Homework 2 Solution

EE263 Stanford University, Fall 2017

Due: Wednesday 10/11/17 11:59pm

1. Quadratic extrapolation of a time series. We are given a series z up to time t. Using a quadratic model, we want to extrapolate, or predict, $z(t+1)$ based on the three previous elements of the series, $z(t)$, $z(t-1)$, and $z(t-2)$. We'll denote the predicted value of $z(t+1)$ by $\hat{z}(t+1)$. More precisely, you will find $\hat{z}(t+1)$ as follows.

a) Find the quadratic function $f(\tau) = a_2\tau^2 + a_1\tau + a_0$ which satisfies $f(t) = z(t)$, $f(t-1) = z(t-1)$, and $f(t-2) = z(t-2)$. Then the extrapolated value is given by $\hat{z}(t+1) = f(t+1)$. Show that

$$\hat{z}(t+1) = c \begin{bmatrix} z(t) \\ z(t-1) \\ z(t-2) \end{bmatrix},$$

where $c \in \mathbb{R}^{1\times 3}$, and does not depend on t. In other words, the quadratic extrapolator is a linear function. Find c explicitly.

b) Use the following matlab code to generate a time series z:

```
t = 1:1000;
z = 5*sin(t/10 + 2) + 0.1*sin(t) + 0.1*sin(2*t - 5);
```

Use the quadratic extrapolation method from part (a) to find $\hat{z}(t)$ for $t = 4, \ldots, 1000$. Find the relative root-mean-square (RMS) error, which is given by

$$\left(\frac{(1/997)\sum_{j=4}^{1000} (\hat{z}(j) - z(j))^2}{(1/997)\sum_{j=4}^{1000} z(j)^2} \right)^{1/2}.$$

Solution.

a) Setting $f(t) = z(t)$, $f(t-1) = z(t-1)$ and $f(t-2) = z(t-2)$ gives the following system of linear equations:

$$\begin{aligned} a_2t^2 + a_1t + a_0 &= z(t) \\ a_2(t-1)^2 + a_1(t-1) + a_0 &= z(t-1) \\ a_2(t-2)^2 + a_1(t-2) + a_0 &= z(t-2) \end{aligned}$$

with solution

$$a_0 = (0.5t^2 - 1.5t + 1)z(t) + (2t - t^2)z(t-1) + (0.5t^2 - 0.5t)z(t-2)$$
$$a_1 = (1.5 - t)z(t) + (2t - 2)z(t-1) + (0.5 - t)z(t-2)$$
$$a_2 = 0.5z(t) - z(t-1) + 0.5z(t-2).$$

Substituting in $\hat{z}(t+1) = a_2(t+1)^2 + a_1(t+1) + a_0$ and simplifying, we get

$$\hat{z}(t+1) = 3z(t) - 3z(t-1) + z(t-2).$$

Hence,

$$c = \begin{bmatrix} 3 & -3 & 1 \end{bmatrix}.$$

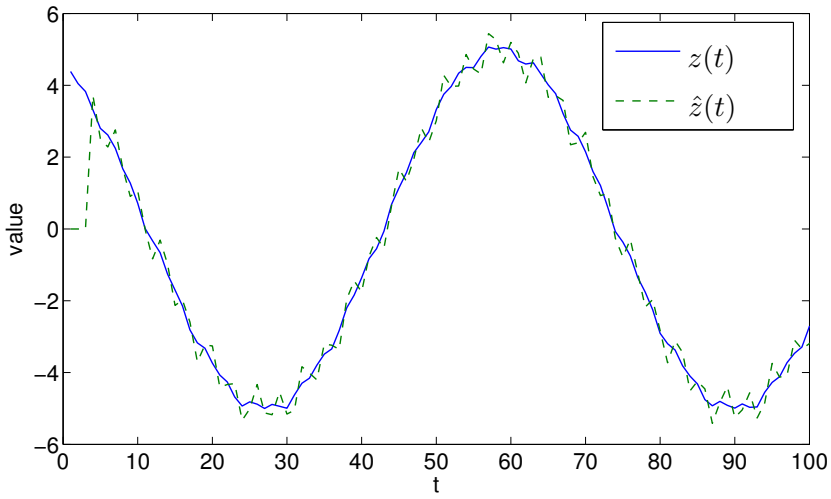
Observe that c does not depend on t, but the coefficients a_0, a_1 and a_2 do. In other words, the quadratic extrapolator f varies between samples, but its value at $t+1$ is always given by the same combination of $z(t)$, $z(t-1)$ and $z(t-2)$.

b) The following matlab code computes the predicted values and finds that the relative RMS error is 0.097:

```
t = 1:1000;
z = 5*sin(t/10 + 2) + 0.1*sin(t) + 0.1*sin(2*t - 5);
c = [3 -3 1];
for j=4:1000
    zhat(j) = c*z(j-1:-1:j-3)';
end
residual = zhat(4:1000)-z(4:1000);
rel_rms = sqrt(mean(residual.^2)/mean(z(4:end).^2))

figure;
plot(t, z, '-', t, zhat, '--');
xlim([0, 100])
xlabel('t'); ylabel('value');
legend('z(t)', 'hatz(t)');
```

In order to get an idea of how good the approximation is, we plot the first 100 samples:



2. Color perception. Human color perception is based on the responses of three different types of color light receptors, called *cones*. The three types of cones have different spectral-response characteristics, and are called L, M, and, S because they respond mainly to long, medium, and short wavelengths, respectively. In this problem we will divide the visible spectrum into 20 bands, and model the cones' responses as follows:

$$L_{\text{cone}} = \sum_{i=1}^{20} l_i p_i, \qquad M_{\text{cone}} = \sum_{i=1}^{20} m_i p_i, \qquad S_{\text{cone}} = \sum_{i=1}^{20} s_i p_i,$$

where p_i is the incident power in the ith wavelength band, and l_i, m_i and s_i are nonnegative constants that describe the spectral responses of the different cones. The perceived color is a complex function of the three cone responses, *i.e.*, the vector $(L_{\text{cone}}, M_{\text{cone}}, S_{\text{cone}})$, with different cone response vectors perceived as different colors. (Actual color perception is a bit more complicated than this, but the basic idea is right.)

a) *Metamers.* When are two light spectra, p and $\tilde{p}$, visually indistinguishable? (Visually identical lights with different spectral power compositions are called *metamers*.)

b) *Visual color matching.* In a color matching problem, an observer is shown a test light, and is asked to change the intensities of three primary lights until the sum of the primary lights looks like the test light. In other words, the observer is asked the find a spectrum of the form

$$p_{\text{match}} = a_1 u + a_2 v + a_3 w,$$

where u, v, w are the spectra of the primary lights, and a_i are the intensities to be found, that is visually indistinguishable from a given test light spectrum p_{test}. Can this always be done? Discuss briefly.

c) *Visual matching with phosphors.* A computer monitor has three phosphors, R, G, and B. It is desired to adjust the phosphor intensities to create a color that looks like a reference test light. Find weights that achieve the match or explain why no such

weights exist. The data for this problem is in `color_perception_data.m`, which contains the vectors `wavelength`, `B_phosphor`, `G_phosphor`, `R_phosphor`, `L_coefficients`, `M_coefficients`, `S_coefficients`, and `test_light`.

d) *Effects of illumination.* An object's surface can be characterized by its reflectance (*i.e.*, the fraction of light it reflects) for each band of wavelengths. If the object is illuminated with a light spectrum characterized by I_i, and the reflectance of the object is r_i (which is between 0 and 1), then the reflected light spectrum is given by $I_i r_i$, where $i = 1, \ldots, 20$ denotes the wavelength band. Now consider two objects illuminated (at different times) by two different light sources, say an incandescent bulb and sunlight. Sally argues that if the two objects look identical when illuminated by a tungsten bulb, then they will look identical when illuminated by sunlight. Beth disagrees: she says that two objects can appear identical when illuminated by a tungsten bulb, but look different when lit by sunlight. Who is right? If Sally is right, explain why. If Beth is right give an example of two objects that appear identical under one light source and different under another. You can use the vectors `sunlight` and `tungsten` defined in the data file as the light sources.

Remark. Spectra, intensities, and reflectances are all nonnegative quantities, which the material of EE263 doesn't address. So just ignore this while doing this problem. These issues can be handled using the material of EE364a, however.

Solution.

a) Let

$$A = \begin{bmatrix} l_1 & l_2 & l_3 & \cdots & l_{20} \\ m_1 & m_2 & m_3 & \cdots & m_{20} \\ s_1 & s_2 & s_3 & \cdots & s_{20} \end{bmatrix}.$$

Now suppose that $c = Ap$ is the cone response to the spectrum p and $\tilde{c} = A\tilde{p}$ is the cone response to spectrum $\tilde{p}$. If the spectra are indistinguishable, then $c = \tilde{c}$ and $Ap = A\tilde{p}$. Solving the last expression for zero gives $A(p - \tilde{p}) = 0$. In other words, p and $\tilde{p}$ are metamers if $(p - \tilde{p}) \in \text{null}(A)$.

b) In symbols, the problem asks if it is always possible to find nonnegative a_1, a_2, and a_3 such that

$$\begin{bmatrix} m_1 \\ m_2 \\ m_3 \end{bmatrix} = Ap_{\text{test}} = A \begin{bmatrix} u & v & w \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}.$$

Let $P = \begin{bmatrix} u & v & w \end{bmatrix}$ and let $B = AP$. If B is invertible, then

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = B^{-1} \begin{bmatrix} m_1 \\ m_2 \\ m_3 \end{bmatrix}.$$

However, B is not necessarily invertible. For example, if $\text{rank}(A) < 3$ or $\text{rank}(P) < 3$ then B will be singular. Physically, A is full rank if the L, M, and S cone responses are linearly independent, which they are. The matrix P is full rank if and only if the spectra

of the primary lights are independent. Even if both A and P are full rank, B could still be singular. Primary lights that generate an invertible B are called *visually independent*. If B is invertible, a_1, a_2, and a_3 exist that satisfy

$$Ap_{\text{test}} = A \begin{bmatrix} u & v & w \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}.$$

but one or more of the a_i may be negative in which case in the experimental setup described, no match would be possible. However, in a more complicated experimental setup that allows the primary lights to be combined either with each other or with p_{test}, a match is always possible if B is invertible. In this case, if $a_i < 0$, the ith light should be mixed with p_{test} instead of the other primary lights. For example, suppose $a_1 < 0$, $a_2, a_3 \geq 0$ and $b_1 = -a_1$, then

$$A(b_1 u + p_{\text{test}}) = A(a_2 v + a_3 w),$$

and each spectrum has a nonnegative weight.

c) Weights can be found as described above. The R, G, and B phosphors should be weighted by 0.4226, 0.0987, and 0.5286 respectively.

