Assignment 9

Benjamin Jakubowski

November 23, 2015

1. True or False

a. Projection of x

The statement "*the projection of a vector on a subspace S is equal to*

$$\mathcal{P}_S x = \sum_{i=1}^{n} \langle x, b_i \rangle b_i$$

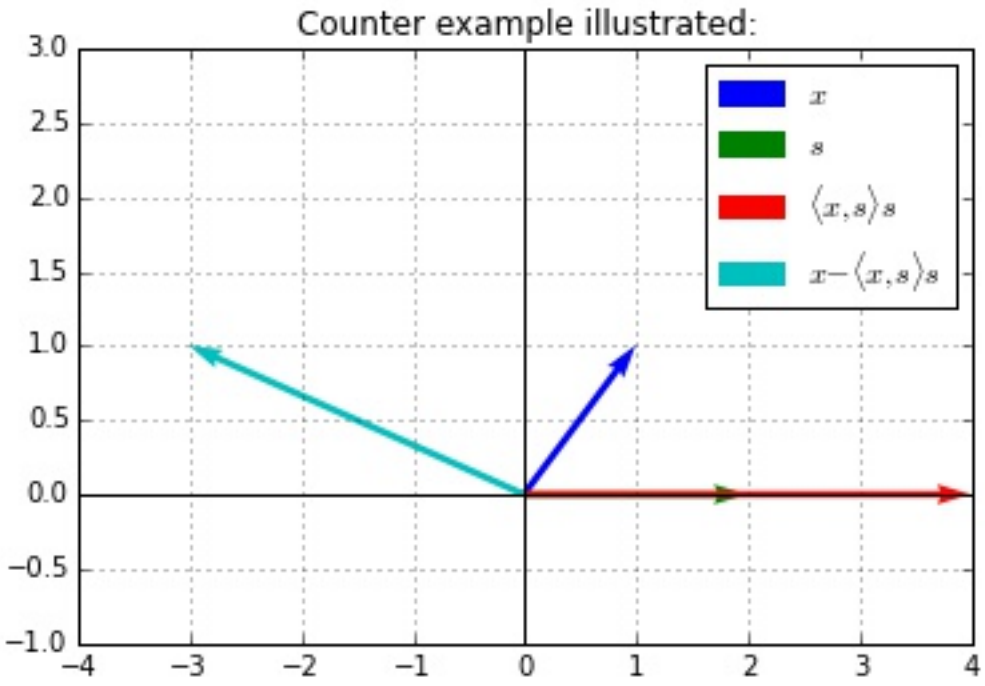
for any basis $b_1, ..., b_n$ of S" is false.
Counterexample:
Let

- $V = R^2$
- $S = \text{span}\,\{s\} = \text{span}\left\{\begin{bmatrix}2\\0\end{bmatrix}\right\}$.
- $x = \begin{bmatrix}1\\1\end{bmatrix}$.

Then $\langle x, s \rangle x = (1 \cdot 2 + 1 \cdot 0) \cdot \begin{bmatrix}2\\0\end{bmatrix} = \begin{bmatrix}4\\0\end{bmatrix}$.
But then

$$x - \langle x, s \rangle s = \begin{bmatrix}1\\1\end{bmatrix} - \begin{bmatrix}4\\0\end{bmatrix} = \begin{bmatrix}-3\\1\end{bmatrix} \notin S^{\perp}$$

We can understand this visually:



Since s is not normal, $\langle x, s\rangle s$ is **not** the projection of x on s. Instead, we'd need to define

$$\mathcal{P}_S x := \langle x, s\rangle \frac{s}{||s||_2}$$

B. Orthogonal complements

The statement "*the orthogonal complement of the orthogonal complement of a subspace S is S*" is true.

Proof:

First, take $y \in S$. By definition, $\langle y, x\rangle = 0$ for all $x \in S^\perp$. Thus (again applying the definition of orthogonal complement) $y \in \left(S^\perp\right)^\perp$, so $S \subseteq \left(S^\perp\right)^\perp$.

