Matrices and Moments: Least Squares Problems

Gene Golub, SungEun Jo, Zheng Su

Comparative study (with David Gleich)

Computer Science Department
Stanford University

Outline of Part I

Least Squares Problems

- Ordinary/Data/Total least squares
- SVD solutions
- Secular equation approaches
- “Inverse” least squares

Gauss Quadrature Rules

- Gauss quadrature theory
- Tri-diagonalization for orthonormal polynomials
- Inverse eigenvalue problem for Gauss-Radau rule

Outline of Part II

Interpolating secular equations

- A common term of secular functions in TLS/DLS
- Variations of Newton's methods
- Seeking the smallest root

Conjugate Gradient method with Quadrature (CGQ)

- Solving TLS and DLS by quadrature and CG method
- Alternative implementations

Comparative study (with David Gleich)

- Large-scale TLS algorithms
- Comparison of large-scale TLS results

Numerical example

- CG-based algorithms

Conclusion

Part I

Review

Approximation problem

- Approximation problems for a linear system:

$$\mathbf{A}\mathbf{x} \approx \mathbf{b}, \quad \mathbf{A} \in \mathbb{R}^{m\times n}, \quad \mathbf{b} \in \mathbb{R}^{m\times 1}, \quad m > n.$$

- Notations:

$\mathbf{A}$ and $\mathbf{b}$	given data
$\mathbf{x}$	solution to determine
^{T}	transpose
$\mathrm{Tr}[\cdot]$	the sum of diagonal entries of matrix
$\|\cdot\|_2$	two-norm of vector
$\|\mathbf{A}\|_F = \sqrt{\mathrm{Tr}[\mathbf{A}^T\mathbf{A}]}$	Frobenius norm of matrix
$\|\cdot\| \equiv \|\cdot\|_2$ or $\|\cdot\|_F$	Euclidean norm
$\Delta\mathbf{A}$ and $\Delta\mathbf{b}$	residual quantities
$_{\text{OLS}}/_{\text{DLS}}/_{\text{TLS}}$	ordinary/data/total least squares

Statements and geometric equivalences

Table: Problem statements and geometric equivalent statements

	Problem statement			[a] Geometric equivalence
OLS	$\min_{\mathbf{x},\Delta\mathbf{b}} \|\Delta\mathbf{b}\|_2$	s.t.	$\mathbf{A}\mathbf{x} = \mathbf{b} + \Delta\mathbf{b}$	$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$
TLS	$\min_{\mathbf{x},\Delta\mathbf{A},\Delta\mathbf{b}} \|[\Delta\mathbf{A},\Delta\mathbf{b}]\|_F$	s.t.	$(\mathbf{A}+\Delta\mathbf{A})\mathbf{x} = \mathbf{b} + \Delta\mathbf{b}$	$\min_{\mathbf{x}} \frac{\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2}{\|\mathbf{x}\|_2^2 + 1}$
DLS	$\min_{\mathbf{x},\Delta\mathbf{A}} \|\Delta\mathbf{A}\|_F$	s.t.	$(\mathbf{A}+\Delta\mathbf{A})\mathbf{x} = \mathbf{b}$	$\min_{\mathbf{x}} \frac{\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2}{\|\mathbf{x}\|_2^2}$

[a] The TLS/DLS equivalent statements are derived by means of the Lagrange method [GV96, DD93, JK05].

- Ordinary Least Squares (OLS): correcting with $\Delta\mathbf{b}$
- Data Least Squares (DLS): correcting with $\Delta\mathbf{A}$
- Total Least Squares (TLS): correcting with $\Delta\mathbf{A}$ and $\Delta\mathbf{b}$
- TLS is also known as Errors-in-Variables modeling.

Singular value decomposition (SVD) approach

Table: Singular value decomposition approach

	TLS: $\min_{\mathbf{x}} \frac{\|\mathbf{Ax}-\mathbf{b}\|^2}{\|\mathbf{x}\|^2+1}$	DLS: $\min_{\mathbf{x}} \frac{\|\mathbf{Ax}-\mathbf{b}\|^2}{\|\mathbf{x}\|^2}$
1	$\sigma_{\min}([\mathbf{A},\mathbf{b}])$ s.t. $\mathbf{v}_{\text{TLS}}(n+1) \neq 0$ $\mathbf{v}_{\text{TLS}}(n+1)$ is the last component of $\mathbf{v}_{\text{TLS}}$.	$\sigma_{\min}(\mathbf{P}_{\mathbf{b}}^{\perp}\mathbf{A})$ s.t. $\mathbf{b}^T\mathbf{A}\mathbf{v}_{\text{DLS}} \neq 0$ $\mathbf{P}_{\mathbf{b}}^{\perp} = \mathbf{I} - \frac{1}{\mathbf{b}^T\mathbf{b}}\mathbf{b}\mathbf{b}^T$
2	$\mathbf{x}_{\text{TLS}} = \frac{-1}{\mathbf{v}_{\text{TLS}}(n+1)}\mathbf{v}_{\text{TLS}}(1:n)$	$\mathbf{x}_{\text{DLS}} = \frac{\mathbf{b}^T\mathbf{b}}{\mathbf{b}^T\mathbf{A}\mathbf{v}_{\text{DLS}}}\mathbf{v}_{\text{DLS}}$
3	$[\Delta\mathbf{A}_{\text{TLS}},\Delta\mathbf{b}_{\text{TLS}}] = -[\mathbf{A},\mathbf{b}]\mathbf{v}_{\text{TLS}}\mathbf{v}_{\text{TLS}}^T$	$\Delta\mathbf{A}_{\text{DLS}} = -\mathbf{P}_{\mathbf{b}}^{\perp}\mathbf{A}\mathbf{v}_{\text{DLS}}\mathbf{v}_{\text{DLS}}^T$
4	$\|[\Delta\mathbf{A}_{\text{TLS}},\Delta\mathbf{b}_{\text{TLS}}]\|_F = \sigma_{\text{TLS}}$	$\|\Delta\mathbf{A}_{\text{DLS}}\|_F = \sigma_{\text{DLS}}$

[1]Equivalent singular value problem and feasibility condition. $\mathbf{v}_{\text{TLS}}$ and $\mathbf{v}_{\text{DLS}}$ are the right singular vectors associated with the smallest singular values, respectively, of $[\mathbf{A},\mathbf{b}]$ and $\mathbf{P}_{\mathbf{b}}^{\perp}\mathbf{A}$. $\sigma_{\min} \equiv$ the minimum singular value; [2]SVD solution; [3]Minimal residual in terms of singular vector; [4]Norm of minimal residuals.