The following Julia code illustrates the steps.

```
# Extraction of the data

include("readJSON263.jl");
mydata = readJSON263("color_perception.json");

L_coefficients = mydata["L_coefficients"]["data"];
M_coefficients = mydata["M_coefficients"]["data"];
S_coefficients = mydata["S_coefficients"]["data"];
R_phosphor = mydata["R_phosphor"]["data"];
G_phosphor = mydata["G_phosphor"]["data"];
B_phosphor = mydata["B_phosphor"]["data"];
test_light = mydata["test_light"]["data"];
tungsten = mydata["tungsten"]["data"];
sunlight = mydata["sunlight"]["data"];

A = [L_coefficients; M_coefficients; S_coefficients];
B = A*[R_phosphor' G_phosphor' B_phosphor'];
weights = B\A*test_light
```

Equivalently, the following matlab code illustrates the steps.

```
close all; clear all;
color_perception;
A = [L_coefficients; M_coefficients; S_coefficients]; B =
A*[R_phosphor' G_phosphor' B_phosphor'];
weights = inv(B)*A*test_light;
```

d) Beth is right. Let r and $\tilde{r}$ be the reflectances of two objects and let p and $\tilde{p}$ be two spectra. Let A be defined as before. Then, the objects will look identical under p if

$$A \underbrace{\begin{bmatrix} r_1 & 0 & \cdots & 0 \\ 0 & r_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & r_{20} \end{bmatrix}}_{R} p = A \underbrace{\begin{bmatrix} \tilde{r}_1 & 0 & \cdots & 0 \\ 0 & \tilde{r}_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \tilde{r}_{20} \end{bmatrix}}_{\tilde{R}} p.$$

This is equivalent to saying $(R - \tilde{R})p \in \text{null}(A)$. The objects will look different under $\tilde{p}$ if, additionally, $AR\tilde{p} \neq A\tilde{R}\tilde{p}$ which means that $(R - \tilde{R})\tilde{p} \notin \text{null}(A)$. The following code shows how to find reflectances r_1 and r_2 for two objects such that the objects will have the same color under tungsten light and will have different colors under sunlight.

```
n = N[:,1];
n = n*10;

for i in 1:20
n[i] = n[i]/tungsten[i];
end

r1 = [0; 0.2; 0.3; 0.7; 0.7; 0.8; 0.8; 0.2; 0.9; 0.8; 0.2; 0.8; 0.9; 0.2; 0.8; 0.3; 0.8;
r2 = r1 - n;

t1 = zeros(20);
t2 = zeros(20);

for i in 1:20
t1[i] = r1[i]*tungsten[i];
t2[i] = r2[i]*tungsten[i];
end

color1_tungsten = A*t1'
color2_tungsten = A*t2'

for i in 1:20
s1[i] = r1[i]*sunlight[i];
s2[i] = r2[i]*sunlight[i];
end

color1_sunlight = A*s1'
color2_sunlight = A*s2'
```

Or, in Matlab:

```
close all; clear all;
```

```
color_perception;
A = [L_coefficients; M_coefficients; S_coefficients]; N = null(A);
n = N(:,1);
n= n*10;
for i = 1:20
n(i) = n(i)/tungsten(i);
end
r1 = [0; .2; .3; .7; .7; .8; .8; .2; .9; .8; .2; .8; .9; .2; .8;
.3; .8; .7; .2; .4];
r2 = r1-n;
for i = 1:20
t1(i) = r1(i)*tungsten(i);
t2(i) = r2(i)*tungsten(i);
end color1_tungsten = A*t1'; color2_tungsten = A*t2';
for i = 1:20
s1(i) = r1(i)*sunlight(i);
s2(i) = r2(i)*sunlight(i);
end color1_sun = A*s1'; color2_sun = A*s2';
253.5187
```

3. Gambler's ruin. Consider a gambling situation involving two players A and B. An example is roulette where, say, player A is a *guest* and player B is the *house*. During any one play of the game there is a probability p, $0 < p < 1$, that player A wins a chip (or coin) from player B, and a probability $q = 1 - p$ that Player B wins a chip from player A. The players begin with initial holdings of a and b chips, respectively. A player wins overall if she obtains all the chips.

a) Find the probability that player A wins.
Hint. This might sound like a problem for a probability and statistics course, but we want you to approach this problem from a linear dynamical systems point of view. Consider the general situation where A has k chips. Denote the probability under these circumstances that player A eventually wins by $u(k)$. Assume $u(k)$ is the state of the system you are analyzing. Can you write a difference equation that describes the dynamics of $u(k)$? To solve your difference equation you can assume the solution has the general form $u(k) = \lambda^k$ (we will see why later in the class). You will also need to come up with two initial conditions to uniquely solve your difference equation. Think of $u(k)$ when player A has no chips, or has $a + b$ chips.

b) As a specific example, suppose you play a roulette wheel that has 37 divisions: 18 are red, 18 are black and one is green. If you bet on either red or black, you win a sum equal to your bet if the outcome is a division of that color (You cannot bet on green). Otherwise you loose your bet. If the bank has 1000 chips and you have 100 chips, what is the chance that you can *break the bank,* betting only one chip on red or black each spin of the wheel?

