} \frac{(u_i, g)}{\mu_i} v_i(t). \tag{6}$$

We stress that f only exists if the right-hand side of (6) indeed converges, which is equivalent to requiring that g is in $\mathcal{N}(K^*)^{\perp}$, the orthogonal complement of the null space of the adjoint of K. It can be shown [29, section 1.3] that if the right-hand side g lies outside $\mathcal{R}(K)$, but still belongs to $\mathcal{R}(K) \oplus \mathcal{R}(K)^{\perp}$ (where $\mathcal{R}(K)$ is the range of K), then the function f given by (6) is a least squares solution to (1). From (6) we see that f is expressed in terms of the singular functions v_i and the corresponding expansion coefficients $(u_i, g)/\mu_i$. One can therefore completely characterize the solution f by an analysis of the coefficients $(u_i, g)/\mu_i$ and the functions v_i.

2.2. *The smoothing property of K*

The overall behaviour of the singular values μ_i and the singular functions u_i and v_i is by no means 'arbitrary'; their behaviour is strongly connected with the properties of the kernel K. The following holds:

- The 'smoother' the kernel K, the faster the singular values μ_i decay to zero (where 'smoothness' is measured by the number of continuous partial derivatives of K) [19].
- The smaller the μ_i, the more oscillations (or zero-crossings) in the singular functions u_i and v_i. This is a consequence of the fact that a Fourier expansion of K must have such oscillation properties [40].

The practical implication of these properties of the triplets $\{\mu_i, u_i, v_i\}$ is that the expression (6) for f can be regarded as a spectral expansion in which the coefficients $(u_i, g)/\mu_i$ describe the spectral properties of the solution f. We see from equation (5) that the integration with K indeed has a smoothing effect: the higher the spectral components in f, the more they are damped in g by the multiplication with μ_i. Moreover, equation (6) shows that the inverse problem, namely that of computing f from g, indeed has the 'reverse' effect on the oscillations in g, namely an amplification of the spectral components (u_i, g) with a factor μ_i^{-1}. This, of course, amplifies the high-frequency components.

2.3. *The Picard condition*

With this behaviour in mind, it is obvious that not every right-hand side g will lead to a 'smooth' square integrable solution f due to the amplification factors μ_i^{-1}. In effect, the right-hand side g must be somewhat 'smoother' than the desired solution f, in order that the right-hand side in (6) actually converges to f. The following *Picard condition* on the right-hand side g is therefore essential (see [29, section 1.2] and [46, theorem 15.18] for more details). In order that there exists a square integrable solution f to the integral equation (1), the right-hand side g must satisfy

$$\sum_{i=1}^{\infty} \left(\frac{(u_i, g)}{\mu_i} \right)^2 < \infty. \tag{7}$$

The Picard condition says that from some point in the summation in (6), the absolute value of the coefficients (u_i, g) must decay faster than the corresponding singular

values μ_i in order that a square integrable solution f exists. Since this condition is so essential in connection with Fredholm integral equations of the first kind, it should—whenever possible—be checked before one tries to solve the integral equation. We return to the numerical aspects of such an investigation in the next section.

The decay of the singular values μ_i is so fundamental for the behaviour of ill-posed problems that it makes sense to use this decay to characterize the degree of ill-posedness of the problem. This was first mentioned by Wahba [69]. Hofmann [44, definition 2.42] has proposed the following definition: if there exists a positive real number ν such that the singular values satisfy $\mu_i = \mathrm{O}(i^{-\nu})$, then ν is called the *degree of ill-posedness*, and the problem is characterized as mildly or moderately ill-posed if $\nu \leqslant 1$ or $\nu > 1$, respectively. On the other hand, if $\mu_i = \mathrm{o}(i^{-\nu})$ for all $\nu > 0$ (i.e. $\nu = \infty$), then the problem is termed severely ill-posed.

Notice that for problems with finite rank, including any discretization of (1), from a purely mathematical point of view the Picard condition is always satisfied, the solution is stable, and there is no need for regularization. However, discrete problems always suffer from some combination of measurement errors, discretization errors, and rounding errors, and the solution to the discretized problem is extremely sensitive to these errors. Hence, from a practical point of view, regularization is still required to filter out the influence of these errors in order to compute a useful solution to the discretized system. One may even introduce a discrete Picard condition for such systems. We return to these aspects in section 5.

3. Analysis of the coefficient matrix

3.1. The singular value decomposition

The superior tool for analysis of the coefficient matrix $\mathbf{A}$ derived from discretization of the kernel K is the *singular value decomposition* (SVD), which is a discrete analogue of the SVE. The SVD of an $m \times n$ matrix $\mathbf{A}$ has the form

$$\mathbf{A} = \sum_{i=1}^{\min(m,n)} \sigma_i \, \boldsymbol{u}_i \, \boldsymbol{v}_i^{\mathrm{T}}, \tag{8}$$

where the singular vectors $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ are orthonormal, i.e. $\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_j = \boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{v}_j = \delta_{ij}$, and where σ_i, the singular values of $\mathbf{A}$, satisfy

$$\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_{\min(m,n)} \geqslant 0.$$

In analogy with the SVE, the pairs $\{\sigma_i^2, \boldsymbol{u}_i\}$ and $\{\sigma_i^2, \boldsymbol{v}_i\}$ are the eigensolutions of the positive semidefinite matrices $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ and $\mathbf{A}^{\mathrm{T}}\mathbf{A}$, respectively. The fundamental definition of the SVD, i.e. the analogue of equation (4) for the SVE, has the form

$$\mathbf{A}\,\boldsymbol{v}_i = \sigma_i \, \boldsymbol{u}_i \qquad i = 1, \ldots, \min(m,n) \tag{9}$$

showing that each vector $\boldsymbol{v}_i$ is mapped onto the corresponding vector $\boldsymbol{u}_i$ with σ_i as the magnification. A wealth of information and details about the SVD can be found in [10, 28].

3.2. Relations between the SVD and SVE

The basic numerical difficulty associated with solving the linear system $\mathbf{A}\,x = b$ or the least squares problem $\min \|\mathbf{A}\,x - b\|_2$ arising from discretization of (1) is that the computed solution x is very sensitive to perturbations of $\mathbf{A}$ and b. This is reflected in the fact that the condition number of $\mathbf{A}$, given by the ratio σ_1/σ_n, is very large. Moreover, the condition number of $\mathbf{A}$ increases with both the order n and the number of data points m. These numerical difficulties are intimately connected with the ill-posedness of the problem (1): the better the discretization, i.e. the better the linear algebraic system 'models' the integral equation, the more the ill-conditioning of the linear system resembles the ill-posedness of (1).

The reason behind the ill-conditioning of the linear system is the fact that the SVD of the matrix $\mathbf{A}$ is very closely related to the SVE of the kernel K. In fact, the singular values σ_i of $\mathbf{A}$ are in many cases approximations to the singular values μ_i of the kernel K, while the singular vectors u_i and v_i yield information about the singular functions of K, see e.g. [1, 17, 33, 70]. In particular, for a Galerkin expansion-method with orthonormal basis functions ψ_i and ϕ_j, it is shown in [33, theorem 2] that

$$0 \leqslant \mu_i - \sigma_i \leqslant \|K - \bar{K}_n\| \qquad i = 1, \ldots, n \tag{10}$$

where the norm of $K - \bar{K}_n$ is given by

$$\|K - \bar{K}_n\| = \left(\int_a^b \int_c^d (K(s,t) - \bar{K}_n(s,t))^2 \,\mathrm{d}s\,\mathrm{d}t \right)^{1/2}$$

and where $\bar{K}_n$ is a degenerate kernel of rank n given by

$$\bar{K}_n(s,t) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}\, \psi_i(s)\, \phi_j(t) \tag{11}$$

in which $a_{ij} = \int_a^b \int_c^d K(s,t)\, \psi_i(s)\, \phi_j(t)\,\mathrm{d}s\,\mathrm{d}t$ are the elements of the matrix $\mathbf{A}$. Similarly, if u_{ij} and v_{ij} denote the elements of the singular vectors u_j and v_j, and if $\bar{u}_j$ and $\bar{v}_j$ denote approximations to the singular functions u_j and v_j, given by

$$\bar{u}_j(s) = \sum_{i=1}^{n} u_{ij}\, \psi_i(s) \qquad \bar{v}_j(x) = \sum_{i=1}^{n} v_{ij}\, \phi_i(x) \tag{12}$$

then the approximation errors in $\bar{u}_j$ and $\bar{v}_j$ are bounded by [33, theorem 5]

$$\max_j \{\|u_j - \bar{u}_j\|, \|v_j - \bar{v}_j\|\} \leqslant \left(\frac{2\|K - \bar{K}_n\|}{\mu_j - \mu_{j+1}} \right)^{1/2}. \tag{13}$$

As a consequence of this tight relation between the SVD and the SVE, it is now obvious that for any discretization of an ill-posed problem we have that:

1. The matrix $\mathbf{A}$ is ill-conditioned, i.e. the condition number σ_1/σ_n is large.
2. The condition number of $\mathbf{A}$ increases with n.

