Numerical Linear Algebra
Lecture 6

Linear Least Squares Problems

- Suppose that we have a matrix A of the size $m \times n$ and the vector b of the size $m \times 1$. The linear least square problem is to find a vector x of the size $n \times 1$ which will minimize $||Ax - b||_2^2$.
- In the case when $m = n$ and the matrix A is nonsingular we can get solution to this problem as $x = A^{-1}b$.
- When $m > n$ (more equations than unknows) the problem is overdetermined
- When $m < n$ (more unknows than equations) the problem is underdetermined
- Applications: curve fitting, statistical modelling.

Matrix Factorizations that Solve the Linear Least Squares Problem

The linear least squares problem has several explicit solutions that we will discuss:

1. normal equations: the fastest but least accurate; it is adequate when the condition number is small.
2. QR decomposition,

   is the standard one and costs up to twice as much as the first method.
3. SVD, is of most use on an ill-conditioned problem, i.e., when A is not of full rank; it is several times more expensive again.
4. Iterative refinement to improve the solution when the problem is ill-conditioned. Can be adapted to deal efficiently with sparse matrices [Å. Björck. Numerical Methods for Least Squares Problems].

We assume initially for methods 1 and 2 that A has full column rank n.

Linear Least Squares Problems

Further we assume that we will deal with overdetermined problems when we have more equations than unknowns. This means that we will be interested in the solution of linear system of equations

$$Ax = b, \tag{1}$$

where A is of the size $m \times n$ with $m > n$, b is vector of the size m, and x is vector of the size n.
In a general case we are not able to get vector b of the size m as a linear combination of the n columns of the matrix A and n components of the vector x, or there is no solution to (1) in the usual case. We will consider methods which can minimize the residual $r = b - Ax$ as a function on x in principle in any norm, but we will use 2-norm because of the convenience from theoretical (relationships of 2-norm with the inner product and orthogonality, smoothness and strict convexity properties) and computational points of view. Also, because of using 2-norm method is called least squares.

We can write the least squares problem as problem of the minimizing of the squared residuals

$$\|r\|_2^2 = \sum_{i=1}^{m} r_i^2 = \sum_{i=1}^{m} (Ax_i - b)^2. \tag{2}$$

In other words, our goal is to find minimum of this residual using least squares:

$$\min_x \|r\|_2^2 = \min_x \sum_{i=1}^{m} r_i^2 = \min_x \sum_{i=1}^{m} (Ax_i - b)^2. \tag{3}$$

Normal Equations

Our goal is to minimize $\|r(x)\|_2^2 = ||Ax - b||_2^2$. To find minimum we derive the *normal equations*: look for the x where the gradient of $||Ax - b||_2^2 = (Ax - b)^T(Ax - b)$ vanishes, or where $\|r'(x)\|_2^2 = 0$. To derive the Fréchet derivative, we consider the difference $\|r(x + e)\|_2^2 - \|r(x)\|_2^2$ and single out the linear part with respect to x:

$$
\begin{aligned}
0 &= \lim_{\|e\|_2 \to 0} \frac{(A(x+e) - b)^T(A(x+e) - b) - (Ax - b)^T(Ax - b)}{||e||_2} \\
&= \lim_{\|e\|_2 \to 0} \frac{((Ax - b) + Ae)^T((Ax - b) + Ae) - (Ax - b)^T(Ax - b)}{||e||_2} \\
&= \lim_{\|e\|_2 \to 0} \frac{\|(Ax - b) + Ae\|_2^2 - \|Ax - b\|_2^2}{||e||_2} \\
&= \lim_{\|e\|_2 \to 0} \frac{\|Ax - b\|_2^2 + 2(Ax - b) \cdot Ae + \|Ae\|_2^2 - \|Ax - b\|_2^2}{||e||_2} \\
&= \lim_{\|e\| \to 0} \frac{2e^T(A^TAx - A^Tb) + e^TA^TAe}{||e||_2}
\end{aligned}
$$

Normal Equations

Thus,

$$0 = \lim_{\|e\|_2 \to 0} \frac{2e^T(A^TAx - A^Tb) + e^TA^TAe}{||e||_2}. \tag{4}$$

The second term in (4) can be estimated as

$$\lim_{\|e\|_2 \to 0} \frac{|e^TA^TAe|}{||e||_2} \leq \lim_{\|e\|_2 \to 0} \frac{||A||_2^2||e||_2^2}{||e||_2} = \lim_{\|e\|_2 \to 0} ||A||_2^2||e||_2 \to 0 \tag{5}$$

Thus, the first term in (4) must also be zero, or

$$A^TAx = A^Tb \tag{6}$$

Equations (10) is a symmetric linear system of the $n \times n$ linear equations for n unknowns called normal equations.

Normal Equations

Using definition of the residual in the functional

$$\frac{1}{2}\|r(x)\|_2^2 = \frac{1}{2}\|Ax - b\|_2^2 \tag{7}$$

can be computed the Hessian matrix $H = A^T A$. If the Hessian matrix $H = A^T A$ is positive definite, then x is indeed a minimum.
We observe first, that $A^T A$ is symmetric since

$$(A^T A)^T = A^T (A^T)^T = A^T A.$$

In the following Lemma we also prove that it is positive definite.

Normal Equations

Lemma

The matrix $A^T A$ is positive definite if and only if the columns of A are linearly independent, or when $rank(A) = n$ (full column rank).

Proof.

We have that $dim(A) = m \times n$, and thus, $dim(A^T A) = n \times n$. Thus, $\forall v \in R^n$ such that $v \neq 0$

$$v^T A^T A v = (Av)^T (Av) = \|Av\|_2^2 \geq 0. \tag{8}$$

For positive definite matrix $A^T A$ we need to show that $v^T A^T A v > 0$. Assume that $v^T A^T A v = 0$. We observe that $Av = 0$ only if the linear combination $\sum_{i=1}^{n} a_{ji} v_i = 0$. Here, a_{ji} are elements of row j in A. This will be true only if columns of A are linearly dependent or when $v = 0$, but this is contradiction with assumption $v^T A^T A v = 0$ since $v \neq 0$ and thus, the columns of A are linearly independent and $v^T A^T A v > 0$. $\square$

Normal Equations

The final conclusion is that if the matrix A has a full rank ($rank(A) = n$) then the system

$$A^T Ax = A^T b$$

is of the size n-by-n and is s.p.d. system of normal equations. It has the same solution x as the least squares problem $\min_x \|Ax - b\|_2^2$ and can be solved efficiently via Cholesky decomposition for $A^T A = LL^T$:

$$\begin{aligned} LL^T x &= A^T b, \\ L^T x &= L^{-1}(A^T b), \\ x &= (L^T)^{-1}(L^{-1}(A^T b)). \end{aligned}$$

Normal Equations

However, in practice the method of normal equations can be inaccurate by two reasons.

- The condition number of $A^T A$ is twice more than twice more than the condition number of the original matrix A:

$$cond(A^T A) = cond(A)^2. \tag{9}$$

  Thus, the method of normal equations can give a squared condition number even when the fit to data is good and the residual is small. This makes the computed solution more sensitive. In this sense the method of normal equations is not stable.

- Information can be lost during computation of the product of $A^T A$.