Finally, take $v \in \left(S^\perp\right)^\perp$. Since $v \in V$ (and S is a subspace), we can write

$$v = \mathcal{P}_{S^\perp} v + \mathcal{P}_S v$$

Since $v \in \left(S^\perp\right)^\perp$, $\mathcal{P}_{S^\perp} v = 0$ so $v = \mathcal{P}_S v$ and $v \in S$. Thus $\left(S^\perp\right)^\perp \subseteq S$, so $\left(S^\perp\right)^\perp = S$.

C. Power method convergence

The statement "*... the power method converges to a vector in the span of $\{v_1\}$ no matter how you initialize it*" is false.

Counterexample:

Recall $A \in \mathbb{R}^n$ has n eigenvalues $\lambda_1 > \lambda_2 > ... > \lambda_n$ corresponding to n linearly independent eigenvectors $v_1, v_2, ..., v_n$.
Now let $x = \sum_{i=1}^{n} \alpha_i Q_{:i}$ be the initializing vector for the power method (where $Q_{:i}$ is the i^{th} column in $A = Q\Lambda Q^{-1}$ or equivalently the i^{th} eigenvector).
Then (by (66) on page 10 of the notes),

$$A^k x = \sum_{i=1}^{n} \alpha_i \lambda_i^k Q_{:i}$$

Now, if $\alpha_1 = 0$, then

$$A^k x = \sum_{i=1}^{n} \alpha_i \lambda_i^k Q_{:i} = \sum_{i=2}^{n} \alpha_i \lambda_i^k Q_{:i}$$

Then, as $k \to \infty$ the term $\alpha_2 \lambda_2^k Q_{:2}$ will dominate. Thus, if you happen to (completely improbably) initialize the power method with $x \in \text{span}\{v_1\}^{\perp}$ it will not converge to a vector in $\text{span}\{v_1\}$ but instead to a vector in $\text{span}\{v_2\}$.

2. Heartbeat

a. Finding a basis for span of Y_1 and Y_2

First, recall we are given:

- $Y_1 = B + M$
- $Y_2 = M + N$
- $E(M) = E(B) = E(N) = 0 \implies E(Y_1) = E(Y_2) = 0$
- $Var(B) = Var(N) = \sigma^2, Var(M) = 3\sigma^s$
- $Cov(B, N) = Cov(N, M) = Cov(B, M) = 0$

We want to find an orthonormal basis for $\text{span}\{Y_1, Y_2\}$.
First note:

$$\begin{aligned} Var(Y_1) &= Var(B + M) = Var(B) + Var(M) \qquad \text{(since uncorrelated)} \\ &= \sigma^2 + 3\sigma^2 = 4\sigma^2 \end{aligned}$$

Similarly,

$$\begin{aligned} Var(Y_2) &= Var(M + N) = Var(M) + Var(N) \qquad \text{(since uncorrelated)} \\ &= 3\sigma^2 + 1\sigma^2 = 4\sigma^2 \end{aligned}$$

Next,

$$\begin{aligned} Cov(Y_1, Y_2) &= E(Y_1Y_2) - E(Y_1)E(Y_2) \\ &= E(Y_1Y_2) \\ &= E\left((B+M)(M+N)\right) \\ &= E\left(BM + M^2 + BN + MN\right) \\ &= E(BM) + E(M^2) + E(BN) + E(MN) \\ &= 0 + E(M^2) + 0 + 0 \\ &= 3\sigma^2 \end{aligned}$$

