TOPIC 2

Spectral and singular value decompositions

6 Further observations

Course: Math 535 (http://www.math.wisc.edu/~roch/mmids/) - Mathematical Methods in Data Science (MMiDS)
Author: Sebastien Roch (http://www.math.wisc.edu/~roch/), Department of Mathematics, University of Wisconsin-Madison

Updated: Sep 21, 2020

We make further observations.

6.1 Condition numbers

In this section we consider condition numbers, a measure of the sensitivity to perturbations of certain numerical problems. We look in particular at the conditioning of the least-squares problem.

6.1.1 Pseudoinverses

We have seen that the least-squares problem provides a solution concept for overdetermined systems. This leads to the following generalization of the matrix inverse.

Definition (Pseudoinverse): Let $A \in \mathbb{R}^{n\times m}$ be a matrix with SVD $A = U\Sigma V^T$ and singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$. A pseudoinverse $A^+ \in \mathbb{R}^{m\times n}$ is defined as

$$A^+ = V\Sigma^+ U^T$$

where Σ^+ is the diagonal matrix with diagonal entries $\sigma_1^{-1}, \ldots, \sigma_r^{-1}$. $\lhd$

While it is not obvious from the definition (*Exercise:* Why? $\lhd$), the pseudoinverse is in fact unique (https://en.wikipedia.org/wiki/Proofs_involving_the_Moore–Penrose_inverse#Proof_of_uniqueness). Note further that

$$AA^+ = U\Sigma V^T V\Sigma^+ U^T = UU^T$$

and

$$A^+ A = V\Sigma^+ U^T U\Sigma V^T = VV^T.$$

Three important cases:

1. In the square nonsingular case, both products are the identity and we necessarily have $A^+ = A^{-1}$ by the Existence of an Inverse Lemma.
2. If A has full column rank $m \leq n$, then $r = m$ and $A^+ A = I_{m\times m}$.
3. If A has full row rank $n \leq m$, then $r = n$ and $AA^+ = I_{n\times n}$.

In the second case above, we recover our solution to the least-squares problem in the overdetermined case.

Lemma (Pseudoinverse and least squares): Let $A \in \mathbb{R}^{n\times m}$ of full column rank m with $m \leq n$. Then

$$A^+ = (A^T A)^{-1} A^T.$$

In particular, the solution to the least-square problem

$$\min_{\mathbf{x}\in\mathbb{R}^m} \|A\mathbf{x} - \mathbf{b}\|$$

is $\mathbf{x}^* = A^+\mathbf{b}$.

Proof idea: We use the SVD definition and check that the two sides are the same.

Proof: We first note that $A^T A$ is nonsingular. Indeed, if $A^T A\mathbf{x} = \mathbf{0}$, then $\mathbf{x}^T A^T A\mathbf{x} = \|A\mathbf{x}\|^2 = 0$ so that $A\mathbf{x} = \mathbf{0}$ by the point-separating property of the vector norm. In turn, since the columns of A are linearly independent by assumption, this implies $\mathbf{x} = \mathbf{0}$. Hence the columns of $A^T A$ are also linearly indepedent.

Now let $A = U\Sigma V^T$ be an SVD of A (with positive singular values). We then note that

$$(A^T A)^{-1} A^T = (V\Sigma U^T U\Sigma V^T)^{-1} V\Sigma U^T = V\Sigma^{-2} V^T V\Sigma U^T = A^+$$

as claimed.

The second claim follows from the normal equations. □

NUMERICAL CORNER In Julia, the pseudoinverse of a matrix can be computed using the function `pinv` (https://docs.julialang.org/en/v1/stdlib/LinearAlgebra/#LinearAlgebra.pinv).

In [1]:
```
#Julia version: 1.5.1
using Plots, LinearAlgebra
```

In [2]:
```
M = [1.5 1.3; 1.2 1.9; 2.1 0.8]
```

Out[2]:
```
3×2 Array{Float64,2}:
 1.5  1.3
 1.2  1.9
 2.1  0.8
```

In [3]:
```
Mp = pinv(M)
```

Out[3]:
```
2×3 Array{Float64,2}:
 0.0930539  -0.311014   0.587446
 0.126271    0.629309  -0.449799
```

In [4]:
```
Mp*M
```

Out[4]:
```
2×2 Array{Float64,2}:
 1.0          -4.05256e-17
 2.5981e-16    1.0
```

6.1.2 Conditioning of matrix-vector multiplication

We define the condition number of a matrix and show that it captures some information about the sensitivity to perturbations of matrix-vector multiplications. First an exercise.