Normal Equations: loss of information in a given floating-point system

Example

$$A = \begin{pmatrix} 1 & 1 \\ \delta & 0 \\ 0 & \delta \end{pmatrix} \tag{10}$$

with $0 < \delta < \sqrt{\varepsilon}$ in a given floating-point system. In floating-point arithmetics we can compute $A^T A$:

$$A^T A = \begin{pmatrix} 1 & \delta & 0 \\ 1 & 0 & \delta \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 \\ \delta & 0 \\ 0 & \delta \end{pmatrix} = \begin{pmatrix} 1+\delta^2 & 1 \\ 1 & 1+\delta^2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \tag{11}$$

which is singular matrix in the working precision.

Data fitting

In this example we present the typical application of least squares called data or curve fitting problem. This problem appears in statistical modelling and experimental engineering when data are generated by laboratory or other measurements.
Suppose that we have data points $(x_i, y_i), i = 1, ..., m$, and our goal is to find the vector of parameters c of the size n which will fit best to the data y_i of the model function $f(x_i, c)$, where $f : R^{n+1} \rightarrow R$, in the least squares sense:

$$\min_c \sum_{i=1}^{m} (y_i - f(x_i, c))^2. \tag{12}$$

If the function $f(x, c)$ is linear then we can solve the problem (12) using least squares method.

The function $f(x, c)$ is linear if we can write it as a linear combination of the functions $\phi_j(x), j = 1, ..., n$ as:

$$f(x, c) = c_1\phi_1(x) + c_2\phi_2(x) + ... + c_n\phi_n(x). \quad (13)$$

Functions $\phi_j(x), j = 1, ..., n$ are called basis functions.
Let now the matrix A will have entries
$a_{ij} = \phi_j(x_i), i = 1, ..., m; j = 1, ..., n$, and vector b will be such that
$b_i = y_i, i = 1, ..., m$. Then a linear data fitting problem takes the form of (1) with $x = c$:

$$Ac \approx b \quad (14)$$

Elements of the matrix A are created by basis functions
$\phi_j(x), j = 1, ..., n$. We will consider now different examples of choosing basis functions $\phi_j(x), j = 1, ..., n$.

Problem of the fitting to a polynomial

In the problem of the fitting to a polynomial

$$f(x, c) = \sum_{i=1}^{d} c_i x^{i-1} \tag{15}$$

of degree $d-1$ to data points $(x_i, y_i), i = 1, ..., m$, basis functions $\phi_j(x), j = 1, ..., n$ can be chosen as $\phi_j(x) = x^{j-1}, j = 1, ..., n$. The matrix A constructed by these basis functions in a polynomial fitting problem is a Vandermonde matrix:

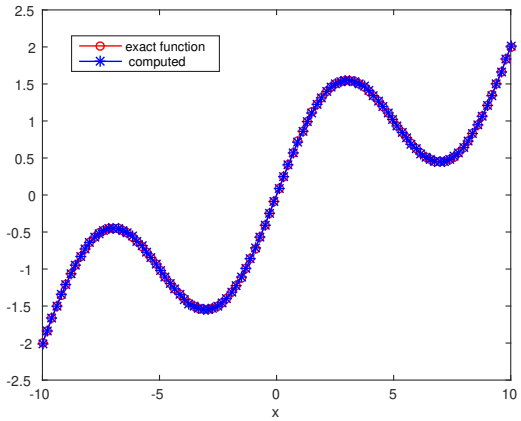
$$A = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{d-1} \\ 1 & x_2 & x_2^2 & \dots & x_2^{d-1} \\ 1 & x_3 & x_3^2 & \dots & x_3^{d-1} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & x_m & x_m^2 & \dots & x_m^{d-1} \end{bmatrix}. \tag{16}$$

Here, $x_i, i = 1, ..., m$ are discrete points on the interval for $x = [x_{left}, x_{right}]$.

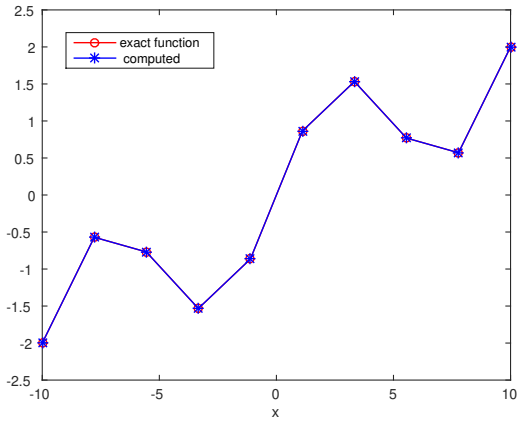
Suppose, that we choose $d = 4$ in (12). Then we can write the polynomial as $f(x, c) = \sum_{i=1}^{4} c_i x^{i-1} = c_1 + c_2 x + c_3 x^2 + c_4 x^3$ and our data fitting problem (14) for this polynomial takes the form

$$\begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ 1 & x_3 & x_3^2 & x_3^3 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_m & x_m^2 & x_m^3 \end{bmatrix} \cdot \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \dots \\ b_m \end{bmatrix}. \tag{17}$$

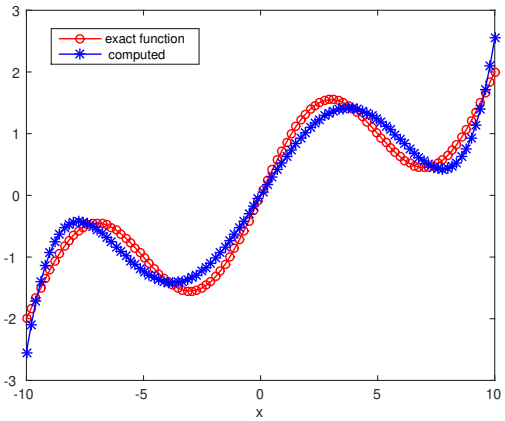
The right hand side of the above system represents measurements or function which we want to fit. Our goal is to find such coefficients $c = \{c_1, c_2, c_3, c_4\}$ which will minimize the residual $r_i = f(x_i, c) - b_i, i = 1..., m$. Since we want minimize squared 2-norm of the residual, or $\|r\|_2^2 = \sum_{i=1}^{m} r_i^2$, then we will solve the linear least squares problem.
Let us consider an example when the right hand side $b_i, i = 1, ...m$ is taken as a smooth function $b = sin(\pi x/5) + x/5$. Figure on the next slide shows polynomial fitting to the function $b = sin(\pi x/5) + x/5$ for different d in (15) on the interval $x \in [-10, 10]$. Using this figure we observe that with increasing of the degree of the polynomial $d - 1$ we have better fit to the exact function $b = sin(\pi x/5) + x/5$. However, for the degree of the polynomial more than 18 we get erratic fit to the function. This happens because matrix A becomes more and more ill-conditioned with increasing of the degree of the polynomial d. And this is, in turn, because of the linear dependence of the columns in the Vandermonde's matrix A.



