# STAT406- Methods of Statistical Learning Lecture 24

Matias Salibian-Barrera

UBC - Sep / Dec 2017

# Principal Components Analysis

- Dimension reduction
- For simplicity of model / interpretation
- Because of model requirements
- How can we find the **best** variables to use?

  (what do we mean by "**best**"?)

  (which variables can I chose from?)

# Principal components analysis

- Two different ways to introduce them...
- but, at the end of the day, they are the same objects –
- – linear combinations of the original variables that:
- are uncorrelated with each other;
- provide the “best lower-dimensional approximation” to the data.

# PCA - Max variance

- Given a centered sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$,
- 1st PC: find $\|\mathbf{a}\| = 1$ such that

$$\text{Var}(Y_1, \ldots, Y_n) = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{a}'\mathbf{X}_i)^2$$

is maximized, $Y_i = \mathbf{a}'\mathbf{X}_i$.

The solution is $\mathbf{a} = \mathbf{a}_0$ the eigenvector of $\mathbf{S}_n$ associated with its largest eigenvalue $\tilde{\lambda}_1$.

# Principal components analysis

- For the 2nd PC: find $\|\mathbf{a}\| = 1$ such that and, if $W_i = \mathbf{a}'\mathbf{X}_i$, $i = 1, \ldots, n$,

$$\text{cov}\,(Y_1, Y_2, \ldots, Y_n; W_1, W_2, \ldots, W_n) = \sum_{i=1}^{n} Y_i W_i = 0$$

and

$$\text{Var}(W_1, \ldots, W_n) = \frac{1}{n-1} \sum_{i=1}^{n} (W_i)^2$$

is maximized.

# Principal components analysis

- The solution is $\mathbf{a} = \mathbf{a}_1$ the eigenvector of $\mathbf{S}_n$ associated with its **2nd largest** eigenvalue $\tilde{\lambda}_2$.
- etc.

# PCA

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^p$ be a sample.

