Numerical linear algebra and some problems in computational statistics

Zdeněk Strakoš

Academy of Sciences and Charles University, Prague
http://www.cs.cas.cz/˜strakos

IASC2008, Yokohama, Japan, December 2008.

This lecture is devoted to the memory

of Tomáš Havránek

Motivation – Computational mathematics

E. Study (1862-1930, Leipzig, Marburg, Göttingen, Greifswald, Bonn, successor of Lipschitz) :

Mathematics is neither the art of calculation
nor the art of avoiding calculations.
Mathematics, however, entails the art
of avoiding superflous calculations
and conducting the necessary ones skilfully.
In this respect one could have learned
from the older authors.

Motivation – Computational mathematics

B.J.C. Baxter, A. Iserles, *On the foundations of computational math.,* in Handbook of Numerical Analysis XI (P.G. Ciarlet and F. Cucker, eds), North-Holland, Amsterdam (2003), 3-34 :

The purpose of computation is not
to produce a solution with least error
but to produce *reliably, robustly and affordably*
a solution which is within a user-specified tolerance.

It should be emphasized that there is really no boundary
between computational mathematics and statistics
see Section 2.5 of the paper above.

Examples

- Singular value decomposition, numerically stable algorithms for computing orthogonal decompositions and projections, various direct and iterative methods and algorithms applicable to problems in computational statistics.
- Linear regression and ordinary least squares, collinearity (Stewart, Marquardt, Belsley, Thisted, Hadi and Velleman (1987)).
- Errors-in-variables modeling, orthogonal regression and total least squares.
- Stochastic partial differential equations and their numerical solution.
- Statistical tools in solving ill-posed problems, information retrieval and data mining, signal processing.
- . . . . . .

Outline

1. Common roots: Moments
2. Linear approximation problems
3. Singular value decomposition and model reduction
4. Matching moments model reduction and Krylov subspace methods
5. Bidiagonalization and linear regression
6. Bidiagonalization and orthogonal regression
7. Back to the roots (matching moments and model reduction)
8. Conclusions

Chebyshev (1855), Heine (1861), Markov (1884)

Let $p(\lambda)$ be a nonnegative function in $(-\infty, \infty)$. Given

$$\int_{-\infty}^{\infty} p(\lambda)\, \lambda^k \, d\lambda \;=\; \int_{-\infty}^{\infty} e^{-\lambda^2}\, \lambda^k \, d\lambda\,, \quad k = 0, 1, \ldots\,,$$

can we conclude that $p(\lambda) = e^{-\lambda^2}$ or, as we say now, that the distribution characterized by the function

$$\int_{-\infty}^{x} p(\lambda)\, d\lambda$$

is a normal one?

See Shohat and Tamarkin (1943), Akhiezer (1965).

Stieltjes (1883-4)

Given a sequence of numbers ξ_k, $k = 0, 1, \ldots$, a non-decreasing distribution function $\omega(\lambda)$ is sought such that

$$\int_0^\infty \lambda^k \, d\omega(\lambda) \;=\; \xi_k \,, \quad k = 0, 1, \ldots \,,$$

where

$$\int_0^\infty \lambda^k \, d\omega(\lambda)$$

represents the k-th (generalized) mechanical moment of the distribution of positive mass on the half line $\lambda \geq 0$. Stieltjes based his investigation on continued fractions; cf. Gantmacher and Krein (1950).

Another formulation

Consider a non-decreasing distribution function $\omega(\lambda)$, $\lambda \geq 0$ with the moments given by the Riemann-Stieltjes integral

$$\xi_k \;=\; \int_0^\infty \lambda^k \, d\omega(\lambda)\,, \quad k = 0, 1, \ldots\,.$$

Find the distribution function $\omega^{(n)}(\lambda)$ with n points of increase $\lambda_i^{(n)}$ $i = 0, 1, \ldots$, which matches the first $2n$ moments for the distribution function $\omega(\lambda)$,

$$\int_0^\infty \lambda^k \, d\omega^{(n)}(\lambda) \;\equiv\; \sum_{i=1}^{n} \omega_i^{(n)} (\lambda_i^{(n)})^k \;=\; \xi_k, \quad k = 0, 1, \ldots, 2n-1\,.$$

This moment problem plays in modern NLA a fundamental role.

1 : Gauss-Christoffel quadrature

Clearly,

$$\int_0^\infty \lambda^k \, d\omega(\lambda) \;=\; \sum_{i=1}^{n} \omega_i^{(n)} \, (\lambda_i^{(n)})^k \,, \quad k = 0, 1, \ldots, 2n-1$$

represents the n-point Gauss-Christoffel quadrature, see

C. F. Gauss, *Methodus nova integralium valores per approximationem inveniendi,* (1814),

C. G. J. Jacobi, *Über Gauss' neue Methode, die Werthe der Integrale näherungsweise zu finden,* (1826),

and the description given in H. H. J. Goldstine, *A History of Numerical Analysis from the 16th through the 19th Century,* (1977).

With no loss of generality we consider $\xi_0 = 1$.

Outline

1. Common roots: Moments
2. Linear approximation problems
3. Singular value decomposition and model reduction
4. Matching moments model reduction and Krylov subspace methods
5. Bidiagonalization and linear regression
6. Bidiagonalization and orthogonal regression
7. Back to the roots
8. Conclusions

Forms of $A\,x \approx b$

- LE: A is square and numerically nonsingular, then $A\,x = b$.
- OLS (linear regresion): A is generally error free rectangular N by M matrix and the right hand side (the observation vector) is significantly affected by errors. Then

$$A\,x = b + r\,, \quad \min \|r\|\,.$$

- TLS (orthogonal regression): Significant errors contained both in the generally rectangular N by M matrix A and the right hand side b. Then

$$(A + E)\,x = b + r\,, \quad \min \|[r, E]\|_F\,,$$

where $\|\cdot\|_F$ means the Frobenius norm of the given matrix, see Rao and Toutenburg (1999), Section 3.12.