Exercise: Let $A \in \mathbb{R}^{n\times n}$ be nonsingular with SVD $A = U\Sigma V^T$ where the singular values satisfy $\sigma_1 \geq \cdots \geq \sigma_n > 0$. Show that

$$\min_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \min_{\mathbf{y}\neq\mathbf{0}} \frac{\|\mathbf{y}\|}{\|A^{-1}\mathbf{y}\|} = \sigma_n = 1/\|A^+\|_2.$$

$\lhd$

Definition (Condition number of a matrix): The condition number (in the induced 2-norm) of a matrix $A \in \mathbb{R}^{n\times m}$ is defined as

$$\kappa_2(A) = \|A\|_2\|A^+\|_2.$$

$\lhd$

In the square nonsingular case, this reduces to

$$\kappa_2(A) = \|A\|_2\|A^{-1}\|_2 = \frac{\sigma_1}{\sigma_n}$$

where we used the exercise above. In words, $\kappa_2(A)$ is the ratio of the largest to the smallest stretching under A.

Theorem (Relative conditioning of matrix-vector multiplication): Let $A \in \mathbb{R}^{n\times n}$ be nonsingular. Then, for any $\mathbf{x} \in \mathbb{R}^n$,

$$\lim_{\delta\to 0} \sup_{0<\|\mathbf{d}\|\leq\delta} \frac{\|A(\mathbf{x}+\mathbf{d}) - A\mathbf{x}\|/\|A\mathbf{x}\|}{\|\mathbf{d}\|/\|\mathbf{x}\|} = \max_{\mathbf{d}\neq\mathbf{0}} \frac{\|A(\mathbf{x}+\mathbf{d}) - A\mathbf{x}\|/\|A\mathbf{x}\|}{\|\mathbf{d}\|/\|\mathbf{x}\|} \leq \kappa_2(A)$$

and the inequality is tight.

The ratio above measures the worst rate of relative change in $A\mathbf{x}$ under infinitesimal perturbations of $\mathbf{x}$. The theorem says that when $\kappa_2(A)$ is large, a case referred to as ill-conditioning, large relative changes in $A\mathbf{x}$ can be obtained from relatively small perturbations to $\mathbf{x}$. In words, a matrix-vector product is potentially sensitive to perturbations when the matrix is ill-conditioned.

Proof: Write

$$\frac{\|A(\mathbf{x}+\mathbf{d}) - A\mathbf{x}\|/\|A\mathbf{x}\|}{\|\mathbf{d}\|/\|\mathbf{x}\|} = \frac{\|A\mathbf{d}\|/\|A\mathbf{x}\|}{\|\mathbf{d}\|/\|\mathbf{x}\|} = \frac{\|A(\mathbf{d}/\|\mathbf{d}\|)\|}{\|A(\mathbf{x}/\|\mathbf{x}\|)\|} \leq \frac{\sigma_1}{\sigma_n}$$

where we used the exercise above.

In particular, we see that the ratio can achieve its maximum by taking $\mathbf{d}$ and $\mathbf{x}$ to be the right singular vectors corresponding to σ_1 and σ_n respectively.□

If we apply the theorem to the inverse instead, we get that the relative conditioning of the nonsingular linear system $A\mathbf{x} = \mathbf{b}$ to perturbations in $\mathbf{b}$ is also $\kappa_2(A)$. The latter can be large in particular when the columns of A are close to linearly dependent.

NUMERICAL CORNER In Julia, the condition number of a matrix can be computed using the function [`cond`](https://docs.julialang.org/en/v1/stdlib/LinearAlgebra/#LinearAlgebra.cond) (https://docs.julialang.org/en/v1/stdlib/LinearAlgebra/#LinearAlgebra.cond).

