8

Total least squares

In atmospheric remote sensing, near real-time software processors frequently use approximations of the Jacobian matrix in order to speed up the calculation. If the forward model $\mathbf{F}(\mathbf{x})$ depends on the state vector $\mathbf{x}$ through some model parameters $\mathbf{b}_k$,

$$\mathbf{F}(\mathbf{x}) = \mathbf{F}(\mathbf{b}_1(\mathbf{x}), \ldots, \mathbf{b}_N(\mathbf{x})),$$

then, an approximate expression of the Jacobian matrix

$$\mathbf{K} = \sum_{k=1}^{N} \frac{\partial \mathbf{F}}{\partial \mathbf{b}_k} \frac{\partial \mathbf{b}_k}{\partial \mathbf{x}},$$

can be obtained by assuming that some $\mathbf{b}_k$ are insensitive to $\mathbf{x}$, i.e., $\partial \mathbf{b}_k / \partial \mathbf{x} = \mathbf{0}$. For example, the limb radiance measured by a detector in the ultraviolet or visible spectral domains can be expressed as

$$I(\lambda, \mathbf{x}) = I_{\mathtt{ss}}(\lambda, \mathbf{x}) + I_{\mathtt{ms}}(\lambda, \mathbf{x}) = I_{\mathtt{ss}}(\lambda, \mathbf{x})[1 + c_{\mathtt{ms}}(\lambda, \mathbf{x})], \tag{8.1}$$

where $I_{\mathtt{ss}}$ and $I_{\mathtt{ms}}$ are the single and multiple scattering terms, λ is the wavelength, and $c_{\mathtt{ms}}$ is a correction factor accounting for the multiple scattering contribution. As the computation of the derivative of $c_{\mathtt{ms}}$ is quite demanding, the Jacobian matrix calculation may involve only the derivative of $I_{\mathtt{ss}}$. Similarly, in a line-by-line model, the absorption coefficient $C_{\mathtt{abs}m}$ of the gas molecule m is the product of the line strength S_{ml} and the normalized line shape function g_{ml} (cf. (1.12)),

$$C_{\mathtt{abs}m}(\nu, T) = \sum_{l} S_{ml}(T)\, g_{ml}(\nu, T),$$

where ν is the wavenumber, T is the temperature, and the summation is over all lines l. As the most important temperature dependence stems from the line strength, the derivative of the line shape function with respect to the temperature is sometimes ignored.

The total least squares (TLS) method is devoted to the solution of linear problems in which both the coefficient matrix and the data are subject to errors. The linear data model can be expressed as

$$\mathbf{y}^{\delta} = (\mathbf{K}_{\Lambda} - \mathbf{\Lambda})\,\mathbf{x} + \boldsymbol{\delta},$$

where the matrix $\mathbf{K}_\Lambda$ is a perturbation of the exact (unknown) matrix $\mathbf{K}$, $\mathbf{K}_\Lambda = \mathbf{K} + \mathbf{\Lambda}$, and the data are affected by the instrumental noise $\boldsymbol{\delta}$.

The TLS method was independently derived in several bodies of work by Golub and Van Loan (1980, 1996), and Van Huffel and Vanderwalle (1991). This literature has advanced the algorithmic and theoretical understanding of the method, as well as its application for computing stable solutions of linear systems of equations with highly ill-conditioned coefficient matrices. In this section we review the truncated and the regularized TLS methods for solving linear ill-posed problems, and reveal the similarity with the Tikhonov regularization. We then present a first attempt to extend the regularized TLS to nonlinear ill-posed problems.

8.1 Formulation

The linear model which encapsulates the uncertainties in the data vector and the coefficient matrix is of the form $\mathbf{K}_\Lambda \mathbf{x} \approx \mathbf{y}^\delta$. To sketch the TLS method, we introduce the augmented matrix $\left[\begin{array}{cc} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{array}\right]$ and consider the homogeneous system of equations

$$\left[\begin{array}{cc} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{array}\right] \left[\begin{array}{c} \mathbf{x} \\ -1 \end{array}\right] = \mathbf{0}. \tag{8.2}$$

We then assume a singular value decomposition of the $m \times (n+1)$ matrix,

$$\left[\begin{array}{cc} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{array}\right] = \bar{\mathbf{U}} \bar{\Sigma} \bar{\mathbf{V}}^T, \tag{8.3}$$

and partition the matrices $\bar{\mathbf{V}}$ and $\bar{\Sigma}$ as follows:

$$\bar{\mathbf{V}} = [\bar{\mathbf{v}}_1, \ldots, \bar{\mathbf{v}}_{n+1}] = \left[\begin{array}{cc} \bar{\mathbf{V}}_{11} & \bar{\mathbf{v}}_{12} \\ \bar{\mathbf{v}}_{21}^T & \bar{v}_{22} \end{array}\right], \quad \bar{\mathbf{V}}_{11} \in \mathbb{R}^{n \times n}, \quad \bar{\mathbf{v}}_{12}, \bar{\mathbf{v}}_{21} \in \mathbb{R}^n, \tag{8.4}$$

and

$$\bar{\Sigma} = \left[\begin{array}{cc} \bar{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \bar{\sigma}_{n+1} \\ \mathbf{0} & \mathbf{0} \end{array}\right], \quad \bar{\Sigma}_1 = \left[\mathrm{diag}\,(\bar{\sigma}_i)_{n \times n}\right],$$

respectively. If $\bar{\sigma}_{n+1} \neq 0$, then rank $\left(\left[\begin{array}{cc} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{array}\right]\right) = n+1$, and the solution of the homogeneous system of equations (8.2) is the trivial solution. Thus, the last component of the solution vector is not -1, and to solve (8.2) it is necessary to reduce the rank of the augmented matrix from $n+1$ to n. This can be achieved by approximating the rank-$(n+1)$ matrix $\left[\begin{array}{cc} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{array}\right]$ by a rank-n matrix $\left[\begin{array}{cc} \mathbf{K}_n & \mathbf{y}_n \end{array}\right]$. As rank $\left(\left[\begin{array}{cc} \mathbf{K}_n & \mathbf{y}_n \end{array}\right]\right) = n$, we may assume that the last column vector of the matrix $\left[\begin{array}{cc} \mathbf{K}_n & \mathbf{y}_n \end{array}\right]$ is a linear combination of the first n column vectors, i.e.,

$$\mathbf{y}_n = \sum_{i=1}^{n} x_i \mathbf{k}_i,$$

with $\mathbf{K}_n = [\mathbf{k}_1, \ldots, \mathbf{k}_n]$, or equivalently that,

$$\mathbf{K}_n \mathbf{x} = \mathbf{y}_n,$$

with $\mathbf{x} = [x_1, \ldots, x_n]^T$. The (matrix) approximation problem can be expressed as the constrained minimization problem

$$\min_{[\tilde{\mathbf{K}}\ \tilde{\mathbf{y}}] \in \mathbb{R}^{m \times (n+1)}} \left\| \begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix} - \begin{bmatrix} \tilde{\mathbf{K}} & \tilde{\mathbf{y}} \end{bmatrix} \right\|_{\mathrm{F}}^2 \tag{8.5}$$

$$\text{subject to } \tilde{\mathbf{K}}\mathbf{x} = \tilde{\mathbf{y}},$$

where the Frobenius norm of the $m \times n$ matrix $\mathbf{A}$ is defined by

$$\|\mathbf{A}\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} [\mathbf{A}]_{ij}^2}.$$

It should be pointed out that the ordinary least squares method minimizes the norm of the residual vector $\mathbf{y}^\delta - \tilde{\mathbf{y}}$ under the assumption that $\mathbf{K}_\Lambda = \tilde{\mathbf{K}}$.

The solution to the minimization problem (8.5) is given by the Eckart–Young–Mirsky theorem (Golub and Van Loan, 1996): the matrix

$$\begin{bmatrix} \mathbf{K}_n & \mathbf{y}_n \end{bmatrix} = \sum_{i=1}^{n} \bar{\sigma}_i \bar{\mathbf{u}}_i \bar{\mathbf{v}}_i^T \tag{8.6}$$

is the closest rank-n matrix to $\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix}$, and we have

$$\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix} - \begin{bmatrix} \mathbf{K}_n & \mathbf{y}_n \end{bmatrix} = \bar{\sigma}_{n+1} \bar{\mathbf{u}}_{n+1} \bar{\mathbf{v}}_{n+1}^T,$$

yielding

$$\left\| \begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix} - \begin{bmatrix} \mathbf{K}_n & \mathbf{y}_n \end{bmatrix} \right\|_{\mathrm{F}} = \bar{\sigma}_{n+1}.$$

The homogeneous system of equations (8.2) is then replaced by a homogeneous system of equations involving the rank-n matrix $\begin{bmatrix} \mathbf{K}_n & \mathbf{y}_n \end{bmatrix}$, that is,

$$\begin{bmatrix} \mathbf{K}_n & \mathbf{y}_n \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ -1 \end{bmatrix} = \mathbf{0}. \tag{8.7}$$

Since (cf. (8.6))

$$\begin{bmatrix} \mathbf{K}_n & \mathbf{y}_n \end{bmatrix} \bar{\mathbf{v}}_{n+1} = \sum_{i=1}^{n} \bar{\sigma}_i \left(\bar{\mathbf{v}}_i^T \bar{\mathbf{v}}_{n+1} \right) \bar{\mathbf{u}}_i = \mathbf{0}, \tag{8.8}$$

we see that the vector $a\bar{\mathbf{v}}_{n+1}$ is the general solution of the homogeneous system of equations (8.7) and that the scalar a is (uniquely) determined by imposing that the last component of the solution vector is -1. We obtain

$$\begin{bmatrix} \mathbf{x}_\Lambda^\delta \\ -1 \end{bmatrix} = -\frac{1}{[\bar{\mathbf{v}}_{n+1}]_{n+1}} \bar{\mathbf{v}}_{n+1}, \tag{8.9}$$

provided that $[\bar{\mathbf{v}}_{n+1}]_{n+1} \neq 0$. From (8.4), we find that the TLS solution can be expressed as

$$\mathbf{x}_\Lambda^\delta = -\frac{1}{\bar{v}_{22}} \bar{\mathbf{v}}_{12}. \tag{8.10}$$

Note that if $\bar{\sigma}_{n+1}$ is a simple singular value, we have (cf. (8.8)) $\mathcal{N}\left(\begin{bmatrix} \mathbf{K}_n & \mathbf{y}_n \end{bmatrix} \right) = \text{span}\{\bar{\mathbf{v}}_{n+1}\}$, and the TLS solution is unique.

