Pr. 1. (sol/hs121a)

(a) Suppose $\boldsymbol{A}$ is $p \times m$, $\boldsymbol{X}$ is $m \times n$, and $\boldsymbol{B}$ is $q \times n$. Let us represent these matrices as

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_n \end{bmatrix}, \quad \boldsymbol{B}^T = \begin{bmatrix} \boldsymbol{b}_1 & \dots & \boldsymbol{b}_q \end{bmatrix}.$$

Using these definitions, we have

$$\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}^T = \boldsymbol{A} \begin{bmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_n \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_1 & \dots & \boldsymbol{b}_q \end{bmatrix} = \boldsymbol{A} \left[\left(\sum_{k=1}^{n} b_{1k} \boldsymbol{x}_k \right) \quad \dots \quad \left(\sum_{k=1}^{n} b_{qk} \boldsymbol{x}_k \right) \right]$$

$$= \left[\boldsymbol{A} \left(\sum_{k=1}^{n} b_{1k} \boldsymbol{x}_k \right) \quad \dots \quad \boldsymbol{A} \left(\sum_{k=1}^{n} b_{qk} \boldsymbol{x}_k \right) \right] \Rightarrow \operatorname{vec}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}^T) = \begin{bmatrix} \boldsymbol{A} \sum_{k=1}^{n} b_{1k} \boldsymbol{x}_k \\ \vdots \\ \boldsymbol{A} \sum_{k=1}^{n} b_{qk} \boldsymbol{x}_k \end{bmatrix}.$$

On the other hand,

$$(\boldsymbol{B} \otimes \boldsymbol{A}) \operatorname{vec}(\boldsymbol{X}) = \begin{bmatrix} b_{11}\boldsymbol{A} & \dots & b_{1n}\boldsymbol{A} \\ \vdots & \ddots & \vdots \\ b_{q1}\boldsymbol{A} & \dots & b_{qn}\boldsymbol{A} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_n \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{n} b_{1k} \boldsymbol{A}\boldsymbol{x}_k \\ \vdots \\ \sum_{k=1}^{n} b_{qk} \boldsymbol{A}\boldsymbol{x}_k \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} \sum_{k=1}^{n} b_{1k} \boldsymbol{x}_k \\ \vdots \\ \boldsymbol{A} \sum_{k=1}^{n} b_{qk} \boldsymbol{x}_k \end{bmatrix}.$$

This establishes the vec trick equality.

(b) Multiplying $(\boldsymbol{A}\boldsymbol{X})\boldsymbol{B}^T$ requires $N^3 + N^3 = 2N^3$ scalar multiplications for $N \times N$ dense matrices.

In contrast, computing the Kronecker product $\boldsymbol{Y} = \boldsymbol{B} \otimes \boldsymbol{A}$ requires N^4 multiples, and then the product $(\boldsymbol{B} \otimes \boldsymbol{A}) \operatorname{vec}(\boldsymbol{X}) = \boldsymbol{Y} \operatorname{vec}(\boldsymbol{X})$ needs another N^4 multiplies, for a total of $2N^4$. Clearly the LHS is more efficient. But the RHS often is useful for paper analysis.

Tuesday October 26 2021 21:52
yuzhanj@umich.edu

Pr. 2. (sol/hs029)
(This solution considers a more general case with some $\theta > 0$. Note that $\theta = 1$ in the problem statement.)

(a) Method 1. An eigenvalue λ of $\boldsymbol{B} = \boldsymbol{A} + \theta \boldsymbol{x}\boldsymbol{x}'$ satisfies $\det(\boldsymbol{B} - \lambda \boldsymbol{I}) = \det(\boldsymbol{A} + \theta \boldsymbol{x}\boldsymbol{x}' - \lambda \boldsymbol{I}) = 0$. Note that when $\lambda \neq A_{ii}$ for any i then $\boldsymbol{A} - \lambda \boldsymbol{I}$ is invertible, so for now we focus on such values of λ. Thus

$$\det(\boldsymbol{A} + \theta \boldsymbol{x}\boldsymbol{x}' - \lambda \boldsymbol{I}) = \det((\boldsymbol{A} - \lambda \boldsymbol{I}) \cdot (\boldsymbol{I} + (\boldsymbol{A} - \lambda \boldsymbol{I})^{-1} \theta \boldsymbol{x}\boldsymbol{x}') = \det(\boldsymbol{A} - \lambda \boldsymbol{I}) \cdot \det(\boldsymbol{I} + \underbrace{\theta (\boldsymbol{A} - \lambda \boldsymbol{I})^{-1} \boldsymbol{x}}_{\triangleq \boldsymbol{w}} \boldsymbol{x}')$$

$$= \det(\boldsymbol{A} - \lambda \boldsymbol{I}) \cdot (1 + \theta \boldsymbol{x}' \underbrace{(\boldsymbol{A} - \lambda \boldsymbol{I})^{-1} \boldsymbol{x}}_{= \boldsymbol{w}}). \qquad \text{(From from HW 1, } \det(\boldsymbol{I} + \boldsymbol{w}\boldsymbol{x}') = 1 + \boldsymbol{x}'\boldsymbol{w}.)$$

Thus an eigenvalue of $\boldsymbol{B}$ that is not an eigenvalue of $\boldsymbol{A}$ must satisfy the equation

$$\boldsymbol{x}'(\lambda \boldsymbol{I} - \boldsymbol{A})^{-1} \boldsymbol{x} = \frac{1}{\theta}.$$

Since $\boldsymbol{A}$ is diagonal, a_{ii} are the eigenvalues of $\boldsymbol{A}$, and we obtain the (implicit) relationship

$$\frac{|x_1|^2}{\lambda - a_{11}} + \frac{|x_2|^2}{\lambda - a_{22}} + \ldots + \frac{|x_n|^2}{\lambda - a_{nn}} = \frac{1}{\theta}. \tag{1}$$

Because $x_i \neq 0$ is given, this equation will have n solutions for λ whenever $\boldsymbol{A}$ does not have repeated eigenvalues. When $\boldsymbol{A}$ has repeated eigenvalues, some of the eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{B}$ will coincide because the degree of the above polynomial will not be n. The problem statement says $\boldsymbol{A}$ has distinct entries so this technicality is avoided.