Solution.

a) Assuming player A has k chips, at the conclusion of the next play she will have either $k+1$ or $k-1$ chips, depending on whether she wins or loses that play. The probabilities of eventually winning must therefore satisfy the difference equation:

$$\begin{aligned} u(k) &= u(k|\text{win } k\text{th play})P(\text{win } k\text{th play}) + u(k|\text{loose } k\text{th play})P(\text{loose } k\text{th play}) \\ &= pu(k+1) + qu(k-1) \end{aligned}$$

In addition, we have the two initial (boundary) conditions:

$$u(0) = 0 \qquad u(a+b) = 1$$

This difference equation for $u(k)$ is linear and has constant coefficients. Assuming a general solution of the form $u(k) = \lambda^k$ and plugging it into the difference equation, after some simplification we get:

$$-p\lambda^2 + \lambda - q = 0$$

The corresponding roots are $\lambda = 1$, $\lambda = q/p$. Accordingly, the general solution assuming $q \neq p$ is:

$$u(k) = c_1 + c_2(q/p)^k$$

The two initial conditions give the equations:

$$\begin{aligned} 0 &= c_1 + c_2 \\ 1 &= c_1 + c_2(q/p)^{a+b} \end{aligned}$$

After solving for c_1 and c_2 and substituting the results into the general solution, we get:

$$u(k) = \frac{1-(q/p)^k}{1-(q/p)^{a+b}}$$

Finally, at the original position where player A has a chips, the corresponding probability of winning is:

$$u(a) = \frac{1-(q/p)^a}{1-(q/p)^{a+b}}$$

b) In this case:

$$p = \frac{18}{37} \qquad q = \frac{19}{37} \qquad a = 100 \qquad b = 1000$$

Thus:

$$u(100) = \frac{1-(19/18)^{100}}{1-(19/18)^{1100}} = 3.29 \times 10^{-24}$$

4. Identifying a point on the unit sphere from spherical distances. In this problem we consider the *unit sphere* in $\mathbb{R}^n$, which is defined as the set of vectors with norm one: $S^n = \{x \in \mathbb{R}^n \mid normx = 1\}$. We define the *spherical distance* between two vectors on the unit sphere as the distance between them, measured along the sphere, *i.e.*, as the angle between the vectors, measured in radians: If $x,\ y \in S^n$, the spherical distance between them is

$$\text{sphdist}(x, y) = \angle(x, y),$$

where we take the angle as lying between 0 and π. (Thus, the maximum distance between two points in S^n is π, which occurs only when the two points x, y are *antipodal*, which means $x = -y$.) Now suppose $p_1, \ldots, p_k \in S^n$ are the (known) positions of some beacons on the unit sphere, and let $x \in S^n$ be an unknown point on the unit sphere. We have exact measurements of the (spherical) distances between each beacon and the unknown point x, *i.e.*, we are given the numbers

$$\rho_i = \text{sphdist}(x, p_i), \quad i = 1, \ldots, k.$$

We would like to determine, without any ambiguity, the exact position of x, based on this information. Find the conditions on $p_1, \ldots, p_k$ under which we can unambiguously determine x, for any $x \in S^n$, given the distances ρ_i. You can give your solution algebraically, using any of the concepts used in class (*e.g.*, nullspace, range, rank), or you can give a geometric condition (involving the vectors p_i). You must justify your answer.

Solution. From

$$\rho_i = \arccos\left(\frac{p_i^T x}{\|p_i\|\|x\|}\right) = \arccos(p_i^T x),$$

we find that

$$p_i^T x = \cos\rho_i, \quad i = 1, \ldots, k.$$

Rewriting these equations in matrix form yields $Ax = b$, where $A \in \mathbb{R}^{k\times n}$ is the matrix with rows $p_1^T, \ldots, p_k^T$, and b is the vector with entries $\cos\rho_1, \ldots, \cos\rho_k$. Now if the matrix A has rank n, we can unambiguously find x given ρ. So a condition under which we can always recover x from the spherical distances is that A has rank n, which means nothing more than the vectors $p_1, \ldots, p_k$ span $\mathbb{R}^n$. In fact, that's exactly the condition. If it doesn't hold, *i.e.*, if the vectors $p_1, \ldots, p_k$ don't span $\mathbb{R}^n$, then there is a nonzero vector q such that

$$p_i^T q = 0, \quad i = 1, \ldots, k.$$

Now choose $x = q/\|q\|$, so x belongs to S^n. Then x has a spherical distance of $\pi/2$ to all the vectors $p_1, \ldots, p_k$, but the same is true for the point $-x$, which also belongs to S^n. It follows that we cannot unambiguously determine x (or $-x$) from the distances; x and $-x$ are indistinguishable. We can also give a very nice geometrical condition. The condition that $p_1, \ldots, p_k$ span $\mathbb{R}^n$, is the same as the condition that there exists no nonzero x with $p_i^T x = 0$. This can be interpreted geometrically as stating that the points $p_1, \ldots, p_k$ do not lie in a common plane (with normal vector x) passing through zero. We can restate this as saying that the points $p_1, \ldots, p_k$ do not lie on a common great circle.

5. Some true/false questions. Determine if the following statements are true or false. No justification or discussion is needed for your answers. What we mean by "true" is that the statement is true for all values of the matrices and vectors given. You can't assume anything about the dimensions of the matrices (unless it's explicitly stated), but you can assume that the dimensions are such that all expressions make sense. For example, the statement "$A+B = B+A$" is true, because no matter what the dimensions of A and B (which must, however, be the same), and no matter what values A and B have, the statement holds. As another example, the statement $A^2 = A$ is false, because there are (square) matrices for which this

doesn't hold. (There are also matrices for which it does hold, *e.g.*, an identity matrix. But that doesn't make the statement true.)

a) If all coefficients (*i.e.*, entries) of the matrix A are positive, then A is full rank.

b) If A and B are onto, then $A + B$ must be onto.

c) If A and B are onto, then so is the matrix $\begin{bmatrix} A & C \\ 0 & B \end{bmatrix}$.

d) If A and B are onto, then so is the matrix $\begin{bmatrix} A \\ B \end{bmatrix}$.

e) If the matrix $\begin{bmatrix} A \\ B \end{bmatrix}$ is onto, then so are the matrices A and B.

f) If A is full rank and skinny, then so is the matrix $\begin{bmatrix} A \\ B \end{bmatrix}$.