Miminum norm OLS solution

Let b be orthogonally decomposed into parts $b|_{\mathcal{R}(A)}$ in the range of A and $b|_{\mathcal{N}(A^T)}$ in the nullspace of A^T ,

$$b = b|_{\mathcal{R}(A)} + b|_{\mathcal{N}(A^T)} .$$

Then the minimum norm solution x is given by

$$A\,x = b|_{\mathcal{R}(A)} , \quad x \in \mathcal{R}(A^T) , \quad r = -\,b|_{\mathcal{N}(A^T)} .$$

The errors in b are assumed to be orthogonal to the subspace generated by the columns of A. If A has full column rank,

$$A^T A\,x = A^T b .$$

In computational statistics $x = (A^T A)^{-1} A^T b .$

Damped OLS solution

Let A represent a discrete ill-posed problem and the right hand side is significantly affected by errors (noise). Then the OLS solution is useless. Instead,

$$A\,x \;=\; b \,+\, r, \quad \min\left\{\|r\|^2 + \alpha^2\,\|x\|^2\right\},$$

which is nothing but the Tikhonov regularization (1963). Equivalently,

$$x \;=\; \operatorname{argmin} \left\| \begin{pmatrix} A \\ \alpha\, I \end{pmatrix} x \,-\, \begin{pmatrix} b \\ 0 \end{pmatrix} \right\|.$$

Example - discretized Fredholm integral equations of the first order.

Ridge regression

Using the normal equations,

$$(A^T A + \alpha^2 I)\, x \;=\; A^T b\,.$$

In computational statistics this is known as the ridge regression (RR),
Rao and Toutenburg (1999), Section 3.10.2, Čížek (2004), Section 8.1.6,

$$x \;=\; (A^T A + \alpha^2 I)^{-1}\, A^T b\,.$$

Caution. 'Ill-posed problems' does not mean the same as
'ill-conditioned problems'.

Outline

1. Common roots: Moments
2. Linear approximation problems
3. Singular value decomposition and model reduction
4. Matching moments model reduction and Krylov subspace methods
5. Bidiagonalization and linear regression
6. Bidiagonalization and orthogonal regression
7. Back to the roots
8. Conclusions

Spectral decomposition of a symmetric matrix

If A is symmetric N by N matrix, then $A = U\Lambda U^T$, $\Lambda = \mathrm{diag}\,(\lambda_j)$, $UU^T = U^TU = I$, $U = [u_1, \ldots, u_N]$.

$$A$$

$$\begin{array}{ccc} u_1 & \stackrel{\lambda_1}{\longrightarrow} & u_1 \\ u_2 & \stackrel{\lambda_2}{\longrightarrow} & u_2 \\ \vdots & & \\ u_N & \stackrel{\lambda_N}{\longrightarrow} & u_N \end{array}$$

One dimensional mutually orthogonal invariant subspaces.

Singular value decomposition (SVD)

Consider an N by M matrix A, with no loss of generality $N \geq M$. Then

$$A \;=\; S\,\Sigma\,W^T \;=\; S_r\,\Sigma_r\,W_r^T \;=\; \sum_{\ell=1}^{r} s_\ell\,\sigma_\ell\,w_\ell^T\,,$$

$$SS^T = S^TS = I,\; W^TW = WW^T = I,\; \Sigma = \mathrm{diag}\,(\sigma_1, \ldots, \sigma_r,\, 0)\,,$$

$$\sigma_1 \geq \sigma_2 \geq \;\ldots\; \geq \sigma_r > 0\,,$$

$$S = [\,S_r,\,\ldots],\; W = [\,W_r,\,\ldots],\; \Sigma_r = \mathrm{diag}\,(\sigma_1,\,\ldots\,,\sigma_r)\,.$$

Singular value decomposition

$$
\begin{array}{ccccc}
 & A & & A^T & \\
 & & \mathcal{R}(A) & & \mathcal{R}(A^T) \\
w_1 & \xrightarrow{\sigma_1} & s_1 & \xrightarrow{\sigma_1} & w_1 \\
w_2 & \xrightarrow{\sigma_2} & s_2 & \xrightarrow{\sigma_2} & w_2 \\
\vdots & \vdots & \dots & \vdots & \vdots \\
w_r & \xrightarrow{\sigma_r} & s_r & \xrightarrow{\sigma_r} & w_r
\end{array}
$$

$$
\mathcal{N}(A) \left\{ \begin{array}{c} w_{r+1} \\ \vdots \\ w_M \end{array} \right\} \to 0\,, \qquad \mathcal{N}(A^T) \left\{ \begin{array}{c} s_{r+1} \\ \vdots \\ s_N \end{array} \right\} \to 0\,,
$$

Singular value decomposition

Distance of the full rank matrix to a nearest singular matrix (rank-deficient matrix):

$$\|A - A_{M-1}\| = \sigma_M\,, \quad \frac{\|A - A_{M-1}\|}{\|A\|} = \frac{\sigma_M}{\sigma_1} = 1/\kappa(A)\,.$$

Ill-conditioning means $\kappa(A)$ large, ill-posedness means much more.

SVD model reduction and PCR

When

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k \quad \gg \quad \sigma_{k+1} \geq \cdots \geq \sigma_r \,,$$

we have a good rank-k approximation

$$A \;\approx\; A_k \;=\; \sum_{1}^{k} s_i \sigma_i w_i^T \,.$$

Please recall the Principal Components Regression (PCA, PCR), where the solution x of the OLS problem is approximated by the so called Truncated SVD approximation (TSVD)

$$x \;=\; \sum_{\ell=1}^{r} \frac{s_\ell^T b}{\sigma_\ell}\, w_\ell \;\approx\; x_k^{\mathrm{PCR}} \;=\; \sum_{\ell=1}^{k} \frac{s_\ell^T b}{\sigma_\ell}\, w_\ell, \quad k \ll r \,.$$

What is wrong with $x = (A^T A)^{-1} A^T b$??

In theory almost nothing. If the computation is done that way, then, apart from some very special cases, almost everything.

See the analysis of the formula using the SVD of A:

$$x = (A^T A)^{-1} A^T b = (W \Sigma^2 W^T)^{-1} W \Sigma S^T b = W \Sigma^{-1} S^T b .$$

If the normal matrix is formed and then inverted, things will not cancel out so nicely. Results computed by inverting the explicitly formed normal matrix are generally expensive and inaccurate; in the worst case they can be a total garbage. The requirements of Baxter and Iserles (2003). - reliably, robustly and affordably - are violated.

What is wrong with forming $A^T A$??