Moreover, the singular vectors u_j and v_j have an increasing number of sign changes as j increases—and it is proved in [40] that this is in fact also true for large j where the bound in (13) becomes less tight.

3.3. SVD analysis

The conclusion from this analysis is that any classical method for solving the systems $\mathbf{A}x = b$ and $\min \|\mathbf{A}x - b\|_2$, such as Cholesky, LU or QR factorization, cannot be used to compute a numerical solution to the first kind of Fredholm integral equation (1). For one thing, the condition number of $\mathbf{A}$ is always so large that rounding errors prevent the computation of an accurate numerical solution x. Moreover, since the discrete problem is always perturbed by approximation and/or discretization errors having components along all the singular vectors, then—even if we were able to solve the algebraic system without rounding errors—we would not obtain a 'smooth' solution x because of the oscillations in the singular vectors.

There are several reasons for actually computing the SVD of the matrix $\mathbf{A}$. Wahba proposed this in 1980 in her influential report [69, section 7], where the availability of numerical software is emphasized. Once the SVD is computed, it lets the user study the 'spectral' properties of the operator K in terms of the SVE, as mentioned in section 2, by means of the singular values and vectors.

One can also use the SVD to check whether the Picard condition (7) for the integral equation *seems to be satisfied* for the underlying, unperturbed problem, by plotting the quantities σ_i, $|u_i^{\mathrm{T}} b|$, and $|u_i^{\mathrm{T}} b|/\sigma_i$. As mentioned above, the errors in the right-hand side b usually have components along all the singular vectors. Provided that the Picard condition is satisfied, the errors will therefore only dominate the coefficients $u_i^{\mathrm{T}} b$ corresponding to small singular values. Hence, if the coefficients $|u_i^{\mathrm{T}} b|$ in average decay faster than the singular values σ_i for small indices i, then it is indeed reasonable to assume that the Picard condition is satisfied. Moreover, if the coefficients $|u_i^{\mathrm{T}} b|$ level off at a plateau for increasing i, then this plateau is an estimate of the error-level in the right-hand side. In this connection, it is important to realize that the approximation errors in the computed singular functions are smallest for small i, cf. equation (13).

4. Regularization and filtering

4.1. Regularization

A reasonable way to compute a meaningful 'smooth' solution to the integral equation (1), i.e. a solution which has some useful properties in common with the exact solution to the underlying—and unknown—unperturbed problem, is to somehow filter out the high-frequency components associated with the small singular values. This approach is, of course, only useful when the required solution indeed has some inherent smoothness. Although this is not always so, e.g. in image restoration where edges and discontinuities are sometimes wanted in the solution, there is a large class of problems for which it is reasonable to seek a 'smooth' approximate solution in which high-frequency components are filtered out.

The classical way to filter out the high-frequency components associated with the small singular values is to apply regularization to the problem. The term regularization was originally associated with a certain technique proposed by Tikhonov [64]; but it is standard terminology today to classify any method that seeks to compute a 'smooth' solution as a regularization method [8]. In its original framework, regularization is applied directly to the integral equation (1). However, the regularization may as well be applied to the linear system derived from discretization of (1). This is often much

simpler to carry out in practice and, due to the tight connection between the original problem and the discrete problem (as mentioned in the previous section), the effect of the regularization on the computed 'smooth' solution is basically the same. In this presentation, we shall therefore restrict our discussion to regularization of the linear algebraic problems $\mathbf{A}\,\boldsymbol{x} = \boldsymbol{b}$ and $\min \|\mathbf{A}\,\boldsymbol{x} - \boldsymbol{b}\|_2$.

The fundamental difficulty with these systems is *instability*, in the sense that the solution $\boldsymbol{x}$ is dominated by contributions corresponding to the smallest singular values, and these contributions consist mainly of errors. The small amount of information in the right-hand side associated with all the small singular values is lost due to the errors. Hence, one can say that the systems are essentially underdetermined, because we are only able to recover information associated with the large singular values of $\mathbf{A}$. In order to compute a unique solution, one must therefore impose stability by specifying some additional information which:

1. Singles out exactly one solution.
2. Seeks to single out a solution which is, in some sense, close to the desired, but unknown, exact solution.

Given some *a priori* estimate $\boldsymbol{x}_0$ of the solution, it is natural to seek to minimize a seminorm of the difference between the computed solution $\boldsymbol{x}$ and the estimate $\boldsymbol{x}_0$:

$$\min \|\mathbf{L}\,(\boldsymbol{x} - \boldsymbol{x}_0)\|_2 \tag{14}$$

where the matrix $\mathbf{L}$ is some appropriately chosen matrix. Typically, $\mathbf{L}$ is either the identity matrix $\mathbf{I}_n$ or a discrete approximation to a derivative operator. If no particular knowledge is available about the wanted solution, then it is reasonable to use $\boldsymbol{x}_0 = \mathbf{0}$, while it is more difficult to advocate a particular matrix $\mathbf{L}$. Experiments with various $\mathbf{L}$ may be necessary.

4.2. Tikhonov's method

There are many ways to combine the side constraint (14) with the original linear algebraic problem. Perhaps the most well-known approach is that proposed by Tikhonov [64] and, independently, by Phillips [56], namely to define the regularized solution $\boldsymbol{x}_\lambda$ as the solution to the following problem

$$\min \left\{ \|\mathbf{A}\,\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \lambda^2 \|\mathbf{L}\,(\boldsymbol{x} - \boldsymbol{x}_0)\|_2^2 \right\}. \tag{15}$$

Here, the quantitiy λ is the regularization parameter which controls the weight given to minimization of the side constraint relative to the minimization of the residual norm. If the null spaces of $\mathbf{A}$ and $\mathbf{L}$ intersect trivially, then the solution $\boldsymbol{x}_\lambda$ to (15) is unique, and it is formally given by

$$\boldsymbol{x}_\lambda = (\mathbf{A}^{\mathrm{T}}\mathbf{A} + \lambda^2 \mathbf{L}^{\mathrm{T}}\mathbf{L})^{-1}(\mathbf{A}^{\mathrm{T}}\boldsymbol{b} + \lambda^2 \mathbf{L}^{\mathrm{T}}\mathbf{L}\,\boldsymbol{x}_0).$$

This formula should not be used for actual computation of $\boldsymbol{x}_\lambda$. For one thing, there is an unavoidable loss of information involved in forming the cross-product matrix $\mathbf{A}^{\mathrm{T}}\mathbf{A}$, due to the finite precision arithmetic. More important, however, is that the work involved in computing the GSVD of the pair $(\mathbf{A}, \mathbf{L})$ is not much larger than that involved in forming the matrix $\mathbf{A}^{\mathrm{T}}\mathbf{A} + \lambda^2 \mathbf{L}^{\mathrm{T}}\mathbf{L}$ and solving the corresponding system of equations. Hence, for approximately the same computational 'cost', the GSVD approach gives much more information about the problem to be solved.

Notice that the Tikhonov regularization in (15) applies to square systems ($m = n$) as well as overdetermined systems ($m > n$). Notice also that the fundamental idea in Tikhonov regularization is to introduce a trade-off between the size of the residual norm and the side constraint. Putting a large weight on the side constraint means that one must accept a larger residual, and vice versa. By choosing a suitable regularization parameter λ, one can single out a satisfactory solution for which the two constraints are balanced. We return to the choice of λ in section 6.

4.3. Truncated SVD

Another well-known regularization method is to simply truncate the SVE-based expansion (6) before the small singular values start to dominate. For the algebraic systems, the similar technique is called truncated SVD [32, 35, 67], and the associated regularized solution $\boldsymbol{x}_k$ is defined by

$$\boldsymbol{x}_k \equiv \sum_{i=1}^{k} \frac{\boldsymbol{u}_i^{\mathrm{T}} \boldsymbol{b}}{\sigma_i} \boldsymbol{v}_i. \tag{16}$$

Here, the integer k plays the role of the regularization parameter. In [37], it is shown that truncated SVD is essentially equivalent to Tikhonov regularization with $\mathbf{L} = \mathbf{I}_n$, in the sense that for any k there exists a λ such that $\|\boldsymbol{x}_\lambda - \boldsymbol{x}_k\|_2 = \mathrm{O}(|\boldsymbol{u}_k^{\mathrm{T}} \boldsymbol{b}|/\sigma_k)$, which is small when the Picard condition is satisfied. In section 5, we shall comment on the general case when $\mathbf{L} \neq \mathbf{I}_n$.