Secular equation approach

Table: Secular equation approach in the generic case

	TLS: $\sigma_{\mathrm{TLS}}^2 = \min_{\mathbf{x}} \frac{\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2}{\|\mathbf{x}\|^2+1}$	DLS: $\sigma_{\mathrm{DLS}}^2 = \min_{\mathbf{x}} \frac{\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2}{\|\mathbf{x}\|^2}$
1	$\mathbf{A}^T(\mathbf{A}\mathbf{x}-\mathbf{b}) = \sigma_{\mathrm{TLS}}^2\mathbf{x}$ $\mathbf{b}^T(\mathbf{b}-\mathbf{A}\mathbf{x}) = \sigma_{\mathrm{TLS}}^2$	$\mathbf{A}^T(\mathbf{A}\mathbf{x}-\mathbf{b}) = \sigma_{\mathrm{DLS}}^2\mathbf{x}$ $\mathbf{b}^T(\mathbf{b}-\mathbf{A}\mathbf{x}) = 0$
2	$\sigma_{\min}(\mathbf{A}) > \sigma_{\mathrm{TLS}}$	$\sigma_{\min}(\mathbf{A}) > \sigma_{\mathrm{DLS}}$
3	$\mathbf{b}^T\mathbf{b} - \mathbf{b}^T\mathbf{A}(\mathbf{A}^T\mathbf{A} - \sigma_{\mathrm{TLS}}^2\mathbf{I})^{-1}\mathbf{A}^T\mathbf{b} = \sigma_{\mathrm{TLS}}^2$	$\mathbf{b}^T\mathbf{b} - \mathbf{b}^T\mathbf{A}(\mathbf{A}^T\mathbf{A} - \sigma_{\mathrm{DLS}}^2\mathbf{I})^{-1}\mathbf{A}^T\mathbf{b} = 0$
4	$\mathbf{x}_{\mathrm{TLS}} = (\mathbf{A}^T\mathbf{A} - \sigma_{\mathrm{TLS}}^2\mathbf{I})^{-1}\mathbf{A}^T\mathbf{b}$	$\mathbf{x}_{\mathrm{DLS}} = (\mathbf{A}^T\mathbf{A} - \sigma_{\mathrm{DLS}}^2\mathbf{I})^{-1}\mathbf{A}^T\mathbf{b}$
5	$[\Delta\mathbf{A}_{\mathrm{TLS}}, \Delta\mathbf{b}_{\mathrm{TLS}}] = \frac{\mathbf{r}_{\mathrm{TLS}}[\mathbf{x}_{\mathrm{TLS}}^T, -1]}{\|\mathbf{x}_{\mathrm{TLS}}\|^2+1}$	$\Delta\mathbf{A}_{\mathrm{DLS}} = \frac{\mathbf{r}_{\mathrm{DLS}}\mathbf{x}_{\mathrm{DLS}}^T}{\|\mathbf{x}_{\mathrm{DLS}}\|^2}$

[1]Normal equations (for stationary points); [2]Generic condition; [3]Secular equation; [4]De-regularized solution; [5]Residuals in terms of solution $\mathbf{x}$ and $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$.

When is the secular equation approach more preferable than the SVD approach?

- The problem is sensitive: $\sigma_{\text{TLS}/\text{DLS}} \approx \sigma_{\min}(\mathbf{A})$
- Least squares solution or de-regularized form is needed: OLS provides a good initial guess of solution. The de-regularized form can be easily calculated by adjusting the amount of negative shift.
- The problem is large: The SVD of a large matrix is very expensive. Instead, we can approximate the secular equation in the large-scaled problem by Gauss quadrature rules.

Riemann-Stieltjes integral

$$\mathbf{M} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T, \qquad \mathbf{\Lambda} = diag(\lambda_i) \qquad 0 < \lambda_1 \le \lambda_2 \le \ldots \le \lambda_n.$$

- $\mathbf{M}$ is symmetric positive definite. $\mathbf{Q}$ is orthonormal.

$$\mathbf{u}^T f(\mathbf{M})\mathbf{u} \;=\; \boldsymbol{\alpha}^T f(\mathbf{\Lambda})\,\boldsymbol{\alpha} \;=\; \sum_{i=1}^{n} f(\lambda_i)\alpha_i^2 = I[f]$$

- $f(\mathbf{M})$ is an analytic function of $\mathbf{M}$ that is defined on $(0,\infty)$.
- $\boldsymbol{\alpha} = \mathbf{Q}^T\mathbf{u}$ for an arbitrary vector $\mathbf{u}$.
- Riemann-Stieltjes integral $I[f]$:

$$I[f] \equiv \int_a^b f(\lambda)\,d\alpha(\lambda), \qquad \alpha(\lambda) = \begin{cases} 0 & \text{if } \lambda < a = \lambda_1 \\ \sum_{j=1}^{i} \alpha_j^2 & \text{if } \lambda_i \le \lambda < \lambda_{i+1} \\ \sum_{j=1}^{n} \alpha_j^2 & \text{if } b = \lambda_n \le \lambda \end{cases}$$

where the measure $\alpha(\lambda)$ is piecewise constant.

Bounds for Riemann-Stieltjes integral

The Gauss quadrature theory is formulated in terms of finite summations:

$$\int_a^b f(\lambda)\,d\alpha(\lambda) = \sum_{j=1}^{N} w_j f(t_j) + \sum_{k=1}^{M} v_k f(z_k) + R[f]$$

- Unknown weights: $[w_j]_{j=1}^N, [v_k]_{k=1}^M$; Unknown nodes: $[t_j]_{j=1}^N$.
- Prescribed nodes: $[z_k]_{k=1}^M$

The remainder term $R[f]$ is given by

$$R[f] = \frac{f^{(2N+M)}(\xi)}{(2N+M)!}\int_a^b \prod_{k=1}^{M}(\lambda - z_k)\left[\prod_{j=1}^{N}(\lambda - t_j)\right]^2 d\alpha(\lambda),\ a < \xi < b.$$

Golub and Meurant [GM93] showed that the sign of the remainder term $R[f]$ can be adjusted by the prescribed nodes. Setting $M = 1$, we will use the Gauss-Radau formula to get the bounds of (a part of) secular function.