Assume that $\bar{\mathbf{X}}_n = \mathbf{0}$

For a subspace $\mathbf{L}$ of dimension $k$ let

$$P(\mathbf{X}_i, \mathbf{L})$$

be the orthogonal projection of $\mathbf{X}_i$ onto $\mathbf{L}$

# Principal components analysis

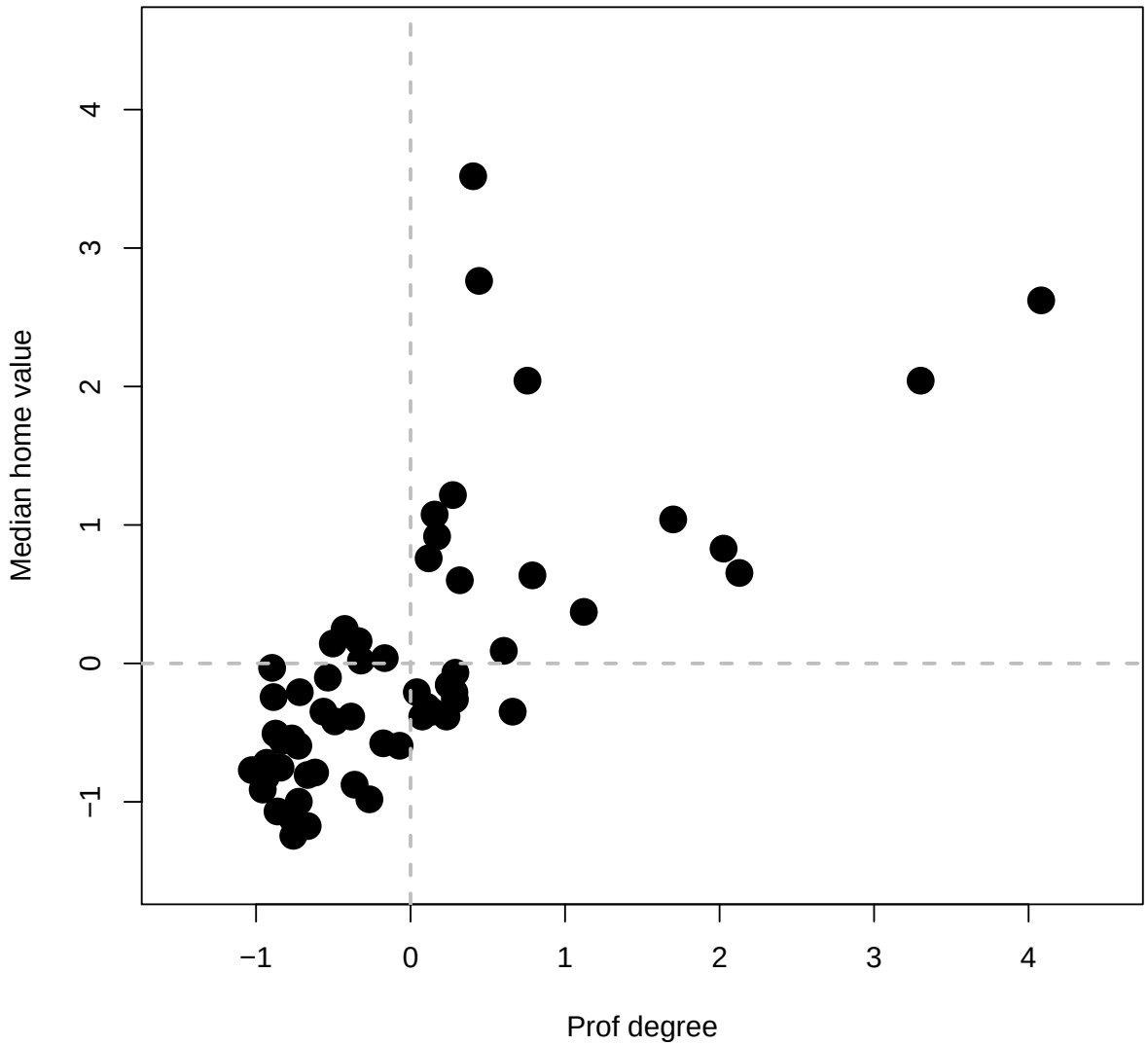