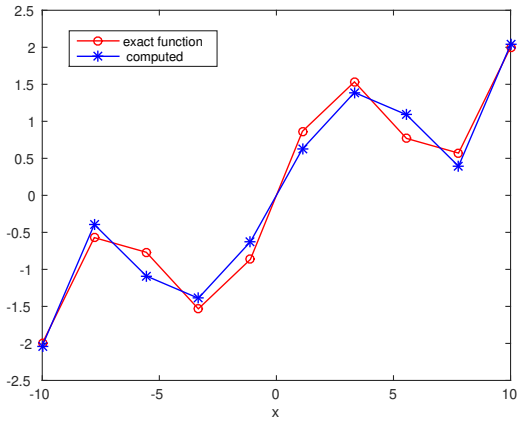
a) d=10



b) d=10



c) d=5



d) d=5

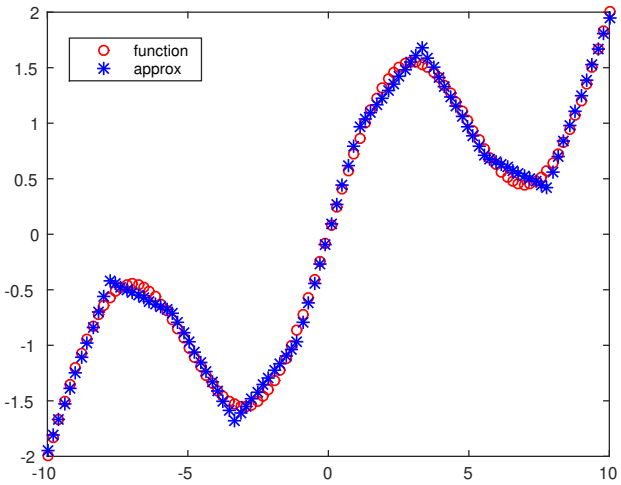
Figure: *Polynomial fitting for different d in (15) to the function* $b = \sin(\pi x/5) + x/5$ *on the interval* $x \in [-10, 10]$ *using the method of normal equations. On the left figures: fit to the 100 points* x_i, $i = 1, \ldots, 100$; *on the right figures: fit to the 10 points* x_i, $i = 1, \ldots, 10$. *Lines with blue stars represent computed function and with red circles – exact one.*

Approximation using linear splines

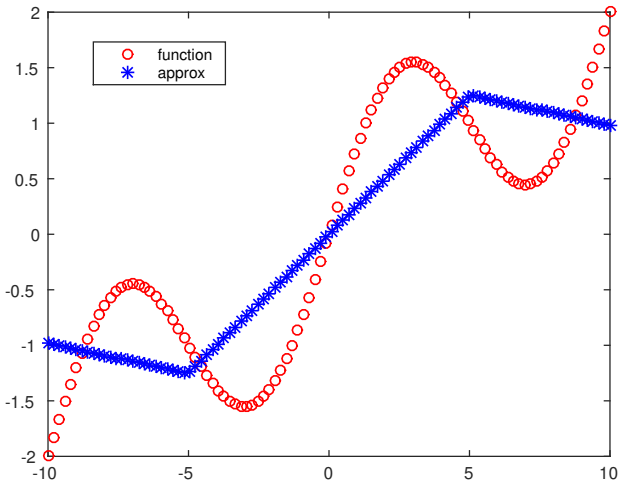
When we want to solve the problem (12) of the approximation to the data vector $y_i, i = 1, ..., m$ with linear splines we use following basis functions $\phi_j(x), j = 1, ..., n$, in (13) which are called also hat functions:

$$\phi_j(x) = \begin{cases} \frac{x - T_{j-1}}{T_j - T_{j-1}}, & T_{j-1} \leq x \leq T_j, \\ \frac{T_{j+1} - x}{T_{j+1} - T_j}, & T_j \leq x \leq T_{j+1}. \end{cases} \tag{18}$$

Here, the column j in the matrix A is constructed by the given values of $\phi_j(x)$ at points $T_j, j = 1, .., n$, which are called conjunction points and are chosen by the user. Using (18) we can conclude that the first basis function is $\phi_1(x) = \frac{T_2 - x}{T_2 - T_1}$ and the last one is $\phi_n(x) = \frac{x - T_{n-1}}{T_n - T_{n-1}}$.
Figure on the next slide shows approximation of a function $b = sin(\pi x/5) + x/5$ on the interval $x \in [-10, 10]$ using linear splines with different number n of conjunction points $T_j, j = 1, ..., n$.



a) n=10



b) n=5

Figure: *Polynomial fitting to the function* $b = sin(\pi x/5) + x/5$ *on the interval* $x \in [-10, 10]$ *using linear splines with different number n of conjunction points* $T_j, j = 1, \ldots, n$ *in (18). Blue stars represent computed function and red circles - exact one.*

Approximation using bellsplines

In the case when we want to solve the problem (12) using bellsplines, the number of bellsplines which can be constructed are $n+2$, and the function $f(x,c)$ in (12) is written as

$$f(x,c) = c_1\phi_1(x) + c_2\phi_2(x) + ... + c_{n+2}\phi_{n+2}(x). \qquad (19)$$

We define

$$\phi_j^0(x) = \begin{cases} 1, & T_j \leq x \leq T_{j+1}, \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

Then all other basis functions, or bellsplines, $\phi_j^k(x), j = 1, ..., n+2; k = 1, 2, 3$ are defined as follows:

$$\phi_j^k(x) = (x - T_k)\frac{\phi_j^{k-1}(x)}{T_{j+k} - T_j} + (T_{j+k+1} - x)\frac{\phi_{j+1}^{k-1}(x)}{T_{j+k+1} - T_{j+1}}. \tag{21}$$

Here, the column j in the matrix A is constructed by the given values of $\phi_j(x)$ at conjunction points $T_j, j = 1, .., n$ which are chosen by the user. If in (21) we obtain ratio $0/0$, then we assign $\phi_j^k(x) = 0$. We define additional three points T_{-2}, T_{-1}, T_0 at the left side of the input interval as $T_{-2} = T_{-1} = T_0 = T_1$, and correspondingly three points $T_{n+1}, T_{n+2}, T_{n+3}$ on the right side of the interval as $T_n = T_{n+1} = T_{n+2} = T_{n+3}$. All together we have $n+6$ conjunction points $T_j, j = 1, ..., n+6$. Number of bellsplines which can be constructed are $n+2$.

If conjunction points T_j are distributed uniformly, then we can introduce the mesh size $h = T_{k+1} - T_k$ and bellsplines can be written explicitly as

$$\phi_j(x) = \begin{cases} \frac{1}{6}t^3 & \text{if } T_{j-2} \le x \le T_{j-1},\ t = \frac{1}{h}(x - T_{j-2}), \\ \frac{1}{6} + \frac{1}{2}(t + t^2 - t^3) & \text{if } T_{j-1} \le x \le T_j,\ t = \frac{1}{h}(x - T_{j-1}), \\ \frac{1}{6} + \frac{1}{2}(t + t^2 - t^3) & \text{if } T_j \le x \le T_{j+1},\ t = \frac{1}{h}(T_{j+1} - x), \\ \frac{1}{6}t^3 & \text{if } T_{j+1} \le x \le T_{j+2}, t = \frac{1}{h}(T_{j+2} - x). \end{cases} \tag{22}$$

In the case of uniformly distributed bellsplines we place additional points at the left side of the input interval as
$T_0 = T_1 - h, T_{-1} = T_1 - 2h, T_{-2}T_1 - 3h$, and correspondingly on the right side of the interval as
$T_{n+1} = T_n + h, T_{n+2} = T_n + 2h, T_{n+3} = T_n + 3h$. Then the function $f(x, c)$ in (12) will be the following linear combination of $n + 2$ functions $\phi_j(x)$ for indices $j = 0, 1, ..., n + 1$:

$$f(x, c) = c_1\phi_0(x) + c_2\phi_1(x) + ... + c_{n+2}\phi_{n+1}(x). \tag{23}$$

Figure on the next slide shows approximation of a function $b = sin(\pi x/5) + x/5$ on the interval $x \in [-10, 10]$ using bellsplines.