Orthonormal polynomials

Define a sequence of polynomials $p_0(\lambda)$, $p_1(\lambda),\ldots$ that are orthonormal with respect to $\alpha(\lambda)$:

$$\int_a^b p_i(\lambda)p_j(\lambda)\,d\alpha(\lambda) = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{otherwise,} \end{cases}$$

and $p_k(\lambda)$ is of exact degree k. Moreover, the roots of $p_k(\lambda)$ are distinct, real and lie in the interval $[a,b]$.

Three-term recurrence relationship

If $\int d\alpha = 1$, the set of orthonormal polynomials satisfies:

$$\lambda\,\mathbf{p}(\lambda) = \mathbf{T}_N\,\mathbf{p}(\lambda) + \gamma_N p_N(\lambda)\,\mathbf{e}_N,$$

where

$$\mathbf{p}(\lambda) = [p_0(\lambda)\ p_1(\lambda)\ \cdots\ p_{N-1}(\lambda)]^T,$$

$$\mathbf{e}_N = (0\ \ \cdots\ 0\ 1)^T \in \mathbb{R}^N,$$

$$\mathbf{T}_N = \begin{pmatrix} \omega_1 & \gamma_1 & & & \\ \gamma_1 & \omega_2 & \gamma_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \gamma_{N-2} & \omega_{N-1} & \gamma_{N-1} \\ & & & \gamma_{N-1} & \omega_N \end{pmatrix}.$$

Lanczos algorithm for quadratures

To obtain the tri-diagonal matrix and hence the Gauss-Radau rule, we will use the Lanczos algorithm with $\mathbf{p}_1 = \mathbf{u}/\|\mathbf{u}\|_2$ as a starting vector:

$$\mathbf{M} \approx \mathbf{P}\,\mathbf{T}_N\,\mathbf{P}^T$$

Eigenvalue decomposition of $\mathbf{T}_N$:

$$\mathbf{T}_N = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T$$

Function of matrix:

$$\mathbf{u}^T f(\mathbf{M})\mathbf{u} \approx \mathbf{u}^T f(\mathbf{P}\,\mathbf{T}_N\,\mathbf{P}^T)\mathbf{u} = \|\mathbf{u}\|_2^2 \mathbf{e}_1^T\,\mathbf{Q} f(\boldsymbol{\Lambda})\mathbf{Q}^T\,\mathbf{e}_1,$$

where $\mathbf{e}_1 = (1\ 0\ \cdots\ 0)^T \in \mathbb{R}^N$.
Thus, the eigenvalues of $\mathbf{T}_N$ give us the nodes and the squares of the first elements of the eigenvectors give the weights:

$$\sum_{j=1}^{N} w_j f(t_j) = \|\mathbf{u}\|_2^2 \sum_{i=1}^{N} (Q_{1i})^2 f(\lambda_i)$$

Inverse eigenvalue problem

To obtain the Gauss-Radau rule, we extend the matrix $\mathbf{T}_N$ in such a way that it has one prescribed eigenvalue z_1.

Lemma

The extended tri-diagonal matrix

$$\hat{\mathbf{T}}_{N+1} \equiv \begin{pmatrix} \mathbf{T}_N & \gamma_N \mathbf{e}_N \\ \gamma_N \mathbf{e}_N^T & \hat{w}_{N+1} \end{pmatrix}$$

has z_1 as an eigenvalue, where $\hat{w}_{N+1} = z_1 + \delta_N$, and δ_N is the last entry of $\boldsymbol{\delta}$ such that

$$(\mathbf{T}_N - z_1 \mathbf{I})\,\boldsymbol{\delta} = \gamma_N^2 \,\mathbf{e}_N. \qquad (1)$$

Proof of the extended tri-diagonal matrix lemma

Proof.

We can verify that z_1 is an eigenvalue of $\hat{\mathbf{T}}_{N+1}$ by investigating the following relation to get (1):

$$\hat{\mathbf{T}}_{N+1}\mathbf{d} = z_1\,\mathbf{d},$$

where $\mathbf{d}$ is a corresponding eigenvector. $\square$

Now, $\hat{\mathbf{T}}_{N+1}$ gives the weights and nodes of the Gauss-Radau rule such that

$$\sum_{j=1}^{N} w_j f(t_j) + v_1 f(z_1) = \|\mathbf{u}\|^2 \mathbf{e}_1^T f(\hat{\mathbf{T}}_{N+1})\,\mathbf{e}_1.$$

The remainder is

$$R[f] = \|\mathbf{u}\|^2 \frac{f^{(2N+1)}(\xi)}{(2N+1)!}\int_a^b (\lambda - z_1)\left[\prod_{j=1}^{N}(\lambda - t_j)\right]^2 d\alpha(\lambda).$$

Part II

Application to solving secular equations

Secular functions

Recall the secular equations:

$$TLS: \qquad \psi_{\text{TLS}}(\lambda) = \mathbf{b}^T\mathbf{b} - \mathbf{b}^T\mathbf{A}(\mathbf{A}^T\mathbf{A} - \lambda\,\mathbf{I})^{-1}\mathbf{A}^T\mathbf{b} - \lambda = 0,$$

$$DLS: \qquad \psi_{\text{DLS}}(\lambda) = \mathbf{b}^T\mathbf{b} - \mathbf{b}^T\mathbf{A}(\mathbf{A}^T\mathbf{A} - \lambda\,\mathbf{I})^{-1}\mathbf{A}^T\mathbf{b} = 0.$$

- $\psi_{\text{TLS}/\text{DLS}}(\lambda)$ is referred to as secular function.
- λ: an estimate of the minimum squared TLS/DLS distance.
- generic condition: $\lambda < \sigma^2_{\min}(\mathbf{A})$ for $(\mathbf{A}^T\mathbf{A} - \lambda\,\mathbf{I})^{-1}$

Thus, in the domain of $0 \le \lambda < \sigma^2_{\min}(\mathbf{A})$, we need to evaluate a matrix function of λ which is common in ψ_{TLS} and ψ_{DLS}:

$$\phi(\lambda) = \mathbf{b}^T\mathbf{A}(\mathbf{A}^T\mathbf{A} - \lambda\,\mathbf{I})^{-1}\mathbf{A}^T\mathbf{b}$$