For example, orthogonal matrices have condition number 1:

In [5]:
```
Q = [1/sqrt(2) 1/sqrt(2); 1/sqrt(2) -1/sqrt(2)]
```

Out[5]:
```
2×2 Array{Float64,2}:
 0.707107   0.707107
 0.707107  -0.707107
```

In [6]:
```
cond(Q)
```

Out[6]:
```
1.0000000000000002
```

And matrices with nearly linearly dependent columns have large condition numbers:

In [2]:
```
eps = 1e-6
A = [1/sqrt(2) 1/sqrt(2); 1/sqrt(2) 1/sqrt(2)+eps]
```

Out[2]:
```
2×2 Array{Float64,2}:
 0.707107  0.707107
 0.707107  0.707108
```

In [3]:
```
cond(A)
```

Out[3]:
```
2.8284291245366517e6
```

Let's look at the SVD of A.

In [4]:
```
F = svd(A)
```

Out[4]:
```
SVD{Float64,Float64,Array{Float64,2}}
U factor:
2×2 Array{Float64,2}:
 -0.707107  -0.707107
 -0.707107   0.707107
singular values:
2-element Array{Float64,1}:
 1.4142140623732717
 4.999998232605338e-7
Vt factor:
2×2 Array{Float64,2}:
 -0.707107  -0.707107
 -0.707107   0.707107
```

We compute the solution to $A\mathbf{x} = \mathbf{b}$ when $\mathbf{b}$ is the right singular vector corresponding to the largest singular value.

In [7]:
```
b = F.V[:,1]
```

Out[7]:
```
2-element Array{Float64,1}:
 -0.7071065311865034
 -0.7071070311865033
```

In [8]:
```
x = A\b
```

Out[8]:
```
2-element Array{Float64,1}:
 -0.4999996465020582
 -0.49999999994448885
```

We make a small perturbation in the direction of the second right singular vector.

In [13]:
```
delta = 1e-6
bp = b + delta*F.V[2,:]
```

Out[13]:
```
2-element Array{Float64,1}:
 -0.7071072382935345
 -0.7071063240799721
```

The relative change in solution is:

In [14]:
```
xp = A\bp
```

Out[14]:
```
2-element Array{Float64,1}:
 -1.9142142088291174
  0.9142135623822168
```

```
In [15]: (norm(x-xp)/norm(x))/(norm(b-bp)/norm(b))
```

```
Out[15]: 2.828429124665918e6
```

Note that this is exactly the condition number of A.

6.1.3 Back to the least-squares problem [optional]

We return to the least-squares problem

$$\min_{\mathbf{x}\in\mathbb{R}^m} \|A\mathbf{x} - \mathbf{b}\|$$

where

$$A = \begin{pmatrix} | & & | \\ \mathbf{a}_1 & \ldots & \mathbf{a}_m \\ | & & | \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}.$$

We showed that the solution satisfies the normal equations

$$A^T A\mathbf{x} = A^T\mathbf{b}.$$

As we show next, the condition number of $A^T A$ can be much larger than that of A itself.

Lemma (Condition number of $A^T A$): Let $A \in \mathbb{R}^{n\times m}$ have full column rank. The

$$\kappa_2(A^T A) = \kappa_2(A)^2.$$

Proof idea: We use the SVD.

Proof: Let $A = U\Sigma V^T$ be an SVD of A with singular values $\sigma_1 \geq \cdots \geq \sigma_m > 0$. Then

$$A^T A = V\Sigma U^T U\Sigma V^T = V\Sigma^2 V^T.$$

In particular the latter expression is an SVD of $A^T A$, and hence the condition number of $A^T A$ is

$$\kappa_2(A^T A) = \frac{\sigma_1^2}{\sigma_m^2} = \kappa_2(A)^2.$$

☐

NUMERICAL CORNER We give a quick example.

In [15]:
```
A = [1. 101.; 1. 102.; 1. 103.; 1. 104.; 1. 105]
```

Out[15]:
```
5×2 Array{Float64,2}:
 1.0  101.0
 1.0  102.0
 1.0  103.0
 1.0  104.0
 1.0  105.0
```

In [16]:
```
cond(A)
```

Out[16]:
```
7503.817028686117
```

In [17]:
```
cond(A'*A)
```

Out[17]:
```
5.630727000263108e7
```

This observation - and the resulting increased numerical instability - is one of the reasons we previously developed an alternative approach to the least-squares problem. Quoting [Sol, Section 5.1]:

> Intuitively, a primary reason that $\mathrm{cond}(A^T A)$ can be large is that columns of A might look "similar" [...] If two columns $\mathbf{a}_i$ and $\mathbf{a}_j$ satisfy $\mathbf{a}_i \approx \mathbf{a}_j$, then the least-squares residual length $\|\mathbf{b} - A\mathbf{x}\|_2$ will not suffer much if we replace multiples of $\mathbf{a}_i$ with multiples of $\mathbf{a}_j$ or vice versa. This wide range of nearly—but not completely—equivalent solutions yields poor conditioning. [...] To solve such poorly conditioned problems, we will employ an alternative technique with closer attention to the column space of A rather than employing row operations as in Gaussian elimination. This strategy identifies and deals with such near-dependencies explicitly, bringing about greater numerical stability.