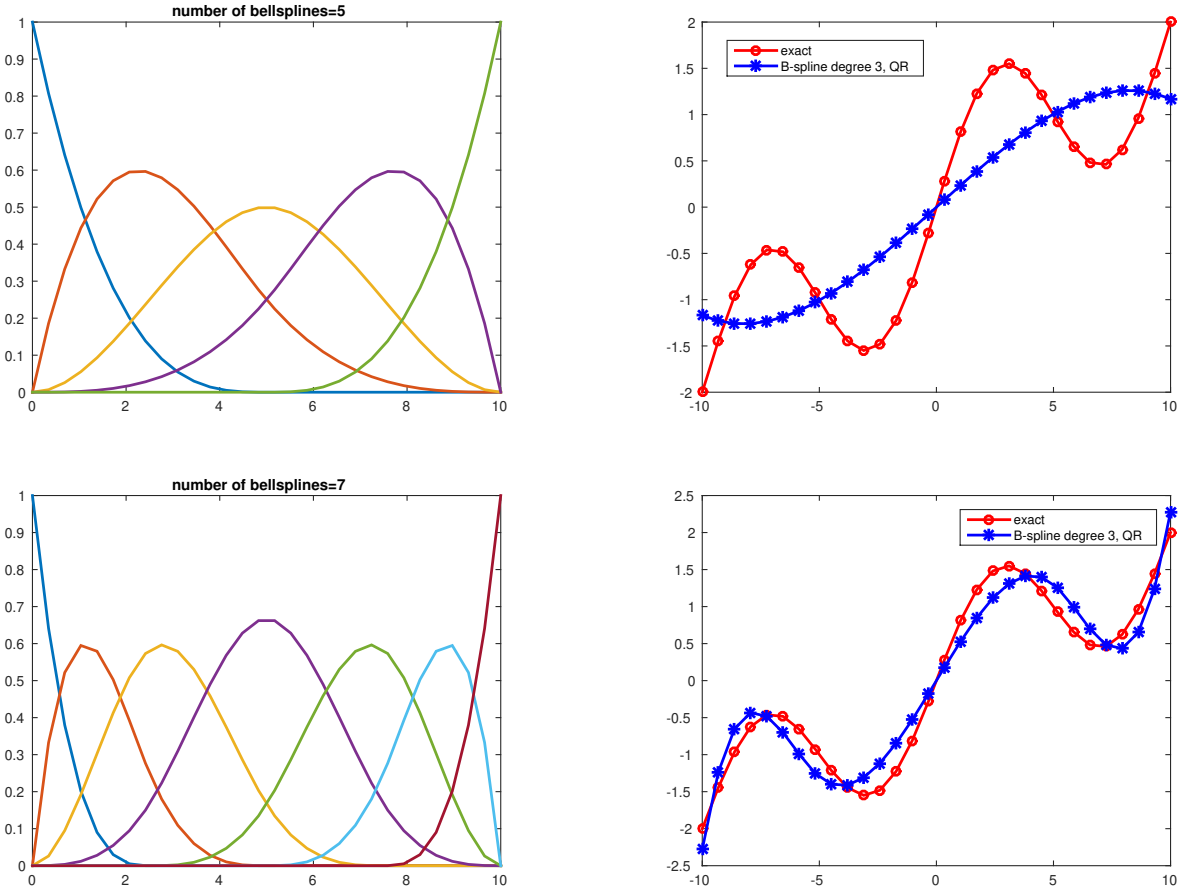


Figure: *Polynomial fitting to the function* $b = sin(\pi x/5) + x/5$ *on the interval* $x \in [-10, 10]$ *with different number of bellsplines. Blue stars represent computed function and red circles - exact one.*

QR Decomposition

THEOREM QR decomposition. *Let A be m-by-n with $m \geq n$. Suppose that A has full column rank. Then there exist a unique m-by-n orthogonal matrix $Q(Q^T Q = I_n)$ and a unique n-by-n upper triangular matrix R with positive diagonals $r_{ii} > 0$ such that $A = QR$.*

Proof. Can be two proofs of this theorem: using the Gram-Schmidt orthogonalization process and using the Hauseholder reflections. The first proof: this theorem is a restatement of the Gram-Schmidt orthogonalization process [P. Halmos. Finite Dimensional Vector Spaces. Van Nostrand, New York, 1958]. If we apply Gram-Schmidt to the columns a_i of $A = [a_1, a_2, \ldots, a_n]$ from left to right, we get a sequence of **orthonormal vectors** (if they are orthogonal and unit vectors) q_1 through q_n spanning the same space: these orthogonal vectors are the columns of Q. Gram-Schmidt also computes coefficients $r_{ji} = q_j^T a_i$ expressing each column a_i as a linear combination of q_1 through q_i: $a_i = \sum_{j=1}^{i} r_{ji} q_j$. The r_{ji} are just the entries of R.

ALGORITHM *The classical Gram-Schmidt (CGS) and modified Gram-Schmidt (MGS) Algorithms for factoring* $A = QR$:

for $i = 1$ *to* n */* compute* i*th columns of* Q *and* R **/*

$q_i = a_i$

for $j = 1$ *to* $i - 1$ */* subtract component in* q_j *direction from* a_i **/*

$$\begin{cases} r_{ji} = q_j^T a_i & CGS \\ r_{ji} = q_j^T q_i & MGS \end{cases}$$

$q_i = q_i - r_{ji} q_j$

end for

$r_{ii} = ||q_i||_2$

if $r_{ii} = 0$ */** a_i *is linearly dependent on* $a_1, \ldots, a_{i-1}$ **/*

quit

end if

$q_i = q_i / rii$

end for

If A has full column rank, r_{ii} will not be zero.

Notes:

- Unfortunately, CGS is numerically unstable in floating point arithmetic when the columns of A are nearly linearly dependent.
- MGS is more stable and will be used in algorithms later in this course but may still result in Q being far from orthogonal ($||Q^T Q - I||$ being far larger than ε) when A is ill-conditioned
- Literature on this subject:

Å. Björck. Solution of Equations volume 1 of Handbook of Numerical Analysis, chapter Least Squares Methods. Elsevier/North Holland, Amsterdam, 1987.

Å. Björck. Least squares methods. Mathematics Department Report, Linkoping University, 1991.

Å. Björck. Numerical Methods for Least Squares Problems. SIAM, Philadelphia, PA, 1996.

N. J. Higham. Accuracy and Stability of Numerical Algorithms. SIAM, Philadelphia, PA, 1996.

We will derive the formula for the x that minimizes $||Ax - b||_2$ using the decomposition $A = QR$ in three slightly different ways. First, we can always choose $m - n$ more **orthonormal vectors** $\tilde{Q}$ so that $[Q, \tilde{Q}]$ is a square orthogonal matrix and thus $\tilde{Q}^T Q = 0$ (for example, we can choose any $m - n$ more independent vectors $\tilde{X}$ that we want and then apply QR Algorithm to the n-by-n nonsingular matrix $[Q, \tilde{X}]$). Then

$$
\begin{aligned}
||Ax - b||_2^2 &= ||[Q, \tilde{Q}]^T (Ax - b)||_2^2 \\
&= \left\| \begin{bmatrix} Q^T \\ \tilde{Q}^T \end{bmatrix} (QRx - b) \right\|_2^2 \\
&= \left\| \begin{bmatrix} I^{n \times n} \\ O^{(m-n) \times n} \end{bmatrix} Rx - \begin{bmatrix} Q^T b \\ \tilde{Q}^T b \end{bmatrix} \right\|_2^2 \\
&= \left\| \begin{bmatrix} Rx - Q^T b \\ -\tilde{Q}^T b \end{bmatrix} \right\|_2^2 \\
&= \left\| Rx - Q^T b \right\|_2^2 + \left\| \tilde{Q}^T b \right\|_2^2 \\
&\geq \left\| \tilde{Q}^T b \right\|_2^2.
\end{aligned}
$$

We can solve $Rx - Q^T b = 0$ for x, since A and R have the same rank, n, and so R is nonsingular. Then $x = R^{-1} Q^T b$, and the minimum value of $||Ax - b||_2$ is $||\tilde{Q}^T b||_2$.