8.2 Truncated total least squares

The truncated TLS method, which in general is devoted to numerically rank deficient problems, is also a suitable regularization method for discrete ill-posed problems. This technique is similar to the truncated SVD that treats small singular values of $\mathbf{K}$ as zeros. In both methods, the redundant information in $\left[\begin{array}{cc} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{array}\right]$ and $\mathbf{K}$, respectively, associated to the small singular values, is discarded and the original ill-posed problem with a full rank matrix is replaced by a well-posed problem with a rank-deficient matrix. This approximation is achieved by means of the Eckart–Young–Mirsky theorem. For example, in the truncated SVD, the matrix $\mathbf{K}$ with rank $(\mathbf{K}) = n$ and singular value decomposition

$$\mathbf{K} = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$

is replaced by the matrix

$$\mathbf{K}_p = \sum_{i=1}^{p} \sigma_i \mathbf{u}_i \mathbf{v}_i^T,$$

with rank $(\mathbf{K}_p) = p$, and the regularized solution takes the form

$$\mathbf{x}_p^\delta = \arg\min_{\mathbf{x}} \left\|\mathbf{y}^\delta - \mathbf{K}_p \mathbf{x}\right\|^2 = \sum_{i=1}^{p} \frac{1}{\sigma_i} \left(\mathbf{u}_i^T \mathbf{y}^\delta\right) \mathbf{v}_i.$$

The major difference between the two methods lies in the way in which the approximation is performed: in the truncated SVD, the modification depends only on $\mathbf{K}$, while in the truncated TLS, the modification depends on both $\mathbf{K}_\Lambda$ and $\mathbf{y}^\delta$. Thus, in the framework of the truncated TLS method we approximate the matrix $\left[\begin{array}{cc} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{array}\right]$ by the rank-p matrix

$$\left[\begin{array}{cc} \mathbf{K}_p & \mathbf{y}_p \end{array}\right] = \sum_{i=1}^{p} \bar{\sigma}_i \bar{\mathbf{u}}_i \bar{\mathbf{v}}_i^T.$$

To determine the number p of large singular values or the truncation index, we may require a user-specified threshold or determine p adaptively. The null space of the approximation matrix is

$$\mathcal{N}\left(\left[\begin{array}{cc} \mathbf{K}_p & \mathbf{y}_p \end{array}\right]\right) = \operatorname{span}\left\{\bar{\mathbf{v}}_{p+1}, \ldots, \bar{\mathbf{v}}_{n+1}\right\},$$

whence accounting for the partition

$$\bar{\mathbf{V}} = [\bar{\mathbf{v}}_1, \ldots, \bar{\mathbf{v}}_{n+1}] = \left[\begin{array}{cc} \bar{\mathbf{V}}_{11} & \bar{\mathbf{V}}_{12} \\ \bar{\mathbf{v}}_{21}^T & \bar{\mathbf{v}}_{22}^T \end{array}\right], \tag{8.11}$$

with $\bar{\mathbf{V}}_{11} \in \mathbb{R}^{n \times p}$, $\bar{\mathbf{V}}_{12} \in \mathbb{R}^{n \times (n-p+1)}$, and

$$\bar{\mathbf{v}}_{21} = \left[\left[\bar{\mathbf{v}}_1\right]_{n+1}, \ldots, \left[\bar{\mathbf{v}}_p\right]_{n+1}\right]^T \in \mathbb{R}^p,$$

$$\bar{\mathbf{v}}_{22} = \left[\left[\bar{\mathbf{v}}_{p+1}\right]_{n+1}, \ldots, \left[\bar{\mathbf{v}}_{n+1}\right]_{n+1}\right]^T \in \mathbb{R}^{n-p+1},$$

we seek the solution as

$$\begin{bmatrix} \mathbf{x}_{\Lambda p}^{\delta} \\ -1 \end{bmatrix} = \sum_{i=p+1}^{n+1} a_i \bar{\mathbf{v}}_i = \begin{bmatrix} \bar{\mathbf{V}}_{12} \\ \bar{\mathbf{v}}_{22}^T \end{bmatrix} \mathbf{a}, \tag{8.12}$$

with $\mathbf{a} = [a_{p+1}, \ldots, a_{n+1}]^T \in \mathbb{R}^{n-p+1}$. From the last equation we find that

$$\bar{\mathbf{v}}_{22}^T \mathbf{a} = -1,$$

or equivalently that

$$\sum_{i=p+1}^{n+1} a_i \left[\bar{\mathbf{v}}_i\right]_{n+1} = -1.$$

Since (cf. (8.12))

$$\left\| \begin{bmatrix} \mathbf{x}_{\Lambda p}^{\delta} \\ -1 \end{bmatrix} \right\|^2 = 1 + \left\| \mathbf{x}_{\Lambda p}^{\delta} \right\|^2 = \sum_{i=p+1}^{n+1} a_i^2, \tag{8.13}$$

we see that the minimum norm solution $\mathbf{x}_{\Lambda p}^{\delta}$ requires a minimum value of $\sum_{i=p+1}^{n+1} a_i^2$. This can be obtained by solving the constrained minimization problem

$$\begin{aligned} &\min_{a_i} \sum_{i=p+1}^{n+1} a_i^2 \\ &\text{subject to } \sum_{i=p+1}^{n+1} a_i \left[\bar{\mathbf{v}}_i\right]_{n+1} = -1. \end{aligned}$$

In the framework of the Lagrange multiplier formalism, the first-order optimality conditions for the Lagrangian function

$$\mathcal{L}(\mathbf{a}, \lambda) = \frac{1}{2} \sum_{i=p+1}^{n+1} a_i^2 + \lambda \left(\sum_{i=p+1}^{n+1} a_i \left[\bar{\mathbf{v}}_i\right]_{n+1} + 1 \right),$$

yield

$$\begin{aligned} &a_i + \lambda \left[\bar{\mathbf{v}}_i\right]_{n+1} = 0, \quad i = p+1, \ldots, n+1, \\ &\sum_{i=p+1}^{n+1} a_i \left[\bar{\mathbf{v}}_i\right]_{n+1} = -1, \end{aligned}$$

and we obtain

$$\mathbf{a} = -\frac{1}{\left\| \bar{\mathbf{v}}_{22} \right\|^2} \bar{\mathbf{v}}_{22}. \tag{8.14}$$

Hence, from (8.12) and (8.14), the minimum norm solution is given by

$$\mathbf{x}_{\Lambda p}^{\delta} = -\frac{1}{\left\| \bar{\mathbf{v}}_{22} \right\|^2} \bar{\mathbf{V}}_{12} \bar{\mathbf{v}}_{22}. \tag{8.15}$$

By (8.13), (8.14) and the Eckart–Young–Mirsky theorem, we have

$$\left\|\mathbf{x}_{\Lambda p}^{\delta}\right\|^2 = \frac{1}{\left\|\bar{\mathbf{v}}_{22}\right\|^2} - 1,$$

and

$$\left\|\mathbf{R}_{\Lambda p}^{\delta}\right\|_{\mathrm{F}}^2 = \left\| \begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix} - \begin{bmatrix} \mathbf{K}_p & \mathbf{y}_p \end{bmatrix} \right\|_{\mathrm{F}}^2 = \bar{\sigma}_{p+1}^2 + \ldots + \bar{\sigma}_{n+1}^2,$$

showing that the solution norm $\left\|\mathbf{x}_{\Lambda p}^{\delta}\right\|$ increases monotonically with p, while the residual norm $\left\|\mathbf{R}_{\Lambda p}^{\delta}\right\|_{\mathrm{F}}$ decreases monotonically with p. These results recommend the discrepancy principle and the L-curve method for computing the truncation index.

In order to demonstrate the regularizing property of the truncated TLS method, we express $\mathbf{x}_{\Lambda p}^{\delta}$ as the filtered sum

$$\mathbf{x}_{\Lambda p}^{\delta} = \sum_{i=1}^{n} f_i \frac{1}{\sigma_i} \left(\mathbf{u}_i^T \mathbf{y}^\delta\right) \mathbf{v}_i, \tag{8.16}$$

where $(\sigma_i; \mathbf{v}_i, \mathbf{u}_i)$ is a singular system of $\mathbf{K}_\Lambda$. In Appendix I it is shown that if $\operatorname{rank}(\mathbf{K}_\Lambda) = n$ and $\operatorname{rank}\left(\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix}\right) = n+1$, and furthermore, if $\mathbf{u}_i^T \mathbf{y}^\delta \neq 0$ for all $i = 1, \ldots, n$, then the filter factors are given by

$$f_i = \frac{1}{\left\|\bar{\mathbf{v}}_{22}\right\|^2} \sum_{j=1}^{p} \frac{\sigma_i^2}{\bar{\sigma}_j^2 - \sigma_i^2} \left[\bar{\mathbf{v}}_j\right]_{n+1}^2, \tag{8.17}$$

and the estimates

$$1 < f_i \leq 1 + \left(\frac{\bar{\sigma}_{p+1}}{\sigma_i}\right)^2 + O\left(\frac{\bar{\sigma}_{p+1}^4}{\sigma_i^4}\right), \quad i = 1, \ldots, p, \tag{8.18}$$

and

$$0 < f_i \leq \frac{1 - \left\|\bar{\mathbf{v}}_{22}\right\|^2}{\left\|\bar{\mathbf{v}}_{22}\right\|^2} \left(\frac{\sigma_i}{\bar{\sigma}_p}\right)^2 \left[1 + O\left(\frac{\sigma_i^2}{\bar{\sigma}_p^2}\right)\right], \quad i = p+1, \ldots, n \tag{8.19}$$

hold. From (8.18), (8.19) and the interlacing property of the singular values of $\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix}$ and $\mathbf{K}_\Lambda$,

$$\bar{\sigma}_1 > \sigma_1 > \ldots > \bar{\sigma}_p > \sigma_p > \bar{\sigma}_{p+1} > \sigma_{p+1} > \ldots > \sigma_n > \bar{\sigma}_{n+1},$$

we see that for $i \ll p$, $\left(\bar{\sigma}_{p+1}/\sigma_i\right)^2 \ll 1$ and the filter factors are close to 1, while for $i \gg p$, $\left(\sigma_i/\bar{\sigma}_p\right)^2 \ll 1$ and the filter factors are very small. Thus, the filter factors of the truncated TLS method resemble the Tikhonov filter factors, and $\mathbf{x}_{\Lambda p}^{\delta}$ is a filtered solution, with the truncation index p playing the role of the regularization parameter.