We quote without proof a theorem from [Ste (https://epubs.siam.org/doi/book/10.1137/1.9781611971408), Theorem 4.2.7] which provides further light on this issue.

Theorem (Accuracy of computed least-squares solutions): Let $\mathbf{x}^*$ be the solution of the least-squares problem $\min_{\mathbf{x}\in\mathbb{R}^m} \|A\mathbf{x} - \mathbf{b}\|$. Let $\mathbf{x}_{\mathrm{NE}}$ be the solution obtained by forming and solving the normal equations in floating-point arithmetic (https://en.wikipedia.org/wiki/Floating-point_arithmetic) with rounding unit ϵ_M. Then $\mathbf{x}_{\mathrm{NE}}$ satisfies

$$\frac{\|\mathbf{x}_{\mathrm{NE}} - \mathbf{x}^*\|}{\|\mathbf{x}^*\|} \leq \gamma_{\mathrm{NE}}\kappa_2^2(A)\left(1 + \frac{\|\mathbf{b}\|}{\|A\|_2\|\mathbf{x}^*\|}\right)\epsilon_M.$$

Let $\mathbf{x}_{\mathrm{QR}}$ be the solution obtained from a QR factorization in the same arithmetic. Then

$$\frac{\|\mathbf{x}_{\mathrm{QR}} - \mathbf{x}^*\|}{\|\mathbf{x}^*\|} \leq 2\gamma_{\mathrm{QR}}\kappa_2(A)\epsilon_M + \gamma_{\mathrm{NE}}\kappa_2^2(A)\frac{\|\mathbf{r}^*\|}{\|A\|_2\|\mathbf{x}^*\|}\epsilon_M$$

where $\mathbf{r}^* = \mathbf{b} - A\mathbf{x}^*$ is the residual vector. The constants γ are slowly growing functions of the dimensions of the problem.

To explain, let's quote [Ste (https://epubs.siam.org/doi/book/10.1137/1.9781611971408), Section 4.2.3] again:

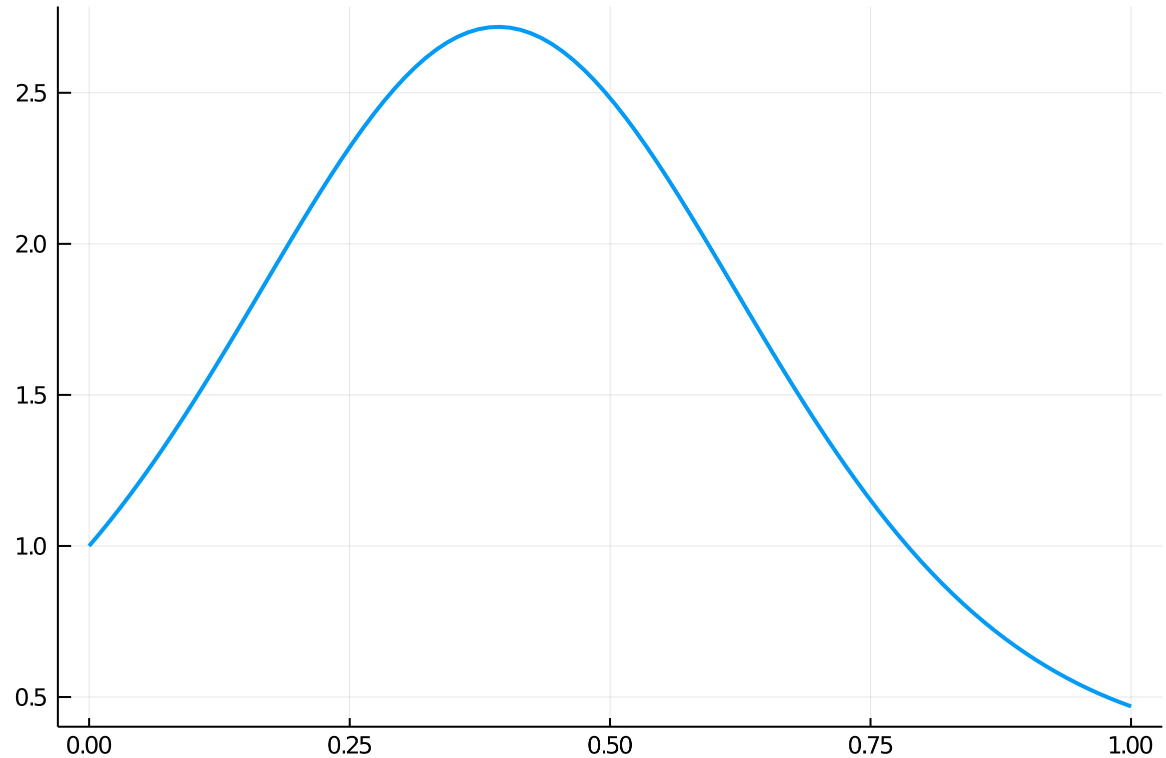
> The perturbation theory for the normal equations shows that $\kappa_2^2(A)$ controls the size of the errors we can expect. The bound for the solution computed from the QR equation also has a term multiplied by $\kappa_2^2(A)$, but this term is also multiplied by the scaled residual, which can diminish its effect. However, in many applications the vector $\mathbf{b}$ is contaminated with error, and the residual can, in general, be no smaller than the size of that error.