Note that the solutions of (1) will *not* correspond to any of the diagonal elements of $\boldsymbol{A}$, so our use of the inverse of $\boldsymbol{A} - \lambda \boldsymbol{I}$ was legitimate.

Method 2. Alternatively, one can use properties 16, 17 in Section 1.4 of Laub (see HW1, Question 3), as follows:

$$\begin{aligned} \det(\boldsymbol{A} - \lambda \boldsymbol{I} + \theta \boldsymbol{x}\boldsymbol{x}') &= \det\left(\begin{bmatrix} \boldsymbol{A} - \lambda \boldsymbol{I} & -\theta \boldsymbol{x} \\ \boldsymbol{x}' & 1 \end{bmatrix} \right) \\ &= \det(\boldsymbol{A} - \lambda \boldsymbol{I}) \det(1 + \theta \boldsymbol{x}'(\boldsymbol{A} - \lambda \boldsymbol{I})^{-1} \boldsymbol{x}). \end{aligned}$$

Method 3. Let z be an eigenvalue of $\boldsymbol{B}$ associated with the eigenvector $\boldsymbol{v}$. Then

$$\begin{aligned} \boldsymbol{B}\boldsymbol{v} = z\boldsymbol{v} &\Rightarrow (\boldsymbol{A} + \theta \boldsymbol{x}\boldsymbol{x}')\boldsymbol{v} = z\boldsymbol{v} \Rightarrow (z\boldsymbol{I} - \boldsymbol{A})\boldsymbol{v} = \theta \boldsymbol{x}\boldsymbol{x}'\boldsymbol{v} \\ &\Rightarrow \boldsymbol{v} = \theta (z\boldsymbol{I} - \boldsymbol{A})^{-1} \boldsymbol{x}\boldsymbol{x}'\boldsymbol{v} \qquad (\text{for } z \neq \lambda(\boldsymbol{A})) \\ &\Rightarrow (\boldsymbol{x}'\boldsymbol{v}) = \theta \boldsymbol{x}'(z\boldsymbol{I} - \boldsymbol{A})^{-1} \boldsymbol{x}(\boldsymbol{x}'\boldsymbol{v}) \\ &\Rightarrow 1 = \theta \boldsymbol{x}'(z\boldsymbol{I} - \boldsymbol{A})^{-1} \boldsymbol{x} \qquad (\text{assuming that the scalar } \boldsymbol{x}'\boldsymbol{v} \neq 0). \end{aligned}$$

The assumption that $\boldsymbol{x}'\boldsymbol{v} \neq 0$ implies that $\boldsymbol{v} \notin \mathcal{N}(\boldsymbol{x})$. More work would be needed to complete this argument, although the equivalence with the previous derivations shows that this assumption is valid whenever $x_i \neq 0$ when considering the eigenvalues of $\boldsymbol{B}$ that are not equal to the eigenvalues of $\boldsymbol{A}$.

Here is a proof that a_{ii} cannot be an eigenvalue of $\boldsymbol{B}$ under the given conditions. When $\lambda = a_{ii}$, clearly the corresponding eigenvector of $\boldsymbol{A}$ is $\boldsymbol{e}_i$. Now suppose that λ is also an eigenvalue of $\boldsymbol{B}$. Then we would have for some vector $\boldsymbol{v}$: $\boldsymbol{B}\boldsymbol{v} = (\boldsymbol{A} + \boldsymbol{x}\boldsymbol{x}')\boldsymbol{v} = \boldsymbol{A}\boldsymbol{v} + \boldsymbol{x}(\boldsymbol{x}'\boldsymbol{v}) = \lambda \boldsymbol{v}$. Now examine the ith element of both sides of that last equality: $a_{ii} v_i + x_i(\boldsymbol{x}'\boldsymbol{v}) = a_{ii} v_i \Rightarrow x_i(\boldsymbol{x}'\boldsymbol{v}) = 0$. Because each element of x_i is nonzero, it must be the case that $\boldsymbol{x}'\boldsymbol{v} = 0$. So we have $\boldsymbol{B}\boldsymbol{v} = \boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v} = a_{ii}\boldsymbol{v}$, so $\boldsymbol{v}$ must be the eigenvector of $\boldsymbol{A}$ corresponding to eigenvalue a_{ii}, which means $\boldsymbol{v} = \boldsymbol{e}_i$. But that would mean $\boldsymbol{x}'\boldsymbol{e}_i = 0$, which would contradict the assumption that all elements of $\boldsymbol{x}$ are nonzero.

Tuesday October 05 2021 21:32 yuzhai@umich.edu

Here is a proof that $\boldsymbol{B}$ can have some eigenvalues that are the same as those of $\boldsymbol{A}$ when $\boldsymbol{A}$ is diagonal with some repeated values. For illustration (and without loss of generality via permutations), suppose $\boldsymbol{A} = \begin{bmatrix} \lambda \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D} \end{bmatrix}$ where the elements of the diagonal matrix $\boldsymbol{D}$ are distinct from λ and where $k \geq 2$. Similarly partition the vector $\boldsymbol{x}$ as $\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix}$ where $\boldsymbol{x}_1$ has length k and $\boldsymbol{x}_2$ has length $n-k$. Then $\boldsymbol{B} = \boldsymbol{A} + \boldsymbol{x}\boldsymbol{x}' = \begin{bmatrix} \lambda \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D} \end{bmatrix} + \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \end{bmatrix}$. Now pick $\boldsymbol{v}_1 \in \mathcal{N}(\boldsymbol{x}_1') \neq \boldsymbol{0}$ which exists because $k \geq 2$ and define $\boldsymbol{v} = \begin{bmatrix} \boldsymbol{v}_1 \\ \boldsymbol{0} \end{bmatrix}$. Then one can see that $\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}$ and $\boldsymbol{B}\boldsymbol{v} = \lambda \boldsymbol{v}$ so in this case $\boldsymbol{A}$ and $\boldsymbol{B}$ have a common eigenvalue (and eigenvector). This is why the problem stated that the elements of $\boldsymbol{A}$ are distinct.