When the dimension of $\mathbf{K}_\Lambda$ is not too large, the singular value decomposition of the augmented matrix $\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix}$ can be computed directly. For large-scale problems, this approach is computationally expensive and an iterative algorithm based on Lanczos bidiagonalization can be used instead (Fierro et al., 1997). The so-called Lanczos truncated TLS

algorithm uses the Lanczos bidiagonalization of the matrix $\mathbf{K}_\Lambda$ to obtain, after p iteration steps, the factorization

$$\mathbf{K}_\Lambda \bar{\mathbf{V}}_p = \bar{\mathbf{U}}_{p+1} \mathbf{B}_p, \tag{8.20}$$

and projects the TLS problem onto the subspace spanned by $\bar{\mathbf{U}}_{p+1} \in \mathbb{R}^{m\times(p+1)}$ and $\bar{\mathbf{V}}_p \in \mathbb{R}^{n\times p}$. The projection is a consequence of the assumption that for a sufficiently large p, all the large singular values of $\mathbf{K}_\Lambda$, which contribute to the regularized solution, have been captured. The projected TLS problem reads as

$$\begin{aligned} &\min_{[\tilde{\mathbf{K}}_p \; \tilde{\mathbf{y}}_p] \in \mathbb{R}^{m\times(n+1)}} \left\| \bar{\mathbf{U}}_{p+1}^T \left(\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix} - \begin{bmatrix} \tilde{\mathbf{K}}_p & \tilde{\mathbf{y}}_p \end{bmatrix} \right) \begin{bmatrix} \bar{\mathbf{V}}_p & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix} \right\|_{\mathrm{F}}^2 \\ &\text{subject to } \bar{\mathbf{U}}_{p+1}^T \tilde{\mathbf{K}}_p \bar{\mathbf{V}}_p \mathbf{z}_p = \bar{\mathbf{U}}_{p+1}^T \tilde{\mathbf{y}}_p, \end{aligned}$$

where we have set $\mathbf{x} = \bar{\mathbf{V}}_p \mathbf{z}_p$ for some $\mathbf{z}_p \in \mathbb{R}^p$. Using the result (cf. (8.20) and (5.36))

$$\bar{\mathbf{U}}_{p+1}^T \begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix} \begin{bmatrix} \bar{\mathbf{V}}_p & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{U}}_{p+1}^T \mathbf{K}_\Lambda \bar{\mathbf{V}}_p & \bar{\mathbf{U}}_{p+1}^T \mathbf{y}^\delta \end{bmatrix} = \begin{bmatrix} \mathbf{B}_p & \beta_1 \mathbf{e}_1^{(p+1)} \end{bmatrix},$$

the constrained minimization problem can be rewritten as

$$\begin{aligned} &\min_{[\tilde{\mathbf{B}}_p \; \tilde{\mathbf{e}}_p] \in \mathbb{R}^{(p+1)\times(p+1)}} \left\| \begin{bmatrix} \mathbf{B}_p & \beta_1 \mathbf{e}_1^{(p+1)} \end{bmatrix} - \begin{bmatrix} \tilde{\mathbf{B}}_p & \tilde{\mathbf{e}}_p \end{bmatrix} \right\|_{\mathrm{F}}^2 \\ &\text{subject to } \tilde{\mathbf{B}}_p \mathbf{z}_p = \tilde{\mathbf{e}}_p, \end{aligned} \tag{8.21}$$

where we have put $\tilde{\mathbf{B}}_p = \bar{\mathbf{U}}_{p+1}^T \tilde{\mathbf{K}}_p \bar{\mathbf{V}}_p$ and $\tilde{\mathbf{e}}_p = \bar{\mathbf{U}}_{p+1}^T \tilde{\mathbf{y}}_p$. Thus, in each Lanczos step, we use the TLS algorithm for the small-scale problem (8.21) to compute a truncated TLS solution $\mathbf{x}_{\Lambda p}^\delta$. More precisely, assuming the singular value decomposition

$$\begin{bmatrix} \mathbf{B}_p & \beta_1 \mathbf{e}_1^{(p+1)} \end{bmatrix} = \bar{\bar{\mathbf{U}}} \bar{\bar{\Sigma}} \bar{\bar{\mathbf{V}}},$$

with

$$\bar{\bar{\mathbf{V}}} = \begin{bmatrix} \bar{\bar{\mathbf{V}}}_{11} & \bar{\bar{\mathbf{v}}}_{12} \\ \bar{\bar{\mathbf{v}}}_{21}^T & \bar{\bar{v}}_{22} \end{bmatrix}, \quad \bar{\bar{\mathbf{V}}}_{11} \in \mathbb{R}^{p\times p}, \quad \bar{\bar{\mathbf{v}}}_{12}, \bar{\bar{\mathbf{v}}}_{21} \in \mathbb{R}^p,$$

the TLS solution to (8.21) is (cf. (8.10))

$$\mathbf{z}_{\Lambda p}^\delta = -\frac{1}{\bar{\bar{v}}_{22}} \bar{\bar{\mathbf{v}}}_{12},$$

and the truncated TLS solution takes the form

$$\mathbf{x}_{\Lambda p}^\delta = \bar{\mathbf{V}}_p \mathbf{z}_{\Lambda p}^\delta = -\frac{1}{\bar{\bar{v}}_{22}} \bar{\mathbf{V}}_p \bar{\bar{\mathbf{v}}}_{12}.$$

In the Lanczos truncated TLS algorithm, the solution norm and the residual norm also possess monotonic behavior, i.e., $\left\|\mathbf{x}_{\Lambda p}^\delta\right\|$ is a increasing function of p, while $\left\|\mathbf{R}_{\Lambda p}^\delta\right\|_{\mathrm{F}}$ is a decreasing function of p (Fierro et al., 1997).

Regularization parameter choice methods for truncated TLS are discrete methods. If explicit knowledge about the errors in $\mathbf{K}_\Lambda$ and $\mathbf{y}^\delta$ is available, the discrepancy principle can be used to compute the truncation index. When the errors in $\mathbf{K}_\Lambda$ and $\mathbf{y}^\delta$ are not available, error-free parameter choice methods can be employed. In this context, we mention that Sima and Van Huffel (2006) formulated the generalized cross-validation in the framework of the Lanczos truncated TLS, while the L-curve method has been applied by Fierro et al. (1997).

The truncated solution $\mathbf{x}_{\Lambda p}^\delta$ is a filtered solution whose main contributions come from the first p singular vectors of $\mathbf{K}_\Lambda$ (Appendix I). Because these vectors are not always the best basis vectors for a regularized solution, we may implicitly include regularization in general form with $\mathbf{L} \neq \mathbf{I}_n$. This is done by transforming the problem involving $\mathbf{K}_\Lambda$ and $\mathbf{L}$ into a standard-form problem with the matrix $\bar{\mathbf{K}}_\Lambda = \mathbf{K}_\Lambda \mathbf{L}^{-1}$. Then, we apply the truncated TLS method to the standard-form problem to obtain a regularized solution $\bar{\mathbf{x}}_{\Lambda p}^\delta$, and finally, we transform $\bar{\mathbf{x}}_{\Lambda p}^\delta$ back to the general-form setting by computing $\mathbf{x}_{\Lambda p}^\delta = \mathbf{L}^{-1}\bar{\mathbf{x}}_{\Lambda p}^\delta$. The conventional and the Lanczos versions of the truncated TLS method are outlined in Algorithms 13 and 14. It should be remarked that Algorithm 13 computes simultaneously the truncated SVD solution and the truncated TLS solution for a fixed value of the truncation index p.

8.3 Regularized total least squares for linear problems

Tikhonov regularization has been recast in the framework of the regularized TLS by Golub et al. (1999). To stress the differences and the similarities between the conventional Tikhonov regularization and the regularized TLS, we first note that Tikhonov regulariza-

Algorithm 13. Algorithm for computing the truncated SVD solution $\mathbf{x}_p^\delta$ and the truncated TLS solution $\mathbf{x}_{\Lambda p}^\delta$ for a fixed value of the truncation index p.

$\bar{\mathbf{K}}_\Lambda \leftarrow \mathbf{K}_\Lambda \mathbf{L}^{-1}$;
{truncated SVD solution}
compute the SVD $\bar{\mathbf{K}}_\Lambda = \mathbf{U}\Sigma\mathbf{V}^T$;
$\bar{\mathbf{x}}_p^\delta \leftarrow \sum_{i=1}^{p} (1/\sigma_i) \left(\mathbf{u}_i^T \mathbf{y}^\delta\right) \mathbf{v}_i$;
$\mathbf{x}_p^\delta \leftarrow \mathbf{L}^{-1}\bar{\mathbf{x}}_p^\delta$;
{truncated TLS solution}
compute the SVD $\begin{bmatrix} \bar{\mathbf{K}}_\Lambda & \mathbf{y}^\delta \end{bmatrix} = \bar{\mathbf{U}}\bar{\Sigma}\bar{\mathbf{V}}^T$;
partition $\bar{\mathbf{V}} = \begin{bmatrix} \bar{\mathbf{V}}_{11} & \bar{\mathbf{V}}_{12} \\ \bar{\mathbf{v}}_{21}^T & \bar{\mathbf{v}}_{22}^T \end{bmatrix}$ with $\bar{\mathbf{V}}_{11} \in \mathbb{R}^{n \times p}$;
$\bar{\mathbf{x}}_{\Lambda p}^\delta \leftarrow -\left(1/\left\|\bar{\mathbf{v}}_{22}\right\|^2\right) \bar{\mathbf{V}}_{12}\bar{\mathbf{v}}_{22}$;
$\mathbf{x}_{\Lambda p}^\delta \leftarrow \mathbf{L}^{-1}\bar{\mathbf{x}}_{\Lambda p}^\delta$;

Algorithm 14. Lanczos truncated TLS algorithm with $p_{\max} > 1$ iterations.