Now let's use Gram-Schmidt to find a basis for span$\{Y_1, Y_2\}$. First we'll find a basis- then we'll normalize the basis vectors to produce an orthonormal basis.
Let $U_1 = Y_1$.
Then let

$$\begin{aligned} U_2 &= Y_2 - \frac{\langle Y_2, U_1\rangle}{\langle U_1, U_1\rangle} U_1 \\ &= Y_2 - \frac{\langle Y_2, Y_1\rangle}{\langle Y_1, Y_1\rangle} Y_1 \\ &= Y_2 - \frac{E(Y_1Y_2)}{E(Y_1^2)} Y_1 \\ &= Y_2 - \frac{Cov(Y_1, Y_2)}{Var(Y_1)} Y_1 \\ &= Y_2 - \frac{3\sigma^2}{4\sigma^2} Y_1 \\ &= Y_2 - \frac{3}{4} Y_1 \end{aligned}$$

So our (non-orthonormal) basis for span$\{Y_1, Y_2\}$ is $\{Y_1, Y_2 - \frac{3}{4}Y_1\}$.
Now we normalize these basis vectors to produce an orthonormal basis:

$$\begin{aligned} V_1 &= \frac{U_1}{\sqrt{\langle U_1, U_1\rangle}} = \frac{Y_1}{\sqrt{Var(Y_1)}} = \frac{Y_1}{2\sigma} \\ V_2 &= \frac{U_2}{\sqrt{\langle U_2, U_2\rangle}} = \frac{Y_2 - \frac{3}{4}Y_1}{\sqrt{Var(Y_2 - \frac{3}{4}Y_1)}} \end{aligned}$$

Thus, to proceed, we first need to find $Var(Y_2 - \frac{3}{4}Y_1)$.

$$\begin{aligned} Var\left(Y_2 - \frac{3}{4}Y_1\right) &= Var(Y_2) + \left(\frac{-3}{4}\right)^2 Var(Y_1) + 2\left(\frac{-3}{4}\right) Cov(Y_1, Y_2) \\ &= 4\sigma^2 + \frac{9}{4}\sigma^2 - \frac{3}{2}\cdot 3\sigma^2 \\ &= \left(4 + \frac{9}{4} - \frac{9}{2}\right)\sigma^2 \\ &= 7/4\sigma^2 \end{aligned}$$

Thus,

$$\begin{aligned} V_1 &= \frac{Y_1}{2\sigma} \\ V_2 &= \frac{Y_2 - \frac{3}{4}Y_1}{\sqrt{Var(Y_2 - \frac{3}{4}Y_1)}} = \frac{Y_2 - \frac{3}{4}Y_1}{\sqrt{7/4\sigma^2}} \\ &= \frac{2}{\sqrt{7}\sigma}Y_2 - \frac{3}{2\sqrt{7}\sigma}Y_1 \end{aligned}$$

So, our final orthonormal basis for span$\{Y_1, Y_2\}$ is $\left\{\frac{1}{2\sigma}Y_1, \frac{2}{\sqrt{7}\sigma}Y_2 - \frac{3}{2\sqrt{7}\sigma}Y_1\right\}$.

B. Best linear estimator

We want to find the best linear estimator $\hat{B}_{LMMSE}$ given Y_1 and Y_2. This is

$$\arg\min_{\hat{B}} E\left((\hat{B} - B)^2\right)$$

subject to $\hat{B} = aY_1 + BY_2 + c$ for $a, b, c \in \mathbb{R}$.

Note: I solved this problem rather circuitously at first. Then I realized I could solve it much more simply by projecting on the basis found in part (a). Instead of just deleting my first attempt (which is also correct), I've decided to present both

B.1. Find $\hat{B}_{LMMSE}$ through explicit minimization

The cost function is

$$\begin{aligned} h(a,b,c) &= E\left((\hat{B} - B)^2\right) \\ &= E\left((aY_1 + bY_2 + c - B)^2\right) \\ &= E\left((a(B+M) + b(M+N) + c - B)^2\right) \\ &= E\left(((a-1)B + (a+b)M + bN + c)^2\right) \\ &= E\left((a-1)^2B^2 + (a+b)^2M^2 + b^2N^2 + c^2\right) \qquad {}^*(\textit{see below for justification}) \\ &= (a-1)^2E(B^2) + (a+b)^2E(M^2) + b^2E(N^2) + c^2 \end{aligned}$$