(b) When $\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$ and $\boldsymbol{x} = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$, then (1) becomes

$$\frac{1}{\lambda - 1} + \frac{1}{\lambda - 2} + \frac{1}{\lambda - 3} = 1.$$

Multiplying through, this becomes a cubic equation:

$$(z-2)(z-3) + (z-1)(z-3) + (z-1)(z-2) = (z-1)(z-2)(z-3).$$

Using `Julia`'s `Polynomials.jl` package, this simplifies to $0 = -17 + 23z - 9z^2 + z^3$, the roots of which are approximately $\{1.32, 2.46, 5.21\}$.

```
using Polynomials: fromroots, roots
fromroot(root::Real) = fromroots([root])
(p1, p2, p3) = fromroot.(1:3)
# solve: (z-2)(z-3) + (z-1)(z-3) + (z-1)(z-2) = (z-1)(z-2)(z-3)
p = p1*p2*p3 - p1*p2 - p2*p3 - p1*p3
roots(p) # basic version that does not generalize easily to n > 3
#q = +(1 .// fromroot.(1:3)..., -1); roots(q) # one line version: generalizes!
```

Tuesday Oct 5 2021 21:32
yuzhanji@umich.edu

Pr. 3. (sol/hs069)

Define $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{q}_1 & \boldsymbol{q}_2 \end{bmatrix}$, and $\boldsymbol{x} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$.

(a) Approach 1 (using normal equations) The columns of $\boldsymbol{A}$ are orthonormal, so the Gram matrix is $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{I}_2$. Thus from the normal equations the unique LLS solution is

$$\hat{\boldsymbol{x}} = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{b} = \boldsymbol{A}'\boldsymbol{b}.$$

This expression is unsurprising, because what we get is precisely the first two "coordinates" of the vector $\boldsymbol{b}$ relative to the basis whose first two basis vectors correspond to the columns of $\boldsymbol{A}$.

Thus the final answer for the optimal linear combination is

$$\boldsymbol{A}\hat{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{A}'\boldsymbol{b}.$$

Approach 2 (using compact SVD)

The columns of $\boldsymbol{A}$ are linearly independent, so the unique **linear least squares estimate** that minimizes $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2$ is given by $\hat{\boldsymbol{x}} = \boldsymbol{A}^+\boldsymbol{b}$. A compact SVD of $\boldsymbol{A}$ is simply:

$$\boldsymbol{A} = \sum_{i=1}^{2} 1 \boldsymbol{q}_i \boldsymbol{e}_i',$$

where $\boldsymbol{e}_i$ denotes the ith unit vector. Thus:

$$\boldsymbol{A}^+ = \sum_{i=1}^{2} 1 \boldsymbol{e}_i \boldsymbol{q}_i',$$

and hence

$$\hat{\boldsymbol{x}} = \boldsymbol{A}^+\boldsymbol{b} = \sum_{i=1}^{2} 1 \boldsymbol{e}_i \boldsymbol{q}_i'\boldsymbol{b} = \boldsymbol{A}'\boldsymbol{b} \Rightarrow \boldsymbol{A}\hat{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{A}'\boldsymbol{b}.$$

(b) Approach 1

Here we have that the residual (or error) vector:

$$\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{A}\hat{\boldsymbol{x}} = \boldsymbol{b} - \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{b}) = (\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}')\boldsymbol{b} \Rightarrow \langle \boldsymbol{r}, \boldsymbol{q}_i \rangle = \boldsymbol{q}_i'\boldsymbol{r} = \boldsymbol{q}_i'(\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}')\boldsymbol{b} = (\boldsymbol{q}_i' - \boldsymbol{e}_i'\boldsymbol{A}')\boldsymbol{b} = (\boldsymbol{q}_i' - \boldsymbol{q}_i')\boldsymbol{b} = \boldsymbol{0}.$$

Approch 2

Here we have that the residual (or error) vector:

$$\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{A}\hat{\boldsymbol{x}} = (\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}')\boldsymbol{b} = \sum_{i=3}^{n} \boldsymbol{q}_i \boldsymbol{q}_i'\boldsymbol{b},$$

where $\boldsymbol{q}_i$ for $i = 3, \ldots, n$ are the $n-2$ (unit norm) basis vectors, orthogonal to $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$ so that $\text{span}(\{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n\}) = \mathbb{R}^n$. Hence $\boldsymbol{q}_1'\boldsymbol{r} = \sum_{i=3}^{n} \boldsymbol{q}_1'\boldsymbol{q}_i\boldsymbol{b} = 0$ and $\boldsymbol{q}_2'\boldsymbol{r} = \sum_{i=3}^{n} \boldsymbol{q}_2'\boldsymbol{q}_i\boldsymbol{b} = 0$.

This property is related to the **projection theorem**.

Tuesday October 05 2021 21:32
wuzhai@umich.edu

Pr. 4. (sol/hs049)

- Approach 1. No SVD, just using properties of **pseudo-inverse**:

$$\begin{aligned}\boldsymbol{x}'(\boldsymbol{I} - \boldsymbol{A}^+\boldsymbol{A})'(\boldsymbol{A}^+\boldsymbol{b}) &= \boldsymbol{x}'(\boldsymbol{I} - (\boldsymbol{A}^+\boldsymbol{A})')(\boldsymbol{A}^+\boldsymbol{b}) = \boldsymbol{x}'(\boldsymbol{I} - (\boldsymbol{A}^+\boldsymbol{A}))(\boldsymbol{A}^+\boldsymbol{b}) \\ &= \boldsymbol{x}'(\boldsymbol{A}^+ - \boldsymbol{A}^+\boldsymbol{A}\boldsymbol{A}^+)\boldsymbol{b}) = \boldsymbol{x}'(\boldsymbol{A}^+ - \boldsymbol{A}^+)\boldsymbol{b}) = \boldsymbol{0}\end{aligned}$$

- Approach 2. **full SVD**:

$$\begin{aligned}(\boldsymbol{A}^+\boldsymbol{b})'(\boldsymbol{I} - \boldsymbol{A}^+\boldsymbol{A})\boldsymbol{x} &= \boldsymbol{b}'(\boldsymbol{A}^+)'(\boldsymbol{I} - \boldsymbol{V}\boldsymbol{\Sigma}^+\boldsymbol{U}'\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}')\boldsymbol{x} = \boldsymbol{b}'\boldsymbol{U}(\boldsymbol{\Sigma}^+)'\boldsymbol{V}'(\boldsymbol{V}\boldsymbol{V}' - \boldsymbol{V}\boldsymbol{\Sigma}^+\boldsymbol{\Sigma}\boldsymbol{V}')\boldsymbol{x} \\ &= \boldsymbol{b}'\boldsymbol{U}((\boldsymbol{\Sigma}^+)' - (\boldsymbol{\Sigma}^+)'\boldsymbol{\Sigma}^+\boldsymbol{\Sigma})\boldsymbol{V}'\boldsymbol{x} = \boldsymbol{0},\end{aligned}$$

  because direct multiplication verifies that $(\boldsymbol{\Sigma}^+)' = (\boldsymbol{\Sigma}^+)'\boldsymbol{\Sigma}^+\boldsymbol{\Sigma}$.