$\beta_1 \leftarrow \left\|\mathbf{y}^\delta\right\|$; $\bar{\mathbf{u}} \leftarrow (1/\beta_1)\,\mathbf{y}^\delta$;
$\mathbf{q} \leftarrow \mathbf{L}^{-T}\mathbf{K}^T\bar{\mathbf{u}}$; $\alpha_1 \leftarrow \|\mathbf{q}\|$; $\bar{\mathbf{v}}_1 \leftarrow (1/\alpha_1)\,\mathbf{q}$;
for $p = 1, p_{\max}$ **do**
  $\mathbf{p} \leftarrow \mathbf{K}\mathbf{L}^{-1}\bar{\mathbf{v}}_p - \alpha_p\bar{\mathbf{u}}$; $\beta_{p+1} \leftarrow \|\mathbf{p}\|$; $\bar{\mathbf{u}} \leftarrow (1/\beta_{p+1})\,\mathbf{p}$;
  if $p > 1$ **then**
    set $\mathbf{A} = \begin{bmatrix} \mathbf{B}_p & \beta_1\mathbf{e}_1^{(p+1)} \end{bmatrix} = \begin{bmatrix} \alpha_1 & 0 & \dots & 0 & \beta_1 \\ \beta_2 & \alpha_2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \alpha_p & 0 \\ 0 & 0 & \dots & \beta_{p+1} & 0 \end{bmatrix}$;
    compute the SVD $\mathbf{A} = \bar{\bar{\mathbf{U}}}\bar{\bar{\Sigma}}\bar{\bar{\mathbf{V}}}$;
    partition $\bar{\bar{\mathbf{V}}} = \begin{bmatrix} \bar{\bar{\mathbf{V}}}_{11} & \bar{\bar{\mathbf{v}}}_{12} \\ \bar{\bar{\mathbf{v}}}_{21}^T & \bar{\bar{v}}_{22} \end{bmatrix}$ with $\bar{\bar{\mathbf{V}}}_{11} \in \mathbb{R}^{p\times p}$;
    $\bar{\mathbf{x}}_{\Lambda p}^\delta \leftarrow -(1/\bar{\bar{v}}_{22})\sum_{j=1}^p \left[\bar{\bar{\mathbf{v}}}_{12}\right]_j \bar{\mathbf{v}}_j$;
    $\mathbf{x}_{\Lambda p}^\delta \leftarrow \mathbf{L}^{-1}\bar{\mathbf{x}}_{\Lambda p}^\delta$;
  end if
  if $p < p_{\max}$ **then**
    $\mathbf{q} \leftarrow \mathbf{L}^{-T}\mathbf{K}^T\bar{\mathbf{u}} - \beta_{p+1}\bar{\mathbf{v}}_p$; $\alpha_{p+1} \leftarrow \|\mathbf{q}\|$; $\bar{\mathbf{v}}_{p+1} \leftarrow (1/\alpha_{p+1})\,\mathbf{q}$;
  end if
end for

tion has an important equivalent formulation as

$$\begin{aligned} &\min_{\mathbf{x}} \left\|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}\right\|^2 \\ &\text{subject to } \|\mathbf{L}\mathbf{x}\| \le \varepsilon, \end{aligned} \tag{8.22}$$

where ε is a positive constant. The constrained least squares problem (8.22) can be solved by using the Lagrange multiplier formalism. Considering the Lagrangian function

$$\mathcal{L}(\mathbf{x}, \alpha) = \left\|\mathbf{y}^\delta - \mathbf{K}\mathbf{x}\right\|^2 + \alpha\left(\|\mathbf{L}\mathbf{x}\|^2 - \varepsilon^2\right),$$

it can be shown that if $\varepsilon \le \left\|\mathbf{L}\mathbf{x}^\delta\right\|$, where $\mathbf{x}^\delta$ is the least squares solution of the equation $\mathbf{K}\mathbf{x} = \mathbf{y}^\delta$, then the solution $\mathbf{x}_\varepsilon^\delta$ to (8.22) is identical to the Tikhonov solution $\mathbf{x}_\alpha^\delta$, with α solving the equation

$$\left\|\mathbf{L}\mathbf{x}_\alpha^\delta\right\|^2 = \varepsilon^2. \tag{8.23}$$

To carry this idea over to the TLS setting, we add the bound $\|\mathbf{L}\mathbf{x}\| \le \varepsilon$ to the ordinary problem (8.5), in which case, the new problem statement becomes

$$\begin{aligned} &\min_{\left[\tilde{\mathbf{K}}\ \tilde{\mathbf{y}}\right] \in \mathbb{R}^{m\times(n+1)}} \left\|\begin{bmatrix} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{bmatrix} - \begin{bmatrix} \tilde{\mathbf{K}} & \tilde{\mathbf{y}} \end{bmatrix}\right\|_{\mathrm{F}}^2 \\ &\text{subject to } \tilde{\mathbf{K}}\mathbf{x} = \tilde{\mathbf{y}} \text{ and } \|\mathbf{L}\mathbf{x}\| \le \varepsilon. \end{aligned} \tag{8.24}$$

The corresponding Lagrangian function is

$$\mathcal{L}\left(\tilde{\mathbf{K}}, \mathbf{x}, \alpha\right) = \left\| \left[\begin{array}{cc} \mathbf{K}_\Lambda & \mathbf{y}^\delta \end{array} \right] - \left[\begin{array}{cc} \tilde{\mathbf{K}} & \tilde{\mathbf{K}}\mathbf{x} \end{array} \right] \right\|_{\mathrm{F}}^2 + \alpha \left(\|\mathbf{L}\mathbf{x}\|^2 - \varepsilon^2 \right),$$

and the Lagrange multiplier α is non-zero if the inequality constraint is active. In fact, the solution $\mathbf{x}_{\Lambda\varepsilon}^\delta$ to (8.24) is different from the TLS solution $\mathbf{x}_\Lambda^\delta$, whenever ε is less than $\left\|\mathbf{L}\mathbf{x}_\Lambda^\delta\right\|$.

To characterize $\mathbf{x}_{\Lambda\varepsilon}^\delta$, we set the partial derivatives of the Lagrangian function to zero. Differentiation with respect to the entries in $\tilde{\mathbf{K}}$ yields

$$\tilde{\mathbf{K}} - \mathbf{K}_\Lambda - \mathbf{r}\mathbf{x}^T = \mathbf{0}, \tag{8.25}$$

with $\mathbf{r} = \mathbf{y}^\delta - \tilde{\mathbf{K}}\mathbf{x}$, while differentiation with respect to the entries in $\mathbf{x}$ gives

$$-\tilde{\mathbf{K}}^T\mathbf{r} + \alpha\mathbf{L}^T\mathbf{L}\mathbf{x} = \mathbf{0}. \tag{8.26}$$

Setting the partial derivative with respect to α to zero also yields

$$\|\mathbf{L}\mathbf{x}\|^2 = \varepsilon^2. \tag{8.27}$$

Making use of the expression of $\mathbf{r}$, we rearrange (8.26) as

$$\left(\tilde{\mathbf{K}}^T\tilde{\mathbf{K}} + \alpha\mathbf{L}^T\mathbf{L}\right)\mathbf{x} = \tilde{\mathbf{K}}^T\mathbf{y}^\delta. \tag{8.28}$$

Now, by (8.25) and (8.26), we have $\mathbf{K}_\Lambda = \tilde{\mathbf{K}} - \mathbf{r}\mathbf{x}^T$ and $\tilde{\mathbf{K}}^T\mathbf{r} = \alpha\mathbf{L}^T\mathbf{L}\mathbf{x}$, respectively, and so, we obtain

$$\mathbf{K}_\Lambda^T\mathbf{K}_\Lambda = \tilde{\mathbf{K}}^T\tilde{\mathbf{K}} - \alpha\mathbf{x}\mathbf{x}^T\mathbf{L}^T\mathbf{L} + \|\mathbf{r}\|^2\,\mathbf{x}\mathbf{x}^T - \alpha\mathbf{L}^T\mathbf{L}\mathbf{x}\mathbf{x}^T \tag{8.29}$$

and

$$\mathbf{K}_\Lambda^T\mathbf{y}^\delta = \tilde{\mathbf{K}}^T\mathbf{y}^\delta - \left(\mathbf{r}^T\mathbf{y}^\delta\right)\mathbf{x}. \tag{8.30}$$

Inserting (8.29) and (8.30) into (8.28), and using the identities (cf. (8.27))

$$\mathbf{x}\mathbf{x}^T\mathbf{L}^T\mathbf{L}\mathbf{x} = \varepsilon^2\mathbf{x}, \quad \|\mathbf{r}\|^2\,\mathbf{x}\mathbf{x}^T\mathbf{x} = \|\mathbf{r}\|^2\,\|\mathbf{x}\|^2\,\mathbf{x},$$

and

$$\mathbf{L}^T\mathbf{L}\mathbf{x}\mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|^2\,\mathbf{L}^T\mathbf{L}\mathbf{x},$$

we arrive at

$$\left(\mathbf{K}_\Lambda^T\mathbf{K}_\Lambda + \alpha_{\mathrm{I}}\mathbf{I}_n + \alpha_{\mathrm{L}}\mathbf{L}^T\mathbf{L}\right)\mathbf{x} = \mathbf{K}_\Lambda^T\mathbf{y}^\delta, \tag{8.31}$$

with

$$\alpha_{\mathrm{I}} = \alpha\varepsilon^2 - \|\mathbf{r}\|^2\,\|\mathbf{x}\|^2 - \mathbf{r}^T\mathbf{y}^\delta \tag{8.32}$$

and

$$\alpha_{\mathrm{L}} = \alpha\left(1 + \|\mathbf{x}\|^2\right). \tag{8.33}$$

The next step of our derivation is the elimination of the Lagrange multiplier α in the expressions of α_{I} and α_{L}. First, we use the relation (cf. (8.25))

$$\mathbf{r} = \mathbf{y}^\delta - \tilde{\mathbf{K}}\mathbf{x} = \mathbf{y}^\delta - \mathbf{K}_\Lambda\mathbf{x} - \|\mathbf{x}\|^2\,\mathbf{r},$$

to obtain

$$\left(1+\|\mathbf{x}\|^2\right)\mathbf{r}=\mathbf{y}^\delta-\mathbf{K}_\Lambda\mathbf{x}, \tag{8.34}$$

and further,

$$\left(1+\|\mathbf{x}\|^2\right)\|\mathbf{r}\|^2=\frac{\left\|\mathbf{y}^\delta-\mathbf{K}_\Lambda\mathbf{x}\right\|^2}{1+\|\mathbf{x}\|^2}. \tag{8.35}$$

On the other hand, scalar multiplying (8.26) by $\mathbf{x}$ gives

$$\alpha=\frac{\mathbf{x}^T\tilde{\mathbf{K}}^T\mathbf{r}}{\|\mathbf{L}\mathbf{x}\|^2}=\frac{1}{\varepsilon^2}\left(\mathbf{r}^T\mathbf{y}^\delta-\|\mathbf{r}\|^2\right). \tag{8.36}$$

Considering the parameter $\alpha_{\mathtt{I}}$, we insert (8.35) and (8.36) into (8.32), and find that

$$\alpha_{\mathtt{I}}=-\frac{\left\|\mathbf{y}^\delta-\mathbf{K}_\Lambda\mathbf{x}\right\|^2}{1+\|\mathbf{x}\|^2}. \tag{8.37}$$

Turning now to the parameter $\alpha_{\mathtt{L}}$, we use (8.33) and (8.36) to get

$$\alpha_{\mathtt{L}}=\alpha\left(1+\|\mathbf{x}\|^2\right)=\frac{1}{\varepsilon^2}\left(\mathbf{r}^T\mathbf{y}^\delta-\|\mathbf{r}\|^2\right)\left(1+\|\mathbf{x}\|^2\right). \tag{8.38}$$