NUMERICAL CORNER Here is a numerical example taken from [TB (https://books.google.com/books/about/Numerical_Linear_Algebra.html?id=JaPtxOytY7kC), Lecture 19]. We will approximate the following function with a polynomial.

In [18]:

```
n = 100
t = (0:n-1)/(n-1)
b = exp.(sin.(4*t))
plot(t, b, legend=false, lw=2)
```

Out[18]:



In [19]:

```
m = 17
A = [t[i].^(j-1) for i=1:n, j=1:m];
```

The condition numbers of A and $A^T A$ are both high in this case.

In [20]:

```
@show cond(A)
@show cond(A'*A);
```

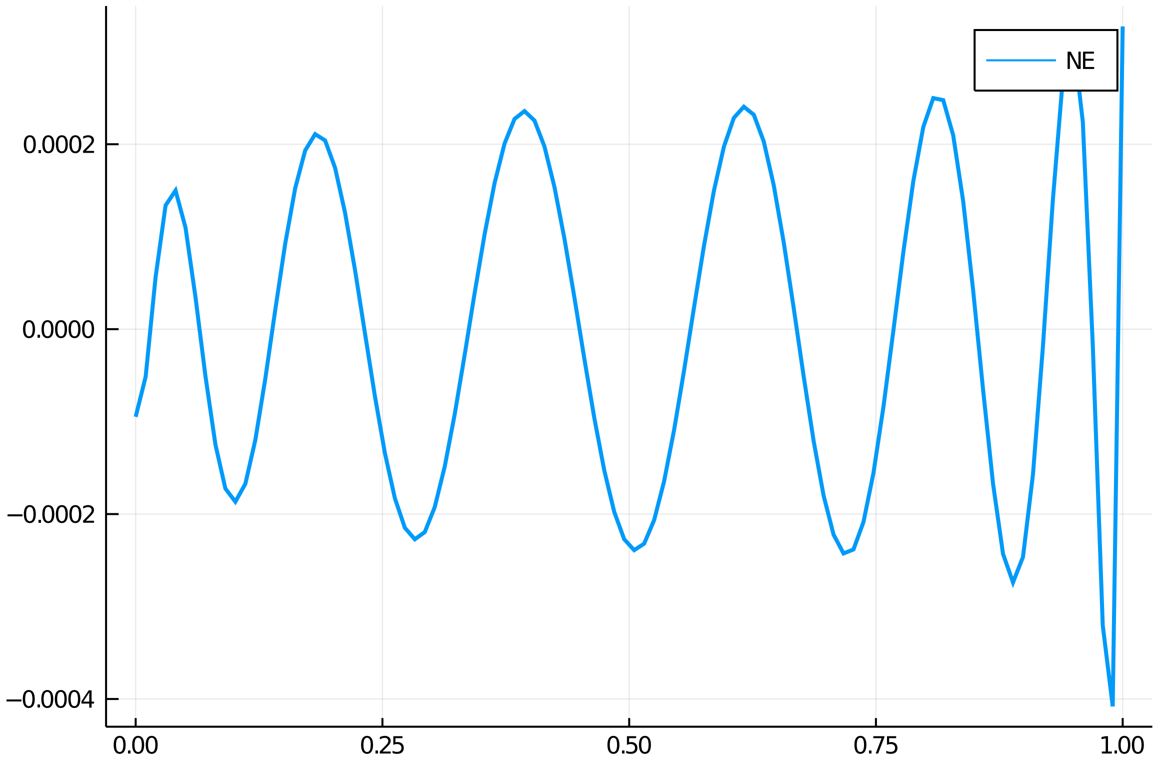
```
cond(A) = 7.558243605585787e11
cond(A' * A) = 6.812930320935918e17
```