- Approach 3. **compact SVD**: $\boldsymbol{A} = \boldsymbol{U}_r\boldsymbol{\Sigma}_r\boldsymbol{V}_r' \Rightarrow \boldsymbol{A}^+ = \boldsymbol{V}_r\boldsymbol{\Sigma}_r^{-1}\boldsymbol{U}_r'$ so:

$$(\boldsymbol{A}^+\boldsymbol{b})'(\boldsymbol{I} - \boldsymbol{A}^+\boldsymbol{A})\boldsymbol{x} = \boldsymbol{b}'\boldsymbol{U}_r\boldsymbol{\Sigma}_r^{-1}\boldsymbol{V}_r'(\boldsymbol{I} - \boldsymbol{V}_r\boldsymbol{V}_r')\boldsymbol{x} = \boldsymbol{b}'\boldsymbol{U}_r\boldsymbol{\Sigma}_r^{-1}(\boldsymbol{V}_r' - \boldsymbol{V}_r'\boldsymbol{V}_r\boldsymbol{V}_r')\boldsymbol{x} = \boldsymbol{0}, \quad \text{because } \boldsymbol{V}_r'\boldsymbol{V}_r = \boldsymbol{I}.$$

  Put another way in terms of **subspaces**: $\boldsymbol{A}^+\boldsymbol{b} \in \mathcal{R}(\boldsymbol{V}_r)$ whereas $(\boldsymbol{I} - \boldsymbol{A}^+\boldsymbol{A})\boldsymbol{x} = (\boldsymbol{I} - \boldsymbol{V}_r\boldsymbol{V}_r')\boldsymbol{x} \in \mathcal{R}(\boldsymbol{V}_r)^\perp$.

Pr. 5. (sol/hs114)

(a) We must show $\forall \alpha \in [0,1]$ that: $\|\boldsymbol{A}(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{z}) - \boldsymbol{b}\|_2 \le \alpha\,\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2 + (1-\alpha)\|\boldsymbol{A}\boldsymbol{z} - \boldsymbol{b}\|_2$.
For any $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^n$:

$$\begin{aligned}f(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{z}) &= \|\boldsymbol{A}(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{z}) - \boldsymbol{b}\|_2 \\ &= \|\boldsymbol{A}(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{z}) - \boldsymbol{b}(\alpha + 1 - \alpha)\|_2 \\ &= \|(\alpha\boldsymbol{A}\boldsymbol{x} - \alpha\boldsymbol{b}) + ((1-\alpha)\boldsymbol{A}\boldsymbol{z} - (1-\alpha)\boldsymbol{b})\|_2 \\ &\le \|\alpha(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})\|_2 + \|(1-\alpha)(\boldsymbol{A}\boldsymbol{z} - \boldsymbol{b})\|_2, \qquad \text{(via the triangle inequality)} \\ &= |\alpha|\,\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2 + |1-\alpha|\,\|\boldsymbol{A}\boldsymbol{z} - \boldsymbol{b}\|_2 \\ &= \alpha\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2 + (1-\alpha)\|\boldsymbol{A}\boldsymbol{z} - \boldsymbol{b}\|_2, \qquad \text{(because } \alpha \in [0,1]\text{)} \\ &= \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{z}).\end{aligned}$$

(b) Next we need to show that, given two matrices $\boldsymbol{A}$, $\boldsymbol{B}$, $\forall \alpha \in [0,1]$:

$$\sigma_1\left(\alpha\boldsymbol{A} + (1-\alpha)\boldsymbol{B}\right) \le \alpha\,\sigma_1(\boldsymbol{A}) + (1-\alpha)\,\sigma_1(\boldsymbol{B}).$$

From the hint:

$$\begin{aligned}\sigma_1(\alpha\boldsymbol{A} + (1-\alpha)\boldsymbol{B}) &= \max_{\|\boldsymbol{u}\|_2 = 1} \|\left(\alpha\boldsymbol{A} + (1-\alpha)\boldsymbol{B}\right)\boldsymbol{u}\|_2 \\ &\le \max_{\|\boldsymbol{u}\|_2 = 1} \left(\|\alpha\boldsymbol{A}\boldsymbol{u}\|_2 + \|(1-\alpha)\boldsymbol{B}\boldsymbol{u}\|_2\right) \quad \text{(by the triangle inequality)} \\ &= \max_{\|\boldsymbol{u}\|_2 = 1} \left(\alpha\|\boldsymbol{A}\boldsymbol{u}\|_2 + (1-\alpha)\|\boldsymbol{B}\boldsymbol{u}\|_2\right) \\ &\le \max_{\|\boldsymbol{u}\|_2 = 1} \alpha\|\boldsymbol{A}\boldsymbol{u}\|_2 + \max_{\|\boldsymbol{u}\|_2 = 1} (1-\alpha)\|\boldsymbol{B}\boldsymbol{u}\|_2 \\ &= \alpha\,\sigma_1(\boldsymbol{A}) + (1-\alpha)\,\sigma_1(\boldsymbol{B}).\end{aligned}$$