4.4. Iterative methods

There is also a whole class of regularization methods associated with iterative methods for solving the linear systems $\mathbf{A}\,\boldsymbol{x} = \boldsymbol{b}$ and $\min \|\mathbf{A}\,\boldsymbol{x} - \boldsymbol{b}\|_2$, of which the most promising are semi-iterative methods and the method of conjugate gradients [12, 31]. For example, if there is a gap between $|\boldsymbol{u}_k^{\mathrm{T}} \boldsymbol{b}|$ and $|\boldsymbol{u}_{k+1}^{\mathrm{T}} \boldsymbol{b}|$, and if one applies k steps of the conjugate gradient algorithm [28, section 10.2] to the system

$$\mathbf{A}^{\mathrm{T}} \mathbf{A}\,\boldsymbol{x} = \mathbf{A}^{\mathrm{T}} \boldsymbol{b}$$

(without actually forming $\mathbf{A}^{\mathrm{T}} \mathbf{A}$), then it is shown in [40] that the such computed solution $\boldsymbol{x}^{(k)}$ is a regularized solution and that $\boldsymbol{x}^{(k)}$ is close to the truncated SVD solution $\boldsymbol{x}_k$ because $\|\boldsymbol{x}^{(k)} - \boldsymbol{x}_k\|_2 = \mathrm{O}(|\boldsymbol{u}_{k+1}^{\mathrm{T}} \boldsymbol{b}/\boldsymbol{u}_k^{\mathrm{T}} \boldsymbol{b}|)$. Truncated SVD, on the other hand, is equivalent to Tikhonov regularization (cf. section 4.3), and termination of the conjugate gradient method is therefore equivalent to Tikhonov regularization with $\mathbf{L} = \mathbf{I}_n$. There are also ways of applying the conjugate gradient algorithm to problems with $\mathbf{L} \neq \mathbf{I}_n$ [11, 31].

5. Analysis of regularized problems

5.1. The generalized SVD

In analogy with the SVE and the SVD, which are the superior tools for analysis of the integral equation and the discretized system, there is a superior tool for analysis of discrete regularization problems with general matrices $\mathbf{L}$. This tool is the *generalized*

singular value decomposition† (GSVD) of the matrix pair $(\mathbf{A}, \mathbf{L})$, originally introduced by Van Loan [66]. For our purpose, it is sufficient to consider the GSVD of $\mathbf{A}$ and $\mathbf{L}$ for the following case where $\mathbf{L}$ has full rank and the dimensions satisfy

$$\mathbf{A} \in \mathbb{R}^{m \times n} \qquad \mathbf{L} \in \mathbb{R}^{p \times n} \qquad m \geqslant n \geqslant p.$$

A more general formulation for all m, n and p makes the notation somewhat more complicated without casting more light on our problem. Define the three matrices

$$\bar{\mathbf{U}} \equiv (\bar{u}_1, \ldots, \bar{u}_m) \in \mathbb{R}^{m \times m}$$

$$\bar{\mathbf{V}} \equiv (\bar{v}_1, \ldots, \bar{v}_p) \in \mathbb{R}^{p \times p}$$

$$\mathbf{W} \equiv (w_1, \ldots, w_n) \in \mathbb{R}^{n \times n}$$

where the vectors $\bar{u}_i$ and $\bar{v}_i$ are orthonormal (i.e. $\bar{\mathbf{U}}^{\mathrm{T}}\bar{\mathbf{U}} = \mathbf{I}_m$ and $\bar{\mathbf{V}}^{\mathrm{T}}\bar{\mathbf{V}} = \mathbf{I}_p$), while the vectors w_i are linearly independent (i.e. $\mathbf{W}$ is non-singular). Then the GSVD of $(\mathbf{A}, \mathbf{L})$ has the form

$$\mathbf{A} = \bar{\mathbf{U}} \begin{pmatrix} \operatorname{diag}(\alpha_i) & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n-p} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{W}^{-1} \qquad \mathbf{L} = \bar{\mathbf{V}} \, (\operatorname{diag}(\beta_i) \quad \mathbf{0}) \, \mathbf{W}^{-1}. \tag{17}$$

Here, $\operatorname{diag}(\alpha_i)$ and $\operatorname{diag}(\beta_i)$ denote diagonal matrices with non-negative diagonal entries $\alpha_1, \ldots, \alpha_p$ and $\beta_1, \ldots, \beta_p$, respectively, whose elements satisfy $\alpha_i^2 + \beta_i^2 = 1$, $i = 1, \ldots, p$. The generalized singular values are defined as the ratios $\gamma_i = \alpha_i / \beta_i$, and they are ordered in non-decreasing order, i.e.

$$\gamma_1 \geqslant \gamma_2 \geqslant \ldots \geqslant \gamma_p \geqslant 0.$$

If, in particular, $\mathbf{L} = \mathbf{I}_n$ then $\bar{u}_i = u_i$, $\bar{v}_i = v_i$, and $\gamma_i = \sigma_i$ for $i = 1, \ldots, n$, and (17) becomes the ordinary SVD of $\mathbf{A}$. There are other related formulations of the GSVD; the one used here is one that is appropriate for our particular purpose, namely to analyse the above-mentioned discrete regularization schemes.

5.2. *Important GSVD relations*

Equipped with the GSVD of $(\mathbf{A}, \mathbf{L})$ it is straightforward to analyse regularized solutions in the same way as we use the SVD to analyse the unregularized solutions. To demonstrate this for general non-zero estimates x_0, it is convenient to define the n-vector

$$\omega = (\omega_1, \ldots, \omega_n)^{\mathrm{T}} \equiv \mathbf{W}^{-1} x_0.$$

Then it can be shown that for any *linear* regularization method there exists a set of associated *filter factors* f_i, $i = 1, \ldots, p$ such that the regularized solution x_{reg} can be written as

$$x_{\mathrm{reg}} = \sum_{i=1}^{p} \left(f_i \frac{\bar{u}_i^{\mathrm{T}} b}{\alpha_i} + (1 - f_i)\, \omega_i \right) w_i + \sum_{i=p+1}^{n} \bar{u}_i^{\mathrm{T}} b \, w_i. \tag{18}$$

† A similar definition of a generalized SVE for general operators K and $\mathcal{L}$ was introduced by Hanke [31].

Using this equation together with (17) it is straightforward to show that the seminorm of $\|\mathbf{L}(\boldsymbol{x}_{\text{reg}} - \boldsymbol{x}_0)\|_2$, i.e. the side constraint, is given by

$$\|\mathbf{L}(\boldsymbol{x}_{\text{reg}} - \boldsymbol{x}_0)\|_2 = \left(\sum_{i=1}^{p} f_i^2 \left(\frac{\bar{\boldsymbol{u}}_i^{\mathrm{T}} \boldsymbol{b}}{\gamma_i} - \beta_i \omega_i \right)^2 \right)^{1/2}. \tag{19}$$

Similarly, the residual norm is given in terms of the GSVD as

$$\|\mathbf{A}\boldsymbol{x}_{\text{reg}} - \boldsymbol{b}\|_2 = \left(\sum_{i=1}^{p} (1 - f_i)^2 \left(\bar{\boldsymbol{u}}_i^{\mathrm{T}} \boldsymbol{b} - \alpha_i \omega_i \right)^2 + \sum_{i=n+1}^{m} (\bar{\boldsymbol{u}}_i^{\mathrm{T}} \boldsymbol{b})^2 \right)^{1/2} \tag{20}$$

The filter factors f_i are functions of the regularization parameter associated with the particular regularization method. For many regularization methods, there are fairly simple expressions for the filter factors. For example, for Tikhonov regularization we have

$$f_i = \frac{\gamma_i^2}{\gamma_i^2 + \lambda^2} \tag{21}$$

while for the conjugate gradient method, stopped after q iterations, the filter factors are

$$f_i = \sigma_i^2 \mathcal{P}_q(\sigma_i^2) \tag{22}$$

where $\mathcal{P}_q$ denotes the so-called Ritz polynomial of order q associated with the conjugate gradient algorithm [43, chapter 4]. Regarding the truncated SVD, it is obvious that the corresponding filter factors are simply 0 and 1. The generalization of truncated SVD to general regularization matrices $\mathbf{L} \neq \mathbf{I}_n$ is truncated GSVD, again with filter factors 0 and 1 [34]. Examples of filter factors for other regularization methods can be found in, e.g., [24, section 4].