Consider

$$A \equiv \begin{pmatrix} 1 & 1 \\ \epsilon & \\ & \epsilon \end{pmatrix}, \quad A^T A = \begin{pmatrix} 1+\epsilon^2 & 1 \\ 1 & 1+\epsilon^2 \end{pmatrix}.$$

Whenever ϵ^2 is smaller than machine precision,
the normal system matrix $A^T A$ is numerically singular!

Therefore the decomposition approach is numerically superior.

This example was given by Läuchli in 1961.
See also Björck (1996), Section 2.2.1 or
Watkins (2002), Example 3.5.25 and Section 4.4, pp. 285–286.

Can normal equations ever be used ?

Yes! Recall, e.g., nonlinear regression and the Levenberg-Marquardt method, see Čížek in Gentle, Härdle, and Mori (2004), Section 8.2.1.3.

For the tall skinny Jacobians J_k , the new direction vectors d_k in the Gauss-Newton method can be efficiently computed by solving

$$(J_k^T J_k + \alpha^2 I)\, d_k = -J_k^T r_k$$

where it can be convenient to form $J_k^T J_k$. Here we need only a rough regularized approximation embedded in the outer iteration process.

Please note that seemingly similar tasks may require in nonlinear and linear regression computations different approaches.

Trouble with solution of the ill-posed problem

Consider the ill-posed problem

$$A\,x \;+\; \eta \;=\; b\,,$$

where x is unknown and the observation vector b is corrupted by the white noise η of the unknown size.

A naive solution, though computed in the most numerically stable way, gives no useful information about x,

$$x^{\mathrm{naive}} \;=\; \sum_{\ell=1}^{r} \frac{s_\ell^T b}{\sigma_\ell}\, w_\ell \;=\; \sum_{\ell=1}^{r} \frac{s_\ell^T (b-\eta)}{\sigma_\ell}\, w_\ell \;+\; \sum_{\ell=1}^{r} \frac{s_\ell^T \eta}{\sigma_\ell}\, w_\ell \;=\; x \;+\; \sum_{\ell=1}^{r} \frac{s_\ell^T \eta}{\sigma_\ell}\, w_\ell\,.$$

For the singular values approaching zero (machine precision) the last term will blow up.

TSVD regularization and spectral filtering

The problem is resolved by truncation of the SVD expansion (TSVD regularization)

$$x \approx \sum_{\ell=1}^{k} \frac{s_\ell^T b}{\sigma_\ell} w_\ell = \sum_{\ell=1}^{k} \frac{s_\ell^T (b-\eta)}{\sigma_\ell} w_\ell + \sum_{\ell=1}^{k} \frac{s_\ell^T \eta}{\sigma_\ell} w_\ell$$

at the price of loosing a part of the useful information about the true solution

$$\sum_{\ell=k+1}^{r} \frac{s_\ell^T (b-\eta)}{\sigma_\ell} w_\ell \, .$$

Spectral filtering description of regularization

More sophisticated methods construct regularized solutions which can be expressed as

$$x^{\mathrm{reg}} = \sum_{\ell=1}^{r} \phi_\ell \, \frac{s_\ell^T b}{\sigma_\ell} \, w_\ell \, ,$$

where the filter factors ϕ_ℓ should be close to one for large singular values and close to zero for small singular values. Such regularized solution is not necessarily computed using the SVD decomposition.

It can be computed, among other techniques, by matching moments model reductions represented by Krylov subspace methods.
For Tikhonov regularization

$$\phi_\ell = \frac{\sigma_\ell^2}{\sigma_\ell^2 + \alpha^2} \, .$$

Example from image deblurring

Example (J. Nagy, Emory University)

Original image (the unknown x)

x = true image

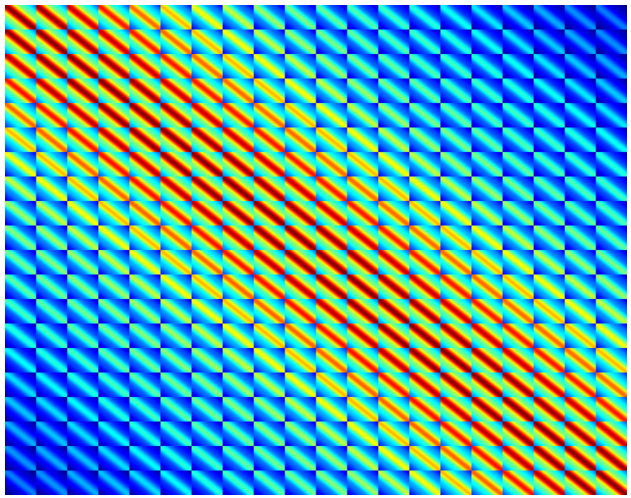
Example from image deblurring

Observed image
(the right hand side b)

b = blurred, noisy image

Matrix A describing
the Point Spread Function

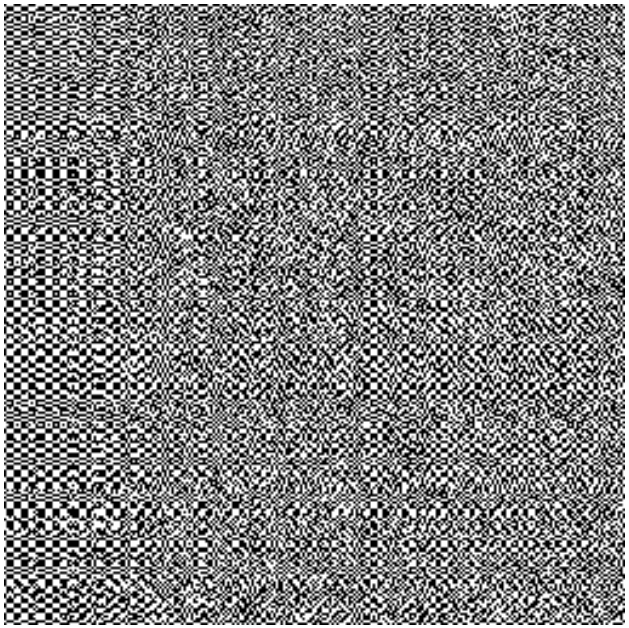
A = matrix



Example from image deblurring

The naive exact solution of $Ax = b$

x = inverse solution



Regularized solution via TSVD or CGLS

659 iterations

Outline

1. Common roots: Moments
2. Linear approximation problems
3. Singular value decomposition and model reduction
4. Matching moments model reduction and Krylov subspace methods
5. Bidiagonalization and linear regression
6. Bidiagonalization and orthogonal regression
7. Back to the roots
8. Conclusions

Example – Linear algebraic equation

Given $Ax = b$ with an SPD matrix A, $r_0 = b - Ax_0$, $v_1 = r_0/\|r_0\|$.