* Note in expanding this quadratic polynomial we can drop all cross terms (of form kXY for $k \in \mathbb{R}$ and $X, Y \in \{B, M, N\}, X \neq Y$) since these random variables are all uncorrelated. We can also drop all terms of the form kX since all the random variables are zero mean.
Note this expression is clearly minimized when $c = 0$. Thus the cost function becomes

$$\begin{aligned} h(a,b) &= (a-1)^2 E(B^2) + (a+b)^2 E(M^2) + b^2 E(N^2) \\ &= (a-1)^2\sigma^2 + (a+b)^2 3\sigma^2 + b^2\sigma^2 \\ &= \left[(a-1)^2 + 3(a+b)^2 + b^2\right]\sigma^2 \end{aligned}$$

Now we find the minimum of this function by setting the partials to zero and evaluating:

$$\begin{aligned} \frac{\delta h}{\delta a} &= \left[2(a-1) + 6(a+b)\right]\sigma^2 = 0 \\ \frac{\delta h}{\delta b} &= \left[6(a+b) + 2b\right]\sigma^2 = 0 \end{aligned}$$

Thus

$$\begin{aligned} 8a + 6b - 2 &= 0 \\ \implies 4a + 3b &= 1 \end{aligned}$$

and

$$\begin{aligned} 6a + 8b &= 0 \\ \implies a &= -4/3b \end{aligned}$$

Substituting yields

$$\begin{aligned} 4(-4/3b) + 3b &= 1 \\ \implies -16b + 9b &= 3 \\ \implies b &= -3/7 \end{aligned}$$

so $a = 4/7$.
Before we proceed, let's check this is in fact a minimum:

$$\begin{aligned} \det \begin{bmatrix} \frac{\delta^2 h}{\delta a^2} & \frac{\delta^2 h}{\delta a \delta b} \\ \frac{\delta^2 h}{\delta b \delta a} & \frac{\delta^2 h}{\delta b^2} \end{bmatrix} &= \det \begin{bmatrix} 8\sigma^2 & 6\sigma^2 \\ 6\sigma^2 & 8\sigma^2 \end{bmatrix} \\ &= 8^2\sigma^4 - 6^2\sigma^4 = 28\sigma^4 > 0 \end{aligned}$$

So it is in fact a minimum. Thus

$$\hat{B}_{LMMSE} = 4/7 \cdot Y_1 - 3/7 \cdot Y_2$$

b.2. Find $\hat{B}_{LMMSE}$ through projection

Now for the better answer. By definition, $\hat{B}_{LMMSE}$ is just the projection of B onto the subspace spanned by Y_1 and Y_2. Hence we can use the orthonormal basis found in part (a) to easily find $\hat{B}_{LMMSE}$:

$$\mathcal{P}_{\text{span}\{Y_1,Y_2\}}B = \langle\frac{1}{2\sigma}Y_1, B\rangle\frac{1}{2\sigma}Y_1 + \langle\frac{2}{\sqrt{7}\sigma}Y_2 - \frac{3}{2\sqrt{7}\sigma}Y_1, B\rangle\left(\frac{2}{\sqrt{7}\sigma}Y_2 - \frac{3}{2\sqrt{7}\sigma}Y_1\right)$$