Solution.

a) If all coefficients (*i.e.*, entries) of the matrix A are positive, then A is full rank.

False. The matrix $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ has all entries positive and is singular, hence not full rank.

b) If A and B are onto, then $A + B$ must be onto.

False. The 1×1 matrix $A = 1$ is full rank, and so is the matrix $B = -1$. But $A + B = 0$ (the 1×1 zero), which is not onto.

c) If A and B are onto, then so is the matrix $\begin{bmatrix} A & C \\ 0 & B \end{bmatrix}$.

True. To show this matrix is onto, we need to show that we can solve the equations

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} A & C \\ 0 & B \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

for any y_1 and y_2. (These are all vectors.) The bottom block row is $y_2 = Bx_2$. Using the fact that B is onto, we can find at least one x_2 such that $y_2 = Bx_2$. The top block row is

$$y_1 = Ax_1 + Cx_2,$$

which we can rewrite as

$$Ax_1 = y_1 - Cx_2.$$

Using the fact that A is onto, we can find at least one x_1 that satisfies this equation. Now we're done.

d) If A and B are onto, then so is the matrix $\begin{bmatrix} A \\ B \end{bmatrix}$.

False. Let A and B both be the 1×1 matrix 1. These are each onto, but $\begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ is not.

e) If the matrix $\begin{bmatrix} A \\ B \end{bmatrix}$ is onto, then so are the matrices A and B.

True. To say that $\begin{bmatrix} A \\ B \end{bmatrix}$ is onto means that for any vector y, we can find at least one x that satisfies

$$y = \begin{bmatrix} A \\ B \end{bmatrix} x.$$

Let's use this to show that A and B are both onto. First let's consider the equation $z = Au$. We can solve this by finding an x that satisfies

$$\begin{bmatrix} z \\ 0 \end{bmatrix} = \begin{bmatrix} A \\ B \end{bmatrix} x.$$

In a similar way can solve the equation $w = Bv$ for any vector w.

f) If A is full rank and skinny, then so is the matrix $\begin{bmatrix} A \\ B \end{bmatrix}$.

True. Since the matrix A is skinny and full rank, its has zero nullspace: whenever we have $Ax = 0$, we can conclude $x = 0$. The matrix $\begin{bmatrix} A \\ B \end{bmatrix}$ is also skinny, so to show it is full rank we must show that it, too, has zero nullspace. To do this suppose that

$$\begin{bmatrix} A \\ B \end{bmatrix} x = 0.$$

This means that $Ax = 0$ and $Bx = 0$. From the first, we conclude that $x = 0$. This shows that $\begin{bmatrix} A \\ B \end{bmatrix}$ is full rank.

6. Temperatures in a multi-core processor. We are concerned with the temperature of a processor at two critical locations. These temperatures, denoted $T = (T_1, T_2)$ (in degrees C), are affine functions of the power dissipated by three processor cores, denoted $P = (P_1, P_2, P_3)$ (in W). We make 4 measurements. In the first, all cores are idling, and dissipate 10W. In the next three measurements, one of the processors is set to full power, 100W, and the other two are idling. In each experiment we measure and note the temperatures at the two critical locations.

P_1	P_2	P_3	T_1	T_2
10W	10W	10W	27°	29°
100W	10W	10W	45°	37°
10W	100W	10W	41°	49°
10W	10W	100W	35°	55°

Suppose we operate all cores at the same power, p. How large can we make p, without T_1 or T_2 exceeding 70°?

You must fully explain your reasoning and method, in addition to providing the numerical solution.

Solution. The temperature vector T is an affine function of the power vector P, *i.e.*, we have $T = AP + b$ for some matrix $A \in \mathbb{R}^{2\times 3}$ and some vector $b \in \mathbb{R}^2$. Once we find A and b, we can predict the temperature T for *any* value of P.

The first approach is to (somewhat laboriously) write equations describing the measurements in terms of the elements of A. Let a_{ij} denote the (i, j) entry of A. We can write out the relations $T = AP + b$ for the 4 experiments listed above as the set of 8 equations

$$\begin{aligned}
10a_{11} + 10a_{12} + 10a_{13} + b_1 &= 27,\\
10a_{21} + 10a_{22} + 10a_{23} + b_2 &= 29,\\
100a_{11} + 10a_{12} + 10a_{13} + b_1 &= 45,\\
100a_{21} + 10a_{22} + 10a_{23} + b_2 &= 37,\\
10a_{11} + 100a_{12} + 10a_{13} + b_1 &= 41,\\
10a_{21} + 100a_{22} + 10a_{23} + b_2 &= 49,\\
10a_{11} + 10a_{12} + 100a_{13} + b_1 &= 35,\\
10a_{21} + 10a_{22} + 100a_{23} + b_2 &= 55.
\end{aligned}$$