Consider the spectral decomposition

$$A = U \,\mathrm{diag}(\lambda_i)\, U^T ,$$

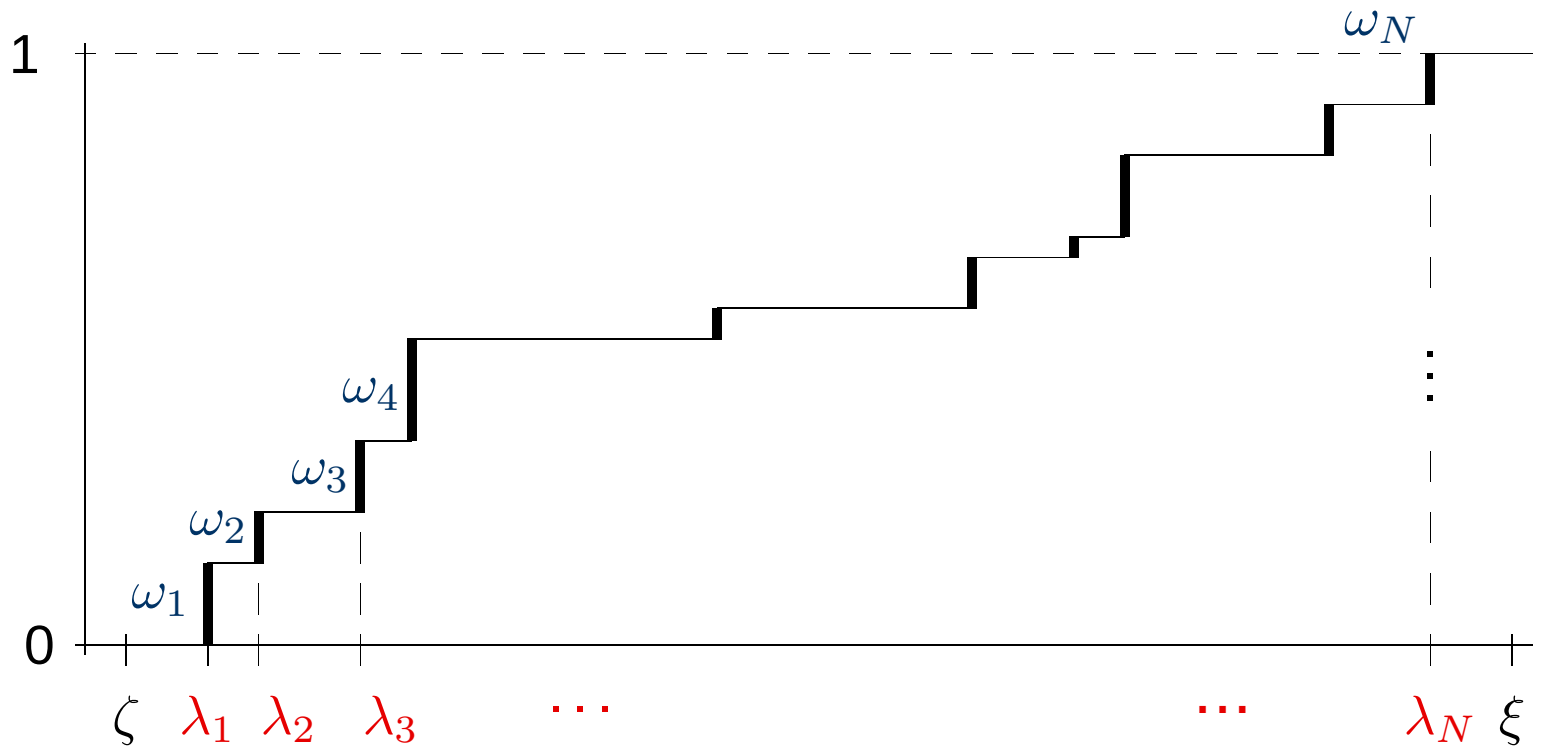
where for clarity of exposition we assume that the eigenvalues are distinct,

$$0 < \lambda_1 < \ldots < \lambda_N , \quad U = [u_1, \ldots , u_N] .$$

A and $v_1(b, x_0)$ determine the distribution function $\omega(\lambda)$ with :

- N points of increase λ_i ,
- weights $\omega_i = |(v_1, u_i)|^2$, $i = 1, \ldots, N$.

Distribution function $\omega(\lambda)$



Stieltjes recurrence for orthogonal polynomials

Let $p_1(\lambda) \equiv 1, p_2(\lambda), \ldots, p_{n+1}(\lambda)$ be the first $n+1$ orthonormal polynomials corresponding to the distribution function $\omega(\lambda)$.

Then, writing $P_n(\lambda) = (p_1(\lambda), \ldots, p_n(\lambda))^T$,

$$\lambda\, P_n(\lambda) \;=\; T_n\, P_n(\lambda) \;+\; \delta_{n+1}\, p_{n+1}(\lambda)\, e_n$$

represents the Stieltjes recurrence (1883-4), with the Jacobi matrix

$$T_n \equiv \begin{pmatrix} \gamma_1 & \delta_2 & & \\ \delta_2 & \gamma_2 & \ddots & \\ & \ddots & \ddots & \delta_n \\ & & \delta_n & \gamma_n \end{pmatrix}, \quad \delta_l > 0\,.$$

Matrix formulation: Lanczos $\equiv$ Stieltjes

In matrix computations, T_n results from the Lanczos process (1951) applied to T_n starting with e_1. Therefore $p_1(\lambda) \equiv 1, p_2(\lambda), \ldots, p_n(\lambda)$ are orthonormal with respect to the inner product

$$(p_s, p_t) \equiv \sum_{i=1}^{n} |(e_1, z_i^{(n)})|^2 \, p_s(\theta_i^{(n)}) \, p_t(\theta_i^{(n)}) \,,$$

where $z_i^{(n)}$ is the orthonormal eigenvector of T_n corresponding to the eigenvalue $\theta_i^{(n)}$, and $p_{n+1}(\lambda)$ has the roots $\theta_i^{(n)}$, $i = 1, \ldots, n$.