Notice that if $p < n$ then there are $n - p$ components of the regularized solution $\boldsymbol{x}_{\text{reg}}$ which are not influenced by the filter factors f_i and, therefore, are independent of the regularization parameter. These components are directed along the vectors $\boldsymbol{w}_{p+1}, \ldots, \boldsymbol{w}_n$ which are the null vectors of the matrix $\mathbf{L}$, i.e.

$$\mathbf{L}\boldsymbol{w}_i = \mathbf{0} \qquad i = p+1, \ldots, n.$$

For all reasonable choices of the regularization matrix $\mathbf{L}$, these null vectors are 'smooth', such that the above-mentioned components of $\boldsymbol{x}_{\text{reg}}$ are also 'smooth' and therefore need no regularization.

5.3. GSVD analysis

We now see that, in analogy with the SVD analysis mentioned previously, we can use the GSVD of $(\mathbf{A}, \mathbf{L})$ to obtain insight into the regularization properties of the particular regularization methods that we are using. For example, examination of the columns $\boldsymbol{w}_i$ of the matrix $\mathbf{W}$ and the corresponding filter factors f_i will tell us about the spectral properties of the regularization process. Typically, we will see that large filter factors are associated with 'smooth' vectors $\boldsymbol{w}_i$ while small filter factors, which

introduce damping into the regularized solution, are associated with vectors w_i with a large amount of high-frequency components. When this is the case, we actually compute a 'smooth' regularized solution.

The concept of a *discrete Picard condition* for discrete regularization problems was suggested in [67] and further elaborated in [36]. This condition says that the coefficients $|\bar{u}_i^T b|$ must decay faster in average than the generalized singular values γ_i in order to ensure:

1. That we are able to compute a 'smooth' regularized solution.
2. That this regularized solution is a reasonable approximation to the exact solution to the unperturbed linear system.

Inspection of the behaviour of the generalized singular values γ_i and the coefficients $|\bar{u}_i^T b|$ therefore yields further insight into the given problem. In particular, we can check whether the coefficients $|\bar{u}_i^T b|$ in average decay faster than the generalized singular values γ_i for small indices i where the errors in b do not dominate $\bar{u}_i^T b$. If this is indeed the case, then we say that the discrete Picard condition is satisfied. For more details, cf [36].

Inspection of the GSVD coefficients also gives a valuable aid in choosing a good regularization parameter. We recall that since we want the filter factors to dampen only those contributions to the regularized solution corresponding to coefficients dominated by the errors, we must choose the regularization parameter such that the filter factors dampen only the desired coefficients. Inspection of the GSVD quantities γ_i, $|\bar{u}_i^T b|$, and $|\bar{u}_i^T b|/\gamma_i$ together with the filter factors f_i will therefore give a good hint to the optimal choice of the regularization parameter.

6. The L-curve

One of the most important problems in connection with the numerical treatment of linear systems derived from ill-posed problems is the choice of the regularization parameter. As mentioned in the introduction, our goal is to 'squeeze out' as much information as possible from the given right-hand side b. Too much regularization leaves out information actually available in b while too little regularization produces a solution dominated by errors. Hence, one should ideally find the regularization parameter that balances the regularization error (i.e. errors introduced by smoothing the data) and the perturbation error from the errors in b. We shall call such a regularization parameter *optimal* (and note that this definition differs from optimal regularization parameters that provide optimal convergence rates as the errors in b tend to zero [24]). The optimal regularization parameter in our sense is closely related to the 'effective rank' defined in [26] as $\sum_{i=1}^{n} f_i$ which measures how much reliable information can be extracted from the system; e.g. for truncated SVD the 'effective rank' is simply the number of non-zero coefficients in the expansion (16). A useful tool for analysis of these aspects is the so-called L-curve [37, 39].

6.1. Definition of the L-curve

The L-curve is a plot of the side constraint $\|\mathbf{L}(x_{\mathrm{reg}} - x_0)\|_2$ versus the residual norm $\|\mathbf{A}x_{\mathrm{reg}} - b\|_2$ for a particular regularization method. Hence, the L-curve is in fact a parametrized curve whose parameter is the regularization parameter, e.g. λ in case of Tikhonov regularization. The name 'L-curve' comes from the fact that for many

problems the curve $(\|\mathbf{A}\,\boldsymbol{x}_{\mathrm{reg}} - \boldsymbol{b}\|_2, \|\mathbf{L}\,(\boldsymbol{x}_{\mathrm{reg}} - \boldsymbol{x}_0)\|_2)$ has an L-shaped 'corner'. The L-curve is such a useful tool in connection with ill-posed problems partly because it provides us with a convenient way to display the inter-relationship between the seminorm $\|\mathbf{L}\,(\boldsymbol{x}_{\mathrm{reg}} - \boldsymbol{x}_0)\|_2$ and the residual norm $\|\mathbf{A}\,\boldsymbol{x}_{\mathrm{reg}} - \boldsymbol{b}\|_2$, and partly because the L-shaped 'corner' corresponds to a regularization parameter which is often near optimal in the above sense.

The use of the L-curve goes back to Miller [49] and Lawson and Hanson [48]; but it has never been widely used as an analysis tool for regularization problems. Nevertheless, the L-curve is a practical means for displaying a lot of information about the regularization problem in a compact form. First of all, the L-curve immediately illustrates the trade-off between minimization of the side constraint $\|\mathbf{L}\,(\boldsymbol{x}_{\mathrm{reg}} - \boldsymbol{x}_0)\|_2$ and minimization of the residual norm $\|\mathbf{A}\,\boldsymbol{x}_{\mathrm{reg}} - \boldsymbol{b}\|_2$. Moreover, the appearance of a 'corner' on the curve shows that there indeed exists a particular regularization parameter which is in some sense optimal because it balances the two minimization goals. This information is easily displayed in the L-curve, but more difficult to display in other plots.

6.2. The 'corner' of the L-curve

To see why a 'corner' on the L-curve exists at all, and why it is associated with a near-optimal regularization parameter, it is reasonable to analyse the appearance of the L-curve in more details. The same idea is illustrated in [58, section 12.6] in connection with optimal (Wiener) filtering.

We shall here summarize the analysis from [37, 39] and refer to the original papers for more details. It is convenient to consider the following model, where the matrices $\mathbf{A}$ and $\mathbf{L}$ are given exactly and where the right-hand side $\boldsymbol{b}$ consists of an unperturbed vector $\hat{\boldsymbol{b}}$ plus a random perturbation $\boldsymbol{e}$ (consisting of measurement errors, approximation errors, etc.). If we assume that:

1. the discrete Picard condition is satisfied,
2. the norm of $\boldsymbol{e}$ satisfies $\|\boldsymbol{e}\|_2 < \|\hat{\boldsymbol{b}}\|_2$,

then the L-curve will have an L-shaped 'corner' as shown in figure 1. There are two distinct parts of the L-curve: a flat part to the right of the 'corner', and a steep part above the 'corner'.

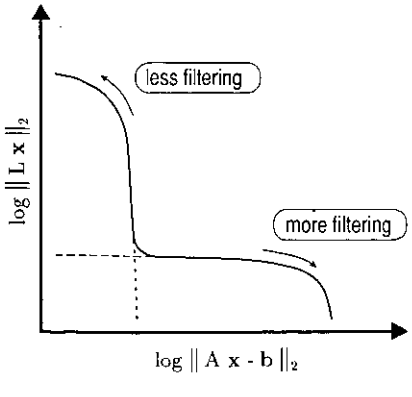

Figure 1. A generic plot of the L-curve for Tikhonov regularization on a log–log scale. The broken curve shows the L-curve for an unperturbed problem, while the dotted curve shows the L-curve for a right-hand side consisting of mere errors.

The flat part of the L-curve to the right of the 'corner' corresponds to regularized solutions where the regularization error dominates, i.e. where the damping is so large

that it successfully dampens most of the influence from the errors e and, to some extent, also filters out information from the underlying unperturbed right-hand side $\hat{b}$. The further to the right on the curve, the more damping is introduced, and the curve therefore eventually bends down until the side constraint is perfectly satisfied: $\|\mathbf{L}(x_{\mathrm{reg}} - x_0)\|_2 = 0$. Assumption 1 above ensures that there exists a reasonably 'smooth' solution to the unperturbed problem, i.e. that the seminorm $\|\mathbf{L}(x_{\mathrm{reg}} - x_0)\|_2$ stays small, such that the L-curve associated with the unperturbed problem (the broken curve in figure 1) has a flat left part as $\lambda \to \infty$.