Next, we define a vector of unknowns, $x = (a_{11}, a_{12}, a_{13}, a_{21}, a_{22}, a_{23}, b_1, b_2) \in \mathbb{R}^8$. We rewrite the 8 equations above as $Cx = d$, where

$$C = \begin{bmatrix}
10 & 10 & 10 & 0 & 0 & 0 & 1 & 0\\
0 & 0 & 0 & 10 & 10 & 10 & 0 & 1\\
100 & 10 & 10 & 0 & 0 & 0 & 1 & 0\\
0 & 0 & 0 & 100 & 10 & 10 & 0 & 1\\
10 & 100 & 10 & 0 & 0 & 0 & 1 & 0\\
0 & 0 & 0 & 10 & 100 & 10 & 0 & 1\\
10 & 10 & 100 & 0 & 0 & 0 & 1 & 0\\
0 & 0 & 0 & 10 & 10 & 100 & 0 & 1
\end{bmatrix} \qquad d = \begin{bmatrix}
27\\ 29\\ 45\\ 37\\ 41\\ 49\\ 35\\ 55
\end{bmatrix}.$$

We solve for x as $x = C^{-1}d$. (It turns out that C is invertible.) Putting the entries of x into the appropriate places in A and b, we have

$$A = \begin{bmatrix}
0.200 & 0.156 & 0.089\\
0.089 & 0.222 & 0.289
\end{bmatrix}, \qquad b = \begin{bmatrix}
22.6\\ 23.0
\end{bmatrix}.$$

At this point we can predict T for any P (assuming we trust the affine model).

Substituting $P = (p, p, p)$ into $T = AP + b$, we get

$$T_1 = 0.444p + 22.6, \qquad T_2 = 0.600p + 23.0.$$

Both of these temperatures are increasing in p (it would be quite surprising if this were not the case). The value of p for which $T_1 = 70$ is $p = (70 - 22.6)/0.444 = 106.8$W. The value of p for which $T_2 = 70$ is $p = (70 - 23)/0.6 = 78.3$W. Thus, the maximum value of p for which both temperatures do not exceed 70° is $p = 78.3$W.

```
%problem data
```

```
C = [10 10 10 0 0 0 1 0
0 0 0 10 10 10 0 1
100 10 10 0 0 0 1 0
0 0 0 100 10 10 0 1
10 100 10 0 0 0 1 0
0 0 0 10 100 10 0 1
10 10 100 0 0 0 1 0
0 0 0 10 10 100 0 1];
d = [27 29 45 37 41 49 35 55]';

%Find affine model
x = C\d;
A = reshape(x(1:6), 3, 2)';
b = x(7:8)';

%Find maximum power
p1 = (70 - b(1))/sum(A(1,:));
p2 = (70 - b(2))/sum(A(2,:));
p = min(p1, p2)
```

Alternative solution. Another way of solving this problem is to directly exploit the fact that T is an affine function of P. This means that if we form any linear combination of the power vectors used in the experiment, *with the coefficients summing to one*, the temperature vector will also be the same linear combination of the temperatures.

By averaging the last three experiments we find if the powers are $P = (40, 40, 40)$, then the temperature vector is $T = (40.33, 47.00)$. (Note that this is really a prediction, based on the observed experimental data and the affineness assumption; it's not a new experiment!)

Now we form a new power vector of the form

$$P = (1-\theta)(10, 10, 10) + \theta(40, 40, 40) = (10 + 30\theta, 10 + 30\theta, 10 + 30\theta),$$

where $\theta \in \mathbb{R}$. The coefficients $1 - \theta$ and θ sum to one, so since T is affine, we find that the corresponding temperature vector is

$$T = (1-\theta)(27, 29) + \theta(40.33, 47.00) = (27 + 13.33\theta, 29 + 18\theta),$$

just as above. The first coefficient hits 70 at $\theta = 3.22$; the second coefficient hits 70 at $\theta = 2.23$. Thus, θ can be as large as $\theta = 2.27$. This corresponds to the powers $P = (78.3, 78.3, 78.3)$.

7. Single sensor failure detection and identification. We have $y = Ax$, where $A \in \mathbb{R}^{m\times n}$ is known, and $x \in \mathbb{R}^n$ is to be found. Unfortunately, up to one sensor may have failed (but you don't know which one has failed, or even whether any has failed). You are given $\tilde{y}$ and not y, where $\tilde{y}$ is the same as y in all entries except, possibly, one (say, the kth entry). If all sensors are operating correctly, we have $y = \tilde{y}$. If the kth sensor fails, we have $\tilde{y}_i = y_i$ for all $i \neq k$.

The file `one_bad_sensor.m`, available on the course web site, defines A and $\tilde{y}$ (as `A` and `ytilde`). Determine which sensor has failed (or if no sensors have failed). You must explain your method, and submit your code.

For this exercise, you can use the matlab code `rank([F g])==rank(F)` to check if $g \in \text{range}(F)$. (We will see later a much better way to check if $g \in \text{range}(F)$.)

Solution. Let $y^{(i)}$ be the measurement vector y with the ith entry removed. Likewise, let $A^{(i)}$ be the measurement matrix with the ith row of A removed. This corresponds to the system without the ith sensor.

If the ith sensor is faulty, we will almost surely have $y \notin \text{range}(A)$ (unless the sensor failure happens to give the same response y_i as that predicted by A, which is highly unlikely). However, once we remove its faulty measurement, we will certainly have $y^{(i)} \in \text{range}(A^{(i)})$.