Consequently,

$$\omega_i^{(n)} = |(e_1, z_i^{(n)})|^2 \,, \quad \lambda_i^{(n)} = \theta_i^{(n)} \,,$$

Golub and Welsh (1969), ... ,
Meurant and S, Acta Numerica (2006).

Lanczos method $\equiv$ matching moments

$$\int_0^\infty \lambda^k \, d\omega(\lambda) \;=\; \sum_{j=1}^{N} \omega_j \, (\lambda_j)^k \;=\; v_1^T \, A^k \, v_1 \,,$$

$$\sum_{i=1}^{n} \omega_i^{(n)} \, (\lambda_i^{(n)})^k \;=\; \sum_{i=1}^{n} \omega_i^{(n)} \, (\theta_i^{(n)})^k \;=\; e_1^T \, T_n^k \, e_1 \,.$$

matching the first $2n$ moments therefore means

$$v_1^T \, A^k \, v_1 \;\equiv\; e_1^T \, T_n^k \, e_1 \,, \quad k = 0, 1, \ldots, 2n-1 \,.$$

Conjugate gradients m. ≡ matching moments

The CG method, Hestenes and Stiefel (1952), constructs the sequence of approximations

$$x_n \in x_0 + \mathcal{K}_n(A, r_0), \quad \mathcal{K}_n(A, r_0) \equiv \operatorname{span}\{r_0, Ar_0, \ldots, A^{n-1}r_0\},$$

such that

$$\|x - x_n\|_A = \min_{u \in x_0 + \mathcal{K}_n(A,r_0)} \|x - u\|_A$$

which is equivalent to the (Galerkin) orthogonality condition

$$A(x - x_n) \perp \mathcal{K}_n(A, r_0).$$

Using the Lanczos orthogonalization process,

$$T_n\, y_n = \|r_0\|\, e_1\,, \quad x_n = x_0 + V_n\, y_n\,.$$

Matching moments model reduction

CG (Lanczos) reduces for A SPD at the step n the original model

$$A\,x \;=\; b\,,\; r_0 = b - Ax_0$$

to

$$T_n\,y_n \;=\; \|r_0\|\,e_1\,,$$

such that the first $2n$ moments are matched,

$$v_1^*\,A^k\,v_1 \;=\; e_1^T\,T_n^k\,e_1\,, \quad k = 0, 1, \ldots, 2n-1\,.$$

Krylov subspace methods in general represent matching moment model reduction.

Outline

1. Common roots: Moments
2. Linear approximation problems
3. Singular value decomposition and model reduction
4. Matching moments model reduction and Krylov subspace methods
5. Bidiagonalization and linear regression
6. Bidiagonalization and orthogonal regression
7. Back to the roots
8. Conclusions

Linear regression and OLS

Recall

$$A\,x \;=\; b|_{\mathcal{R}(A)}\,, \quad x \,\in\, \mathcal{R}(A^T)\,,$$

$$A^T A\,x \;=\; A^T b\,.$$

Apply CG to the system of normal equations with the matrix $A^T A$ and the right hand side $A^T b$?

Hestenes and Stiefel CGLS (1952)

Let $q_1, \ldots, q_k$ be an orthonormal basis of the Krylov subspace

$$\mathcal{K}_k\left(A^TA,\, A^Tb\right) \equiv \operatorname{span}\left\{A^Tb,\, (A^TA)A^Tb,\, \ldots,\, (A^TA)^{k-1}A^Tb\right\}.$$

Considering $x_k = Q_k\, y_k \in \mathcal{K}_k(A^TA, A^Tb)$, we get $x_k \in \mathcal{R}(A^T)$. Then

$$\|x - x_k\|_{A^TA} = \min_{u \in x_0 + \mathcal{K}_k(A^TA, A^Tb)} \|x - u\|_{A^TA}$$

represents CG applied to $A^TAx = A^Tb$. It gives the approximation to the OLS minimum norm solution

$$A\left(Q_k\, y_k\right) = b + \hat{r}_k\,, \quad \min \|\hat{r}_k\|\,.$$

Golub and Kahan bidiagonalization (1965)

Starting from $p_1 = b/\|b\|$, compute two sequences of orthonormal vectors $p_1, p_2, \ldots, p_{k+1}$ and $q_1, \ldots, q_k$ such that, in the matrix form,

$$A^T P_k = Q_k B_k^T , \quad A Q_k = P_{k+1} B_{k+} ,$$

$$B_k = \begin{pmatrix} \alpha_1 & & & \\ \beta_2 & \ddots & & \\ & \ddots & \ddots & \\ & & \beta_k & \alpha_k \end{pmatrix} , \quad B_{k+} = \begin{pmatrix} \alpha_1 & & & \\ \beta_2 & \ddots & & \\ & \ddots & \ddots & \\ & & \ddots & \alpha_k \\ & & & \beta_{k+1} \end{pmatrix} ,$$

where the matrices $P_{k+1} \equiv [p_1, \ldots, p_{k+1}]$ and $Q_k \equiv [q_1, \ldots, q_k]$ have orthonormal columns, and $\alpha_\ell \geq 0$, $\beta_\ell \geq 0$, $\ell = 1, \ldots$.

Paige and Saunders LSQR (1982)

Using the Golub and Kahan iterative bidiagonalization,

$$A\,(Q_k\,y_k) \;=\; b \,+\, \hat{r}_k\,, \quad \min \|\hat{r}_k\|$$

gives

$$P_{k+1}\,(B_{k+}\,y_k \;-\; \|b\|\,e_1) \;\equiv\; P_{k+1}\,r_k \;=\; \hat{r}_k\,, \quad \|r_k\| = \|\hat{r}_k\|\,.$$