We first use the normal equations and plot the residual vector.

In [21]:
```
xNE = (A'*A)\(A'*b)
@show norm(b-A*xNE)
plot(t, b-A*xNE, label="NE", lw=2)
```

```
norm(b - A * xNE) = 0.0017440726708434444
```

Out[21]:

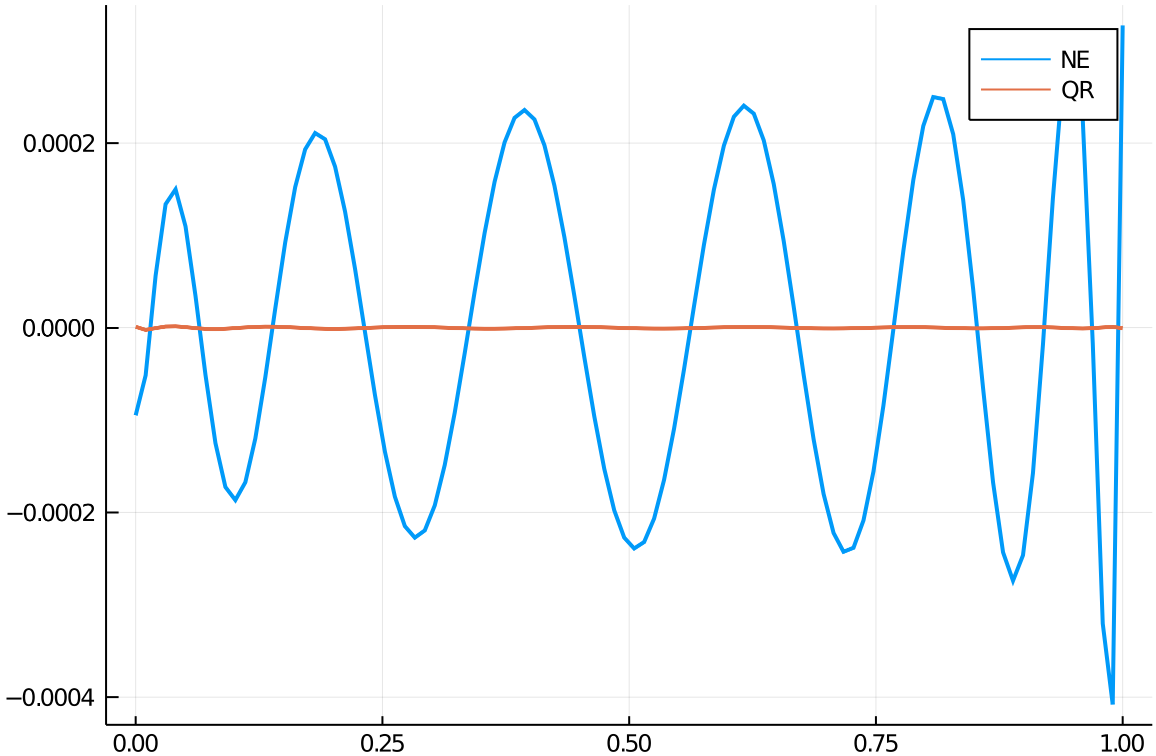


We then use the `qr` (https://docs.julialang.org/en/v1/stdlib/LinearAlgebra/#LinearAlgebra.qr) function of Julia to compute the QR solution instead. Note that `Matrix` (https://docs.julialang.org/en/v1/base/arrays/#Base.Matrix) is used to transform the factors obtained from `qr` into regular arrays.