Dimension reduction

Find the “best” subspace $\mathbf{L}$ of dimension $k$

“Best” means: that best approximates the data

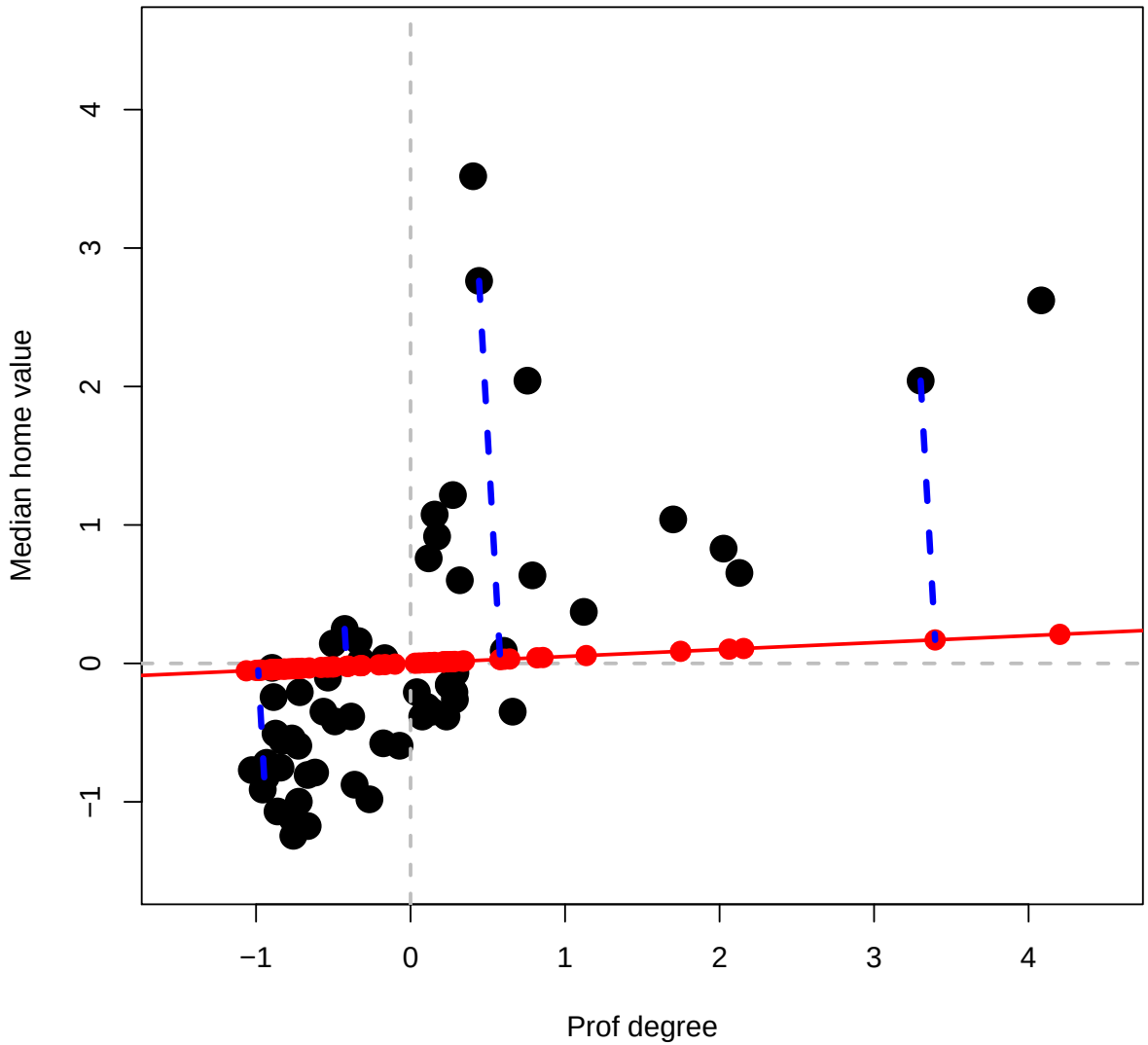
$$\min_{\mathbf{L}} \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{X}_i - P(\mathbf{X}_i, \mathbf{L})\|^2$$