Here is a second, slightly different derivation that does not use the matrix $\tilde{Q}$. Rewrite $Ax - b$ as

$$\begin{aligned} Ax - b &= QRx - b = QRx - (QQ^T + I - QQ^T)b \\ &= Q(Rx - Q^T b) - (I - QQ^T)b. \end{aligned}$$

Note that the vectors $Q(Rx - Q^T b)$ and $(I - QQ^T)b$ are orthogonal, because $(Q(Rx - Q^T b))^T((I - QQ^T)b) =$
$(Rx - Q^T b)^T[Q^T(I - QQ^T)]b = (Rx - Q^T b)^T[0]b = 0$. Therefore, by the Pythagorean theorem,

$$\begin{aligned} \|Ax - b\|_2^2 &= \|Q(Rx - Q^T b)\|_2^2 + \|(I - QQ^T)b\|_2^2 \\ &= \|Rx - Q^T b\|_2^2 + \|(I - QQ^T)b\|_2^2. \end{aligned}$$

where we have used $||Qy||_2^2 = ||y||_2^2$. This sum of squares is minimized when the first term is zero, i.e., $x = R^{-1}Q^T b$.

Finally, here is a third derivation that starts from the normal equations solution:

$$\begin{aligned} x &= (A^T A)^{-1} A^T b \\ &= (R^T Q^T Q R)^{-1} R^T Q^T b = (R^T R)^{-1} R^T Q^T b \\ &= R^{-1} R^{-T} R^T Q^T b = R^{-1} Q^T b. \end{aligned}$$

Singular values

The singular values, or s-numbers of a compact operator $T : X \to Y$ acting between Hilbert spaces X and Y, are the square roots of the eigenvalues of the nonnegative self-adjoint operator $T^*T : X \to X$ (where T^* denotes the adjoint of T).

$$\sigma(T) = \sqrt{\lambda(T^*T)}.$$

The singular values are nonnegative real numbers, usually listed in decreasing order $(s_1(T), s_2(T), ...)$. If T is self-adjoint, then the largest singular value $s1(T)$ is equal to the operator norm of T.
In the case of a normal matrix A (or $A^*A = AA^*$, when A is real then $A^TA = AA^T$), the spectral theorem can be applied to obtain unitary diagonalization of A as $A = U\Lambda U^*$. Therefore, $\sqrt{A^*A} = U|\Lambda|U^*$ and so the singular values are simply the absolute values of the eigenvalues.

Singular Value Decomposition

THEOREM SVD. *Let A be an arbitrary m-by-n matrix with $m \geq n$. Then we can write $A = U\Sigma V^T$, where U is m-by-n and satisfies $U^T U = I$, V is n-by-n and satisfies $V^T V = I$, and $\Sigma = diag(\sigma_1, \ldots, \sigma_n)$, where $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$. The columns $u_1, \ldots, u_n$ of U are called $left\ singular\ vectors$. The columns $v_1, \ldots, v_n$ of V are called $right\ singular\ vectors$. The σ_i are called $singular\ values$. (If $m < n$, the SVD is defined by considering A^T.)*

THEOREM *Let $A = U\Sigma V^T$ be the SVD of the m-by-n matrix A, where $m \geq n$. (There are analogous results for $m < n$.)*

- 1. Suppose that A is symmetric, with eigenvalues λ_i and orthonormal eigenvectors u_i. In other words $A = U\Lambda U^T$ is an eigendecomposition of A, with $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$, and $U = [u_1, \ldots, u_n]$, and $UU^T = I$. Then an SVD of A is $A = U\Sigma V^T$, where $\sigma_i = |\lambda_i|$ and $\upsilon_i = sign(\lambda_i)u_i$, where $sign(0) = 1$.
- 2. The eigenvalues of the symmetric matrix A^TA are σ_i^2. The right singular vectors υ_i are corresponding orthonormal eigenvectors.
- 3. The eigenvalues of the symmetric matrix AA^T are σ_i^2 and $m - n$ zeroes. The left singular vectors u_i are corresponding orthonormal eigenvectors for the eigenvalues σ_i^2. One can take any $m - n$ other orthogonal vectors as eigenvectors for the eigenvalue 0.

- 4. Let $H = \begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix}$, where A is square and $A = U\Sigma V^T$ is the SVD of A. Let $\Sigma = diag(\sigma_1, \ldots, \sigma_n)$, $U = [u_1, \ldots, u_n]$, and $V = [v_1, \ldots, v_n]$. Then the 2n eigenvalues of H are $\pm\sigma_i$, with corresponding unit eigenvectors $\frac{1}{\sqrt{2}} \begin{bmatrix} v_i \\ \pm u_i \end{bmatrix}$.
- 5. If A has full rank, the solution of $\min_x \|Ax - b\|_2$ is $x = V\Sigma^{-1}U^T b$.
- 6. $\|A\|_2 = \sigma_1$. If A is square and nonsingular, then $\|A^{-1}\|_2^{-1} = \sigma_n$ and $\|A\|_2 \cdot \|A^{-1}\|_2 = \frac{\sigma_1}{\sigma_n}$.
- 7. Write $V = [v_1, v_2, \ldots, v_n]$ and $U = [u_1, u_2, \ldots, u_n]$, so $A = U\Sigma V^T = \sum_{i=1}^{n} \sigma_i u_i v_i^T$ (a sum of rank-1 matrices). Then a matrix of rank $k < n$ closest to A (measured with $||\cdot||_2$) is $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$ and $||A - A_k||_2 = \sigma_{k+1}$. We may also write $A_k = U\Sigma_k V^T$ where $\Sigma_k = diag(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$.

Proof.

1. **Suppose that A is symmetric, with eigenvalues λ_i and orthonormal eigenvectors u_i. In other words $A = U\Lambda U^T$ is an eigendecomposition of A, with $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$, and $U = [u_1, \ldots, u_n]$, and $UU^T = I$. Then an SVD of A is $A = U\Sigma V^T$, where $\sigma_i = |\lambda_i|$ and $v_i = sign(\lambda_i)u_i$, where $sign(0) = 1$.**

This is true by the definition of the SVD.

2. **The eigenvalues of the symmetric matrix $A^T A$ are σ_i^2. The right singular vectors v_i are corresponding orthonormal eigenvectors.**
$A^T A = V \Sigma U^T U \Sigma V^T = V \Sigma^2 V^T$. This is an eigendecomposition of $A^T A$, with the columns of V the eigenvectors and the diagonal entries of Σ^2 the eigenvalues.

3. **The eigenvalues of the symmetric matrix AA^T are σ_i^2 and $m-n$ zeroes. The left singular vectors u_i are corresponding orthonormal eigenvectors for the eigenvalues σ_i^2. One can take any $m-n$ other orthogonal vectors as eigenvectors for the eigenvalue 0.**

Choose an m-by-$(m-n)$ matrix $\tilde{U}$ so that $[U, \tilde{U}]$ is square and orthogonal. Then write

$$AA^T = U\Sigma V^T V\Sigma U^T = U\Sigma^2 U^T = \left[U, \tilde{U}\right] \cdot \begin{bmatrix} \Sigma^2 & 0 \\ 0 & 0 \end{bmatrix} \cdot \left[U, \tilde{U}\right]^T.$$

This is an eigendecomposition of AA^T.