The steep part of the L-curve above the 'corner' correspond to regularized solutions where contributions from the errors e dominate, because too little damping is introduced. From equation (19) we see that as less filtering is introduced, the seminorm $\|\mathbf{L}(x_{\mathrm{reg}} - x_0)\|_2$ increases rapidly due to the division by the small generalized singular values γ_i. When still less filtering is introduced, the curve eventually levels off at a plateau when essentially all the influence from the errors has been extracted (this plateau would not arise in infinite-dimensional problems). Assumption 2 above ensures that the L-curve associated merely with the errors e (the dotted curve in figure 1) bends down somewhat to the left of the bend in the L-curve for the unperturbed problem, and intersects the abscissa axis at $\|e\|_2$.

The L-curve associated with a real problem, in which $b = \hat{b} + e$, is of course a combination of the two 'ideal' L-curves associated with $\hat{b}$ and e. Since one of the two contributions to x_{reg} will dominate for most regularization parameters, the L-curve will appear as the solid curve in figure 1, with a steep part where perturbation errors dominate x_{reg} and a flat part where regularization errors dominate x_{reg}. Moreover, there is a small transition region exactly at the 'corner' where both contributions are approximately equal in magnitude. It is proved in [39] that the 'corner' is particularly pronounced for a log–log scale.

6.3. The choice of regularization parameter

From this brief analysis of the L-curve it is clear that the optimal choice of the regularization parameter is one that corresponds to a point on the L-curve near the 'corner', because this point represents a solution with a favorable balance between the two types of errors. If the regularization parameter is smaller than the optimal one then the solution will be influenced too much by the contributions from the errors e, while loss of information occurs if too much filtering is introduced by choosing a regularization parameter much larger than the optimal one.

It is interesting to notice that other methods for choosing an optimal value of the regularization parameter, based on completely different criteria than the L-curve, often lead to a regularization parameter which is close to the regularization parameter chosen from the L-curve. This is, for example, the case for the discrepancy principle [29, section 3.3] and [51, section 10], the quasi-optimality criterion [51, section 27], and generalized cross-validation [27], cf the analysis in [37].

The discrepancy principle requires that the user supplies an upper bound for the norm of the errors, $\epsilon_e \geqslant \|e\|_2$, and then one chooses the regularization parameter such that the residual norm satisfies

$$\|\mathbf{A}\, x_{\mathrm{reg}} - b\|_2 = \epsilon_e .$$

In terms of the L-curve, this simply corresponds to finding the intersection of a vertical line at ϵ_e and the L-curve. If the upper bound ϵ_e is a good estimate of the

norm of the errors then this intersection will be close to the 'corner' of the L-curve, although somewhat to the right with a corresponding danger of oversmoothing (i.e. introducing too much damping).

A common feature of the quasi-optimality criterion and generalized cross-validation is that they do not require any additional information, because they implicitly seek to estimate the error norm $\|e\|_2$ from the data. The underlying principle in cross validation is that if an arbitrary observation is left out and then predicted using the remaining $m - 1$ observations, then the optimal regularization parameter minimizes the sum of squares of these prediction errors. Generalized cross-validation is based on the same principle and, in addition, ensures that the regularization parameter found has some desirable invariance properties, such as being invariant to an orthogonal transformation (which includes permutations) of the data. For Tikhonov regularization, this leads to choosing the regularization parameter as the minimizer of the following function

$$\mathcal{G}(\lambda) \equiv \frac{\|\mathbf{A}\,\boldsymbol{x}_\lambda - b\|_2^2}{\left(\operatorname{trace}\left(\mathbf{I}_m - \mathbf{A}\left(\mathbf{A}^T\mathbf{A} + \lambda^2\mathbf{L}^T\mathbf{L}\right)^{-1}\mathbf{A}^T\right)\right)^2}. \tag{23}$$

The function $\mathcal{G}(\lambda)$ ideally has a minimum where the solution $\boldsymbol{x}_\lambda$ changes from being dominated by regularization errors to be dominated by perturbation errors. Essentially the same is true for the quasi-optimality function

$$\mathcal{Q}(\lambda) \equiv \left\| \lambda \frac{\mathrm{d}\boldsymbol{x}_\lambda}{\mathrm{d}\lambda} \right\|_2 \tag{24}$$

where $\mathrm{d}\boldsymbol{x}_\lambda/\mathrm{d}\lambda$ is the derivative of $\boldsymbol{x}_\lambda$ with respect to λ. From the previous discussion of the L-curve and, in particular, its 'corner' it is clear that generalized cross-validation as well as the quasi-optimality criterion lead to solutions close to this 'corner'.

The L-curve has recently gained more interest as a method in its own right for choosing the regularization parameter. The idea is to select that regularization parameter which corresponds exactly to the corner of the L-curve. Here, we define the corner as the point on the curve which has maximum curvature [39]. If the L-curve is discrete (e.g. for truncated SVD and iterative methods) then we fit a 2D cubic spline curve to the points and compute the curvature of this spline curve. The corner can then be found by a one-dimensional optimization routine such as FMIN from [25, section 8] or BRENT from [58, section 10.2]. This algorithm is at least as robust as the above-mentioned methods and is in fact superior to them in the case of highly correlated errors in the right-hand side. See [37, 39, 62] for more details as well as numerical examples. The L-curve criterion still lacks a thorough convergence analysis like the one carried out in [68] for cross validation.

7. Software and shortcuts

The numerical tools that we have mentioned in the previous sections, in particular the SVD and the GSVD, are of little practical use unless software is easily available for computation of these decompositions. Fortunately, the new Fortran linear algebra library LAPACK [2, 3] includes highly efficient and reliable subroutines for computation of both the SVD and the GSVD. The names of these subroutines are

Table 1. Subroutines for computing the SVD and the QR factorization in some of the most popular numerical libraries. Only LAPACK provides a routine for the GSVD.

Library	SVD subroutine	QR subroutine
ACM TOMS [13]	HYBSVD	—
IMSL [45]	LSVRR	LQRRR
LAPACK [2, 3]	_GESVD	_GEQRF
	_GGSVD (GSVD)	
LINPACK [22]	_SVDC	_QRDC
NAG [52]	F02WEF	F01QCF
Numerical Recipes [58]	SVDCMP	—

listed in table 1. The table also lists several other libraries that include subroutines for computing the SVD as well as the QR factorization.

Once the SVD and/or the GSVD has been computed, it is possible to investigate a number of regularization methods, namely those for which the corresponding filter factors are (simple) functions of the generalized singular values and the regularization parameter. Such an analysis gives important information about the spectral properties of the particular regularization method, and thus it helps the user in selecting a suitable regularization parameter.

Regarding iterative methods, the computation of the filter factors requires the actual execution of the iterative algorithm in order to produce the filter factors. For example, the Ritz polynomial $\mathcal{P}_q$ that appears in equation (22) for the filter factors associated with conjugate gradients requires that q iterations be performed. An implementation of the conjugate gradient algorithm applied to the system $\mathbf{A}^T\mathbf{A}\,x = \mathbf{A}^T b$ which is particularly well suited for discrete regularization problems—where the condition number of $\mathbf{A}$ is very large—is the algorithm LSQR by Paige and Saunders [55]. LSQR is based on the Lanczos bidiagonalization process [28, section 9.3.3 and section 9.3.4] which is mathematically identical to conjugate gradients, and for ill-posed problems it is numerically superior to algorithms based directly on conjugate gradients. An implementation of LSQR is described in [55].

A technique that will save significant computational effort for regularization problems of a general form, i.e. problems with $\mathbf{L} \neq \mathbf{I}_n$, is to avoid the computation of the GSVD of the matrix pair $(\mathbf{A}, \mathbf{L})$. It is always possible to transform such a problem into one in standard form with $\mathbf{L} = \mathbf{I}_n$, and the transformation can be included implicitly in iterative methods [11, 23, 31]. Another way to avoid computation of the GSVD is to use the modified truncated-SVD algorithm [41], which only requires the computation of the SVD of $\mathbf{A}$ plus a QR factorization of a matrix with dimensions smaller than those of $\mathbf{A}$. The key idea is to first compute a truncated SVD solution x_k as described in section 4, and then add a correction x_c to x_k consisting of a linear combination of the vectors $v_{k+1}, \ldots, v_n$, such that $\|\mathbf{L}(x_k + x_c)\|$ is minimized. It is demonstrated in [41] that this solution is in many cases practically the same as the solution x_λ computed by Tikhonov regularization. Subroutines for computing a QR factorization are available from most of the libraries mentioned in table 1.