Since $Cov(M, B) = Cov(N, B) = 0$, this becomes

$$\begin{aligned}\mathcal{P}_{\text{span}\{Y_1,Y_2\}}B &= \langle\frac{1}{2\sigma}B, B\rangle\frac{1}{2\sigma}Y_1 + \langle\frac{-3}{2\sqrt{7}\sigma}B, B\rangle\left(\frac{2}{\sqrt{7}\sigma}Y_2 - \frac{3}{2\sqrt{7}\sigma}Y_1\right)\\ &= \frac{1}{4}Y_1 - \frac{3}{2\sqrt{7}\sigma}\cdot\sigma^2\left(\frac{2}{\sqrt{7}\sigma}Y_2 - \frac{3}{2\sqrt{7}\sigma}Y_1\right)\\ &= \frac{1}{4}Y_1 - \frac{3}{7}Y_2 + \frac{9}{28}Y_1\\ &= \frac{4}{7}Y_1 - \frac{3}{7}Y_2\end{aligned}$$

This is a much more elegant solution, because it immediately lends itself to a geometric interpretation- $\hat{B}_{LMMSE}$ is the closest vector to B in the subspace span$\{Y_1, Y_2\}$.

3. Cold

a. Probability Bob has a cold today

Let's assume the following

1. Bob's age in days is large.
2. This process is well modeled as a time-invariant Markov process.

Additionally let

- $T = s$ if sick today, $T = w$ if well today
- $Y = s$ if sick yesterday, $Y = w$ if well yesterday

Then, the transition matrix is

$$P = \begin{bmatrix}P(T=s|Y=s) & P(T=s|Y=w)\\ P(T=w|Y=s) & P(T=w|Y=w)\end{bmatrix} = \begin{bmatrix}0.8 & 0.1\\ 0.2 & 0.9\end{bmatrix}$$

If Bob is indeed old, we can assume this Markov process has converged to its steady state (corresponding to $\lambda_1 = 1$). Thus,

$$v_1 = \begin{bmatrix}P(T=s)\\ P(T=w)\end{bmatrix} = \begin{bmatrix}P(Y=s)\\ P(Y=w)\end{bmatrix}$$

such that

$$\begin{aligned} Pv_1 &= v_1 \\ (\mathbb{I}_2 - P)v_1 &= 0 \\ \begin{bmatrix} .2 & -.1 \\ -.2 & .1 \end{bmatrix} v_1 &= 0 \end{aligned}$$

Now recalling $v_1 = \begin{bmatrix} P(T=s) \\ P(T=w) \end{bmatrix}$, we know:

$$\begin{aligned} .2P(T=s) - 0.1P(T=w) &= 0 \\ \implies \quad 0.2P(T=s) &= 0.1P(T=w) \\ \implies \quad P(T=s) &= 1/2P(T=w) \end{aligned}$$

Finally, since v_1 is a PMF, we know $P(T=s) + P(T=w) = 1$. Hence, under our assumptions, $P(T=s) = 1/3$.

b. Probability Bob had a cold three days ago

Now we know Bob has a cold today. We're interested in predicting whether he was sick three days ago. First, let

$$D_3 = \begin{cases} s & \text{if sick three days ago} \\ w & \text{if well three days ago} \end{cases}$$

Then we know our MAP estimate for D_3 is

$$\hat{D_{3MAP}} = \arg \max_{x \in \{s,w\}} P(D_3 = x | T = s)$$

But (using Bayes rule)

$$P(D_3 = x | T = s) = c \cdot P(D_3 = x) P(T = s | D_3 = x)$$

where c is the inverse probability of the evidence (a constant that does not effect our MAP estimate).
Next, we find $P(T = s|D_3 = x)$. Note we find this probability by first determining the vector $\begin{bmatrix} P(T=s|D_3=x) \\ P(T=w|D_3=x) \end{bmatrix}$ for $x \in \{s, w\}$.

$$\begin{bmatrix} P(T=s|D_3=x) \\ P(T=w|D_3=x) \end{bmatrix} = \begin{cases} P^3 \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \text{if } x = s \\ P^3 \begin{bmatrix} 0 \\ 1 \end{bmatrix} & \text{if } x = w \end{cases}$$