Finally, a relationship connecting $\alpha_{\mathtt{L}}$ and $\alpha_{\mathtt{I}}$ can be derived as follows: by (8.35) and (8.37), we have $\alpha_{\mathtt{I}}=-\|\mathbf{r}\|^2\left(1+\|\mathbf{x}\|^2\right)$, whence using (8.34), (8.38) becomes

$$\alpha_{\mathtt{L}}=\frac{1}{\varepsilon^2}\left[\mathbf{y}^{\delta T}\left(\mathbf{y}^\delta-\mathbf{K}_\Lambda\mathbf{x}\right)+\alpha_{\mathtt{I}}\right]. \tag{8.39}$$

To evaluate the approximation error $\left\|\left[\begin{array}{cc}\mathbf{K}_\Lambda & \mathbf{y}^\delta\end{array}\right]-\left[\begin{array}{cc}\tilde{\mathbf{K}} & \tilde{\mathbf{y}}\end{array}\right]\right\|_{\mathtt{F}}$, we use the relation (cf. (8.25))

$$\left[\begin{array}{cc}\mathbf{K}_\Lambda & \mathbf{y}^\delta\end{array}\right]-\left[\begin{array}{cc}\tilde{\mathbf{K}} & \tilde{\mathbf{K}}\mathbf{x}\end{array}\right]=\left[\begin{array}{cc}\mathbf{K}_\Lambda-\tilde{\mathbf{K}} & \mathbf{r}\end{array}\right]=\left[\begin{array}{cc}-\mathbf{r}\mathbf{x}^T & \mathbf{r}\end{array}\right]=-\mathbf{r}\left[\begin{array}{c}\mathbf{x}\\ -1\end{array}\right]^T,$$

together with (8.35) and (8.37), to obtain

$$\left\|\left[\begin{array}{cc}\mathbf{K}_\Lambda & \mathbf{y}^\delta\end{array}\right]-\left[\begin{array}{cc}\tilde{\mathbf{K}} & \tilde{\mathbf{y}}\end{array}\right]\right\|_{\mathtt{F}}^2=\left(1+\|\mathbf{x}\|^2\right)\|\mathbf{r}\|^2=-\alpha_{\mathtt{I}}. \tag{8.40}$$

Collecting all results we conclude that $\mathbf{x}_{\Lambda\varepsilon}^\delta$ is the solution of equation (8.31) with $\alpha_{\mathtt{I}}$ and $\alpha_{\mathtt{L}}$ given by (8.37) and (8.39), respectively. The main features of the regularized TLS are presented below (Golub et al., 1999).

(1) If the matrix $\alpha_{\mathtt{I}}\mathbf{I}_n+\alpha_{\mathtt{L}}\mathbf{L}^T\mathbf{L}$ is positive definite, then the regularized TLS solution corresponds to the Tikhonov solution with the penalty term $\alpha_{\mathtt{I}}\|\mathbf{x}\|^2+\alpha_{\mathtt{L}}\|\mathbf{L}\mathbf{x}\|^2$. If the matrix $\alpha_{\mathtt{I}}\mathbf{I}_n+\alpha_{\mathtt{L}}\mathbf{L}^T\mathbf{L}$ is indefinite or negative definite, there is no equivalent interpretation.

(2) For a given ε, there are several pairs of parameters α_{I} and α_{L} and thus several solutions $\mathbf{x}_{\Lambda\varepsilon}^{\delta}$ that satisfy (8.31), (8.37) and (8.39). However, from (8.40), we see that only the solution with the smallest value of $|\alpha_{\mathrm{I}}|$ solves the constrained minimization problem (8.24).

(3) If $\varepsilon < \left\|\mathbf{L}\mathbf{x}_{\Lambda}^{\delta}\right\|$, where $\mathbf{x}_{\Lambda}^{\delta}$ is the TLS solution (8.10), the inequality constraint is binding, the Lagrange multiplier α is positive and by (8.33), it follows that $\alpha_{\mathrm{L}} > 0$. From (8.37) it is apparent that α_{I} is always negative and thus adds some deregularization to the solution. The residual (8.40) is a monotonically decreasing function of ε, and so, α_{I} is a monotonically increasing function of ε. If $\varepsilon = \left\|\mathbf{L}\mathbf{x}_{\Lambda}^{\delta}\right\|$, the Lagrange multiplier α is zero and the regularized TLS solution $\mathbf{x}_{\Lambda\varepsilon}^{\delta}$ coincides with the TLS solution $\mathbf{x}_{\Lambda}^{\delta}$; for larger ε, the constraint is never again binding and so, the solution remains unchanged.

To compute the regularized TLS solution $\mathbf{x}_{\Lambda\varepsilon}^{\delta}$ we have to solve a nonlinear problem, and several techniques have been proposed in the literature. In Golub et al. (1999), α_{L} is considered as free parameter, a corresponding value is computed for α_{I}, and the system of equations (8.31) is solved in an efficient way. The idea is to transform (8.31) into the augmented system of equations

$$\begin{bmatrix} \mathbf{I}_m & \mathbf{0} & \mathbf{K}_\Lambda \\ \mathbf{0} & \mathbf{I}_n & \sqrt{\alpha_{\mathrm{L}}}\mathbf{L} \\ \mathbf{K}_\Lambda^T & \sqrt{\alpha_{\mathrm{L}}}\mathbf{L}^T & -\alpha_{\mathrm{I}}\mathbf{I}_n \end{bmatrix} \begin{bmatrix} \mathbf{r} \\ \mathbf{s} \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{y}^\delta \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix},$$

to reduce $\mathbf{K}_\Lambda$ to an $n \times n$ bidiagonal form by means of orthogonal transformations, to apply Elden's algorithm to annihilate the matrix term containing the factor $\sqrt{\alpha_{\mathrm{L}}}$, and finally, to use a symmetric perfect shuffle reordering to obtain a symmetric, tridiagonal, indefinite matrix of size $2n \times 2n$ containing the parameter α_{I} on the main diagonal.

In Guo and Renault (2002), a shifted inverse power method is used to obtain the eigenpair

$$\left(\lambda, \begin{bmatrix} \mathbf{x} \\ -1 \end{bmatrix}\right)$$

for the problem

$$\mathbf{B}(\mathbf{x}) \begin{bmatrix} \mathbf{x} \\ -1 \end{bmatrix} = \lambda \begin{bmatrix} \mathbf{x} \\ -1 \end{bmatrix}, \tag{8.41}$$

where

$$\mathbf{B}(\mathbf{x}) = \begin{bmatrix} \mathbf{K}_\Lambda^T\mathbf{K}_\Lambda + \alpha_{\mathrm{L}}(\mathbf{x})\,\mathbf{L}^T\mathbf{L} & \mathbf{K}_\Lambda^T\mathbf{y}^\delta \\ \mathbf{y}^{\delta T}\mathbf{K}_\Lambda & -\alpha_{\mathrm{L}}(\mathbf{x})\,\varepsilon^2 + \mathbf{y}^{\delta T}\mathbf{y}^\delta \end{bmatrix}$$

is an $(n+1) \times (n+1)$ matrix, $\lambda = -\alpha_{\mathrm{I}}$, and α_{L} is given by (cf. (8.37) and (8.39))

$$\alpha_{\mathrm{L}}(\mathbf{x}) = \frac{1}{\varepsilon^2}\left[\mathbf{y}^{\delta T}\left(\mathbf{y}^\delta - \mathbf{K}_\Lambda\mathbf{x}\right) - \frac{\left\|\mathbf{y}^\delta - \mathbf{K}_\Lambda\mathbf{x}\right\|^2}{1 + \|\mathbf{x}\|^2}\right]. \tag{8.42}$$

In Renault and Guo (2005), the solution of the eigenproblem (8.41) is considered together with the solution of a nonlinear equation which guarantees the bound $\|\mathbf{L}\mathbf{x}\| = \varepsilon$. To describe the main features of this algorithm, we consider the decomposition

$$\mathbf{B}(\alpha_{\mathrm{L}}) = \mathbf{M} + \alpha_{\mathrm{L}}\mathbf{N},$$

where

$$\mathbf{M} = \begin{bmatrix} \mathbf{K}_\Lambda^T \mathbf{K}_\Lambda & \mathbf{K}_\Lambda^T \mathbf{y}^\delta \\ \mathbf{y}^{\delta T} \mathbf{K}_\Lambda & \mathbf{y}^{\delta T} \mathbf{y}^\delta \end{bmatrix}, \quad \mathbf{N} = \begin{bmatrix} \mathbf{L}^T \mathbf{L} & \mathbf{0} \\ \mathbf{0} & -\varepsilon^2 \end{bmatrix},$$

and denote by

$$\left(\lambda_{\alpha_{\mathrm{L}}}, \begin{bmatrix} \mathbf{x}_{\alpha_{\mathrm{L}}} \\ -1 \end{bmatrix} \right)$$

the eigenpair corresponding to the smallest eigenvalue of $\mathbf{B}(\alpha_{\mathrm{L}})$. For a fixed ε, we introduce the function

$$g(\mathbf{x}) = \frac{\|\mathbf{L}\mathbf{x}\|^2 - \varepsilon^2}{1 + \|\mathbf{x}\|^2},$$

and compute $\hat{\alpha}_{\mathrm{L}}$ such that $\mathbf{x}_{\hat{\alpha}_{\mathrm{L}}}$ solves the equation

$$g(\mathbf{x}_{\alpha_{\mathrm{L}}}) = 0; \tag{8.43}$$

$\mathbf{x}_{\hat{\alpha}_{\mathrm{L}}}$ is then the regularized TLS solution of (8.24). To justify this algorithm, we assume that $\mathbf{x}_{\hat{\alpha}_{\mathrm{L}}}$ satisfies the eigensystem equation

$$\mathbf{B}(\hat{\alpha}_{\mathrm{L}}) \begin{bmatrix} \mathbf{x}_{\hat{\alpha}_{\mathrm{L}}} \\ -1 \end{bmatrix} = \lambda_{\hat{\alpha}_{\mathrm{L}}} \begin{bmatrix} \mathbf{x}_{\hat{\alpha}_{\mathrm{L}}} \\ -1 \end{bmatrix}, \tag{8.44}$$

and is also a solution of equation (8.43). The first block equation of the eigenvalue problem (8.44) gives (8.31) with $\hat{\alpha}_{\mathrm{I}} = -\lambda_{\hat{\alpha}_{\mathrm{L}}}$, while the second block equation yields (8.39). Multiplying the eigensystem equation by $[\mathbf{x}_{\hat{\alpha}_{\mathrm{L}}}^T, -1]$, we find that

$$\lambda_{\hat{\alpha}_{\mathrm{L}}} = \frac{1}{1 + \|\mathbf{x}_{\hat{\alpha}_{\mathrm{L}}}\|^2} \left[\left\| \mathbf{y}^\delta - \mathbf{K}_\Lambda \mathbf{x}_{\hat{\alpha}_{\mathrm{L}}} \right\|^2 + \hat{\alpha}_{\mathrm{L}} \left(\|\mathbf{L}\mathbf{x}_{\hat{\alpha}_{\mathrm{L}}}\|^2 - \varepsilon^2 \right) \right]. \tag{8.45}$$

Since $g(\mathbf{x}_{\hat{\alpha}_{\mathrm{L}}}) = 0$, it follows that $\|\mathbf{L}\mathbf{x}_{\hat{\alpha}_{\mathrm{L}}}\|^2 = \varepsilon^2$, and (8.45) becomes

$$\lambda_{\hat{\alpha}_{\mathrm{L}}} = \frac{\left\| \mathbf{y}^\delta - \mathbf{K}_\Lambda \mathbf{x}_{\hat{\alpha}_{\mathrm{L}}} \right\|^2}{1 + \|\mathbf{x}_{\hat{\alpha}_{\mathrm{L}}}\|^2}; \tag{8.46}$$

thus $\hat{\alpha}_{\mathrm{I}} = -\lambda_{\hat{\alpha}_{\mathrm{L}}}$ satisfies indeed (8.37). In summary, $\mathbf{x}_{\hat{\alpha}_{\mathrm{L}}}$solves equation (8.31) with $\hat{\alpha}_{\mathrm{I}}$ as in (8.37) and $\hat{\alpha}_{\mathrm{L}}$ as in (8.39). Since $\lambda_{\hat{\alpha}_{\mathrm{L}}}$ is the smallest eigenvalue of $\mathbf{B}$, the present approach explicitly computes a solution with the smallest value of $|\alpha_{\mathrm{I}}|$.