For many ill-posed problems all the singular values decay gradually to zero with no particular gap in the spectrum. However, there are also situations where there is a distinct gap in the singular value spectrum, i.e. where there is a distinct cluster of large singular values. In this situation, we say that the matrix $\mathbf{A}$ is rank deficient and the number of large singular values is the numerical rank of $\mathbf{A}$, i.e. the numerical

rank r satisfies $\sigma_r \gg \sigma_{r+1}$. For these problems, it is usually not necessary to compute the SVD of $\mathbf{A}$, because a regularized solution can be computed by alternative, computationally cheaper factorizations that are able to filter out the small singular values. One such factorization is the rank-revealing QR (RRQR) factorization of $\mathbf{A}$, which is a pivoted QR factorization of the form

$$\mathbf{A}\,\mathbf{\Pi} = \mathbf{Q}\,\mathbf{R} = \mathbf{Q}\begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ 0 & \mathbf{R}_{22} \end{pmatrix} \qquad \mathbf{R}_{11} \in \mathbb{R}^{r\times r} \tag{25}$$

which satisfies the followng two conditions:

1. The smallest singular value of $\mathbf{R}_{11}$ is approximately equal to σ_r.
2. The norm $\|\mathbf{R}_{22}\|_2$ is of the order σ_{r+1}.

We see that the RRQR factorization (25) reveals the numerical rank of $\mathbf{A}$ as having a well-conditioned leading $r\times r$ triangular block and a trailing triangular $(n-r)\times(n-r)$ block with small norm. The computational effort involved in computing an RRQR factorization is significantly smaller than that required to compute the SVD. Although the solution computed by means of the RRQR factorization of $\mathbf{A}$ is not identical to a truncated SVD solution, for rank deficient problems the two solutions are usually so close to it that the difference is of no importance. For further details about RRQR factorizations, their properties, and their use for solving rank deficient problems, cf. [14, 15, 16]. Software for computing RRQR factorizations is available from ACM TOMS [59].

For all the numerical tools that we have described in this paper, it is most convenient to have access to a system that combines high-quality algorithms with easy access to computer graphics. Three such systems are Matlab [50], CLAM [60], and Mathematica [71]. The author has developed a set of Matlab routines [38] for doing computations with all the above-mentioned tools, including GSVD and RRQR factorization, choosing Matlab in preference to Mathematica and CLAM partly because Matlab is so well suited for linear algebra computations and partly because Matlab is already widely used for scientific computing within the numerical analysis community. The Matlab routines are available from the author.

8. Numerical examples

8.1. Main example: image restoration

The purpose of this final section is to illustrate the theory from the previous sections with two numerical examples. As our main example we choose a one-dimensional model problem in image reconstruction from [61]. In this model, f is the original signal, the kernel K describes the point spread function of an infinitely long slit, and the right-hand side g is the measured signal which consists of the original signal f integrated with K plus additional noise. The kernel K is given by

$$K(s,t) = (\cos s + \cos t)\left(\frac{\sin u}{u}\right)^2 \qquad u = \pi(\sin s + \sin t) \qquad s,t \in [-\pi/2, \pi/2] \tag{26}$$

and as the unperturbed signal f we choose a simple function with two 'humps'

$$f(t) = 2\exp(-6(t-0.8)^2) + \exp(-2(t+0.5)^2). \tag{27}$$

Both K and f are discretized by means of simple collocation such that the elements of the vectors in the SVD-analysis are samples of approximations to the corresponding functions in the SVE-analysis. We choose $m = n = 64$. This yields $\mathbf{A}$ and x, and then the right-hand side vector b is generated as

$$b = \mathbf{A}x + e \tag{28}$$

where the elements of the perturbation vector e are normally distributed with zero mean and standard deviation $\sigma = 10^{-7}$. The vectors x and b, which represent the functions f and g, are shown in figure 2. Notice that the two 'humps' in f are smoothed out to a great extent in g by the integration with K. The purpose of solving the integral equation is now to reconstruct the exact f as accurately as possible from g.

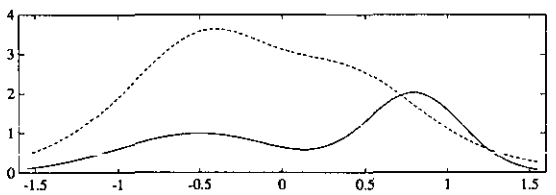

Figure 2. The exact solution f (solid curve) and the corresponding right-hand side g (broken curve) to the model problem (26). The two 'humps' in f are smeared out to a great extent in g.

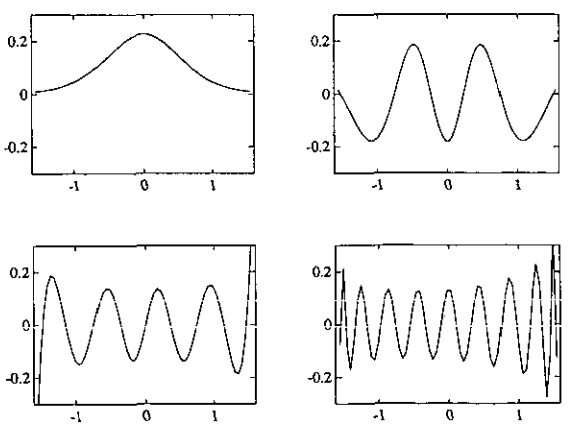

Figure 3. Four right singular functions v_i for $i = 1, 5, 10, 20$ for the model problem. We see that the higher the index, the more high-frequency components are present in v_i.

In figure 3 we show four singular vectors from the SVD of $\mathbf{A}$, namely v_1, v_5, v_{10}, and v_{20}. These vectors are samples of the corresponding right singular functions v_i. It is clear that the higher the index i, i.e. the smaller the singular value, the higher the spectral components in v_i. The same holds for the left singular functions u_i (not shown here).

Figure 4 shows a plot of the singular values σ_i and the quantities $|u_i^T b|$ and $|u_i^T b|/\sigma_i$, as advocated in section 3. We make several remarks here. First, we notice

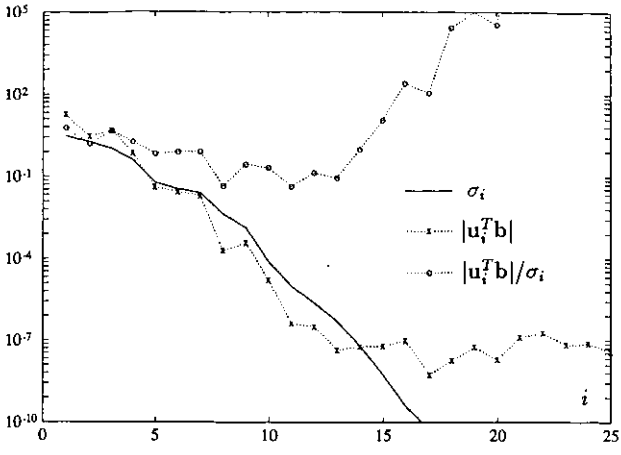

Figure 4. The quantities σ_i, $|u_i^T b|$, and $|u_i^T b|/\sigma_i$ for the model problem. Notice that the coefficients $|u_i^T b|$ decay slightly faster than the singular values σ_i for $i \leqslant 13$.

that all the singular values decay gradually to zero (until they level off at a plateau defined by the machine precision outside of the plot). This is the typical behaviour of the singular values associated with an ill-posed problem. Next, we notice that the coefficients $|u_i^T b|$ also decay for small i, while for larger i they level off at a plateau that reflects the errors e in the right-hand side (28). This is also the typical situation. The error level is approximately equal to $\|e\|_2/\sqrt{n} = 9.0 \times 10^{-8}$. Since the coefficients $|u_i^T b|$ decay slightly faster than the singular values, we assume that the Picard condition is indeed satisfied by the underlying, unperturbed problem.

Finally, we see from figure 4 that we want the filter factors f_i to damp the components $(u_i^T b/\sigma_i)\, v_i$ in the solution for $i > 13$, i.e. for singular values smaller than about 5×10^{-7}. Using filter factors according to this criterion, we will retain all the significant information available in the right-hand side while, simultaneously, we filter out most of the high-frequency noise. Our experiments show that a good choice of the matrix L in the side constraint (14) is the identity matrix, $L = I_n$. Hence, in this example there is no need to compute the GSVD, which makes this presentation slightly simpler.

Let us now consider the filter factors in more details for a few regularization methods. Obviously, if we use truncated SVD then we want to truncate the expansion for x_k in (16) at $k = 13$, hence $f_i = 1$ for $i \leqslant 13$ and $f_i = 0$ for $i > 13$. If we use Tikhonov regularization and choose $\lambda = \sigma_k$, then the Tikhonov filter factors $f_i = \sigma_i^2/(\sigma_i^2 + \lambda^2)$ satisfy $f_k = \frac{1}{2}$ and $f_i = O(\lambda^{-2})$ for $i > k$, which means that the spectral filtering of the singular values effectively sets in at σ_k. A good guess of a suited regularization parameter is therefore $\lambda = \sigma_{13} = 4.7 \times 10^{-7}$ because we want the filter factors to damp the singular values smaller than σ_{13}. The function $\sigma^2/(\sigma^2 + \lambda^2)$ for this choice of λ is shown as the solid curve in figure 5.