Using WolframAlpha, we find $P^3 = \begin{bmatrix} 0.526 & 0.219 \\ 0.438 & 0.781 \end{bmatrix}$. Thus

$$\begin{bmatrix} P(T = s|D_3 = x) \\ P(T = w|D_3 = x) \end{bmatrix} = \begin{cases} \begin{bmatrix} 0.561 \\ 0.438 \end{bmatrix} & \text{if } x = s \\ \begin{bmatrix} 0.219 \\ 0.781 \end{bmatrix} & \text{if } x = w \end{cases}$$

Thus

$$P(T = s|D_3 = x) = \begin{cases} 0.561 & \text{if } x = s \\ 0.219 & \text{if } x = w \end{cases}$$

Finally, note (again under our steady state approximation) $P(D_3 = s) = 1/3$ and $P(D_3 = w) = 2/3$. Hence,

$$c \cdot P(D_3 = x)P(T = s|D_3 = x) = \begin{cases} c \cdot 1/3 \cdot 0.561 = 0.187c & \text{if } x = s \\ c \cdot 2/3 \cdot 0.219 = 0.146c & \text{if } x = w \end{cases}$$

Thus, our MAP estimate for D_3 is

$$\hat{D}_{3MAP} = s$$

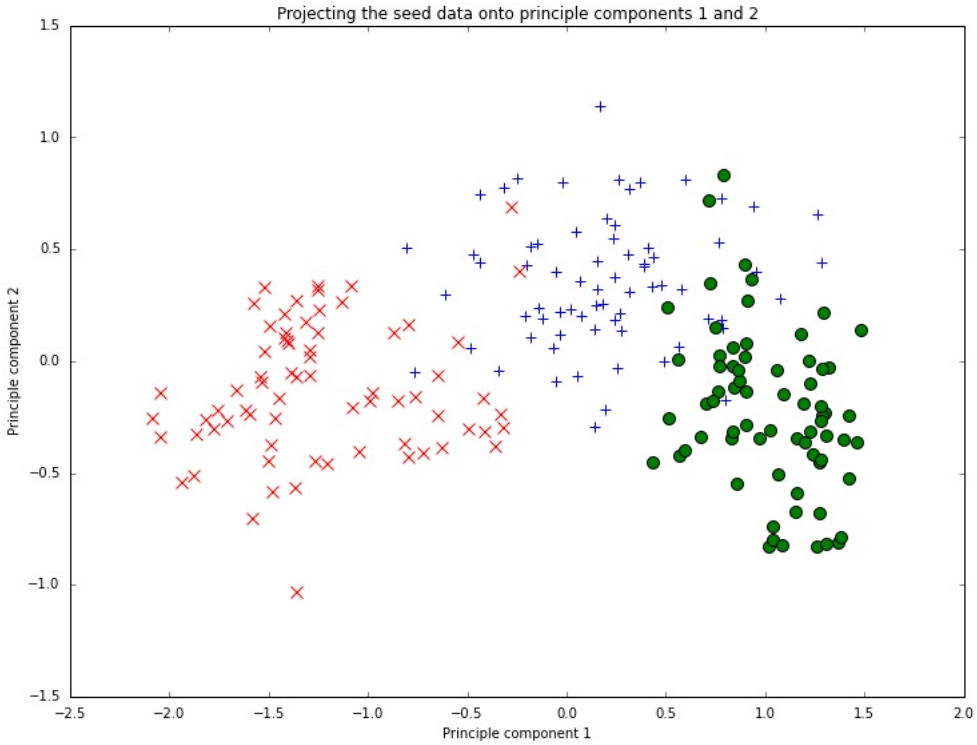
and our probability of error is $\frac{0.146c}{0.146c+0.187c} = 43.8\%$.