4. **Let** $H = \begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix}$, **where** A **is square and** $A = U\Sigma V^T$ **is the SVD of** A. **Let** $\Sigma = diag(\sigma_1, \ldots, \sigma_n)$, $U = [u_1, \ldots, u_n]$, **and** $V = [v_1, \ldots, v_n]$. **Then the 2n eigenvalues of** H **are** $\pm\sigma_i$, **with corresponding unit eigenvectors** $\frac{1}{\sqrt{2}} \begin{bmatrix} v_i \\ \pm u_i \end{bmatrix}$.

We substitute $A = U\Sigma V^T$ into H to get:

$$H = \begin{bmatrix} 0 & V\Sigma U^T \\ U\Sigma V^T & 0 \end{bmatrix}$$

Choose orthogonal matrix G such that

$$G = \frac{1}{\sqrt{2}} \begin{bmatrix} V & V \\ U & -U \end{bmatrix}$$

It is orthogonal since

$$I = GG^T = \frac{1}{2} \begin{bmatrix} VV^T + VV^T & 0 \\ 0 & UU^T + UU^T \end{bmatrix}$$

Then we observe that

$$G \begin{bmatrix} \Sigma & 0 \\ 0 & \Sigma \end{bmatrix} G^T = \begin{bmatrix} 0 & V\Sigma U^T \\ U\Sigma V^T & 0 \end{bmatrix} = H$$

Then using the spectral theorem we can conclude that the $2n$ eigenvalues of H are $\pm\sigma_i$, with corresponding eigenvectors $\frac{1}{\sqrt{2}} \begin{bmatrix} v_i \\ \pm u_i \end{bmatrix}$.

5. **If A has full rank, the solution of $\min_x \|Ax - b\|_2$ is $x = V\Sigma^{-1}U^T b$.**

$\|Ax - b\|_2^2 = \|U\Sigma V^T x - b\|_2^2$. Since A has full rank, so does Σ, and thus Σ is invertible. Now let $[U, \tilde{U}]$ be square and orthogonal as above so

$$\begin{aligned} \|U\Sigma V^T x - b\|_2^2 &= \left\| \begin{bmatrix} U^T \\ \tilde{U}^T \end{bmatrix} (U\Sigma V^T x - b) \right\|_2^2 \\ &= \left\| \begin{bmatrix} \Sigma V^T x - U^T b \\ -\tilde{U}^T b \end{bmatrix} \right\|_2^2 \\ &= \|\Sigma V^T x - U^T b\|_2^2 + \|\tilde{U}^T b\|_2^2. \end{aligned}$$

This is minimized by making the first term zero, i.e.,
$x = V\Sigma^{-1}U^T b$.

6. $\|A\|_2 = \sigma_1$. **If A is square and nonsingular, then $\|A^{-1}\|_2^{-1} = \sigma_n$ and $\|A\|_2 \cdot \|A^{-1}\|_2 = \frac{\sigma_1}{\sigma_n}$.**

It is clear from its definition that the two-norm of a diagonal matrix is the largest absolute entry on its diagonal. Thus, by property of the norm, $\|A\|_2 = \|U^T A V\|_2 = \|U^T U \Sigma V^T V\|_2 = \|\Sigma\|_2 = \sigma_1$ and $\|A^{-1}\|_2 = \|V^T A^{-1} U\|_2 = \|\Sigma^{-1}\|_2 = \sigma_n^{-1}$.

Remark:

$\|A^{-1}\|_2 = \|V^T A^{-1} U\|_2 = \|V^T (U \Sigma V^T)^{-1} U\|_2 = \|\Sigma^{-1}\|_2 = \sigma_n^{-1}$.

7. Write $V = [v_1, v_2, \ldots, v_n]$ and $U = [u_1, u_2, \ldots, u_n]$, so $A = U\Sigma V^T = \sum_{i=1}^{n} \sigma_i u_i v_i^T$ (a sum of rank-1 matrices). Then a matrix of rank $k < n$ closest to A (measured with $||\cdot||_2$) is $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$ and $||A - A_k||_2 = \sigma_{k+1}$. We may also write $A_k = U\Sigma_k V^T$ where $\Sigma_k = diag(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$.

7. Write $V = [v_1, v_2, \ldots, v_n]$ and $U = [u_1, u_2, \ldots, u_n]$, so $A = U\Sigma V^T = \sum_{i=1}^{n} \sigma_i u_i v_i^T$ (a sum of rank-1 matrices). Then a matrix of rank $k < n$ closest to A (measured with $||\cdot||_2$) is $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$ and $||A - A_k||_2 = \sigma_{k+1}$. We may also write $A_k = U\Sigma_k V^T$ where $\Sigma_k = diag(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$.
A_k has rank k by construction and

$$||A - A_k||_2 = \left\| \sum_{i=1}^{n} \sigma_i u_i v_i^T - \sum_{i=1}^{k} \sigma_i u_i v_i^T \right\|$$

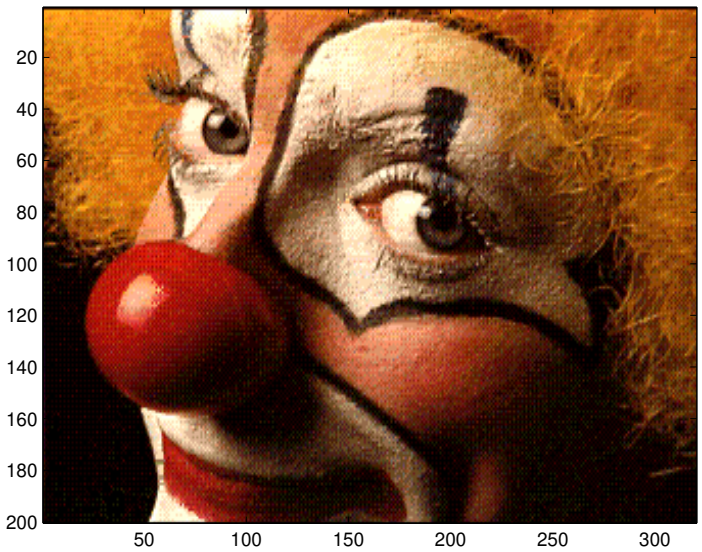
$$= \left\| \sum_{i=k+1}^{n} \sigma_i u_i v_i^T \right\| = \left\| U \begin{bmatrix} 0 & & & \\ & \sigma_{k+1} & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix} V^T \right\|_2 = \sigma_{k+1}.$$

It remains to show that there is no closer rank k matrix to A. Let B be any rank k matrix, so its null space has dimension $n-k$. The space spanned by $\{v_1, ..., v_{k+1}\}$ has dimension $k+1$. Since the sum of their dimensions is $(n-k)+(k+1) > n$, these two spaces must overlap. Let h be a unit vector in their intersection. Then