Bounds of a common function

Now we evaluate the bounds of the scalar quantity ϕ:

$$\phi = \mathbf{g}^T f_{\frac{1}{x}}(\mathbf{M})\,\mathbf{g}, \quad \mathbf{M} = \mathbf{A}^T\mathbf{A} - \lambda\,\mathbf{I}, \quad \mathbf{g} = \mathbf{A}^T\mathbf{b},$$

where $f_{\frac{1}{x}}(x) = \frac{1}{x}$. Then, the quadrature rule $\hat{\phi}_{N+1}(z_1) = \|\mathbf{g}\|^2 \mathbf{e}_1^T f_{\frac{1}{x}}(\hat{\mathbf{T}}_{N+1})\,\mathbf{e}_1$ is described in terms of the remainder:

$$\hat{\phi}_{N+1}(z_1) = I[f_{\frac{1}{x}}] + \|\mathbf{g}\|^2 (\xi)^{-(2N+2)} \int_a^b (\lambda - z_1) \left[\prod_{j=1}^{N} (\lambda - t_j)\right]^2 d\alpha(\lambda).$$

We note that $\frac{f_{\frac{1}{x}}^{(2N+1)}(\xi)}{(2N+1)!} = -(\xi)^{-(2N+2)} < 0, \quad \lambda < a < \xi < b$. Thus, we have the bounds:

$$\hat{\phi}_{N+1}(\zeta_b) < I[f_{\frac{1}{x}}] < \hat{\phi}_{N+1}(\zeta_a), \quad \zeta_a < a < b < \zeta_b.$$

Comments on bounds

- $f_{\frac{1}{x}}(x)$ is well defined on the proper interval (a, b) such that the sign of the derivative function $f_{\frac{1}{x}}^{(2N+1)}(\xi)$ is not changed with the interval $\xi \in (a, b)$.
- Since $\sqrt{||\mathbf{A}||_1||\mathbf{A}||_\infty} > b$, we may use $\zeta_b = \sqrt{||\mathbf{A}||_1||\mathbf{A}||_\infty}$.
- However, the lower bound of a (the smallest eigenvalue of $\mathbf{M}$) is not easily obtainable. ζ_a is determined very roughly.
- This explains why the upper bound of $I[f_{\frac{1}{x}}]$ is usually poorer than the lower bound.

Lanczos process with a shift for efficiency

- The tri-diagonalization is independent of the shift.
- Tri-diag($[\mathbf{g},\ \mathbf{M}]$) $\equiv$ Tri-diag($[\mathbf{g},\ \mathbf{M}+\lambda\,\mathbf{I}]$)
- $(\mathbf{M}+\lambda\,\mathbf{I})\mathbf{Q}_N = \mathbf{Q}_N(\mathbf{T}_N+\lambda\,\mathbf{I}) \ \Leftrightarrow\ \mathbf{M}\mathbf{Q}_N = \mathbf{Q}_N\mathbf{T}_N$

Then we re-define the extended matrix $\hat{\mathbf{T}}_{N+1}$ as

$$\mathbf{J}_{N+1} \equiv \left(\begin{array}{c|c} \mathbf{T}_N - \lambda\,\mathbf{I}_N & \gamma_N\,\mathbf{e}_N \\ \hline \gamma_N\,\mathbf{e}_N^T & w \end{array}\right),$$

where $\mathbf{T}_N$ and γ_N are calculated by Tri-Diag of $[\mathbf{g},\ \mathbf{A}^T\mathbf{A}]$, (not $[\mathbf{g},\ \mathbf{M}]$), and w is determined so that $\mathbf{J}_{N+1}$ has a prescribed eigenvalue z_1. Thus, $w = z_1 + d_N$, where d_N is the last entry of $\mathbf{d}$ such that

$$(\mathbf{T}_N - (z_1+\lambda)\,\mathbf{I}_N)\,\mathbf{d} = \gamma_N^2\,\mathbf{e}_N.$$

Finally, we have the quadrature for bounds:

$$\hat{\phi}_{N+1}(z_1) = \|\mathbf{g}\|^2 \mathbf{e}_1^T f_{\frac{1}{x}}(\mathbf{J}_{N+1})\,\mathbf{e}_1 = \|\mathbf{g}\|^2 \mathbf{e}_1^T (\mathbf{J}_{N+1})^{-1}\,\mathbf{e}_1.$$

Once we solve a tri-diagonal $\mathbf{J}_{N+1}\mathbf{y} = \mathbf{e}_1$ for $\mathbf{y}$, we have

$$\hat{\phi}_{N+1}(z_1) = \|\mathbf{g}\|^2 \mathbf{e}_1^T \mathbf{y} = \|\mathbf{g}\|^2 y_1.$$

For later interpolation, we need to evaluate the derivatives of the matrix function $\phi(\lambda)$ w.r.t. λ by approximating with $f_{\frac{1}{x^2}}(x) = x^{-2}$ and $f_{\frac{1}{x^3}}(x) = x^{-3}$:

$$\hat{\phi}' = \|\mathbf{g}\|^2 \mathbf{e}_1^T f_{\frac{1}{x^2}}(\mathbf{J}_{N+1})\,\mathbf{e}_1 = \|\mathbf{g}\|^2 \mathbf{e}_1^T (\mathbf{J}_{N+1})^{-2}\,\mathbf{e}_1,$$

$$\hat{\phi}'' = 2\,\|\mathbf{g}\|^2 \mathbf{e}_1^T f_{\frac{1}{x^3}}(\mathbf{J}_{N+1})\,\mathbf{e}_1 = 2\,\|\mathbf{g}\|^2 \mathbf{e}_1^T (\mathbf{J}_{N+1})^{-3}\,\mathbf{e}_1.$$

By solving $\mathbf{J}_{N+1}\mathbf{h} = \mathbf{y}$., we have

$$\hat{\phi}' = \|\mathbf{g}\|^2 \|\mathbf{y}\|^2, \qquad \hat{\phi}'' = 2\,\|\mathbf{g}\|^2 \mathbf{y}^T (\mathbf{J}_{N+1})^{-1}\mathbf{y} = 2\,\|\mathbf{g}\|^2 \mathbf{y}^T \mathbf{h}.$$

A symmetric, tri-diagonal, and positive definite system requires $O(N)$ flops [GV96] to be solved.