4. PCA of Wheat data

a. Projecting the seeds data onto different principle components

A.1. Projecting onto principle components [1] and [2]

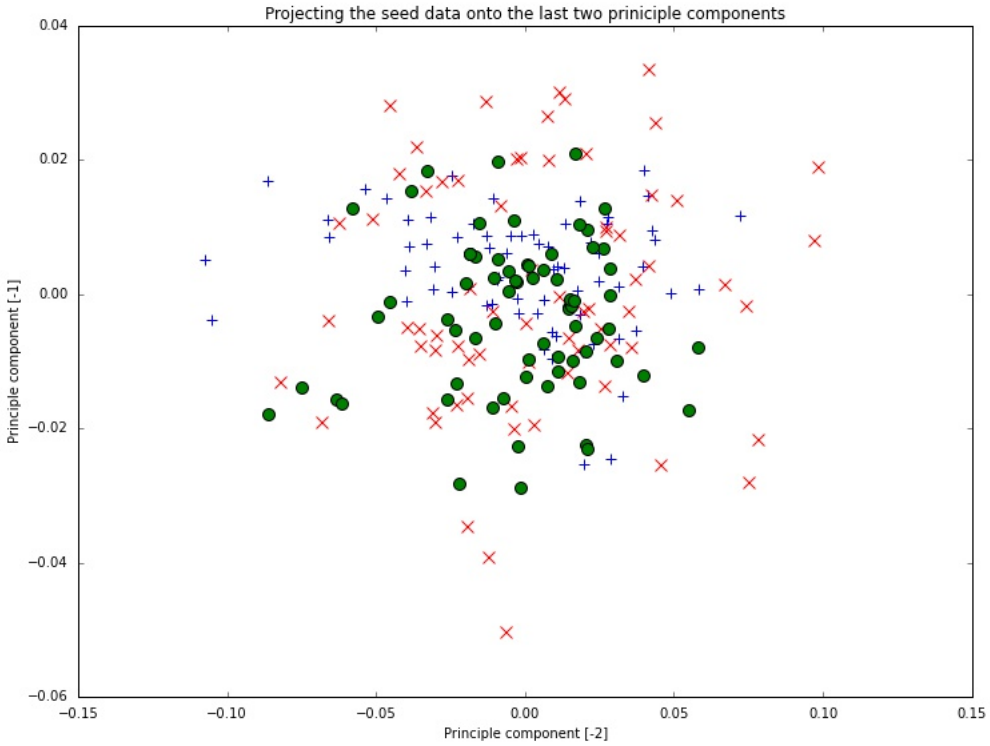
The projection of the seeds data onto principle components [1] and [2] is shown below:



As you can see, the three seed varietals are well-discriminated when projected onto the first two principle components.

A.2. Projecting onto principle components [-2] and [-1]

The projection of the seeds data onto principle components [-2] and [-1] is shown below:



As you can see, the three seed varietals are poorly discriminated when projected onto the last two principle components. This is as expected (assuming the predictors are in fact informative). The first two principle components capture the majority of the variance (energy), while the last two components capture very little (and, as such, are essentially describing noise dimensions in the predictor space).

B. Using SVD for learning

Imagine we have a training set and a test set. To classify the seeds of the test set using dot products and the SVD, I would:

1. Break the training set into k folds (perhaps k=10) for cross validation.
2. Determine the singular value decomposition for the x-validation training set.
3. Iteratively project the hold-out set onto the first $i \in \{1, ..., n\}$ principal components. For each i, classify each observation y in the hold-out x-validation set as

$$\arg \min_{\text{Variety} \in \{\text{Kama, Rosa, Canadian}\}} ||y - \bar{x}_{variety}||_2$$

and determine the accuracy (or performance under some other evaluation metric). Note that here $\bar{x}_{variety}$ is the mean vector for each variety.

4. Pick the number j of principal components that minimize the loss function (i.e. minimize 0-1 loss), averaged over the k-folds.

5. Determine the singular value decomposition for the entire dataset. Then take the first j principal components, project the test set observations on these components, and use the same classification method described above to classify each observation.