# PCA - simple example

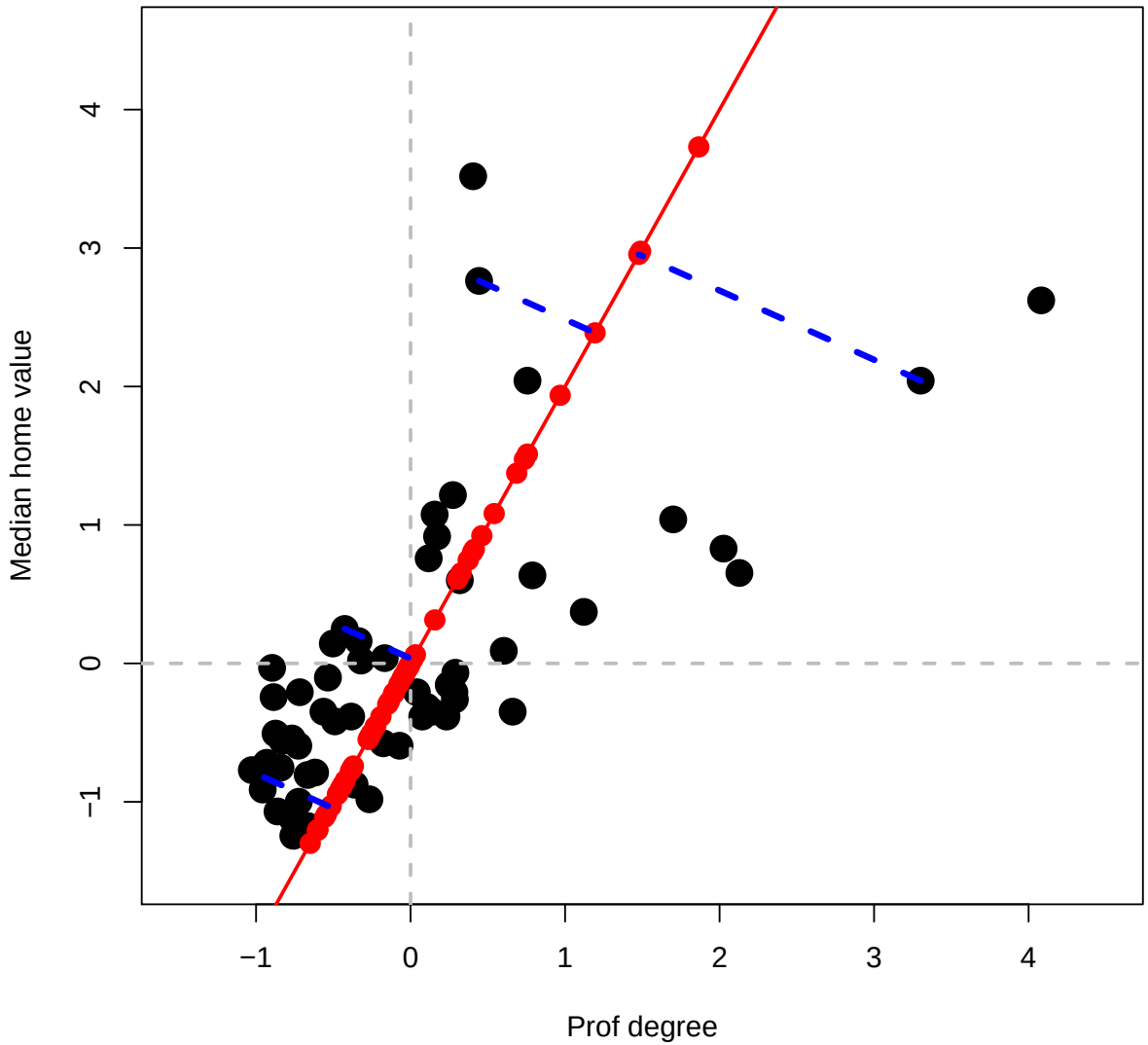



Census data – Percent of pop. with professional degree & Median home value

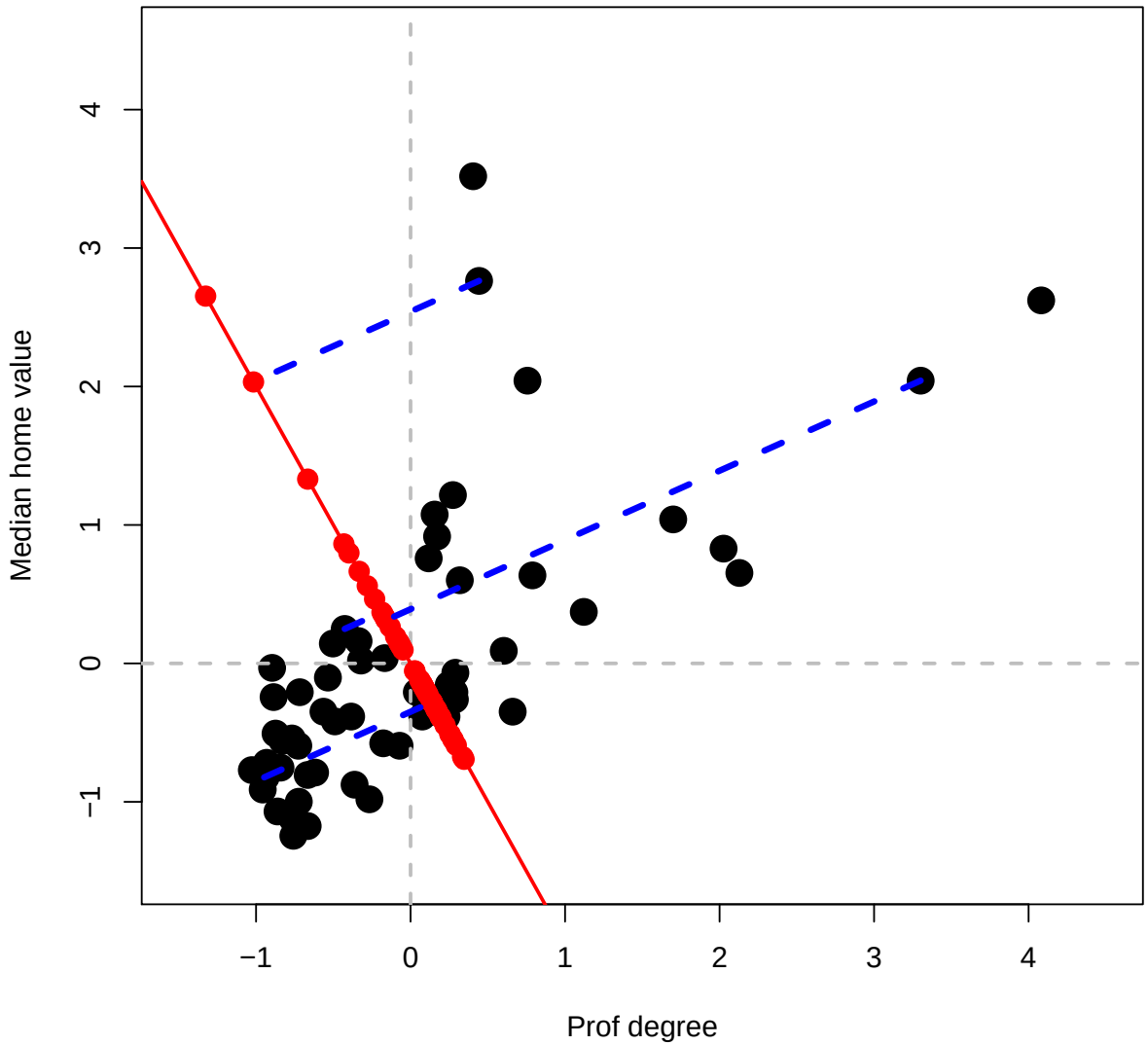
# Principal components analysis



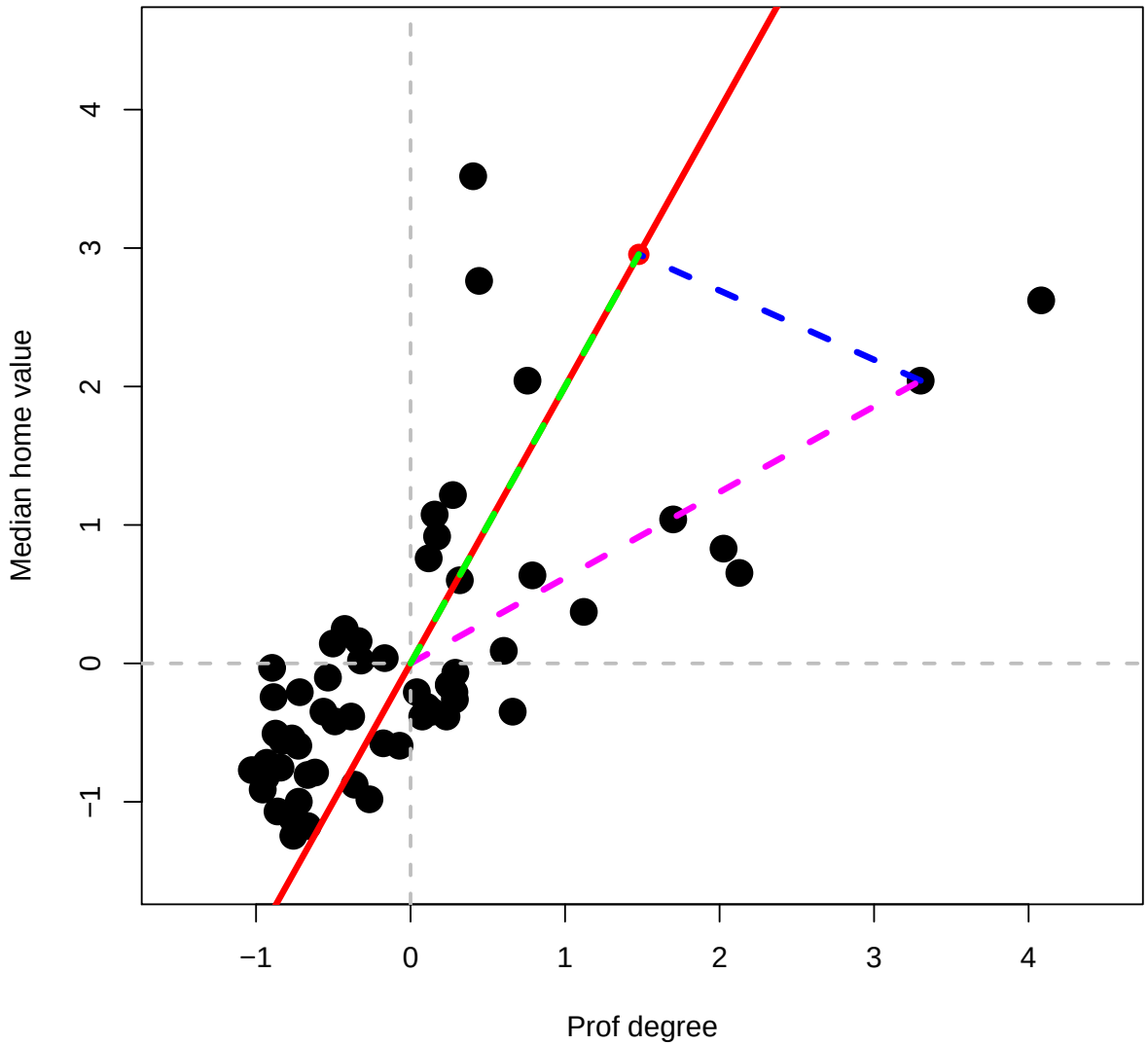
# Principal components analysis



# Principal components analysis



# Principal components analysis



# Principal components analysis

$$\min_{\mathbf{L}} \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{X}_i - P(\mathbf{X}_i, \mathbf{L})\|^2 =$$

$$\min_{\mathbf{L}} \frac{1}{n}\sum_{i=1}^{n} \left[ \|\mathbf{X}_i\|^2 - \|P(\mathbf{X}_i, \mathbf{L})\|^2 \right]$$

We need to find