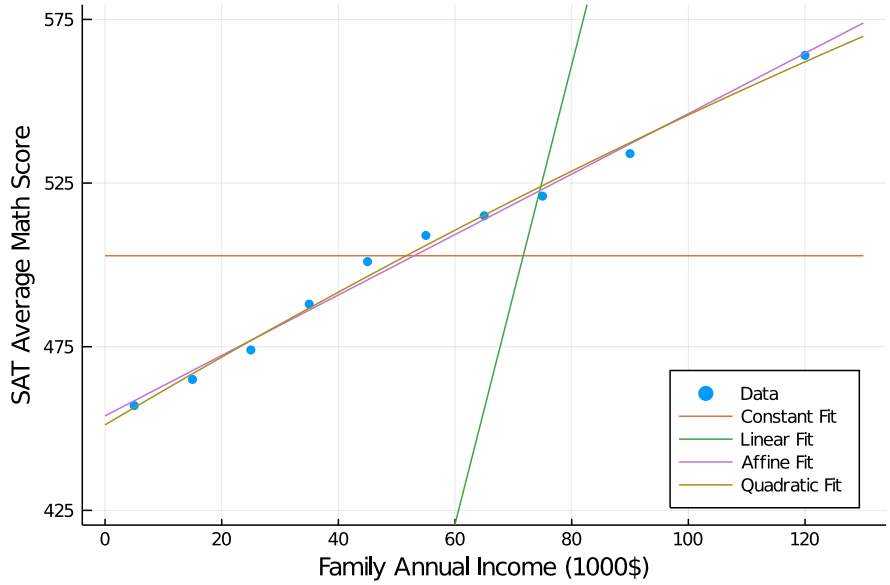
Thus, $\sigma_1(\cdot)$ is a convex function.

Tuesday October 05 2021 21:32 yuzhangjiao@umich.edu

Pr. 6. (sol/hsj4s)
Here is the code for the LS solution for the coefficients.

```
using CSV
using Plots; default(markerstrokecolor=:auto)

data = CSV.File("../prob/hpj4s.csv", normalizenames=true, transpose=true)
y = data.Math # math scores from the file (should be from 457 to 564)
income = [5:10:75; 90; 120]
scatter(income, y, label="Data", ylim=[425,575], ytick=425:50:575)
plot!(xlabel = "Family Annual Income (1000\$)")
plot!(ylabel = "SAT Average Math Score")
plot!(legend = :bottomright)

As = [ones(length(income),1),
    reshape(income, :, 1),
    [income .^ 0 income .^ 1],
    [income .^ 0 income .^ 1 income .^ 2],
]
xh = Array{Any}(undef, 4)
for i=1:4
    xh[i] = As[i] \ y
end
display([[xh[1]; 0; 0] [0; xh[2]; 0] [xh[3]; 0] xh[4]]) # for table

inc = 0:130
x = xh[1]; plot!(inc, x[1] * inc.^0, label="Constant Fit")
x = xh[2]; plot!(inc, x[1] * inc, label="Linear Fit")
x = xh[3]; plot!(inc, x[1] .+ x[2] * inc, label="Affine Fit")
x = xh[4]; plot!(inc, x[1] .+ x[2] * inc + x[3] * inc.^2, label="Quadratic Fit")

#savefig("hsj4s.pdf")
```

(a) Scatter plot and fits:



Note that the "linear" model fits very poorly; often a "linear" model in the regression literature really means an affine model. One must be careful with the terminology: the quadratic model is quadratic in income, but it is a linear function of the coefficients $\{\beta_0, \beta_1, \beta_2\}$, so **linear regression** via a LS fit is still appropriate.

(b) Table of polynomial coefficients:

Constant Fit	Linear Fit	Affine Fit	Quadratic Fit
502.8	0	453.86	451.114
0	7.01	0.923	1.06
0	0	0	-0.0011

(c) The β_1 coefficients of the affine and quadratic fits are both approximately 1, showing that for every \$1000 increase in family income, the SAT Math scores average about 1 point higher. Originally, SAT stood for "Scholastic Aptitude Test" but the dependence on family income raises doubts about how much it measures aptitude.

The β_2 coefficient is very small, but one must keep in mind that the $(\text{income})^2$ values can be very large, so one cannot conclude that it is unimportant solely by looking at its value. However, from the plot we see that the quadratic fit is nearly the same as the linear fit and there is no obvious evidence for a quadratic trend in this data. So we can conclude that average SAT scores have roughly an affine relationship with family income.

Tuesday October 05 2021 21:32
yuzhanji@umich.edu

Pr. 7. (sol/hs068)
Define for integer values d:

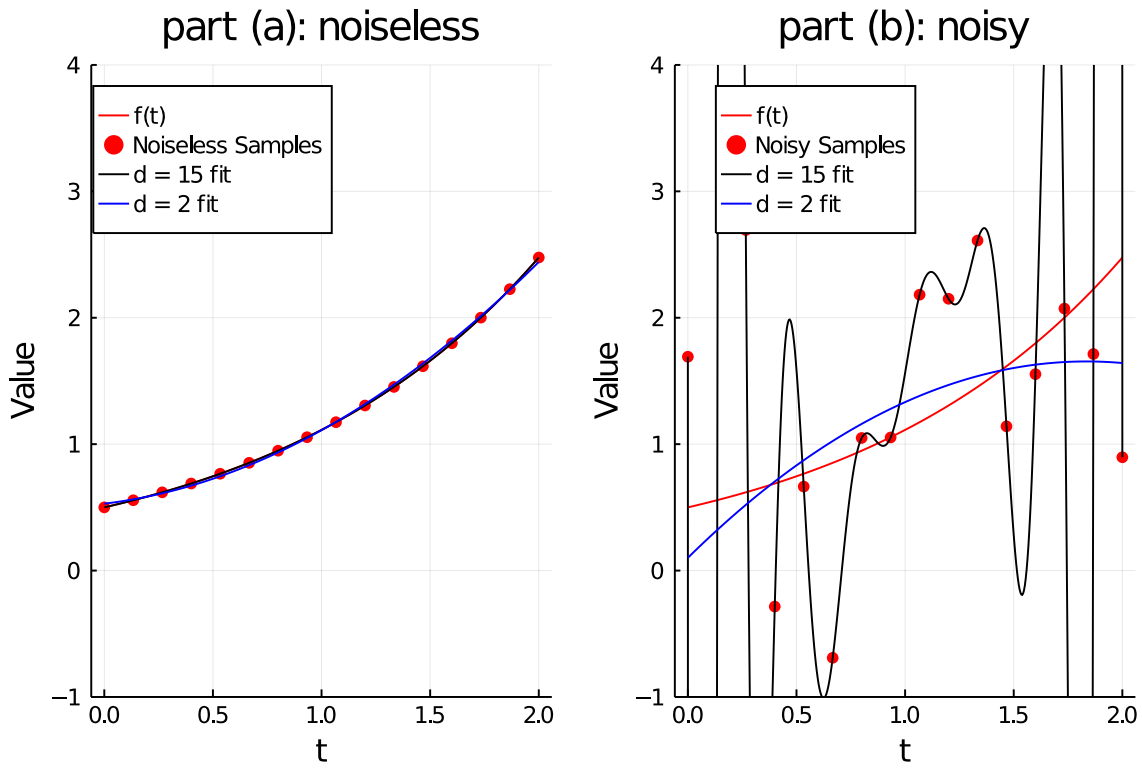
$$\boldsymbol{y} \triangleq \begin{bmatrix} f(t_1) \\ \vdots \\ f(t_{16}) \end{bmatrix} \in \mathbb{R}^{16}, \quad \boldsymbol{A} \triangleq \begin{bmatrix} 1 & t_1 & t_1^2 & \dots & t_1^d \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & t_{16} & t_{16}^2 & \dots & t_{16}^d \end{bmatrix} \in \mathbb{R}^{16 \times (d+1)}, \quad \boldsymbol{x} \triangleq \begin{bmatrix} x_1 \\ \vdots \\ x_{d+1} \end{bmatrix} \in \mathbb{R}^{d+1}.$$