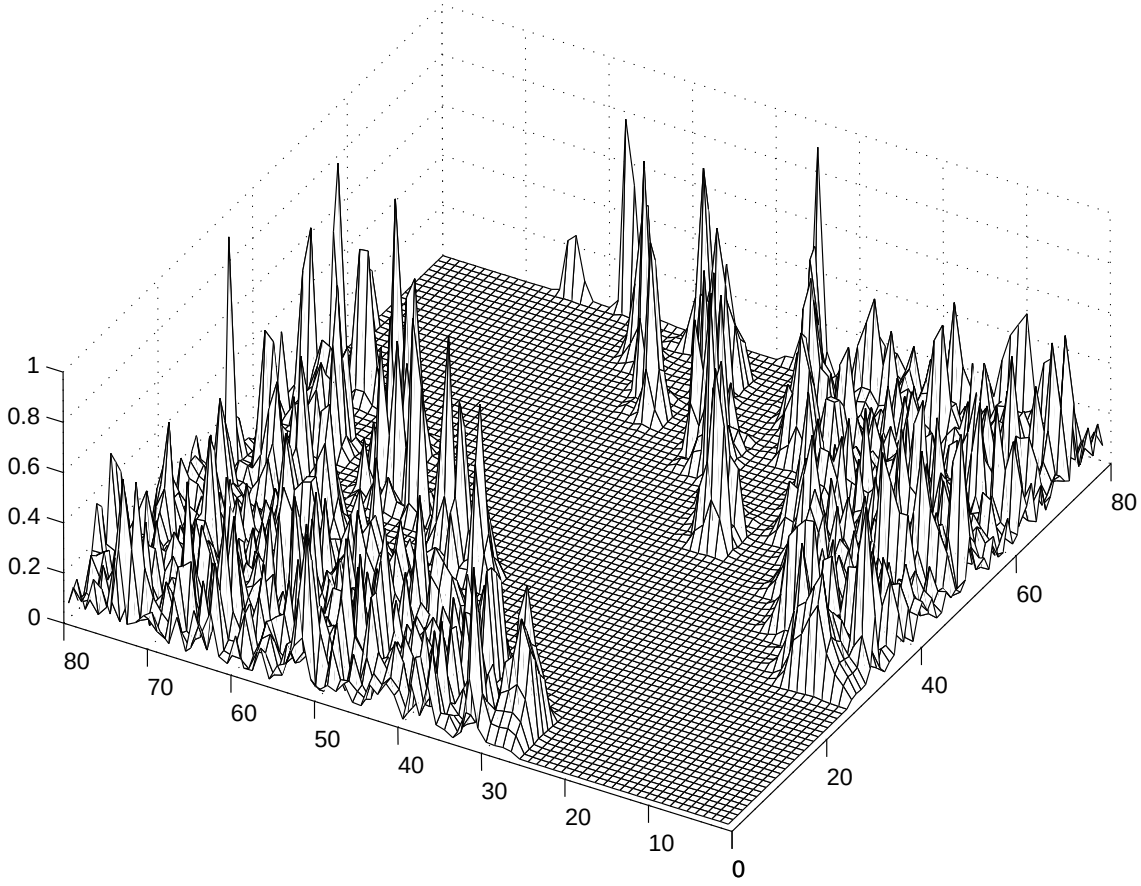
Consequently,

$$B_{k+}\,y_k \;=\; \|b\|\,e_1 \,+\, r_k\,, \quad \min \|r_k\|\,, \quad x_k \;=\; Q_k\,y_k\,.$$

CGLS (1952) $\equiv$ LSQR (1982) $\equiv$ PLS of Wold (1975)

Does the implementation matter? It does!

Loss of orthogonality among the computed Lanczos vectors



Outline

1. Common roots: Moments
2. Linear approximation problems
3. Singular value decomposition and model reduction
4. Matching moments model reduction and Krylov subspace methods
5. Bidiagonalization and linear regression
6. Bidiagonalization and orthogonal regression
7. Back to the roots
8. Conclusions

Orthogonal regression may not have a solution!

The data A, b can suffer from

- multiplicities – the solution may not be unique. Look for the solution minimal in norm.
- conceptual difficulties – when there are stronger colinearities among the columns of A than between the columnspace of A and the right hand side b, the OR (TLS) solution does not exist.

  An extreme example: A not of full column rank and $b \notin \mathbf{R}(A)$.

We need a clear concept of the TLS problem and of its solution which covers all cases, see Paige and S (2002, 2006). It would be ideal to separate the information necessary and sufficient for solving the problem from the rest.

Fundamental block structure

Orthogonal invariance gives $P^T A Q\,(Q^T x) \approx P^T b\,.$ Assume the structure

$$P^T\,[\,b\,,\,A\,Q\,] \;=\; \left[\begin{array}{c||c|c} b_1 & A_{11} & 0 \\ \hline\hline 0 & 0 & A_{22} \end{array}\right].$$

The problem $A\,x \approx b$ can be rewritten as two independent approximation problems

$$\begin{aligned} A_{11}\,x_1 &\approx b_1\,, \\ A_{22}\,x_2 &\approx 0\,, \end{aligned}$$

with the solution $x = Q \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$

A meaningful solution

But $A_{22}\, x_2 \approx 0$ says x_2 lies approximately in the null space of A_{22}, and no more. Thus, unless there is a reason not to, we can set $x_2 = 0$.

Now since we have obtained b with the intent to estimate x , and since x_2 does not contribute to b in any way,

the best we can do is estimate x_1 from $A_{11}\, x_1 \approx b_1$, giving

$$x = Q \begin{bmatrix} x_1 \\ 0 \end{bmatrix} .$$

The transformation (compatible case)

Such an orthogonal transformation is given by the Golub-Kahan bidiagonalization. In fact, A_{22} need not be bidiagonalized, $[b_1, A_{11}]$ has nonzero bidiagonal elements and is either

$$[b_1,\ A_{11}] = \left[\begin{array}{c|ccc} \beta_1 & \alpha_1 & & \\ & \beta_2 & \alpha_2 & \\ & & \ddots & \ddots \\ & & & \beta_p \quad \alpha_p \end{array}\right], \quad \beta_i \alpha_i \neq 0, \quad i = 1, \ldots, p$$

if $\beta_{p+1} = 0$ or $p = N$, (where A is $N \times M$), or

The transformation (incompatible case)

$$[b_1,\ A_{11}] = \left[\begin{array}{c|cccc} \beta_1 & \alpha_1 & & & \\ & \beta_2 & \alpha_2 & & \\ & & \cdot & \cdot & \\ & & & \beta_p & \alpha_p \\ & & & & \beta_{p+1} \end{array}\right], \quad \beta_i \alpha_i \neq 0\,,\ \beta_{p+1} \neq 0$$

if $\alpha_{p+1} = 0$ or $p = M$ (where A is $N \times M$).

Core problem theorem

(a) A_{11} has no zero or multiple singular values, so any zero singular values or repeats that A has must appear in A_{22} .