$$\max_{\mathbf{L}} \frac{1}{n}\sum_{i=1}^{n} \|P(\mathbf{X}_i, \mathbf{L})\|^2$$

# Principal components analysis

To fix ideas take $k = 1$. In this case $\mathbf{L} = < \mathbf{a} >$ for some vector $\mathbf{a} \in \mathbb{R}^p$ with $\|\mathbf{a}\| = 1$. Thus

$$P(\mathbf{X}_i, \mathbf{L}) = \mathbf{a}\,(\mathbf{a}'\mathbf{X}_i)$$

hence

$$\sum_{i=1}^{n} \| P(\mathbf{X}_i, \mathbf{L}) \|^2 = \sum_{i=1}^{n} \| \mathbf{a}\,(\mathbf{a}'\mathbf{X}_i) \|^2$$

# Principal components analysis

$$\max_{\mathbf{L}} \frac{1}{n}\sum_{i=1}^{n} \| P(\mathbf{X}_i, \mathbf{L}) \|^2 =$$

$$\max_{\|\mathbf{a}\|=1} \frac{1}{n}\sum_{i=1}^{n} \| \mathbf{a}\,(\mathbf{a}'\mathbf{X}_i) \|^2 =$$

$$\max_{\|\mathbf{a}\|=1} \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i'\mathbf{a}\,\mathbf{a}'\mathbf{a}\,\mathbf{a}'\mathbf{X}_i =$$

$$\max_{\|\mathbf{a}\|=1} \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i'\mathbf{a}\,\mathbf{a}'\mathbf{X}_i$$

# Principal components analysis

$$\max_{\mathbf{L}} \frac{1}{n}\sum_{i=1}^{n} \| P(\mathbf{X}_i, \mathbf{L}) \|^2 =$$

$$\max_{\|\mathbf{a}\|=1} \frac{1}{n}\sum_{i=1}^{n} \mathbf{a}'\mathbf{X}_i\mathbf{X}_i'\mathbf{a} =$$

$$\max_{\|\mathbf{a}\|=1} \mathbf{a}' \left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i\mathbf{X}_i'\right] \mathbf{a} =$$