Lemma (Monotonicity of bound sequences)

Along with Lanczos processes, a sequence of bound estimates of $\phi = \mathbf{g}^T f_{\frac{1}{x}}(\mathbf{M})\,\mathbf{g}$ with full-rank symmetric $\mathbf{M} \in \mathbb{R}^{n\times n}$ and $f_{\frac{1}{x}}(x) = \frac{1}{x}$ is generated by Gauss quadrature rules. Then the estimated sequence $\hat{\phi}_{N+1}$ is necessarily monotonic.
In other words, given each prescribed node ζ_a or ζ_b such that

$$\zeta_a < \sigma_{\min}(\mathbf{M}) < \sigma_{\max}(\mathbf{M}) < \zeta_b,$$

the lower and upper bound sequences for $I[f_{\frac{1}{x}}]$ satisfy

$$\cdots < \hat{\phi}_N(\zeta_b) < \hat{\phi}_{N+1}(\zeta_b) < \cdots < I[f_{\frac{1}{x}}] < \cdots < \hat{\phi}_{N+1}(\zeta_a) < \hat{\phi}_N(\zeta_a) < \cdots.$$

Note that the complete Lanczos processes yield the exact evaluation:

$$\phi = \|\mathbf{g}\|^2 \mathbf{e}_1^T f_{\frac{1}{x}}(\mathbf{T}_n - \lambda\,\mathbf{I}_n)\,\mathbf{e}_1.$$

Interpolating the root of secular equations

We approximate the common function $\phi(\lambda)$ by using Lanczos processes combined with Gauss quadrature rule, where we proceed the processes until the upper- and lower- bounds of $\phi(\lambda)$ match within a tolerance.

Suppose λ_k is the current estimate of the minimum distance. In order to interpolate the root of the secular equation $\psi(\lambda_k)$, we need to evaluate the followings:

$$
\begin{aligned}
\psi_{\text{TLS}}(\lambda_k) &= \|\mathbf{b}\|^2 - \lambda_k - \phi(\lambda_k), & \psi_{\text{DLS}}(\lambda_k) &= \|\mathbf{b}\|^2 - \phi(\lambda_k).\\
\psi'_{\text{TLS}}(\lambda_k) &= -1 - \phi'(\lambda_k), & \psi'_{\text{DLS}}(\lambda_k) &= -\phi'(\lambda_k).\\
\psi''_{\text{TLS}}(\lambda_k) &= -\phi''(\lambda_k), & \psi''_{\text{DLS}}(\lambda_k) &= -\phi''(\lambda_k).
\end{aligned}
$$

Then, consider one-point interpolating methods to obtain λ_{k+1} such that

$$\psi(\lambda_{k+1}) = 0.$$

Note that roots of secular equation consist of the stationary points of the geometrically equivalent cost function. We want to find the smallest root $\lambda_{k+1} \in [0, \sigma^2_{\min}(\mathbf{A}))$ from the definition of TLS/DLS problem. However, we can not achieve it without using additional information on the locations of poles such as $\sigma^2_{\min}(\mathbf{A}) \le \min_j \sum_i |a_{ij}|^2$ and $\sigma^2_{\max}(\mathbf{A}) \le \|\mathbf{A}\|_1 \cdot \|\mathbf{A}\|_\infty$. In the following sections, we will discuss how to use bisection and the upper-bound of the smallest pole.

Variations of Newton's method take the form:

$$\lambda_{k+1} = \lambda_k - \frac{\psi(\lambda_k)}{\psi'(\lambda_k)} \cdot C_k,$$

where C_k denotes a convergence factor [Gan78, Gan85] according to methods such as the Newton's method, the Halley's variation, and simple rational approximation in the following Table .

Table: Variations of Newton's method

	Newton's	SRA[a]	Halley's
Interpolating function $h(\lambda) \approx \psi(\lambda)$	$h(\lambda) = c_0 + c_1\lambda$	$h(\lambda) = \|\mathbf{b}\|^2 - \frac{c_1}{c_2-\lambda}$	$h(\lambda) = c_0 - \frac{c_1}{c_2-\lambda}$
Convergence factor C_k	1	$\frac{\|\mathbf{b}\|^2-\psi(\lambda_k)}{\|\mathbf{b}\|^2}$	$1/\left(1-\frac{\psi(\lambda_k)\,\psi''(\lambda_k)}{2(\psi'(\lambda_k))^2}\right)$
Rate of (local) convergence	Quadratic	Quadratic	Cubic
Convergence region[b]	Narrow	Wide	Wider
Algebraic interpretation[c]	$g(\lambda) = \psi(\lambda)$	$g(\lambda) = 1 - \frac{\|\mathbf{b}\|^2}{\|\mathbf{b}\|^2-\psi(\lambda)}$	$g(\lambda) = \frac{\psi(\lambda)}{\sqrt{\psi'(\lambda)}}$

[a] Simple Rational Approximation

[b] Global convergence in root-finding of secular equations

[c] Equivalently, solve $g(\lambda) = 0$ by Newton's method.

Mixing with bisection

Whenever we detect that the root estimate is larger than $\sigma^2_{\min}(\mathbf{A})$, we bisect the estimate to assure that it is less than $\sigma^2_{\min}(\mathbf{A})$. The monotonicity of the sequence of estimate bounds of Gauss quadratures (GQ) is utilized based on the following scenario:

1. With the initial guess of root, we obtain the sequence of bounds of secular function by means of GQ.
2. If the sequences are not monotonic, we conclude that the root estimate is larger than the squared smallest singular value of $\mathbf{A}$. Then the root estimate is cut by half, and go to Step 1 with the modified estimate of root.
3. Otherwise, we interpolate the root of secular equation by using the estimate of the secular function and its derivatives.
4. If the new root estimate is close to the previous one within a tolerance, then we calculate the de-regularized solution, and stop the algorithm. Otherwise, go to step 1.

Although the violation of monotonicity is only a necessary condition, our numerical simulation works well.