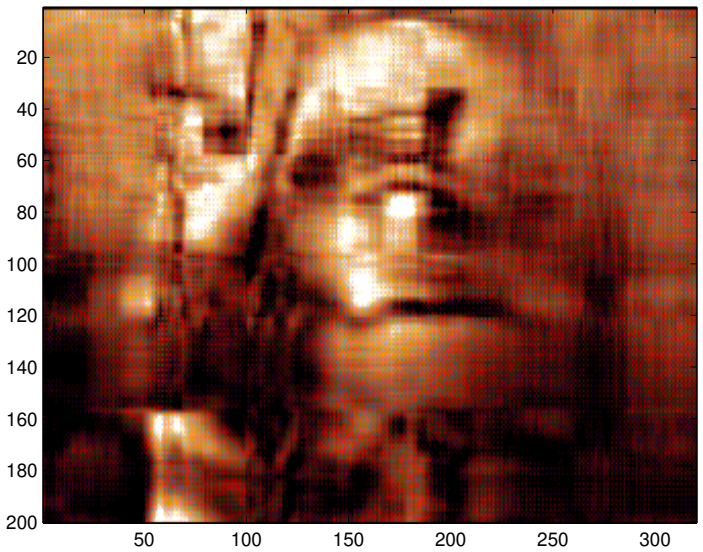
$$\begin{aligned}\|A-B\|_2^2 \geq \|(A-B)h\|_2^2 = \|Ah\|_2^2 = \left\|U\Sigma V^T h\right\|_2^2 \\ = \left\|\Sigma(V^T h)\right\|_2^2 \geq \sigma_{k+1}^2 \left\|V^T h\right\|_2^2 = \sigma_{k+1}^2.\end{aligned}$$

□

Example of application of linear systems: image compression using SVD



a) Original image



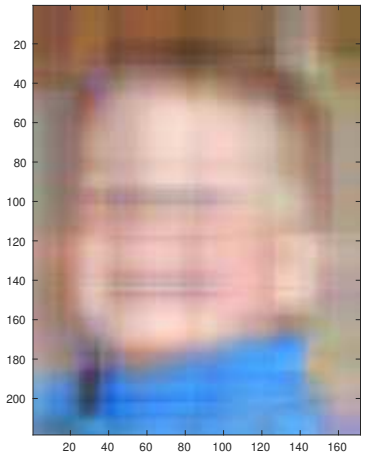
b) Rank k=20 approximation

Example of application of linear systems: image compression using SVD in Matlab

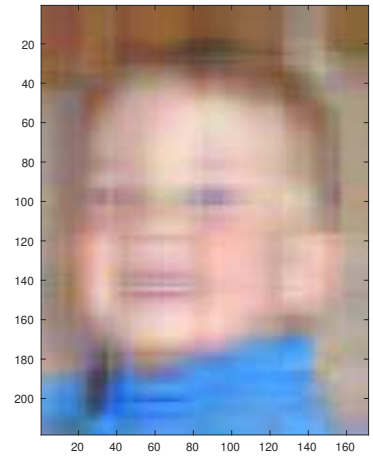
See path for other pictures:
/matlab-2012b/toolbox/matlab/demos
load clown.mat;
Size(X) = $m \times n = 320 \times 200$ pixels.
[U,S,V] = svd(X);
colormap(map);
k=20;
image(U(:,1:k)*S(1:k,1:k)*V(:,1:k)');
Now: size(U)= $m \times k$, size(V)= $n \times k$.

Image compression using SVD in Matlab

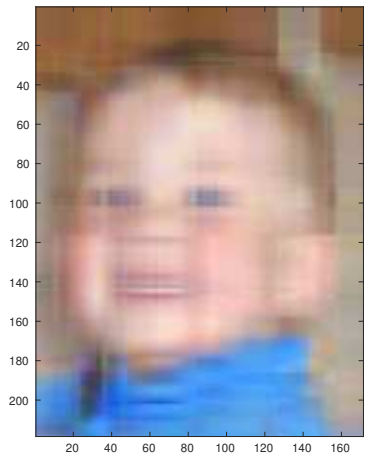
a) Original image



b) Rank k=4 approximation



b) Rank k=5 approximation



c) Rank k=6 approximation

d) Rank k=10 approximation

d) Rank k=15 approximation

Example of application of linear systems: image compression using SVD for arbitrary image

To get image on the previous slide, I took picture in jpg-format and loaded it in matlab. You can also try to use following matlab code for your own pictures:

```
A = imread('Child.jpg');
Real size of A: size(A) ans= 218 171 3
figure(1); image(DDA);
DDA=im2double(A);
[U1,S1,V1] = svd(DDA(:,:,1)); [U2,S2,V2] = svd(DDA(:,:,2));
[U3,S3,V3] = svd(DDA(:,:,3));
k=15;
svd1 = U1(:,1:k)*S1(1:k,1:k)*V1(:,1:k)';
svd2 = U2(:,1:k)*S2(1:k,1:k)*V2(:,1:k)';
svd3 = U3(:,1:k)*S3(1:k,1:k)*V3(:,1:k)';
DDAnew = zeros(size(DDA));
DDAnew(:,:,1) = svd1; DDAnew(:,:,2) = svd2; DDAnew(:,:,3) = svd3;
figure(2); image(DDAnew);
```

Perturbation Theory for the Least Squares Problem

When A is not square, we define its condition number with respect to the 2-norm to be $k_2(A) \equiv \sigma_{max}(A)/\sigma_{min}(A)$. This reduces to the usual condition number when A is square. The next theorem justifies this definition.

THEOREM *Suppose that A is m-by-n with $m \geq n$ and has full rank. Suppose that x minimizes $\|Ax - b\|_2$. Let $r = Ax - b$ be the residual. Let $\tilde{x}$ minimize $\|(A + \delta A)\tilde{x} - (b + \delta b)\|_2$. Assume* $\epsilon \equiv \max(\frac{\|\delta A\|_2}{\|b\|_2}, \frac{\|\delta b\|_2}{\|b\|_2}) < \frac{1}{k_2(A)} = \frac{\sigma_{min}(A)}{\sigma_{max}(A)}$. *Then*

$$\frac{\|\tilde{x} - x\|}{\|x\|} \leq \epsilon \cdot \left\{ \frac{2 \cdot k_2(A)}{\cos\theta} + \tan\theta \cdot k_2^2(A) \right\} + O(\epsilon^2) \equiv \epsilon \cdot k_{LS} + O(\epsilon^2),$$

where $\sin\theta = \frac{\|r\|_2}{\|b\|_2}$. *In other words, θ is the angle between the vectors b and Ax and measures whether the residual norm $\|r\|_2$ is large (near $\|b\|$) or small (near 0). k_{LS} is the condition number for the least squares problem.*

Sketch of Proof. Expand $\tilde{x} = ((A + \delta A)^T(A + \delta A))^{-1}(A + \delta A)^T(b + \delta b)$ in powers of δA and δb. Then remove all non-linear terms, leave the linear terms for δA and δb. $\square$

Nonlinear least squares problems

Suppose that for our data points $(x_i, y_i), i = 1, ..., m$ we want to find the vector of parameters $c = (c_1, ..., c_n)$ which will fit best to the data $y_i, i = 1, ..., m$ of the model function $f(x_i, c), i = 1, ..., m$. We consider the case when the model function $f : R^{n+1} \to R$ is nonlinear now. Our goal is to find minimum of the residual $r = y - f(x, c)$ in the least squares sense:

$$\min_c \sum_{i=1}^{m} (y_i - f(x_i, c))^2. \tag{24}$$

To solve problem (24) we can still use the linear least squares method if we can transform the nonlinear function $f(x, c)$ to the linear one. This can be done if the function $f(x, c)$ can be represented in the form $f(x, c) = A \exp^{cx}, A = const$. Then taking logarithm of $f(x, c)$ we get: $\ln f = \ln A + cx$, which is already linear function. Then linear least squares problem after this transformation can be written as

$$\min_c \sum_{i=1}^{m} (\ln y_i - \ln f(x_i, c))^2. \tag{25}$$

Computer exercise 1 (1 p.)