In [22]:
```
F = qr(A)
Q, R = Matrix(F.Q), Matrix(F.R)
xQR = R\(Q'*b)
@show norm(b-A*xQR)
plot!(t, b-A*xQR, label="QR", lw=2)
```

```
norm(b - A * xQR) = 7.359747057852724e-6
```

Out[22]:



6.2 Proof of Spectral theorem [optional]

When A is symmetric, that is, $a_{ij} = a_{ji}$ for all i, j, a remarkable result is that A is similar to a diagonal matrix by an orthogonal transformation. Put differently, there exits an orthonormal basis of $\mathbb{R}^d$ made of eigenvectors of A.

Theorem (Spectral Theorem): Let $A \in \mathbb{R}^{d \times d}$ be a symmetric matrix, that is, $A^T = A$. Then A has d orthonormal eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_d$ with corresponding (not necessarily distinct) real eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$. In matrix form, this is written as the matrix factorization

$$A = Q \Lambda Q^T = \sum_{i=1}^{d} \lambda_i \mathbf{q}_i \mathbf{q}_i^T$$

where Q has columns $\mathbf{q}_1, \ldots, \mathbf{q}_d$ and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$. We refer to this factorization as a spectral decomposition of A.

One might hope that the SVD of a symmetric matrix would produce identical left and right singular vectors, thereby providing the desired eigendecomposition. However that is not the case.

Exercise: Compute the SVD of $A = [-1]$. $\lhd$

However, the same roadmap that led to the SVD can be made to produce an eigendecomposition.

Exercise: Consider the block matrices

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix}, \qquad \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

where $\mathbf{y} \in \mathbb{R}^{d_1}$, $\mathbf{z} \in \mathbb{R}^{d_2}$, $X_{11}, A \in \mathbb{R}^{d_1 \times d_1}$, $X_{12}, B \in \mathbb{R}^{d_1 \times d_2}$, $X_{21}, C \in \mathbb{R}^{d_2 \times d_1}$, and $X_{22}, D \in \mathbb{R}^{d_2 \times d_2}$. Show that

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix}^T \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix} = \mathbf{y}^T A\mathbf{y} + \mathbf{y}^T B\mathbf{z} + \mathbf{z}^T C\mathbf{y} + \mathbf{z}^T D\mathbf{z}.$$

$\lhd$

Proof idea (Spectral Theorem): Use a greedy sequence, as in the SVD derivation, this time maximizing the quadratic form (https://en.wikipedia.org/wiki/Quadratic_form#Real_quadratic_forms) $\langle \mathbf{v}, A\mathbf{v} \rangle$. How is this quadratic form is related to eigenvalues? Note that, for a unit eigenvector $\mathbf{v}$ with eigenvalue λ, we have $\langle \mathbf{v}, A\mathbf{v} \rangle = \langle \mathbf{v}, \lambda \mathbf{v} \rangle = \lambda$.

Proof (Spectral Theorem): We proceed by induction.

A first eigenvector: Let $A_1 = A$ and

$$\mathbf{v}_1 \in \arg\max\{\langle \mathbf{v}, A_1 \mathbf{v} \rangle : \|\mathbf{v}\| = 1\}$$

and

$$\lambda_1 = \max\{\langle \mathbf{v}, A_1 \mathbf{v} \rangle : \|\mathbf{v}\| = 1\}.$$

Complete $\mathbf{v}_1$ into an orthonormal basis of $\mathbb{R}^d$, $\mathbf{v}_1$, $\hat{\mathbf{v}}_2, \ldots, \hat{\mathbf{v}}_d$, and form the block matrix

$$\hat{W}_1 = \begin{pmatrix} \mathbf{v}_1 & \hat{V}_1 \end{pmatrix}$$

where the columns of $\hat{V}_1$ are $\hat{\mathbf{v}}_2, \ldots, \hat{\mathbf{v}}_d$. Note that $\hat{W}_1$ is orthogonal by construction.