Stabilizing with the estimated smallest pole

- Although the bisection scheme almost always achieves the smallest root, it may suffer from a ‘bi-stability’ problem which means the estimates are alternating between two values.
- To get around this, we employ the estimation of the smallest pole by modifying the previous scenario. If we detect the current estimate of root is larger than $\sigma^2_{\min}(\mathbf{A})$, we cut the estimate by half and set the upper-bound of the smallest pole to the current estimate of root as well.

$$\hat{\sigma}^2_{\min}(\mathbf{A}) = \lambda_k, \qquad \lambda_{k+1} = \frac{1}{2}\lambda_k.$$

Then, when we interpolate the next estimate of root, we take a harmonic sum between the Newton-based step $\delta_k = \frac{\psi(\lambda_k)}{\psi'(\lambda_k)} \cdot C_k$ and the distance from the upper-bound of the smallest pole estimation.

$$\lambda_{k+1} = \lambda_k + \frac{1}{\frac{-1}{\delta_k} + \frac{1}{\hat{\sigma}^2_{\min}(\mathbf{A}) - \lambda_k}} = \lambda_k - \delta_k + \delta_k^2/(\delta_k + \lambda_k - \hat{\sigma}^2_{\min}(\mathbf{A}))$$

CGQ as a secular equation approach

1. Find the smallest root of secular equation for TLS or DLS.
   1.1 Evaluate the bounds of secular function by *Gauss-Radau quadrature rule*.
   1.2 Interpolate the zero of the function by *a variation of Newton method*.
   1.3 Determine a proper interval for the smallest zero by *bisection* and *harmonic-summation with the upper-bound of the smallest pole*.
2. Solve a de-regularized system with a shift of the smallest root.
   - Solve the symmetric, positive-definite system by *the conjugate gradient (CG) method*,
   - Or, solve the tri-diagonal system with shift.

Alternative implementations

- Reuse of Lanczos vectors with sufficient memory
- Regeneration of Lanczos vectors with knowledge of tri-diagonal entries
- Avoiding of explicit multiplication of $\mathbf{A}^T\mathbf{A}$
- Shifting into Lanczos bi-diagonalization
- Using backward perturbations

Backward perturbations for linear least squares I

$$\min_x \|b - Ax\|_2, \qquad A : m \times n, \qquad b : m \times 1.$$

ξ: arbitrary vector, calculated.

$$\hat{x} = \xi + e; \quad \hat{x} = A^+ b.$$

$$\xi = (A + \delta A)^+ b.$$

$$\rho = b - A\xi.$$

$$\mu(\xi) = \min_{\delta A} \|\delta A\|_F.$$

$$\boxed{\mu(\xi) = \min\{\tfrac{\|\rho\|_2}{\|\xi\|_2}, \sigma_{\min}(A, B)\}}$$

(Karlson, Waldén & Sun)

$$B = \frac{\|\rho\|_2}{\|\xi\|_2}\left(I - \frac{\rho\rho^T}{\|\rho\|_2}\right).$$

Backward perturbations for linear least squares II

$$\tilde{\mu}^2(\xi) = \rho^T A(\alpha A^T A + \beta I)^{-1} A^T \rho$$

$$\alpha = \|\xi\|_2^2, \quad \beta = \|\rho\|_2^2$$

$$\boxed{\tilde{\mu}(\xi) \sim \mu(\xi)}$$

$$\lim_{\xi \to \hat{x}} \frac{\tilde{\mu}(\xi)}{\mu(\hat{x})} = 1$$

(Grcar)

Björck's algorithm

Solve the system of nonlinear equations

$$\begin{bmatrix} A^T A & A^T b \\ b^T A & b^T b \end{bmatrix} \begin{bmatrix} x \\ -1 \end{bmatrix} = \lambda \begin{bmatrix} x \\ -1 \end{bmatrix},$$

or equivalently, the system

$$\begin{bmatrix} f(x,\lambda) \\ g(x,\lambda) \end{bmatrix} = \begin{bmatrix} -A^T r - \lambda x \\ -b^T r + \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

with $r = b - Ax$ using a Rayleigh-quotient iteration (RQI). This algorithm will always converge to a singular value/vector pair, but we might not get $\lambda = \sigma_{n+1}^2$. Björck suggested one initial inverse iteration (i.e. $\lambda = 0$) to move closer to the desired λ, and then apply the RQI procedure.

Björck's algorithm details

After some manipulations, Björck's algorithm greatly simplifies. The following presentation emphasizes the computationally intensive steps.

1. Solve the least squares system in A, b for x_{LS}.
2. Perform one inverse iteration (solve $A^T A x^{(0)} = x_{\text{LS}}$) to get the initial $x^{(0)}$.
3. While not converged... solve two systems in $(A^T A - \lambda I)x = b$ to iterate to $x^{(k+1)}$, but if we detect $A^T A - \lambda I$ is negative definite, decrease λ and repeat the iteration.

Björck suggests using the PCGTLS algorithm to solve each linear system with the Cholesky factor R or $A^T A$ as a preconditioner. Inside the CG procedure, we detect $A^T A - \lambda I$ is negative definite and use the CG vectors to compute a new value of λ.

Notes on our implementation of Björck's algorithm

- Instead of using PCGTLS with the Cholesky factor of $A^T A$, we use a matrix-free approach for large scale $m > 800$ and apply the unpreconditioned CGTLS algorithm. The tolerance used in the CG method is 10^{-12}.
- To compute the initial least squares solution, we use the LSQR algorithm.
- We detect convergence when λ changes by less than 10^{-12} or the normalized residual increases (theory states the normalized residual always decreases).
- After we detect convergence, we run one more iteration of the algorithm to ensure that we compute an x "for" the λ.

Details of the matrix moments based algorithm

- Algorithm 2 uses the Golub-Kahan bidiagonalization of A and applies the moment algorithm to $T = B^T B$ instead of computing T directly from the Lanczos process on $A^T A$.
- Algorithm 1 restarts the Lanczos process at each iteration.
- Algorithm 2 never restarts the bidiagonalization process and simply continues the process at each iteration.

Problems

- Jo's problems, 15×8 and 750×400
- Björck's problem 1: 30×15 matrix
- Large scale problems with 10000×5000 and 100000×60000 matrices.

The large scale problems were generated using random Householder matrices to build the SVD of $\begin{bmatrix} A & b \end{bmatrix}$ in product form. Each large-scale matrix was available solely as an operator to all of the algorithms. The singular values of

$$\begin{bmatrix} A & b \end{bmatrix}$$

are

$$\sigma_i = \log(i) + |N(0,1)|,$$

where $N(0,1)$ is a standard normal random variable.