Consider the nonlinear equation

$$y(T) = A \cdot \exp^{-\frac{E}{T - T_0}}$$

presenting one of the models of the viscosity of glasses (see paper G. S. Fulcher, "ANALYSIS OF RECENT MEASUREMENTS OF THE VISCOSITY OF GLASSES" on the course homepage). Here, T is the known temperature, $y(T)$ is the known output data. Determine parameters A, E, T_0 which are positive constants by knowing T and output data $y(T)$. Determine parameters A, E, T_0 which are positive constants by knowing T and output data $y(T)$.

Hints:

1. Transform first the nonlinear function $y(T)$ to the linear one and solve then linear least squares problem. Discretize T by N points and compute discrete values of $y(T)$ as $y_i = y(T_i)$ for the known values of parameters A, E, T_0. Then forget about these parameters (we will call them exact parameters A^*, E^*, T_0^*) and solve the linear least squares problem to recover these exact parameters.

2. You can choose exact parameters A^*, E^*, T_0^* as well as T as some positive constants. For example, take $E^* = 6 \cdot 10^3, A^* = exp^{-2.64}, T_0^* = 400, T = 750 + 10 * i, i = 1, ..., N$, where N is the number of discretization points. See Table II in the paper G. S. Fulcher, "ANALYSIS OF RECENT MEASUREMENTS OF THE VISCOSITY OF GLASSES" for some other possible choises of these constants.

3. Add random noise δ to data $y(T)$ using the formula

$$y_\delta(T) = y(T)(1 + \delta\alpha), \tag{26}$$

where $\alpha \in (-1, 1)$ is randomly distributed number and $\delta \in [0, 1]$ is the noise level. For example, if noise in data is 5%, then $\delta = 0.05$.

- You can use several Matlab's functions to test adding of the noise, for example, use

```
r = randi([-1 1],size(y),1)
for j=1:n
  for i=1:n
    ydelta(n*(i-1)+j) = y(n*(i-1)+j)*(1 + 0.1*r(n*(i-1)+j));
  end
end
```

- Try also add normally distributed Gaussian noise

$$N(y|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(y-\mu)^2}{2\sigma^2}}.$$

Here, μ is mean, σ^2 is variance, σ is standard deviation.

Below is example how to add Gaussian noise $N(y|\mu,\sigma^2)$ with mean $\mu = 0$ and variance $\sigma^2 = 0.01$ to matrix A in MATLAB:

```
Anoise = A + 0.01*randn(size(A)) + 0;
```

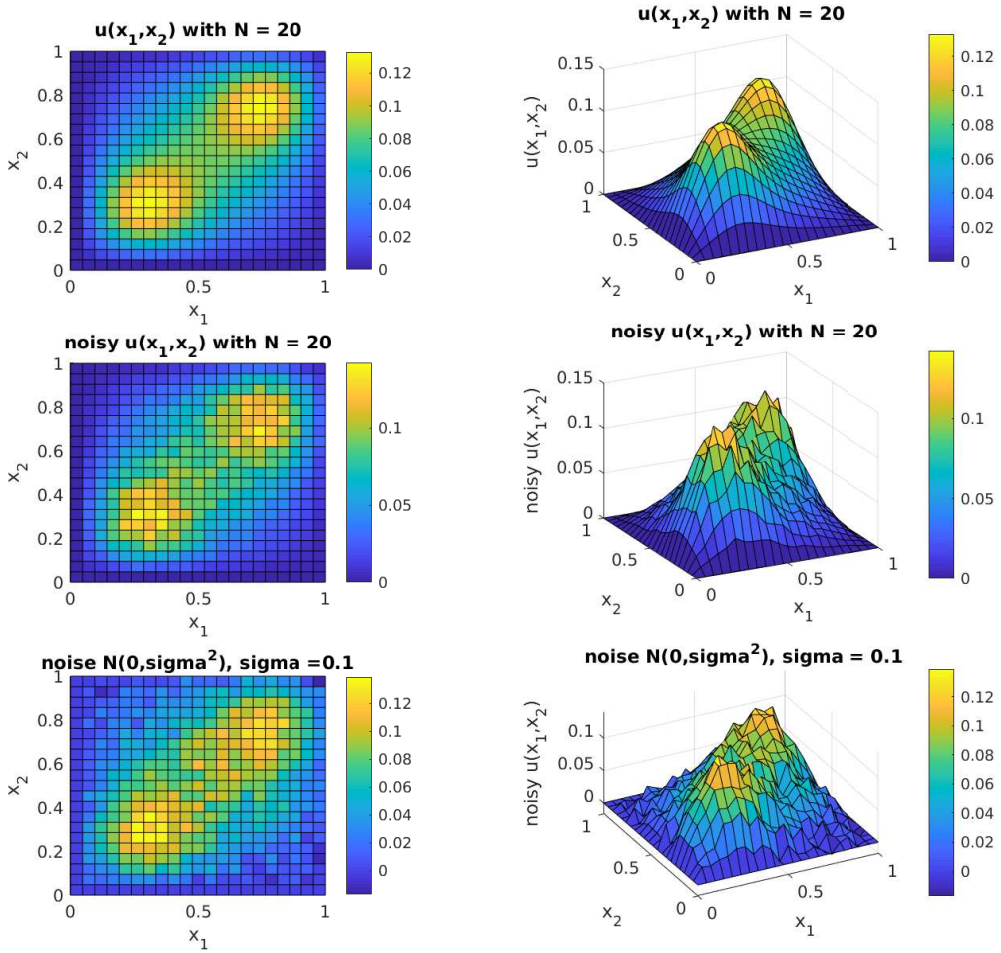


Figure: Top figures: Solution of Poisson's equation (example of section 8.4.4 of the course book). Middle figures: Noisy solution obtained via (26). Bottom figures: noisy solution obtained via adding normally distributed Gaussian noise $N(y|0, 0.01)$.

4. Solve the linear least squares problem using the method of normal equations, QR and then SVD decompositions. Analyze obtained results by computing the relative errors e_A, e_E, e_{T_0} in the computed parameters depending on the different noise level $\delta \in [0, 1]$ in data $y_\sigma(T)$ for every method.

   The relative errors e_A, e_E, e_{T_0} in the computed parameters A, E, T_0 are given by:

$$
\begin{aligned}
e_A &= \frac{|A - A^*|}{|A^*|}, \\
e_E &= \frac{|E - E^*|}{|E^*|}, \\
e_{T_0} &= \frac{|T_0 - T_0^*|}{|T_0^*|}.
\end{aligned}
\tag{27}
$$

- Here, A^*, E^*, T_0^* are exact values and A, E, T_0 are computed one. Present results how relative errors (27) depend on the relative noise $\delta \in [0, 1]$ in graphical form and in the corresponding table.

5. Choose different number of discretization points N and present results of computations in graphical form and in the corresponding table. More precisely, present how relative errors (27) depend on the number of measurements N if you solve the linear least squares problem using the method of normal equations, QR and then SVD decomposition.

6. Using results obtained in items 4 and 5, analyze, what is the minimal number of observations N should be chosen to get reasonable reconstruction of parameters A, E, T_0 within the noise level σ ?