The solution $\hat{\boldsymbol{x}} = \boldsymbol{A}^+ \boldsymbol{y}$ yields the the desired optimal least-squares estimate of the coefficients of the degree-d polynomial $p_d(t) = \sum_{i=1}^{d+1} x_i t^{i-1}$ that minimizes the residual norm $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2$. The $\boldsymbol{A}$ of the form above is called a **Vandermonde matrix**.
Here is `Julia` code:

```
using LinearAlgebra: norm
using Random: seed!
using Plots; default(markerstrokecolor=:auto)

T = LinRange(0, 2, 16)
seed!(3); e = randn(length(T))
b = 0.5 * exp.(0.8*T) # part a
y = b + e # part b
d = length(T)-1

A15 = [tt^j for tt in T, j=0:d]
A2  = A15[:,1:3]
x15b = A15 \ b # pinv also acceptable due to small problem size
x2b  = A2 \ b
x15y = A15 \ y
x2y  = A2 \ y
t = LinRange(0,2,500)

@show norm(A2 * x2b - b), norm(A15 * x15b - b)
@show norm(A2 * x2y - y), norm(A15 * x15y - y)
@show norm(A2 * x2y - b), norm(A15 * x15y - b)

a15 = [tt^j for tt in t, j=0:d]
a2 = a15[:,1:3]

plot(t, 0.5*exp.(0.8*t), color =:red, label = "f(t)", ylim=(-1,4))
scatter!(T,b, marker=(:circle,:red), label = "Noiseless Samples")
plot!(t, a15*x15b, color =:black, label = "d = 15 fit", legend=:top)
plot!(t, a2*x2b, color =:blue, label = "d = 2 fit", xaxis = "t", yaxis = "Value")
plot!(title = "part (a): noiseless")
pa = current()

plot(t, 0.5*exp.(0.8*t), color =:red, label = "f(t)", ylim=(-1,4))
scatter!(T, y, marker=(:circle,:red), label = "Noisy Samples")
plot!(t, a15*x15y, color =:black, label = "d = 15 fit", legend=:top)
plot!(t, a2*x2y, color =:blue, label = "d = 2 fit", xaxis = "t", yaxis = "Value")
plot!(title = "part (b): noisy")
pb = current()
plot(pa, pb)
#savefig("hs068.pdf")
```

(a) For the noiseless case, this yielded the left figure below. Note that the error for $d = 15$ is much smaller than for $d = 2$. In fact the $d = 15$ error is exactly zero (to within numerical precision) because $\text{rank}(\boldsymbol{A}) = 16 = \dim(\boldsymbol{y})$ so the solution is exact! Comparing the least-squares coefficients obtained for $d = 2$ and $d = 15$ reveals that the first three terms are about the same.

(b) When noise is added we get the right figure below. The polynomial for $d = 15$ passes *exactly* through the noisy samples. This property follows from the previous argument because $\text{rank}(\boldsymbol{A}) = 16$ so that error will be zero, relative to the samples, but large relative to the function. This example illustrates why choosing large model orders ($= d$) can make things worse by "over fitting" noise.



Use the `cond(A)` command in `Julia` to obtain the condition number: the ratio of the largest and smallest singular value of a matrix. The condition number is important because the measurement error manifests as a $\delta \boldsymbol{y}$ so that the error in the least squares solution (relative to the noise-less case) is given by exactly by $\delta \hat{\boldsymbol{x}} = \boldsymbol{A}^+(\boldsymbol{y} + \delta \boldsymbol{y}) - \boldsymbol{A}^+ \boldsymbol{y} = \boldsymbol{A}^+ \delta \boldsymbol{y}$. A large condition number means that small $\delta \boldsymbol{y}$ can get amplified, producing large $\delta \boldsymbol{x}$ and hence instabilities as manifested in the overfitting.

An estimate of the degree to which $\delta \boldsymbol{x}$ can change as a function of the condition number κ is given by the relation:

$$\frac{\|\delta \boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \kappa \frac{\|\delta \boldsymbol{y}\|}{\|\boldsymbol{y}\|}.$$

Here $\|\delta \boldsymbol{y}\|$ is the norm of the noise. The condition number $\kappa = \sigma_1(\boldsymbol{A}) / \sigma_{\min(m,n)}(\boldsymbol{A})$ is large when $d = 15$ (check it) compared to when $d = 2$.