Parameter choices

		Algorithm 1			Algorithm 2		
		$\lambda^{(0)}=0$	$\lambda^{(0)}=1$	$\lambda^{(0)}=\rho$	$\lambda^{(0)}=0$	$\lambda^{(0)}=1$	$\lambda^{(0)}=\rho$
1	newton	6	4	5	6	4	5
	sra	5	5	5	5	5	5
	halley	5	5	6	5	5	6
2	newton	++	++	–	*8	*8	*7
	sra	++	–	++	*12	*24	*7
	halley	++	++	++	*14	*23	*6
3	newton	–	–	–	*20	*7	*10
	sra	–	–	–	*20	25	*64
	halley	–	–	–	*55	55	*12
4	newton	–	–	–	*15	*11	*11
	sra	++	–	–	*15	*25	*14
	halley	–	–	–	*20	*57	*11
5	newton	100	–	–	++	++	++
	sra	100	-5	–	++	++	–
	halley	100	–	–	++	++	–

* wrong root; ++ correct w/o convergence; – no convergence

Convergence

TEST	ALG	ITERS	ERROR	TIME	LANZ.
jo	björck	6	1.0×10^{-14}	0	
$(15,8)$	Alg 1	5	4.4×10^{-16}	0	
$\sigma^2=5.6\times10^{-1}$	Alg 2	5	3.3×10^{-14}	0	12
jo	björck	7	8.5×10^{0}	0.2	
$(750,400)$	Alg 1	>100	8.5×10^{-14}	52.5	
$\sigma^2=1.8\times10^{1}$	Alg 2	23	5.0×10^{-1}	0.7	163
large-scale	björck	8	1.1×10^{-16}	0.5	
$(10000,5000)$	Alg 1	>100	1.0×10^{-3}	36.1	
$\sigma^2=1.9\times10^{-1}$	Alg 2	55	8.3×10^{-16}	1.5	152
large-scale	björck	5	3.9×10^{-17}	5.1	
$(100000,60000)$	Alg 1	>100	5.5×10^{-7}	324.9	
$\sigma^2=3.5\times10^{-3}$	Alg 2	57	5.3×10^{-8}	14.6	155
björck	björck	7	2.6×10^{-19}	0	
$(30,15)$	Alg 1	>100	σ^2	0.3	
$\sigma^2=9.9\times10^{-12}$	Alg 2	18	2.9×10^{4}	0	33

Numerical data generation

The numerical data are generated as follows [GvM91]. Let the error-free data matrix $\mathbf{A}_s = \mathbf{U}_s \mathbf{\Sigma}_s \mathbf{V}_s^T$, where $\mathbf{U}_s = \mathbf{I}_m - 2\frac{\mathbf{u}_s \mathbf{u}_s^T}{\mathbf{u}_s^T \mathbf{u}_s}$ and $\mathbf{V}_s = \mathbf{I}_n - 2\frac{\mathbf{v}_s \mathbf{v}_s^T}{\mathbf{v}_s^T \mathbf{v}_s}$ are Householder matrices and $\mathbf{\Sigma}_s = \mathrm{diag}(\sigma_k)$ is a m-by-n diagonal matrix with $[\sigma_1, \cdots, \sigma_n] = [\sqrt{n}, \cdots, 1]$. The vectors $\mathbf{u}_s$ and $\mathbf{v}_s$ consist of pseudo-random numbers generated by a function 'randn' providing uniformly distributed random numbers in Matlab^{TM}. An error-free observation vector $\mathbf{b}_s$ is computed by $\mathbf{b}_s = \mathbf{A}_s \mathbf{x}_s$. The error-prone data $\mathbf{A}$ and $\mathbf{b}$ are generated with a given deviation σ_n as follows:

$$
\begin{aligned}
&\mathrm{randn}('\mathrm{state}', 108881); \text{\% seed for random number} \\
&\mathbf{u}_s = \mathrm{randn}(m,\ 1); \\
&\mathbf{v}_s = \mathrm{randn}(n,\ 1); \\
&\mathbf{A}_s = \mathbf{U}_s \mathbf{\Sigma}_s \mathbf{V}_s^T; \\
&\mathbf{x}_s = (1./(1:n))'; \text{\%}\ [1,\ 1/2,\ \cdots,\ 1/n]^T \\
&\mathbf{b}_s = \mathbf{A}_s \mathbf{x}_s; \\
&\sigma_n = 0.3; \\
&\mathbf{b} = \mathbf{b}_s + \sigma_n * \mathrm{randn}(m,\ 1); \\
&\mathbf{A} = \mathbf{A}_s + \sigma_n * \mathrm{randn}(m,\ n);
\end{aligned}
$$

CG-based algorithms

We can categorize algorithms in the following example as below.

Table: CG-based algorithms for solving secular equation.

	CGQ-BS-KP	CGQ-BS	CG-BS-KP	CG-BS
Gauss Quadrature rules	Used	Used	Not	Not
Bisection	Used	Used	Used	Used
Upper-bound of the smallest pole	Used	Not	Used	Not

The previous CG-based approaches in [GJK06] or [BHM00] employ matrix-vector multiplications to evaluate the secular function and derivative from the intermediate solution $\mathbf{x}(\lambda)$ such that

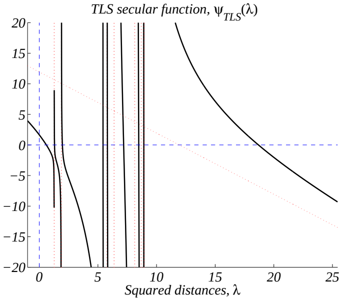
$$(\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I})\mathbf{x}(\lambda) = \mathbf{A}^T\mathbf{b}.$$

Then the secular functions and derivatives are obtained as follows.