Figure 5 also shows the filter factors associated with the solution $x^{(q)}$ after $q = 13$ steps of the LSQR algorithm with reorthogonalization in each step. Without reorthogonalization we would need more steps, but we would arrive at almost the same solution. The filter factors are computed by means of equation (22). Notice that these filter factors qualitatively behave almost like the Tikhonov filter factors. As a consequence, the LSQR and Tikhonov solutions are practically identical for these choices of λ and q, in that $\|x_\lambda - x^{(q)}\|_2/\|x_\lambda\|_2 = 5.8 \times 10^{-3}$. The same holds for the truncated SVD solution x_k: $\|x_\lambda - x_k\|_2/\|x_\lambda\|_2 = 5.8 \times 10^{-3}$. This illustrates that

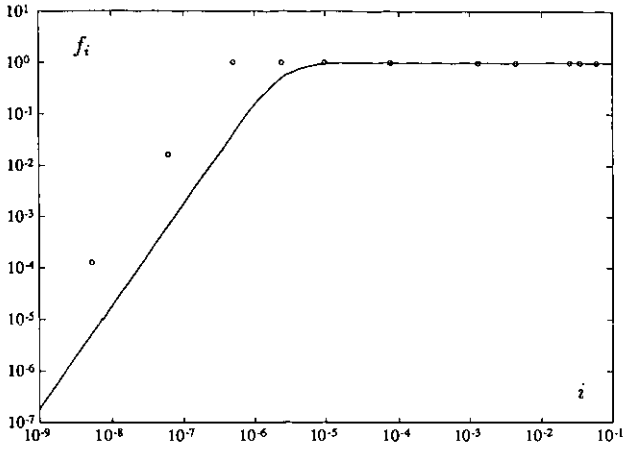

Figure 5. The filter factors for Tikhonov regularization (solid curve) and for the LSQR method with reorthogonalization (broken curve). The filter factors are essentially identical.

regularization methods with similar filter factors tend to produce similar solutions. A proof of this can be found in [37, theorem 4].

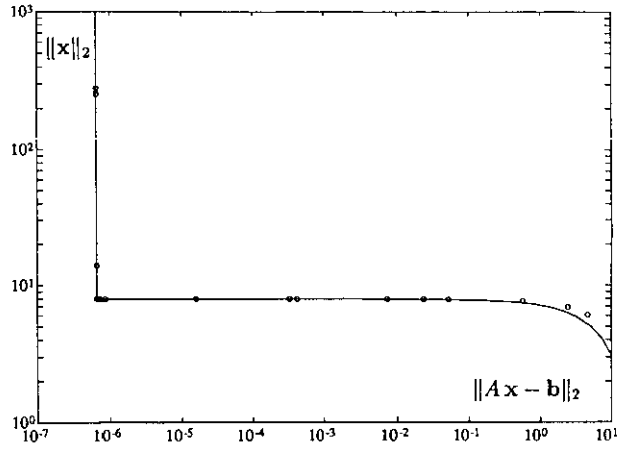

Figure 6. Two L-curves for the model problem: the continuous L-curve for Tikhonov regularization (solid curve) and the discrete L-curve for LSQR (circles).

The L-curve for Tikhonov regularization is shown as the solid curve in figure 6, together with the discrete L-curve (shown as circles) for LSQR. The circles are very close to the solid line, confirming that the LSQR solutions $\boldsymbol{x}^{(q)}$ are indeed very similar to the solutions $\boldsymbol{x}_\lambda$ computed by Tikhonov regularization. For this particular model problem, the L-curve has a very pronounced corner at $\|\mathbf{A}\,\boldsymbol{x}_\lambda - b\|_2 = 6.8 \times 10^{-7}$ and $\|\boldsymbol{x}_\lambda\|_2 = 8.0$. The corresponding value of the regularization parameter λ, computed by finding the maximum curvature of the L-curve, is $\lambda^* = 1.1 \times 10^{-7}$. It is no surprise that this λ^* is close to the regularization parameter $\lambda = \sigma_{13}$ that we chose based on inspection of the σ_i and the $|\boldsymbol{u}_i^{\mathrm{T}} b|$ in figure 4, because the corner corresponds to a good balance between the residual norm $\|\mathbf{A}\,\boldsymbol{x}_\lambda - b\|_2$ and the side constraint $\|\boldsymbol{x}_\lambda\|_2$.

We also applied GCV to this model problem and obtained a regularization parameter $\lambda = 7.4 \times 10^{-7}$, which is not much different from the one chosen by the L-curve method.

This analysis of the model problem (26) has demonstrated how the plots in figures 2–6 reveal a lot of important information about the integral equation, which

helps in computing a reasonable regularized solution. The analysis has also shown how different regularization methods can be analysed and compared easily, for example by comparing their filter factors and their associated L-curves for the particular problem.

8.2. Second example: no solution exists

We conclude this section with a second example which clearly illustrates that one should always be very careful when dealing with ill-posed problems. Consider the integral equation

$$\int_0^1 \frac{f(t)}{1+s+t}\,\mathrm{d}t = 1 \qquad 0 \leqslant s \leqslant 1. \tag{29}$$

This integral equation has no square integrable solution [21, p 7]! Yet, when we discretize the integral equation and apply regularization to the discrete problem, then we will always compute some vector which is likely to be mistaken for the *non-existing* solution to (29). However, if we care to plot the quantities σ_i and $|u_i^{\mathrm{T}}b|$ as shown in figure 7, then we see that the coefficients $|u_i^{\mathrm{T}}b|$ *never* decay faster than the singular values σ_i—not even for the smallest i. (In addition, we see that σ_i and $|u_i^{\mathrm{T}}b|$ level off at approximately ϵ_M and $\sqrt{\epsilon_M}$, respectively, as we would expect from equations (10) and (13).) Figure 7 strongly suggests that the Picard condition is *not* satisfied for this integral equation, and therefore one cannot compute a solution by any numerical method whatsoever.

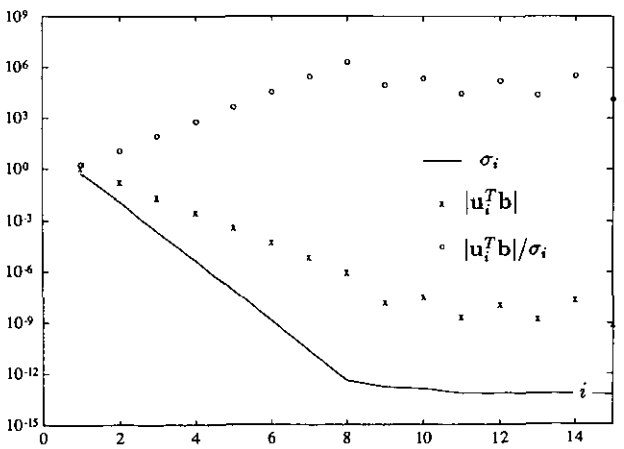

Figure 7. A plot of the singular values σ_i and the coefficients $|u_i^{\mathrm{T}}b|$ for the integral equation (29) which has no square integrable solution. Notice that all the coefficients $|u_i^{\mathrm{T}}b|$ decay *slower* than the singular values.

Acknowledgment

I wish to thank the referee for constructive comments and criticism that helped to improve the presentation of this paper.

References

[1] Allen R C Jr, Boland W R, Faber V and Wing G M 1985 Singular values and condition numbers of Galerkin matrices arising from linear integral equations of the first kind *J. Math. Anal. Appl.* **109** 564–90

[2] Anderson E, Bai Z, Bischof C, Demmel J, Dongarra J, DuCroz J, Greenbaum A, Hammarling S, McKenney A, Ostrouchow O and Sorensen D 1992 *LAPACK: Users' Guide* (Philadelphia, PA: SIAM)

[3] Anderson E, Bischof C, Demmel J, Dongarra J, DuCroz J, Hammarling S and Kahan W 1990 Prospectus for an extension to LAPACK *Report CS-90-118* Computer Science Department, University of Tennessee

[4] Andrews H C and Hunt B R 1977 *Digital Image Restoration* (Englewood Cliffs, NJ: Prentice Hall)

[5] Baker C T H 1977 *The Numerical Treatment of Integral Equations* (Oxford: Clarendon)

[6] Baumeister J 1987 *Stable Solution of Inverse Problems* (Braunschweig: Vieweg)

[7] Bertero M, De Mol C and Pike E R 1985 Linear inverse problems with discrete data: I. General formulation and singular system analysis *Inverse Problems* **1** 301–30

[8] Bertero M, De Mol C and Pike E R 1988 Linear inverse problems with discrete data: II. Stability and regularization *Inverse Problems* **4** 573–94