(c)

polynomial degree:	$d = 2$	$d = 15$
Residual norm $\|\boldsymbol{A}\hat{\boldsymbol{x}}(\boldsymbol{b}) - \boldsymbol{b}\|_2$ noiseless (a)	0.071	7.6e-16
Residual norm $\|\boldsymbol{A}\hat{\boldsymbol{x}}(\boldsymbol{y}) - \boldsymbol{y}\|_2$ noisy (b)	4.438	8.3e-5
Fitting error $\|\boldsymbol{A}\hat{\boldsymbol{x}}(\boldsymbol{y}) - \boldsymbol{b}\|_2$	1.271	4.641

Grader: the numerical values for $d = 15$ are unreliable; accept anything small (below `1e-4`) for the first two and anything larger than 1 for the last one.

(d) The residual norm for degree 15 is smaller than that of degree 2 because $\boldsymbol{A}_{15} = \begin{bmatrix} \boldsymbol{A}_2 & \boldsymbol{A}_{3:15} \end{bmatrix}$ so $\mathcal{R}(\boldsymbol{A}_2) \subseteq \mathcal{R}(\boldsymbol{A}_{15})$.

Non-graded problem(s) below

Pr. 8. (sol/hs136)

(a)
- $\boldsymbol{x}_{\mathsf{opt}} = z\boldsymbol{v}_1$, $\boldsymbol{y}_{\mathsf{opt}} = z\boldsymbol{u}_1$, where $|z| = 1$
- $\boldsymbol{x}_{\mathsf{opt}}$ and $\boldsymbol{y}_{\mathsf{opt}}$ are unique to within a sign ambiguity when $N = 1$ or when $N > 1$ and $\sigma_1(\boldsymbol{A}) > \sigma_2(\boldsymbol{A})$.
- Negating $\boldsymbol{A}$ does not change the answer (aside from a potential sign flip), as the singular values are unchanged, but one of the left/right singular vectors will have a sign-flip.
- Adding an orthogonality constraint of $\boldsymbol{x} \perp \text{span}(\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_{r-1})$ suffices. Equivalently, projecting the columns of $\boldsymbol{A}$ onto the orthogonal complement of $\text{span}(\boldsymbol{u}_1, \dots, \boldsymbol{u}_{r-1})$ also suffices, i.e., we replace $\boldsymbol{A}$ with $\left(\boldsymbol{I} - \boldsymbol{P}_{\text{span}(\boldsymbol{u}_1,\dots,\boldsymbol{u}_{r-1})}\right)\boldsymbol{A}$.
  We may also iterate this algorithm: given $\boldsymbol{u}_1$ through $\boldsymbol{u}_{k-1}$, we may seek $\boldsymbol{u}_k$ that is orthogonal to $\text{span}(\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_{k-1})$, or, replace $\boldsymbol{A}$ with $\left(\boldsymbol{I} - \boldsymbol{P}_{\text{span}(\boldsymbol{u}_1,\dots,\boldsymbol{u}_{k-1})}\right)\boldsymbol{A}$.
  Indeed, given $\boldsymbol{B} = \left(\boldsymbol{I} - \boldsymbol{P}_{\text{span}(\boldsymbol{u}_1,\dots,\boldsymbol{u}_{k-2})}\right)\boldsymbol{A}$, we may form $\left(\boldsymbol{I} - \boldsymbol{P}_{\text{span}(\boldsymbol{u}_1,\dots,\boldsymbol{u}_{k-1})}\right)\boldsymbol{A} = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{u}_{k-1}})\boldsymbol{B}$.

(b)
- Consider the vector $\boldsymbol{z} = \boldsymbol{V}'\boldsymbol{x}$. We know that $\|\boldsymbol{x}\|_2 = 1 \iff \|\boldsymbol{z}\|_2 = 1$. Now,

$$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}'\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}'\boldsymbol{x} = \boldsymbol{z}'\boldsymbol{\Lambda}\boldsymbol{z} = \sum_{k=1}^{N} \lambda_k |z_k|^2 \le \lambda_1 \sum_{k=1}^{N} |z_k|^2 = \lambda_1 \|\boldsymbol{z}\|^2 = \lambda_1.$$

  This is an equality if $\boldsymbol{z} = z\boldsymbol{e}_1$ for $|z| = 1$. Hence, $\boldsymbol{x}_{\mathsf{opt}} = z\boldsymbol{v}_1$.
  If $\lambda_1 = \lambda_2$ then we also achieve equality when $\boldsymbol{z} = z\boldsymbol{e}_2$, so in general the solution is not unique.
- $\boldsymbol{x}_{\mathsf{opt}}'\boldsymbol{A}\boldsymbol{x}_{\mathsf{opt}} = z\boldsymbol{v}_1'\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}'(z^*\boldsymbol{v}_1) = \lambda_1$. (This equality holds even in the non-unique case.)
- The solution $\boldsymbol{x}_{\mathsf{opt}}$ will be unique (to within $|z| = 1$) when $\lambda_1 > \lambda_2$.

(c)
- Use the same projection as the previous problem. If $\boldsymbol{x} \perp \boldsymbol{v}_1 \perp \cdots \perp \boldsymbol{v}_{K-1}$, then $\boldsymbol{z} = \boldsymbol{V}'\boldsymbol{x}$ yields that $z_1 = z_2 = \cdots = z_{K-1} = 0$. Hence, we have

$$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \sum_{k=1}^{N} \lambda_k |z_k|^2 = \sum_{k=K}^{N} \lambda_k |z_k|^2 \le \lambda_K \sum_{k=1}^{N} |z_k|^2 = \lambda_K.$$

  Equality holds if $\boldsymbol{z} = z\boldsymbol{e}_K$, i.e., $\boldsymbol{x}_{\mathsf{opt}} = z\boldsymbol{v}_K$.
- $\boldsymbol{x}_{\mathsf{opt}}'\boldsymbol{A}\boldsymbol{x}_{\mathsf{opt}} = \lambda_K$
- The solution $\boldsymbol{x}_{\mathsf{opt}}$ is unique (to within $|z| = 1$) when $\lambda_{K-1} > \lambda_K > \lambda_{K+1}$.