For a practical implementation of the method of Renault and Guo we note the following results:

(1) if $\lambda_{n+1} > 0$ is the smallest eigenvalue of the matrix $\mathbf{B}$ and $\mathbf{v}_{n+1}$ is the corresponding eigenvector, then $\lambda_{\alpha_{\mathrm{L}}} = \lambda_{n+1}$ and

$$\begin{bmatrix} \mathbf{x}_{\alpha_{\mathrm{L}}} \\ -1 \end{bmatrix} = -\frac{1}{[\mathbf{v}_{n+1}]_{n+1}} \mathbf{v}_{n+1};$$

(2) $g(\mathbf{x}_{\alpha_{\mathrm{L}}})$ is a monotonically decreasing function of α_{L}, and there exists only one solution $\hat{\alpha}_{\mathrm{L}}$ of the equation $g(\mathbf{x}_{\alpha_{\mathrm{L}}}) = 0$.

Algorithm 15 computes the Tikhonov solution and the regularized TLS solution for a fixed value of the parameter α corresponding to the method of Tikhonov regularization Both solutions are related to each other through the constraint norms. The input parameter α is used to determine the bound ε and to estimate a bisection interval for $\alpha_{\mathtt{L}}$. The algorithm also computes the 'equivalent' regularization matrix defined as

$$\alpha \mathbf{L}_{\mathtt{eq}}^T \mathbf{L}_{\mathtt{eq}} = \hat{\alpha}_{\mathtt{I}} \mathbf{I}_n + \hat{\alpha}_{\mathtt{L}} \mathbf{L}^T \mathbf{L}. \tag{8.47}$$

This factorization is performed by using the Cholesky method with added multiple of identity, which takes into account that for large negative values of $\hat{\alpha}_{\mathtt{I}}$, the matrix $\hat{\alpha}_{\mathtt{I}} \mathbf{I}_n + \hat{\alpha}_{\mathtt{L}} \mathbf{L}^T \mathbf{L}$ may not be positive definite. Note that strategies based on modifying a Cholesky factorization or a symmetric indefinite factorization of a non-positive definite Hessian are standard approaches in the framework of Newton's method (Nocedal and Wright, 2006).

Algorithm 15. Algorithm for computing the regularized TLS solution by solving the eigenvalue problem (8.41). The regularization parameter α corresponds to the method of Tikhonov regularization. The algorithm computes the solution $\mathbf{x}_{\hat{\alpha}_{\mathtt{L}}}$, the regularization parameters $\hat{\alpha}_{\mathtt{L}}$ and $\hat{\alpha}_{\mathtt{I}}$, and the equivalent regularization matrix $\mathbf{L}_{\mathtt{eq}}$.

compute the Tikhonov solution $\mathbf{x}_\alpha^\delta$ for α, i.e., $\mathbf{x}_\alpha^\delta = \left(\mathbf{K}_\Lambda^T \mathbf{K}_\Lambda + \alpha \mathbf{L}^T \mathbf{L}\right)^{-1} \mathbf{K}_\Lambda^T \mathbf{y}^\delta$;
$\varepsilon \leftarrow \left\| \mathbf{L} \mathbf{x}_\alpha^\delta \right\|$;
compute the matrices $\mathbf{M}$ and $\mathbf{N}$;
estimate a bisection interval $[\alpha_{\mathtt{L}\min}, \alpha_{\mathtt{L}\max}]$ for $\alpha_{\mathtt{L}}$ around α;
solve $g(\alpha_{\mathtt{L}}) = 0$ in $[\alpha_{\mathtt{L}\min}, \alpha_{\mathtt{L}\max}]$ using $FuncEval(\alpha_{\mathtt{L}}, \varepsilon, \mathbf{M}, \mathbf{N};\ g, \mathbf{x}_{\alpha_{\mathtt{L}}}, \alpha_{\mathtt{I}})$;
store the solution $\hat{\alpha}_{\mathtt{L}}$ and the corresponding $\mathbf{x}_{\hat{\alpha}_{\mathtt{L}}}$ and $\hat{\alpha}_{\mathtt{I}}$;
{regularization matrix using Cholesky factorization with added multiple of identity}
choose the tolerance ε_α, e.g., $\varepsilon_\alpha = 0.001$;
$\triangle\alpha \leftarrow \varepsilon_\alpha \left|\hat{\alpha}_{\mathtt{I}}\right|$; $stop \leftarrow$ false;
while $stop =$ false **do**
    attempt to apply the Cholesky factorization to obtain $\mathbf{L}_{\mathtt{eq}}^T \mathbf{L}_{\mathtt{eq}} = \hat{\alpha}_{\mathtt{I}} \mathbf{I}_n + \hat{\alpha}_{\mathtt{L}} \mathbf{L}^T \mathbf{L}$;
    if factorization is successful **then**
        $stop \leftarrow$ true;
    else
        $\hat{\alpha}_{\mathtt{I}} \leftarrow \hat{\alpha}_{\mathtt{I}} + \triangle\alpha$;
    end if
end while
$\mathbf{L}_{\mathtt{eq}} \leftarrow (1/\sqrt{\alpha})\, \mathbf{L}_{\mathtt{eq}}$.

{for given $\alpha_{\mathtt{L}}$, the routine computes $g(\alpha_{\mathtt{L}})$, $\mathbf{x}_{\alpha_{\mathtt{L}}}$ and $\alpha_{\mathtt{I}}$}
routine $FuncEval(\alpha_{\mathtt{L}}, \varepsilon, \mathbf{M}, \mathbf{N};\ g, \mathbf{x}_{\alpha_{\mathtt{L}}}, \alpha_{\mathtt{I}})$
$\mathbf{B} \leftarrow \mathbf{M} + \alpha_{\mathtt{L}} \mathbf{N}$;
compute the smallest eigenvalue λ_{n+1} of $\mathbf{B}$ and the eigenvector $\mathbf{v}_{n+1}$;
compute $\mathbf{x}_{\alpha_{\mathtt{L}}}$ as $\begin{bmatrix} \mathbf{x}_{\alpha_{\mathtt{L}}} \\ -1 \end{bmatrix} = -\left(1/\left[\mathbf{v}_{n+1}\right]_{n+1}\right) \mathbf{v}_{n+1}$;
$\alpha_{\mathtt{I}} \leftarrow -\lambda_{n+1}$;
$g \leftarrow \left(\left\|\mathbf{L}\mathbf{x}_{\alpha_{\mathtt{L}}}\right\|^2 - \varepsilon^2\right) / \left(1 + \left\|\mathbf{x}_{\alpha_{\mathtt{L}}}\right\|^2\right)$.

In Sima et al. (2003), the objective function is the so-called orthogonal distance, and the constrained minimization problem takes the form (cf. (8.37) and (8.40))

$$\min_{\mathbf{x}} \frac{\left\|\mathbf{y}^{\delta}-\mathbf{K}_{\Lambda}\mathbf{x}\right\|^{2}}{1+\|\mathbf{x}\|^{2}} \quad \text{subject to } \|\mathbf{L}\mathbf{x}\| \leq \varepsilon.$$

The first-order optimality conditions for the Lagrangian function

$$\mathcal{L}(\mathbf{x},\lambda)=\frac{\left\|\mathbf{y}^{\delta}-\mathbf{K}_{\Lambda}\mathbf{x}\right\|^{2}}{1+\|\mathbf{x}\|^{2}}+\lambda\left(\|\mathbf{L}\mathbf{x}\|^{2}-\varepsilon^{2}\right),$$

yield

$$\mathbf{D}(\mathbf{x})\,\mathbf{x}+\lambda\mathbf{L}^{T}\mathbf{L}\mathbf{x}=\mathbf{d}(\mathbf{x}),\ \ \|\mathbf{L}\mathbf{x}\|^{2}=\varepsilon^{2}, \tag{8.48}$$

with

$$\mathbf{D}(\mathbf{x})=\frac{\mathbf{K}_{\Lambda}^{T}\mathbf{K}_{\Lambda}}{1+\|\mathbf{x}\|^{2}}-\frac{\left\|\mathbf{y}^{\delta}-\mathbf{K}_{\Lambda}\mathbf{x}\right\|^{2}}{\left(1+\|\mathbf{x}\|^{2}\right)^{2}}\mathbf{I}_{n},\ \ \mathbf{d}(\mathbf{x})=\frac{\mathbf{K}_{\Lambda}^{T}\mathbf{y}^{\delta}}{1+\|\mathbf{x}\|^{2}}.$$