[9] Bertero M, Poggio T A and Torre V 1988 Ill-posed problems in early vision *Proc. IEEE* **76** 869–89

[10] Björck Å 1990 Least squares methods *Handbook of Numerical Analysis* vol I, ed P G Ciarlet and J L Lions (Amsterdam: Elsevier)

[11] Björck Å 1988 A bidiagonalization algorithm for solving large and sparse ill-posed systems of linear equations *BIT* **28** 659–70

[12] Brakhage H 1987 On ill-posed problems and the method of conjugate gradients *Inverse and Ill-Posed Problems* ed H W Engl and C W Groetsch (New York: Academic)

[13] Chan T F 1982 An improved algorithm for computing the singular value decomposition *ACM Trans. Math. Software* **8** 72–83

[14] Chan T F and Hansen P C 1990 Computing truncated SVD least squares solutions by rank revealing QR-factorizations *SIAM J. Sci. Stat. Comput.* **11** 519–30

[15] Chan T F and Hansen P C 1992 Some applications of the rank revealing QR-factorization *SIAM J. Sci. Stat. Comput.* **13** 727–41

[16] Chan T F and Hansen P C 1991 Low-rank revealing QR factorizations *CAM Report 91-08* Department of Mathematics, UCLA (to appear in J. Num. Lin. Alg. Appl.)

[17] Chatelain F 1983 *Spectral Approximation of Linear Operators* (New York: Academic)

[18] Craig I J D and Brown J C 1986 *Inverse Problems in Astronomy* (Bristol: Adam Hilger)

[19] de Hoog F R 1980 Review of Fredholm equations of the first kind *The Application and Numerical Solution of Integral Equations* ed R S Anderssen, F R de Hoog and M A Lukas (Leyden: Sijthoff and Noordhoff)

[20] Delves L M and Mohamed J L 1985 *Computational Methods for Integral Equations* (Cambridge: Cambridge University Press)

[21] Delves L M and Walsh J 1974 *Numerical Solution of Integral Equations* (Oxford: Clarendon)

[22] Dongarra J J, Bunch J R, Moler B and Stewart G W 1979 *Linpack Users' Guide* (Philadelphia, PA: SIAM)

[23] Eldén L 1977 Algorithms for the regularization of ill-conditioned least squares problems *BIT* **17** 134–45

[24] Engl H W and Gfrerer H 1988 A posteori parameter choice for general regularization methods for solving linear ill-posed problems *Appl. Num. Math.* **4** 395–417

[25] Forsythe G E, Malcolm M A and Moler C B 1977 *Computer Methods for Mathematical Computations* (Englewood Cliffs, NJ: Prentice Hall)

[26] Gilliam D S, Lund J R and Vogel C R 1990 Quantifying information content for ill-posed problems *Inverse Problems* **6** 725–36

[27] Golub G H, Heath M T and Wahba G 1979 Generalized cross validation as a method for choosing a good ridge parameter *Technometrics* **21** 215–24

[28] Golub G H and Van Loan C F 1989 *Matrix Computations* 2nd edn (Baltimore, MD: Johns Hopkins University Press)

[29] Groetsch C W 1984 *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind* (Boston, MA: Pitman)

[30] Groetsch C W and Engl H W (ed) 1987 *Inverse and Ill-Posed Problems* (New York: Academic)

[31] Hanke M 1992 Regularization with differential operators. An iterative approach *Numer. Funct. Anal. Optim.* **13** 523–40
[32] Hansen P C 1987 The truncated SVD as a method for regularization *BIT* **27** 354–553
[33] Hansen P C 1987 Computation of the singular value expansion *Computing* **40** 185–99
[34] Hansen P C 1989 Regularization, GSVD and truncated GSVD *BIT* **29** 491–504
[35] Hansen P C 1990 Truncated SVD solutions to discrete ill-posed problems with ill-determined numerical rank *SIAM J. Sci. Stat. Comput.* **11** 503–18
[36] Hansen P C 1990 The discrete Picard condition for discrete ill-posed problems *BIT* **30** 658–72
[37] Hansen P C 1990 Analysis of discrete ill-posed problems by means of the L-curve *SIAM Review* in press
[38] Hansen P C 1992 Regularization tools, a Matlab package for analysis and solution of discrete ill-posed problems *Report UNIC-92-03* UNI•C
[39] Hansen P C and O'Leary D P 1991 The use of the L-curve in the regularization of discrete ill-posed problems *Report UMIACS-TR-91-142* Deptartment of Computer Science, University of Maryland (submitted to *SIAM J. Sci. Stat. Comput.*)
[40] Hansen P C, O'Leary D P and Stewart G W Regularizing properties of conjugate gradient iterations, in preparation
[41] Hansen P C, Sekii T, and Shibahashi H 1992 The modified truncated-SVD method for regularization in general form *SIAM J. Sci. Stat. Comput.* **13** in press
[42] Harrington R F 1968 *Field Computation by Moment Methods* (New York: Macmillan)
[43] Hestenes M R 1980 *Conjugate Direction Methods in Optimization* (Berlin: Springer)
[44] Hofmann B 1986 *Regularization for Applied Inverse and Ill-Posed Problems* (Stuttgart: Teubner)
[45] IMSL 1989 *IMSL Math/Library*
[46] Kress R 1989 *Linear Integral Equations* (Berlin: Springer)
[47] Larsen J, Lund-Andersen H and Krogsaa B 1983 Transient transport across the blood-retina barrier *Bull. Math. Biology* **45** 749–58
[48] Lawson C L and Hanson R J 1974 *Solving Least Squares Problems* (Englewood Cliffs, NJ: Prentice-Hall)
[49] Miller K 1970 Least squares methods for ill-posed problems with a prescribed bound *SIAM J. Math. Anal.* **1** 52–74
[50] Moler C B, Little J N and Bangert S 1987 *Pro-Matlab User's Guide* (Massachusetts: The MathWorks)
[51] Morozov V A 1984 *Methods for Solving Incorrectly Posed Problems* (Berlin: Springer)
[52] NAG 1990 *NAG Fortran Library Manual, Mark 14* (Oxford: NAG Ltd)
[53] Natterer F 1986 *The Mathematics of Computerized Tomography* (New York: Wiley)
[54] Osaki Y and Shibahashi H (ed) 1990 *Progress of Seismology of the Sun and Stars* (Berlin: Springer)
[55] Paige C C and Saunders M A 1982 LSQR: an algorithm for sparse linear equations and sparse least squares *ACM Trans. Math. Software* **8** 43–71
[56] Phillips D L 1962 A technique for the numerical solution of certain integral equations of the first kind *J. ACM* **9** 84–97
[57] Parker R L 1977 Understanding inverse theory *Ann. Rev. Earth Planet. Sci.* **5** 35–64
[58] Press W H, Flannery B P, Teukolski S A and Vetterling W T 1986 *Numerical Recipes* (Cambridge: Cambridge University Press)
[59] Reichel L and Gragg W B 1990 Algorithm 686: Fortran subroutines for updating the QR decomposition *ACM Trans. Math. Software* **16** 369–77
[60] Scientific Computing Associates 1988 *An Introduction to CLAM*
[61] Shaw C B Jr 1972 Improvement of the resolution of an instrument by numerical solution of an integral equation *J. Math. Anal. Appl.* **37** 83–112
[62] Smith R C, Bowers K L and Vogel C R 1991 Numerical recovery of material parameters in Euler–Bernoulli beam models *ICASE Report 91-14* NASA Langley Research Center
[63] Smithies F 1958 *Integral Equations* (Cambridge: Cambridge University Press)
[64] Tikhonov A N and Arsenin V Y 1977 *Solutions of Ill-Posed Problems* (New York: Wiley)
[65] Tikhonov A N and Goncharsky A V (ed) 1987 *Ill-Posed Problems in the Natural Sciences* (Moscow: MIR)
[66] Van Loan C F 1976 Generalizing the singular value decomposition *SIAM J. Numer. Anal.* **13** 76–83
[67] Varah J M 1979 A practical examination of some numerical methods for linear discrete ill-posed problems *SIAM Rev.* **21** 100–11
[68] Wahba G 1977 Practical approximate solutions to linear operator equations when the data are noisy *SIAM J. Numer. Anal.* **14** 651–67

[69] Wahba G 1980 Ill-posed problems: numerical and statistical methods for mildly, moderately and severely ill-posed problems with noisy data *Technical Report 595* Department of Statistics, University of Wisconsin

[70] Wing G M 1985 Condition numbers of matrices arising from the numerical solution of linear integral equations of the first kind *J. Integral Equations* (Suppl.) **9** 191–204

[71] Wolfram S 1988 *Mathematica* (Reading, MA: Addison-Wesley)