(b) A_{11} has minimal dimensions, and A_{22} maximal dimensions, over all orthogonal transformations of the form given above.

(c) All components of b_1 in the left singular vector subspaces of A_{11} are nonzero. Consequently, the solution of the TLS problem $A_{11}x_1 \approx b_1$ can be obtained by the standard algorithm of Golub and Van Loan (1980), see also Rao and Toutenburg (1999).

Core problem significance

Any left upper part of the core problem can be seen as a result of the moment matching model reduction.

All information contained in the reduced model is necessary for solving the original problem.

The full core problem contains necessary and sufficient information for solving the original problem.

Outline

1. Common roots: Moments
2. Linear approximation problems
3. Singular value decomposition and model reduction
4. Matching moments model reduction and Krylov subspace methods
5. Bidiagonalization and linear regression
6. Bidiagonalization and orthogonal regression
7. Back to the roots
8. Conclusions

Main points of the lecture

(1) Moments, moment matching model reduction.
Krylov subspace methods represent the matrix formulation of the moment problem.

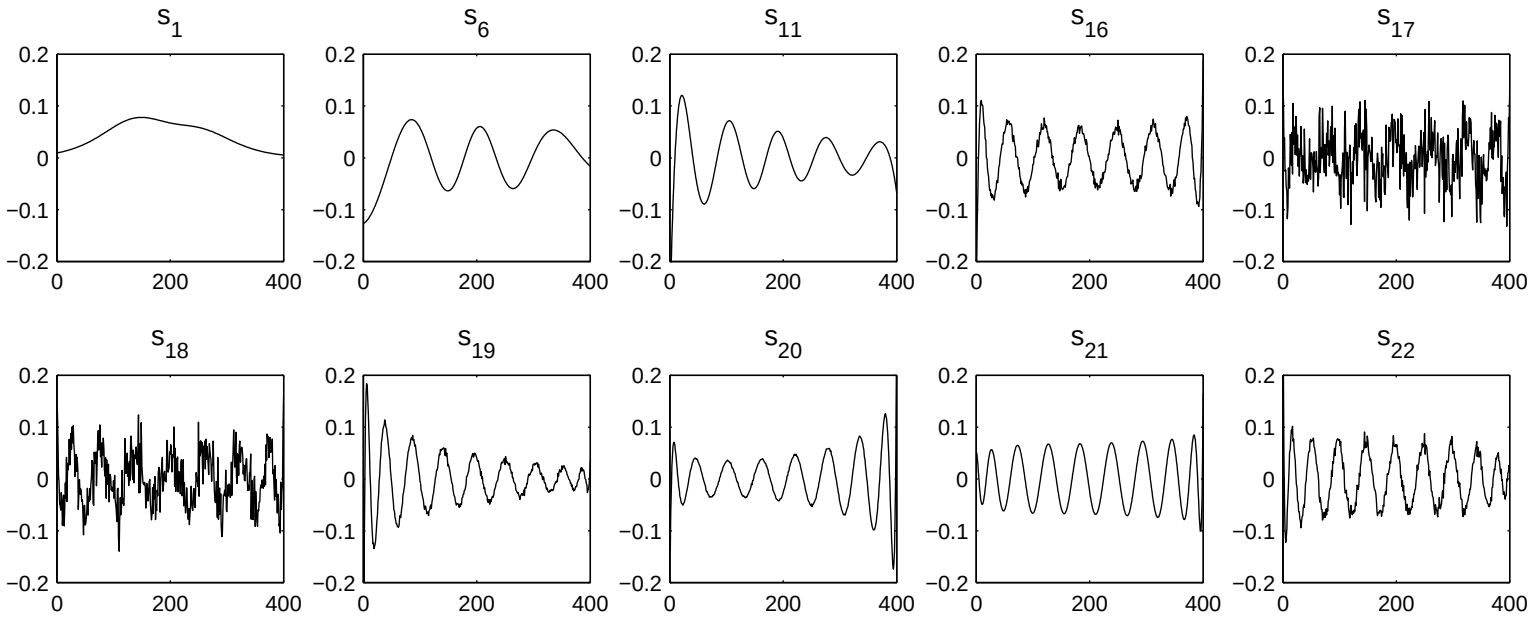
(2) Convergence – accuracy of the approximation.
Numerical stability – reliability of the result.
Complexity – how much does it cost.

(3) Golub-Kahan orthogonal bidiagonalization.
Decomposition of data, OLS, TLS, regularization,
noise revealing, see Hnětynková, Plešinger and S (2008).

(4) Orthogonality as a fundamental principle.
Theoretical and computational.

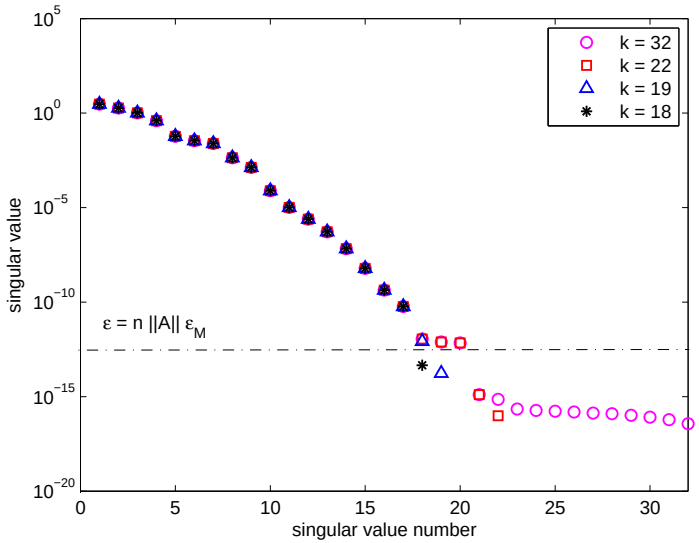
Noise revealing in G-K bidiagonalization