The problem (8.48) is first transformed into the standard form and then solved iteratively by using a fixed point iteration method. Assuming that $\mathbf{L}$ is square and nonsingular, the transformation to the standard form gives

$$(\mathbf{W}+\lambda\mathbf{I}_{n})\,\bar{\mathbf{x}}=\mathbf{h},\ \ \|\bar{\mathbf{x}}\|^{2}=\varepsilon^{2}, \tag{8.49}$$

with $\bar{\mathbf{x}}=\mathbf{L}\mathbf{x}$, $\mathbf{W}=\mathbf{L}^{-T}\mathbf{D}\mathbf{L}^{-1}$ and $\mathbf{h}=\mathbf{L}^{-T}\mathbf{d}$. Note that since $\mathbf{D}$ is a symmetric matrix, $\mathbf{W}$ is also a symmetric matrix. Let us now consider the problem

$$(\mathbf{W}+\lambda\mathbf{I}_{n})^{2}\,\mathbf{u}=\mathbf{h},\ \ \mathbf{h}^{T}\mathbf{u}=\varepsilon^{2} \tag{8.50}$$

for $\mathbf{u}\in\mathbb{R}^{n}$. Setting

$$\bar{\mathbf{x}}=(\mathbf{W}+\lambda\mathbf{I}_{n})\,\mathbf{u},$$

and taking into account that, due to the symmetry of $\mathbf{W}+\lambda\mathbf{I}_{n}$, there holds

$$\varepsilon^{2}=\mathbf{h}^{T}\mathbf{u}=\mathbf{u}^{T}(\mathbf{W}+\lambda\mathbf{I}_{n})^{2}\,\mathbf{u}=\|\bar{\mathbf{x}}\|^{2},$$

we see that the problems (8.49) and (8.50) are equivalent. Further, using the identity

$$\mathbf{h}=\frac{1}{\varepsilon^{2}}\left(\mathbf{h}^{T}\mathbf{u}\right)\mathbf{h}=\frac{1}{\varepsilon^{2}}\mathbf{h}\mathbf{h}^{T}\mathbf{u},$$

we deduce that (8.50) can be transformed into the quadratic eigenvalue problem

$$\left(\lambda^{2}\mathbf{I}_{n}+2\lambda\mathbf{W}+\mathbf{W}^{2}-\frac{1}{\varepsilon^{2}}\mathbf{h}\mathbf{h}^{T}\right)\mathbf{u}=\mathbf{0}. \tag{8.51}$$

This quadratic eigenvalue problem is solved in order to find the largest eigenvalue λ and the corresponding eigenvector $\mathbf{u}$ scaled so that $\mathbf{h}^{T}\mathbf{u}=\varepsilon^{2}$. As all matrices in (8.51) are

real and symmetric, the quadratic eigenvalues are real and come in complex conjugate pairs. Moreover, the special form of the quadratic eigenvalue problem (8.51) implies that the rightmost (largest real) eigenvalue is real and positive. The solution of the original problem is then recovered by first computing $\bar{\mathbf{x}} = (\mathbf{W} + \lambda \mathbf{I}_n)\,\mathbf{u}$ and then $\mathbf{x} = \mathbf{L}^{-1}\bar{\mathbf{x}}$.

Algorithm 16. Algorithm for computing the regularized TLS solution by solving the quadratic eigenvalue problem (8.51). The regularization parameter α corresponds to the method of Tikhonov regularization. The algorithm computes the solution $\mathbf{x}$, the regularization parameters $\alpha_{\mathtt{L}}$ and $\alpha_{\mathtt{I}}$, and the equivalent regularization matrix $\mathbf{L}_{\mathtt{eq}}$.

---

choose the tolerances $\varepsilon_{\mathtt{1}}$ and $\varepsilon_{\mathtt{x}}$ for the convergence test;
compute the Tikhonov solution $\mathbf{x}_\alpha^\delta$ for α, i.e., $\mathbf{x}_\alpha^\delta = \left(\mathbf{K}_\Lambda^T\mathbf{K}_\Lambda + \alpha\mathbf{L}^T\mathbf{L}\right)^{-1}\mathbf{K}_\Lambda^T\mathbf{y}^\delta$;
$\varepsilon \leftarrow \left\|\mathbf{L}\mathbf{x}_\alpha^\delta\right\|$;
$\bar{\mathbf{K}}_\Lambda \leftarrow \mathbf{K}_\Lambda\mathbf{L}^{-1}$;
$stop \leftarrow$ false; $k \leftarrow 0$; $\mathbf{x} \leftarrow \mathbf{x}_\alpha^\delta$; {starting vector}
while $stop =$ false **do**
  $r \leftarrow \left\|\mathbf{y}^\delta - \mathbf{K}_\Lambda\mathbf{x}\right\|^2 / \left(1 + \|\mathbf{x}\|^2\right)$; $c \leftarrow 1/\left(1 + \|\mathbf{x}\|^2\right)$;
  $\mathbf{W} \leftarrow c\bar{\mathbf{K}}_\Lambda^T\bar{\mathbf{K}}_\Lambda - rc\mathbf{L}^{-T}\mathbf{L}^{-1}$; $\mathbf{h} \leftarrow c\bar{\mathbf{K}}_\Lambda^T\mathbf{y}^\delta$;
  set $\mathbf{A} = \begin{bmatrix} -2\mathbf{W} & -\mathbf{W}^2 + \varepsilon^{-2}\mathbf{h}\mathbf{h}^T \\ \mathbf{I}_n & \mathbf{0} \end{bmatrix}$;
  compute the largest eigenvalue λ and the corresponding eigenvector $\begin{bmatrix} \mathbf{v} \\ \mathbf{u} \end{bmatrix}$ of $\mathbf{A}$;
  $\mathbf{u} \leftarrow \left(\varepsilon^2/\mathbf{h}^T\mathbf{u}\right)\mathbf{u}$; {scale $\mathbf{u}$}
  $\mathbf{W} \leftarrow \mathbf{W} + \lambda\mathbf{I}_n$;
  $\mathbf{x} \leftarrow \mathbf{L}^{-1}\mathbf{W}\mathbf{u}$;
  {convergence test}
  if $k > 0$ **and** $\left|\lambda - \lambda_{\mathtt{prv}}\right| \le \varepsilon_{\mathtt{1}}\lambda$ **and** $\left\|\mathbf{x} - \mathbf{x}_{\mathtt{prv}}\right\| \le \varepsilon_{\mathtt{x}}\|\mathbf{x}\|$ **then**
    $stop \leftarrow$ true;
  else
    $\lambda_{\mathtt{prv}} \leftarrow \lambda$; $\mathbf{x}_{\mathtt{prv}} \leftarrow \mathbf{x}$;
    $k \leftarrow k + 1$;
  end if
end while
$\alpha_{\mathtt{L}} \leftarrow \lambda\left(1 + \|\mathbf{x}\|^2\right)$; $\alpha_{\mathtt{I}} \leftarrow -\left\|\mathbf{y}^\delta - \mathbf{K}_\Lambda\mathbf{x}\right\|^2 / \left(1 + \|\mathbf{x}\|^2\right)$;
compute $\mathbf{L}_{\mathtt{eq}}$ as in Algorithm 15

The quadratic eigenvalue problem (8.51) is equivalent to the linear eigenvalue problem

$$\begin{bmatrix} -2\mathbf{W} & -\mathbf{W}^2 + \frac{1}{\varepsilon^2}\mathbf{h}\mathbf{h}^T \\ \mathbf{I}_n & \mathbf{0} \end{bmatrix}\begin{bmatrix} \mathbf{v} \\ \mathbf{u} \end{bmatrix} = \lambda\begin{bmatrix} \mathbf{v} \\ \mathbf{u} \end{bmatrix},$$

and this can be solved by using for example, the routine DGEEV from the LAPACK library (Anderson et al., 1995), or the routine DNAUPD from the ARPACK library (Maschhoff and Sorensen, 1996). The DNAUPD routine is more efficient because it calculates only the largest eigenvalue and the corresponding eigenvector by using Arnoldi's method (Arnoldi,

1951). The Algorithm 16 generates a sequence $\{(\lambda_k, \mathbf{x}_k)\}$ by solving the quadratic eigenvalue problem (8.51) at each iteration step k. From the analysis of Sima et al. (2003) we infer the following results:

(1) $\mathbf{x}_k$ should correspond to the largest eigenvalue $\lambda_k > 0$ since only then the algorithm converges;
(2) the orthogonal distance decreases at each iteration step;
(3) any limit point of the sequence $\{(\lambda_k, \mathbf{x}_k)\}$ solves equation (8.48) .

The last result suggests that instead of requiring the convergence of the sequence $\{(\lambda_k, \mathbf{x}_k)\}$ we may check if equation (8.48) is satisfied within a prescribed tolerance at each iteration step.

8.4 Regularized total least squares for nonlinear problems

As stated in Chapter 6, the solution of a nonlinear ill-posed problem by means of Tikhonov regularization is equivalent to the solution of a sequence of ill-posed linearizations of the forward model about the current iterate. Essentially, at the iteration step k, we solve the linearized equation

$$\mathbf{K}_{\alpha k} \triangle \mathbf{x} = \mathbf{y}_k^{\delta}, \tag{8.52}$$

with $\triangle \mathbf{x} = \mathbf{x} - \mathbf{x}_{\mathtt{a}}$, $\mathbf{K}_{\alpha k} = \mathbf{K}\left(\mathbf{x}_{\alpha k}^{\delta}\right)$, and

$$\mathbf{y}_k^{\delta} = \mathbf{y}^{\delta} - \mathbf{F}\left(\mathbf{x}_{\alpha k}^{\delta}\right) + \mathbf{K}_{\alpha k}\left(\mathbf{x}_{\alpha k}^{\delta} - \mathbf{x}_{\mathtt{a}}\right),$$

via Tikhonov regularization with the penalty term $\|\mathbf{L}\triangle\mathbf{x}\|^2$ and the regularization parameter α. If $\triangle\mathbf{x}_{\alpha k}^{\delta}$ is the minimizer of the Tikhonov function

$$\mathcal{F}_{1\alpha k}\left(\triangle\mathbf{x}\right) = \left\|\mathbf{y}_k^{\delta} - \mathbf{K}_{\alpha k}\triangle\mathbf{x}\right\|^2 + \alpha\left\|\mathbf{L}\triangle\mathbf{x}\right\|^2, \tag{8.53}$$

the new iterate is given by $\mathbf{x}_{\alpha k+1}^{\delta} = \mathbf{x}_{\mathtt{a}} + \triangle\mathbf{x}_{\alpha k}^{\delta}$, and the constraint norm can be readily computed as

$$\varepsilon = \left\|\mathbf{L}\triangle\mathbf{x}_{\alpha k}^{\delta}\right\|. \tag{8.54}$$

In the framework of the regularized TLS, we assume that $\mathbf{K}_{\alpha k}$ is contaminated by errors, and instead of minimizing (8.53) we solve the problem

$$\min_{\left[\tilde{\mathbf{K}}\ \tilde{\mathbf{y}}\right]\in\mathbb{R}^{m\times(n+1)}} \left\|\left[\begin{array}{cc}\mathbf{K}_{\alpha k} & \mathbf{y}_k^{\delta}\end{array}\right] - \left[\begin{array}{cc}\tilde{\mathbf{K}} & \tilde{\mathbf{y}}\end{array}\right]\right\|_{\mathtt{F}}^2 \tag{8.55}$$

$$\text{subject to } \tilde{\mathbf{K}}\triangle\mathbf{x} = \tilde{\mathbf{y}} \text{ and } \|\mathbf{L}\triangle\mathbf{x}\| \leq \varepsilon,$$

with ε being given by (8.54). The free parameter of the method is the Tikhonov regularization parameter α, and the Algorithms 15 and 16 can be used to compute both the Tikhonov solution and the regularized TLS solution. Although the numerical implementation of the regularized TLS is very similar to that of Tikhonov regularization, the use of a step-length procedure is problematic. In principle it can be applied for the objective function

$$\mathcal{F}_{\alpha}\left(\mathbf{x}\right) = \frac{1}{2}\left\|\mathbf{f}_{\alpha}\left(\mathbf{x}\right)\right\|^2, \quad \mathbf{f}_{\alpha}\left(\mathbf{x}\right) = \left[\begin{array}{c}\mathbf{F}\left(\mathbf{x}\right) - \mathbf{y}^{\delta} \\ \sqrt{\alpha}\mathbf{L}_{\mathtt{eq}}\left(\mathbf{x} - \mathbf{x}_{\mathtt{a}}\right)\end{array}\right], \tag{8.56}$$

but solving (8.55) is not equivalent to minimizing (8.56) at the iteration step k because $\mathbf{L}_{\mathtt{eq}}$ may not be the exact Cholesky factor of $\hat{\alpha}_{\mathrm{I}}\mathbf{I}_n + \hat{\alpha}_{\mathrm{L}}\mathbf{L}^T\mathbf{L}$ (cf. (8.47)).