$$\max_{\|\mathbf{a}\|=1} \mathbf{a}' Q_n \mathbf{a}$$

# Principal components analysis

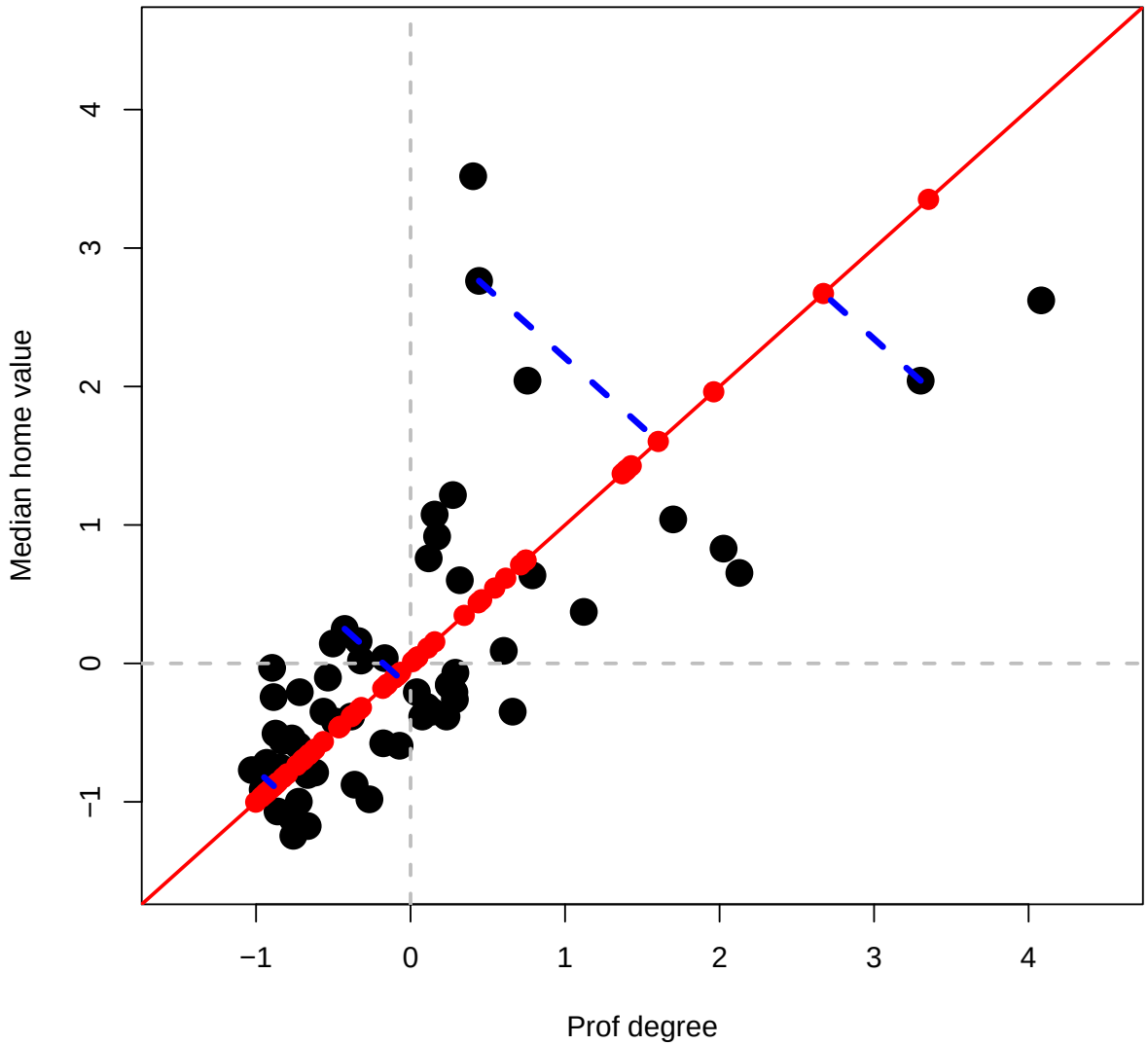
- the “best” subspace **L** of dimension **1** is generated by the eigenvector $\mathbf{a}_0$ of $Q_n$ associated with its largest eigenvalue $\lambda_1$.
- The “scores” are the coefficients of the projections

$$y_i = \mathbf{X}_i' \mathbf{a}_0 , \qquad i = 1, \ldots, n$$

- The projections are

$$\mathbf{v}_i = \mathbf{a}_0 y_i = \mathbf{a}_0 \left(\mathbf{X}_i' \mathbf{a}_0\right) , \qquad i = 1, \ldots, n$$

# Principal components analysis



# Principal components analysis

There is nothing special about **L** being of dimension $k = 1$

In general, let $\mathbf{B} \in \mathbb{R}^{p \times k}$ be an orthonormal ($\mathbf{B}'\mathbf{B} = \mathbf{I}_k$) basis for **L**, then

$$P(\mathbf{X}_i, \mathbf{L}) = \mathbf{B}\,(\mathbf{B}'\mathbf{X}_i)$$

and

$$\frac{1}{n}\sum_{i=1}^{n} \|P(\mathbf{X}_i, \mathbf{L})\|^2 = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{B}\,(\mathbf{B}'\mathbf{X}_i)\|^2 =$$

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i'\mathbf{B}\,\mathbf{B}'\mathbf{X}_i$$

# Principal components analysis

Furthermore

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{B}\,\mathbf{B}'\mathbf{X}_i = \frac{1}{n}\sum_{i=1}^{n}\text{trace}\,(\mathbf{B}'\mathbf{X}_i\mathbf{X}_i'\mathbf{B}) =$$