To test if a vector z is in range(C), we can use matlab and compare `rank([C z]) == rank(C)`. If they are equal, $z \in \text{range}(C)$. Otherwise `rank([C z]) == rank(C) + 1`. To find a faulty sensor, we remove one row of A at a time, and use the above test.

The following matlab code solves the problem

```
one_bad_sensor

for k=1:m
    withoutk=[1:k-1 k+1:m];
    Atent = A(withoutk,:);
    ytent = ytilde(withoutk);
    if rank([ Atent ytent ]) == rank(Atent)
        k
    end
end
```

The 11th sensor is faulty.

8. Projection matrices. A matrix $P \in \mathbb{R}^{n\times n}$ is called a *projection matrix* if $P = P^\mathsf{T}$ and $P^2 = P$.

a) Show that if P is a projection matrix then so is $I - P$.

b) Suppose that the columns of $U \in \mathbb{R}^{n\times k}$ are orthonormal. Show that UU^T is a projection matrix. (Later we will show that the converse is true: every projection matrix can be expressed as UU^T for some U with orthonormal columns.)

c) Suppose $A \in \mathbb{R}^{n\times k}$ is full rank, with $k \leq n$. Show that $A(A^\mathsf{T}A)^{-1}A^\mathsf{T}$ is a projection matrix.

d) If $S \subseteq \mathbb{R}^n$ and $x \in \mathbb{R}^n$, the point y in S closest to x is called the *projection of x on S*. Show that if P is a projection matrix, then $y = Px$ is the projection of x on range(P). (Which is why such matrices are called projection matrices ...)

Solution.

a) To show that $I - P$ is a projection matrix we need to check two properties:

i. $I - P = (I - P)^\mathsf{T}$

ii. $(I-P)^2 = I - P$.

The first one is easy: $(I-P)^\mathsf{T} = I - P^\mathsf{T} = I - P$ because $P = P^\mathsf{T}$ (P is a projection matrix.) The show the second property we have

$$\begin{aligned}(I-P)^2 &= I - 2P + P^2 \\ &= I - 2P + P \qquad (\text{since } P = P^2) \\ &= I - P\end{aligned}$$

and we are done.

b) Since the columns of U are orthonormal we have $U^\mathsf{T}U = I$. Using this fact it is easy to prove that UU^T is a projection matrix, *i.e.*, $(UU^\mathsf{T})^\mathsf{T} = UU^\mathsf{T}$ and $(UU^\mathsf{T})^2 = UU^\mathsf{T}$. Clearly, $(UU^\mathsf{T})^\mathsf{T} = (U^\mathsf{T})^\mathsf{T}U^\mathsf{T} = UU^\mathsf{T}$ and

$$\begin{aligned}(UU^\mathsf{T})^2 &= (UU^\mathsf{T})(UU^\mathsf{T}) \\ &= U(U^\mathsf{T}U)U^\mathsf{T} \\ &= UU^\mathsf{T} \qquad (\text{since } U^\mathsf{T}U = I).\end{aligned}$$

c) First note that $\left(A(A^\mathsf{T}A)^{-1}A^\mathsf{T}\right)^\mathsf{T} = A(A^\mathsf{T}A)^{-1}A^\mathsf{T}$ because

$$\begin{aligned}\left(A(A^\mathsf{T}A)^{-1}A^\mathsf{T}\right)^\mathsf{T} &= (A^\mathsf{T})^\mathsf{T}\left((A^\mathsf{T}A)^{-1}\right)^\mathsf{T}A^\mathsf{T} \\ &= A\left((A^\mathsf{T}A)^\mathsf{T}\right)^{-1}A^\mathsf{T} \\ &= A(A^\mathsf{T}A)^{-1}A^\mathsf{T}.\end{aligned}$$

Also $\left(A(A^\mathsf{T}A)^{-1}A^\mathsf{T}\right)^2 = A(A^\mathsf{T}A)^{-1}A^\mathsf{T}$ because

$$\begin{aligned}\left(A(A^\mathsf{T}A)^{-1}A^\mathsf{T}\right)^2 &= \left(A(A^\mathsf{T}A)^{-1}A^\mathsf{T}\right)\left(A(A^\mathsf{T}A)^{-1}A^\mathsf{T}\right) \\ &= A\left((A^\mathsf{T}A)^{-1}A^\mathsf{T}A\right)(A^\mathsf{T}A)^{-1}A^\mathsf{T} \\ &= A(A^\mathsf{T}A)^{-1}A^\mathsf{T} \qquad (\text{since } (A^\mathsf{T}A)^{-1}A^\mathsf{T}A = I).\end{aligned}$$

d) To show that Px is the projection of x on range(P) we verify that the "error" $x - Px$ is orthogonal to *any* vector in range(P). Since range(P) is nothing but the span of the columns of P we only need to show that $x - Px$ is orthogonal to the columns of P, or in other words, $P^\mathsf{T}(x - Px) = 0$. But

$$\begin{aligned}P^\mathsf{T}(x - Px) &= P(x - Px) \qquad (\text{since } P = P^\mathsf{T}) \\ &= Px - P^2x \\ &= 0 \qquad (\text{since } P^2 = P)\end{aligned}$$

and we are done.