In our numerical analysis, we consider the O_3 retrieval test problem and compute the Jacobian matrix $\mathbf{K}_{\mathtt{ss}}$ by assuming only the single scattering contribution (cf. (8.1)). Furthermore, at each iteration step, we perturb this matrix as

$$[\mathbf{K}_{k\alpha}]_{ij} = [\mathbf{K}_{\mathtt{ss}k\alpha}]_{ij} + \sigma_\Lambda \varepsilon_{ij} [\mathbf{K}_{\mathtt{ss}k\alpha}]_{ij} ,$$

where the elements ε_{ij} are from a normal distribution with zero mean and unit variance. Figure 8.1 shows the relative errors in the Tikhonov and the regularized TLS solutions for four values of the standard deviation σ_Λ, namely 0, 0.01, 0.02 and 0.03. In all situations, the minimum solution error for the regularized TLS is clearly smaller than that for Tikhonov regularization. Even in the case $\sigma_\Lambda = 0$ there is a solution improvement due to the approximate Jacobian calculation. The plots also show that the minima of the TLS errors are flat and this situation is beneficial for the inversion process.

In Figure 8.2 we plot the Tikhonov and the regularized TLS solutions, corresponding to the minimizers of the error curves in Figure 8.1. In fact, the improvement of the TLS error as compared to the Tikhonov error is due to the additional term $\alpha_{\mathrm{I}}\mathbf{I}_n$ in Eq. (8.31).

From the point of view of their accuracy, the regularized TLS algorithms solving the eigenvalue problem (8.41) and the quadratic eigenvalue problem (8.51) are completely

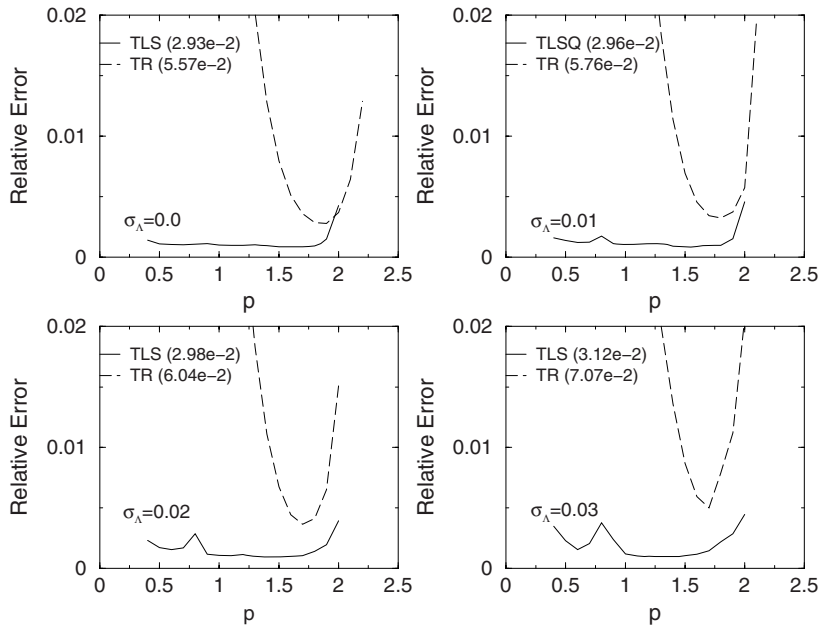

Fig. 8.1. Relative errors in the Tikhonov and the regularized TLS solutions as a function of the exponent p, where $\alpha = \sigma^p$ and σ is the noise standard deviation. The results correspond to the O_3 retrieval test problem and are computed with the regularized TLS algorithm solving the quadratic eigenvalue problem (8.51). The numbers in parentheses indicate the minimum values of the relative solution error.

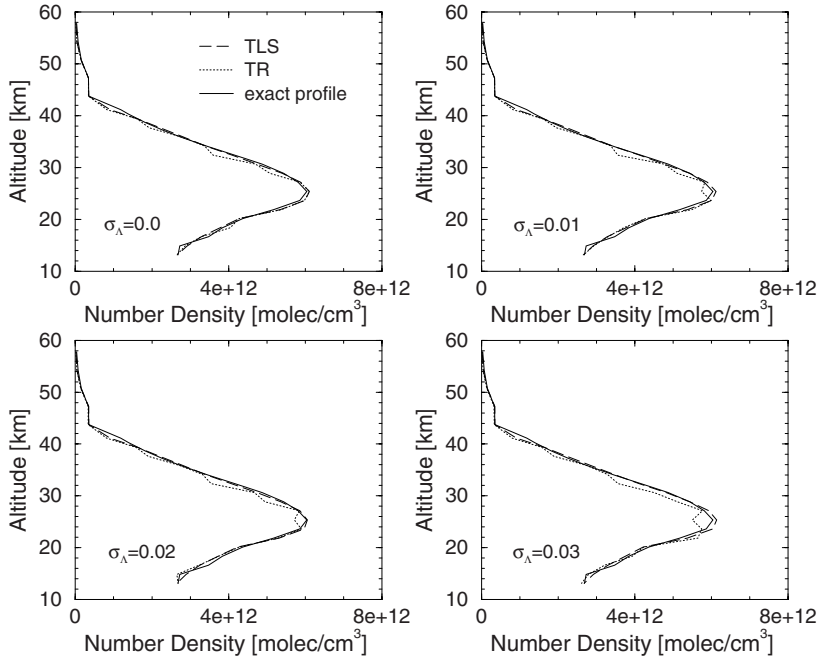

Fig. 8.2. Tikhonov (TR) and regularized TLS solutions corresponding to the minimizers of the error curves in Figure 8.1.

equivalent. However, the computation time of the algorithm based on a quadratic eigenvalue problem is on average 6 times smaller (Table 8.1). The main drawback of the regularized TLS is the extraordinarily large number of iteration steps (and so, computation time) as compared to Tikhonov regularization. The decrease of the solution error by a factor of 4–5 is accompanied by an increase of the computation time by a factor of 7–8.

The large number of iteration steps is also a consequence of the fact that we do not use a step-length procedure to guarantee a monotonic decrease of the residual norm (Figure 8.3). A step-length algorithm stops the iterative process too early (because the search direction is not a descent direction for the Tikhonov function), and as a result, the solu-

Table 8.1. Computation time in min:ss format. The numbers in parentheses indicate the number of iteration steps for Tikhonov regularization (TR) and the regularized TLS algorithms solving the eigenvalue problem (8.41) (TLS-EP) and the quadratic eigenvalue problem (8.51) (TLS-QEP).

	Standard deviation σ_Λ			
Method	0	0.01	0.02	0.03
TR	0:14 (4)	0:15 (6)	0:18 (8)	0:24 (16)
TLS-QEP	1:24 (108)	1:37 (124)	2:23 (202)	2:58 (243)
TLS-EP	8:01 (108)	9:57 (124)	13:13 (202)	19:17 (243)

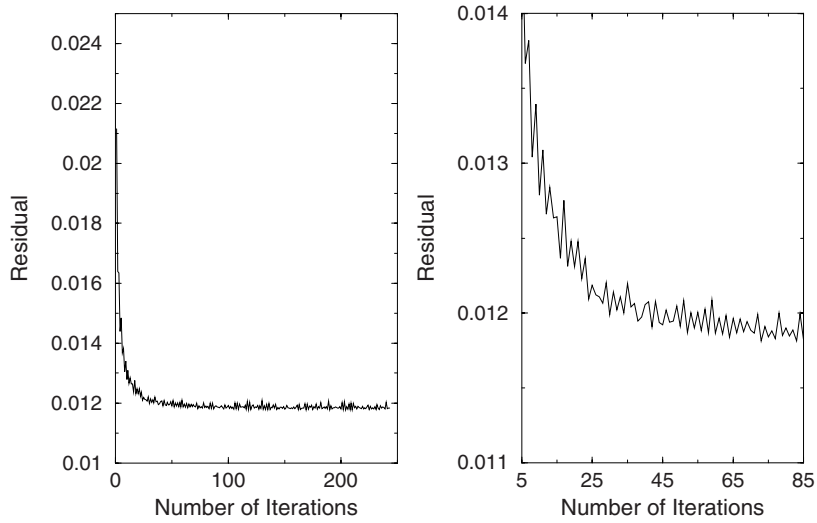

Fig. 8.3. History of the residual norm in the case $\sigma_\Lambda = 0.03$. In the left panel the curves are plotted for all iteration steps, while in the right panel, the y-axis is zoomed out.

tion errors are not sufficiently small. For example, in the case $\sigma_\Lambda = 0.03$, the regularized TLS with a step-length algorithm terminates after 19 iteration steps with a solution error of $1.56 \cdot 10^{-2}$, and if the step-length algorithm is turned off, it terminates after 243 iteration steps with a solution error of $9.77 \cdot 10^{-4}$.

The design of an efficient regularized TLS algorithm for nonlinear problems is far from being complete. The selection of an optimal value of the regularization parameter by an a posteriori method will dramatically increase the computational effort, while the use of a variable regularization parameter computed for example, by using the L-curve method, is also problematic. In our numerical simulations, the L-curve either does not have a distinctive L-shape, or it predicts values of the regularization parameter that are too small.

The regularized TLS has been applied to atmospheric trace gas profile retrievals by Koner and Drummond (2008). In this work, the regularized TLS algorithm solving the quadratic eigenvalue problem (8.51) is used for the automatic determination of the regularization strength.