Getting one step closer to diagonalization: We show next that W_1 gets us one step closer to a diagonal matrix by similarity transformation. Note first that

$$\hat{W}_1 A_1 \hat{W}_1^T = \begin{pmatrix} \lambda_1 & \mathbf{w}_1^T \\ \mathbf{w}_1 & A_2 \end{pmatrix}$$

where $\mathbf{w}_1 = \hat{V}_1^T A_1 \mathbf{v}_1$ and $A_2 = \hat{V}_1^T A_1 \hat{V}_1$. The key claim is that $\mathbf{w}_1 = \mathbf{0}$. This follows from a contradiction argument. Indeed, suppose $\mathbf{w}_1 \neq \mathbf{0}$ and consider the unit (Exercise: Why?) vector

$$\mathbf{z} = \hat{W}_1^T \times \frac{1}{\sqrt{1 + \delta^2 \|\mathbf{w}_1\|^2}} \begin{pmatrix} 1 \\ \delta \mathbf{w}_1 \end{pmatrix}$$

which, by the exercise above, achieves objective value

$$\begin{aligned} \mathbf{z}^T A_1 \mathbf{z} &= \frac{1}{1 + \delta^2 \|\mathbf{w}_1\|^2} \begin{pmatrix} 1 \\ \delta \mathbf{w}_1 \end{pmatrix}^T \begin{pmatrix} \lambda_1 & \mathbf{w}_1^T \\ \mathbf{w}_1 & A_2 \end{pmatrix} \begin{pmatrix} 1 \\ \delta \mathbf{w}_1 \end{pmatrix} \\ &= \frac{1}{1 + \delta^2 \|\mathbf{w}_1\|^2} \left(\lambda_1 + 2\delta \|\mathbf{w}_1\|^2 + \delta^2 \mathbf{w}_1^T A_2 \mathbf{w}_1 \right). \end{aligned}$$

In the proof of the *Left Singular Vectors are Orthogonal Lemma*, we showed that for $\epsilon \in (0, 1)$

$$\frac{1}{\sqrt{1 + \epsilon^2}} \geq 1 - \epsilon^2/2.$$

So for δ small enough

$$\begin{aligned} \mathbf{z}^T A_1 \mathbf{z} &\geq (\lambda_1 + 2\delta \|\mathbf{w}_1\|^2 + \delta^2 \mathbf{w}_1^T A_2 \mathbf{w}_1)(1 - \delta^2 \|\mathbf{w}_1\|^2/2) \\ &\geq \lambda_1 + 2\delta \|\mathbf{w}_1\|^2 + C\delta^2 \\ &> \lambda_1 \end{aligned}$$

where $C \in \mathbb{R}$ depends on $\mathbf{w}_1$ and A_2. That gives the desired contradiction. So, letting $W_1 = \hat{W}_1$,

$$W_1 A_1 W_1^T = \begin{pmatrix} \lambda_1 & \mathbf{0} \\ \mathbf{0} & A_2 \end{pmatrix}.$$

Finally note that $A_2 = \hat{V}_1^T A_1 \hat{V}_1$ is symmetric

$$A_2^T = (\hat{V}_1^T A_1 \hat{V}_1)^T = \hat{V}_1^T A_1^T \hat{V}_1 = \hat{V}_1^T A_1 \hat{V}_1 = A_2$$

by the symmetry of A_1 itself.

Next step of the induction: Apply the same argument to the symmetric matrix $A_2 \in \mathbb{R}^{(d-1)\times(d-1)}$, let $\hat{W}_2 \in \mathbb{R}^{(d-1)\times(d-1)}$ be the corresponding orthogonal matrix, and define λ_2 and A_3 through the equation

$$\hat{W}_2 A_2 \hat{W}_2^T = \begin{pmatrix} \lambda_2 & \mathbf{0} \\ \mathbf{0} & A_3 \end{pmatrix}.$$

Define the block matrix

$$W_2 = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \hat{W}_2 \end{pmatrix}$$

and observe that

$$W_2 W_1 A_1 W_1^T W_2^T = W_2 \begin{pmatrix} \lambda_1 & \mathbf{0} \\ \mathbf{0} & A_2 \end{pmatrix} W_2^T = \begin{pmatrix} \lambda_1 & \mathbf{0} \\ \mathbf{0} & \hat{W}_2 A_2 \hat{W}_2^T \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 & \mathbf{0} \\ 0 & \lambda_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & A_3 \end{pmatrix}.$$

Proceeding similarly by induction gives the claim. □