9. Groups of equivalent statements. In the list below there are 11 statements about two square matrices A and B in $\mathbb{R}^{n\times n}$.

a) $\text{range}(B) \subseteq \text{range}(A)$.

b) there exists a matrix $Y \in \mathbb{R}^{n\times n}$ such that $B = YA$.

c) $AB = 0$.

d) $BA = 0$.

e) $\text{rank}(\begin{bmatrix} A & B \end{bmatrix}) = \text{rank}(A)$.

f) $\text{range}(A) \perp \text{null}(B^{\mathsf{T}})$.

g) $\text{rank}(\begin{bmatrix} A \\ B \end{bmatrix}) = \text{rank}(A)$.

h) $\text{range}(A) \subseteq \text{null}(B)$.

i) there exists a matrix $Z \in \mathbb{R}^{n\times n}$ such that $B = AZ$.

j) $\text{rank}(\begin{bmatrix} A & B \end{bmatrix}) = \text{rank}(B)$.

k) $\text{null}(A) \subseteq \text{null}(B)$.

Your job is to collect them into (the largest possible) groups of equivalent statements. Two statements are equivalent if each one implies the other. For example, the statement 'A is onto' is equivalent to '$\text{null}(A) = \{0\}$' (when A is square, which we assume here), because every square matrix that is onto has zero nullspace, and vice versa. Two statements are not equivalent if there exist (real) square matrices A and B for which one holds, but the other does not. A group of statements is equivalent if any pair of statements in the group is equivalent.

We want *just* your answer, which will consist of lists of mutually equivalent statements; we do not need any justification.

Put your answer in the following specific form. List each group of equivalent statements on a line, in (alphabetic) order. Each new line should start with the first letter not listed above. For example, you might give your answer as

a, c, d, h
b, i
e
f, g, j, k.

This means you believe that statements a, c, d, and h are equivalent; statements b and i are equivalent; and statements f, g, j, and k are equivalent. You also believe that the first group of statements is not equivalent to the second, or the third, and so on.

Solution. Let b_i be the ith column of B.

$$\begin{aligned}\text{range}(B) \subseteq \text{range}(A) &\Leftrightarrow \text{every column of } B \text{ is in the range of } A \\ &\Leftrightarrow \text{there exists a vector } z_i \text{ such that } b_i = Az_i \\ &\Leftrightarrow \text{there exists a matrix } Z \in \mathbb{R}^{n\times n} \text{ such that } B = AZ \\ &\Leftrightarrow \text{rank}(\begin{bmatrix} A & B \end{bmatrix}) = \text{rank}(A). \end{aligned} \tag{1}$$

This shows that statements a, e and i are equivalent.

$$\begin{aligned}\text{null}(A) \subseteq \text{null}(B) &\Leftrightarrow \text{null}(A)^\perp \supseteq \text{null}(B)^\perp \\ &\Leftrightarrow \text{range}(B^\mathsf{T}) \subseteq \text{range}(A^\mathsf{T}) \\ &\Leftrightarrow \text{there exists a matrix } \tilde{Y} \in \mathbb{R}^{n\times n} \text{ such that } B^\mathsf{T} = A^\mathsf{T}\tilde{Y} \\ &\Leftrightarrow \text{there exists a matrix } Y \in \mathbb{R}^{n\times n} \text{ such that } B = YA \\ &\Leftrightarrow \text{rank}(\begin{bmatrix} A^\mathsf{T} & B^\mathsf{T} \end{bmatrix}) = \text{rank}(A^\mathsf{T}) \\ &\Leftrightarrow \text{rank}(\begin{bmatrix} A \\ B \end{bmatrix}) = \text{rank}(A). \end{aligned} \tag{2}$$

This shows that statements b, g and k are equivalent.

$$\begin{aligned}\text{range}(A) \subseteq \text{null}(B) &\Leftrightarrow \text{for all } z \in \mathbb{R}^n,\ B(Az) = 0 \\ &\Leftrightarrow BA = 0. \end{aligned} \tag{3}$$

This shows that statements d and h are equivalent.

$$\begin{aligned}\text{range}(A) \perp \text{null}(B^\mathsf{T}) &\Leftrightarrow \text{range}(A) \subseteq \text{null}(B^\mathsf{T})^\perp \\ &\Leftrightarrow \text{range}(A) \subseteq \text{range}(B) \\ &\Leftrightarrow \text{rank}(\begin{bmatrix} A & B \end{bmatrix}) = \text{rank}(B). \end{aligned} \tag{4}$$

This shows that statements f and j are equivalent.

None of these groups of statements is equivalent to any other, or to c. This is demonstrated by the following counterexamples.

Take

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}.$$

Since $AB = 0$ but $BA \neq 0$, then group (3) and statement c are not equivalent. Furthermore since

$$\text{rank}(\begin{bmatrix} A \\ B \end{bmatrix}) = \text{rank}(A) = \text{rank}(B) = 1$$

but $\text{rank}([A \quad B]) = 2$, groups (2) and (1) are not equivalent. Groups (2) and (4) are not either.

When $A = B \neq 0$, $\text{null}(A) = \text{null}(B)$ but $AB = BA = A^2 \neq 0$. Hence groups (2) and (3) are not equivalent. Group (2) and statement c are not equivalent either.

Take

$$A = I, \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}.$$

Since $\text{rank}([AB]) = \text{rank}(A) = 2$ but $\text{rank}(B) = 1$, groups (1) and (4) are not equivalent. Furthermore since $BA \neq 0$ groups (1) and (3) are not equivalent. Since $AB \neq 0$, group (1) and statement c aren't either.

In a similar fashion, taking

$$A = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad B = I,$$

shows that groups (3) and (4) are not equivalent and that statement c and group (4) aren't either.

Thus, the final answer is

a, e, i

b, g, k

c

d, h

f, j.