Pr. 9. (sol/hs050)
Let $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}'$. If $\text{rank}(\boldsymbol{A}) = N$, then $\boldsymbol{A}$ has N linearly independent columns, $N \le M$, and $\boldsymbol{A}'\boldsymbol{A}$ is invertible. Then:

$$(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}' = (\boldsymbol{V}\boldsymbol{\Sigma}'\boldsymbol{U}'\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}')^{-1}\boldsymbol{V}\boldsymbol{\Sigma}'\boldsymbol{U}' = (\boldsymbol{V}\boldsymbol{\Sigma}'\boldsymbol{\Sigma}\boldsymbol{V}')^{-1}\boldsymbol{V}\boldsymbol{\Sigma}'\boldsymbol{U}'$$

$$= \boldsymbol{V}(\boldsymbol{\Sigma}'\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}'\boldsymbol{U}' = \boldsymbol{V} \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \dots & 0 \\ & \ddots & \ddots & \\ 0 & 0 & \dots & \frac{1}{\sigma_n^2} \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \dots & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 & \dots & 0 \\ & \ddots & & & \ddots & \\ 0 & 0 & \dots & \sigma_n & \dots & 0 \end{bmatrix} \boldsymbol{U}'$$

$$= \boldsymbol{V} \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \dots & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2} & \dots & 0 & \dots & 0 \\ & \ddots & & & \ddots & \\ 0 & 0 & \dots & \frac{1}{\sigma_n} & \dots & 0 \end{bmatrix} \boldsymbol{U}' = \boldsymbol{V}\boldsymbol{\Sigma}^{+}\boldsymbol{U}' = \boldsymbol{A}^{\dagger}.$$

Tuesday October 05 2021 2:32
yuzhanli@umich.edu

Pr. 10. (sol/hs067)

(a) The problem of finding the line $y = \alpha x + b$ that best fits the points $(1,2)$, $(2,1)$ and $(3,3)$ is equivalent to the problem: Find $\boldsymbol{\beta}$ that minimizes $\|\boldsymbol{z} - \boldsymbol{A}\boldsymbol{\beta}\|_2$, with $\boldsymbol{z} \triangleq \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}$, $\boldsymbol{A} \triangleq \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{bmatrix}$, $\boldsymbol{\beta} = \begin{bmatrix} \alpha \\ b \end{bmatrix}$.

Thus the optimal LLS solution is $\widehat{\boldsymbol{\beta}} = \boldsymbol{A}^+ \boldsymbol{z} = \boldsymbol{V}\boldsymbol{\Sigma}^+\boldsymbol{U}'\boldsymbol{z} = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{z}$.

Here $(\boldsymbol{A}'\boldsymbol{A})^{-1} = \begin{bmatrix} 14 & 6 \\ 6 & 3 \end{bmatrix}^{-1} = \frac{1}{14 \cdot 3 - 6 \cdot 6} \begin{bmatrix} 3 & -6 \\ -6 & 14 \end{bmatrix} = \frac{1}{6} \begin{bmatrix} 3 & -6 \\ -6 & 14 \end{bmatrix}$, and $\boldsymbol{A}'\boldsymbol{z} = \begin{bmatrix} 13 \\ 6 \end{bmatrix}$,

so $\widehat{\boldsymbol{\beta}} = \frac{1}{6} \begin{bmatrix} 3 & -6 \\ -6 & 14 \end{bmatrix} \begin{bmatrix} 13 \\ 6 \end{bmatrix} = \frac{1}{6} \begin{bmatrix} 3 \\ 6 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}$.

We repeat this calculation using the SVD. Let $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}'$ denote a SVD of $\boldsymbol{A}$. Numerically computing an SVD yields:

$$\boldsymbol{U} = \begin{bmatrix} -0.3231 & 0.8538 & 0.4082 \\ -0.5475 & 0.1832 & -0.8165 \\ -0.7719 & -0.4873 & 0.4082 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 4.0791 & 0 \\ 0 & 0.6005 \\ 0 & 0 \end{bmatrix}, \quad \boldsymbol{V} = \begin{bmatrix} -0.9153 & -0.4027 \\ -0.4027 & 0.9153 \end{bmatrix},$$

so $\widehat{\boldsymbol{\beta}} = \boldsymbol{V}\boldsymbol{\Sigma}^+\boldsymbol{U}' = \boldsymbol{V} \begin{bmatrix} 0.2451 & 0 & 0 \\ 0 & 1.6653 & 0 \end{bmatrix} \boldsymbol{U}'\boldsymbol{z} = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}$.

Unsurprisingly we get the same answer since $\boldsymbol{A}^+ = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}$ when $\boldsymbol{A}$ has full column rank.

(b) Here we are asked to find the line $y = \alpha x + b$ that best fits the points $(2,1)$, $(1,2)$ and $(3,3)$. We could repeat the above computation, or we can recognize that all that has changed is that the first two points are swapped between (a) and (b). Thus the solution $\boldsymbol{\beta}$ of the two problems will be identical because the least-squares criterion is the same for any ordering of the (x_i, y_i) points. A more formal way of seeing is to first note that the new $\boldsymbol{A}$ matrix for (b) is exactly same as the $\boldsymbol{A}$ matrix in (a) except for the first two rows swapping places. Denote the "$\boldsymbol{A}$" matrix in part (a) by $\boldsymbol{A}_a$ and the "$\boldsymbol{A}$" matrix in part (b) by $\boldsymbol{A}_b$. Then $\boldsymbol{A}_b = \underbrace{\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\triangleq \boldsymbol{P}} \boldsymbol{A}_a$.

Note that $\boldsymbol{P}$ is a orthogonal matrix because $\boldsymbol{P}\boldsymbol{P}' = \boldsymbol{I}$ Such a $\boldsymbol{P}$ is called a permutation matrix. We also have that the new $\boldsymbol{z}_b = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \boldsymbol{P}\boldsymbol{z}_a$. Thus the new problem is Find $\boldsymbol{\beta}_b$ that minimizes $\|\boldsymbol{z}_b - \boldsymbol{A}_b\beta\|_2 = \|\boldsymbol{P}\boldsymbol{z}_a - \boldsymbol{P}\boldsymbol{A}_a\beta\|_2 = \|\boldsymbol{z}_a - \boldsymbol{A}_a\beta\|_2$, because $\boldsymbol{P}$ is an orthogonal matrix. Thus we get the same solution $\boldsymbol{\beta}$.

Tuesday, October 05, 2021 2:12:22 PM
yuzhanli@umich.edu