Entries of the left bidiagonalization vectors p_ℓ

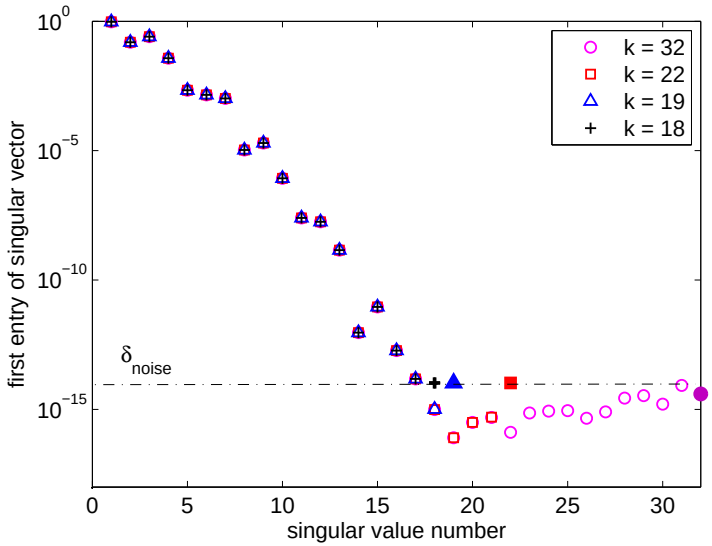


Noise level determination based on moments!

Singular values
of the reduced model

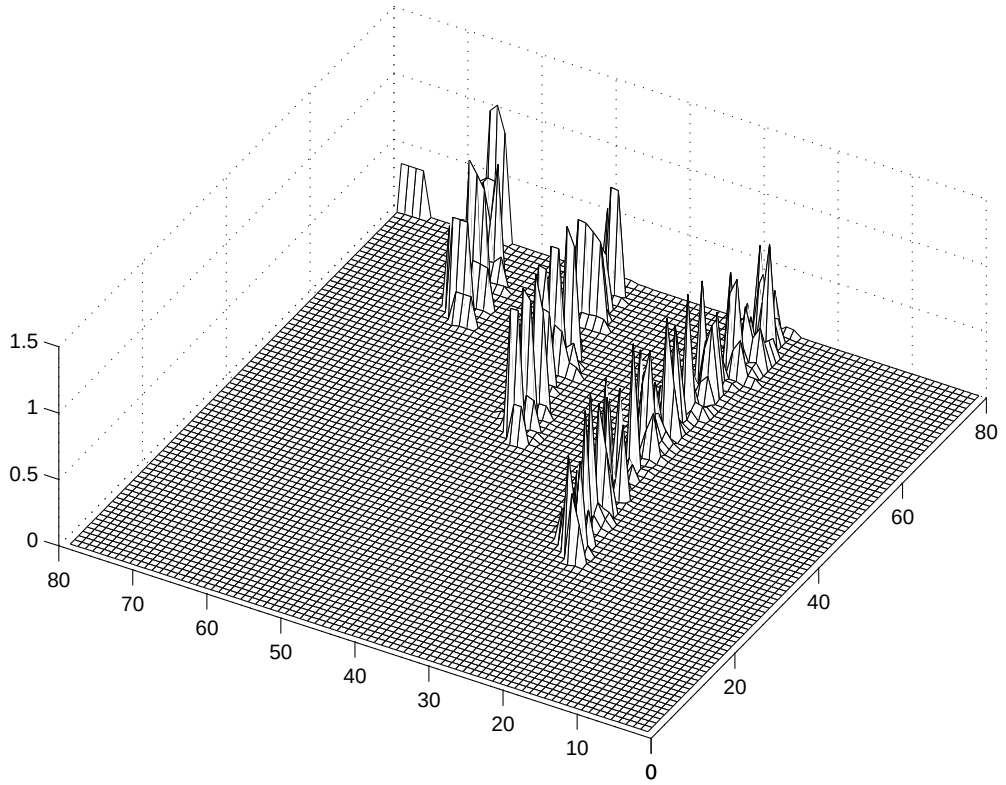


The corresponding weights
in the reduced model



Must orthogonality be always preserved ? **No!**

Loss of orthogonality between the computed Lanczos vectors and the computed Ritz vectors



Outline

1. Common roots: Moments
2. Linear approximation problems
3. Singular value decomposition and model reduction
4. Matching moments model reduction and Krylov subspace methods
5. Bidiagonalization and linear regression
6. Bidiagonalization and orthogonal regression
7. Back to the roots
8. Conclusions

Software

B.J.C. Baxter, A. Iserles, *On the foundations of computational mathematics,* in Handbook of Numerical Analysis XI (P.G. Ciarlet and F. Cucker, eds), North-Holland, Amsterdam (2003), 3-34 :

The attitude of “don’t think, the software will do it for you”,
comforting as it might be to some, will not do.

If one wish to compute,
probably the best initial step is to learn the underlying mathematics,
rather than rushing to a book of numerical recipes.

Even the best software can fail to produce good results if used improperly. Computational modeling requires mastering necessary computational mathematics.

Literature

Rao, R. C. and Toutenburg, H. (1999). *Linear models: least squares and alternatives, second edition,* Springer.

Givens, G. H. and Hoeting, J. A. (2005). *Computational Statistics,* J. Wiley.

Martinez, W. L. and Martinez, A. R. (2002). *Computational Statistics Handbook with Matlab,* Chapman&Hall.

Gentle, J. E., Härdle, W. and Mori, Y. (eds) (2004). *Handbook of Computational statistics, Concepts and Methods,* Springer.

Basics of numerical linear algebra is included, but is seems somehow isolated from description of particular topics in computational statistics. References to relevant NLA literature are very rare. Parallel developments lasting for decades without a single reference to the other field. Some serious computational misconceptions can be found even in very recent monographs.

The philosophical message

- Our world seems to prefer fast and shallow to slow but deep. General trends, also in science, lead to narrow specialization and fragmentation, even within individual disciplines.

- True interdisciplinary approaches mean a new quality which is deeply rooted in different disciplines and which makes bridges between them. A bridge with shallow foundations will not stay for long.

- All fields need mutual transfer of knowledge, which is impossible without building up deep mutual understanding all across the mathematical landscape. This is not always supported by the way the science is financed these days, but it is worth the struggle.

Overlap in our talk

- Moments in statistics, moments in modern NLA.
- PCA and SVD model reduction.
- Ridge regression and regularization.
- Nonlinear regression and optimization.
- PLS and LSQR.
- Orthogonal regression and TLS.

Despite their different focus, a context for application and data analysis on one side and development of generally applicable, reliable, robust and efficient methods and algorithms on the other, computational statistics and numerical linear algebra can enormously benefit from recalling their common roots and developing further their mutual overlap.

Thank you!