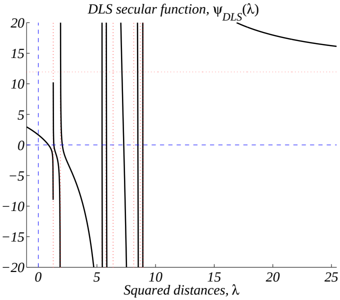
$$\psi_{\text{TLS}}(\lambda) = \|\mathbf{b}\|^2 - \mathbf{b}^T\mathbf{A}\mathbf{x}(\lambda) - \lambda \qquad \psi_{\text{DLS}}(\lambda) = \|\mathbf{b}\|^2 - \mathbf{b}^T\mathbf{A}\mathbf{x}(\lambda)$$

$$\psi'_{\text{TLS}}(\lambda) = -\|\mathbf{x}(\lambda)\|^2 - 1 \qquad \psi'_{\text{DLS}}(\lambda) = -\|\mathbf{x}(\lambda)\|^2$$

Secular function plots with $\mathbf{A} \in \mathbb{R}^{15\times 8}$

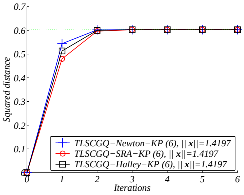


(a) TLS secular function plot

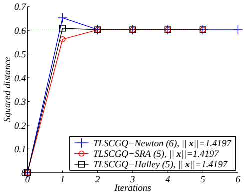


(b) DLS secular function plot

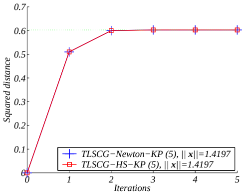
Convergent curves in TLS with $\mathbf{A} \in \mathbb{R}^{15\times 8}$



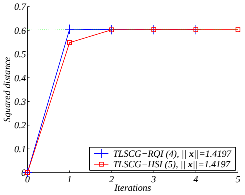
(a) Variations of CGQ with the bisection and the smallest pole estimation.



(b) Variations of CGQ with the bisection only.

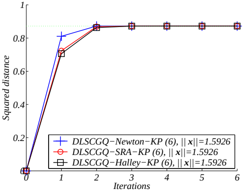


(c) CG with two different interpolations, the bisection and the smallest pole estimation.

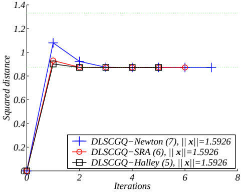


(d) CG with two different inverse iterations, and the bisection only.

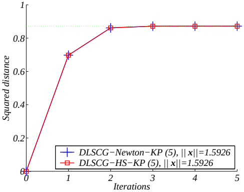
Convergent curves in DLS with $\mathbf{A} \in \mathbb{R}^{15\times 8}$



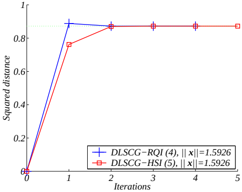
(a) Variations of CGQ with the bisection and the smallest pole estimation.



(b) Variations of CGQ with the bisection only.

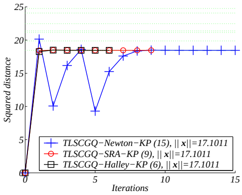


(c) CG with two different interpolations, the bisection and the smallest pole estimation.

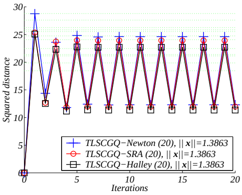


(d) CG with two different inverse iterations, and the bisection only.

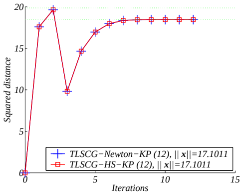
Convergent curves in TLS with $\mathbf{A} \in \mathbb{R}^{750 \times 400}$



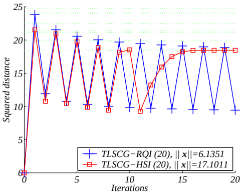
(a) Variations of CGQ with the bisection and the smallest pole estimation.



(b) Variations of CGQ with the bisection only.

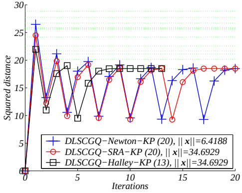


(c) CG with two different interpolations, the bisection and the smallest pole estimation.

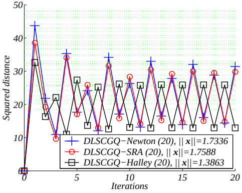


(d) CG with two different inverse iterations, and the bisection only.

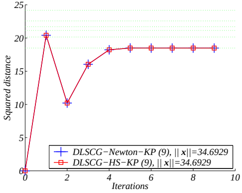
Convergent curves in DLS with $\mathbf{A} \in \mathbb{R}^{750\times 400}$



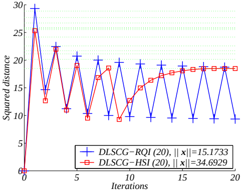
(a) Variations of CGQ with the bisection and the smallest pole estimation.



(b) Variations of CGQ with the bisection only.



(c) CG with two different interpolations and the smallest pole estimation.



(d) CG with two different inverse iterations, and the bisection only.

Conclusion

- The presented Conjugate Gradient (CG) with Quadrature (CGQ) approximates secular functions by means of Gauss Quadrature (GQ) rules.
  - The previous CG based approaches are exhaustively calculating *intermediate solutions* to evaluate secular functions.
- Interpolating the smallest root by variations of Newton method with stabilized modification:
  - Bisection to assure of the convergence to the smallest root.
  - Approximating with rational function of the estimated smallest pole to get around bi-stability problem.
- CGQ does not pursue the intermediate solution until the root estimate converges.
- The overall computational complexity can be reduced by adjusting accuracy of GQ.

Åke Björck, P. Heggernes, and P. Matstoms.
Methods for large scale total least squares problems.
SIAM J. Matrix Anal. Appl., 22(3):413–429, 2000.

R. D. DeGroat and E. M. Dowling.
The data least squares problem and channel equalization.
41(1):407–411, January 1993.

Walter Gander.
On the linear least squares problem with a quadratic constraint.
Technical Report STAN-CS-78-697, Stanford University, 1978.

W. Gander.
On halley's iteration method.
The American mathematical monthly, 92(2):131 – 134, 1985.

Gene H. Golub, SungEun Jo, and Sang Woo Kim.
Solving secular equations for total/data least squares problems.
SIAM J. Matrix Anal. Appl., submitted for publication, 2006.

G. H. Golub and G. Meurant.
Matrices, moments and quadrature.
Report SCCM-93-07, Computer Science Department, Stanford University, 1993.

References

Gene H. Golub and Charles. F. Van Loan.
Matrix Computations.
Johns Hopkins Univ. Pr., Baltimore, MD, third edition, November 1996.

G. H. Golub and U. von Matt.
Quadratically constrained least squares and quadratic problems.
Numerische Mathematik, 59(1):561–580, December 1991.

SungEun Jo and Sang Woo Kim.
Consistent normalized least mean square filtering with noisy data matrix.
(6):2112– 2123, June 2005.