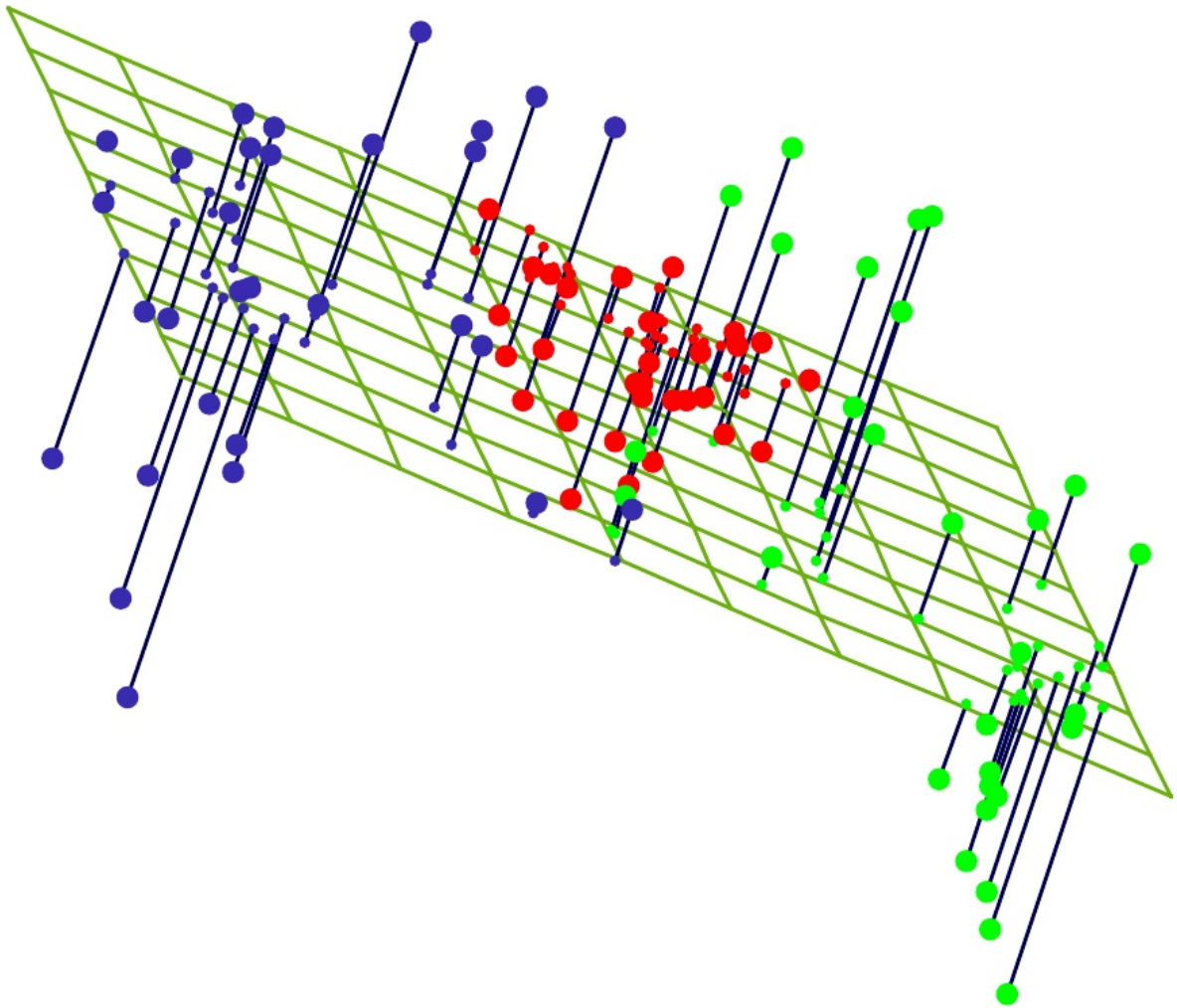
$$\text{trace}\,(\mathbf{B}'Q_n\,\mathbf{B}) \leq \sum_{j=1}^{k}\lambda_j$$

and since

$$\text{trace}\,(\mathbf{U}_k'Q_n\,\mathbf{U}_k) = \sum_{j=1}^{k}\lambda_j$$

the “best” subspace is generated by the $k$ “largest” eigenvectors of $Q_n$

# Principal components analysis



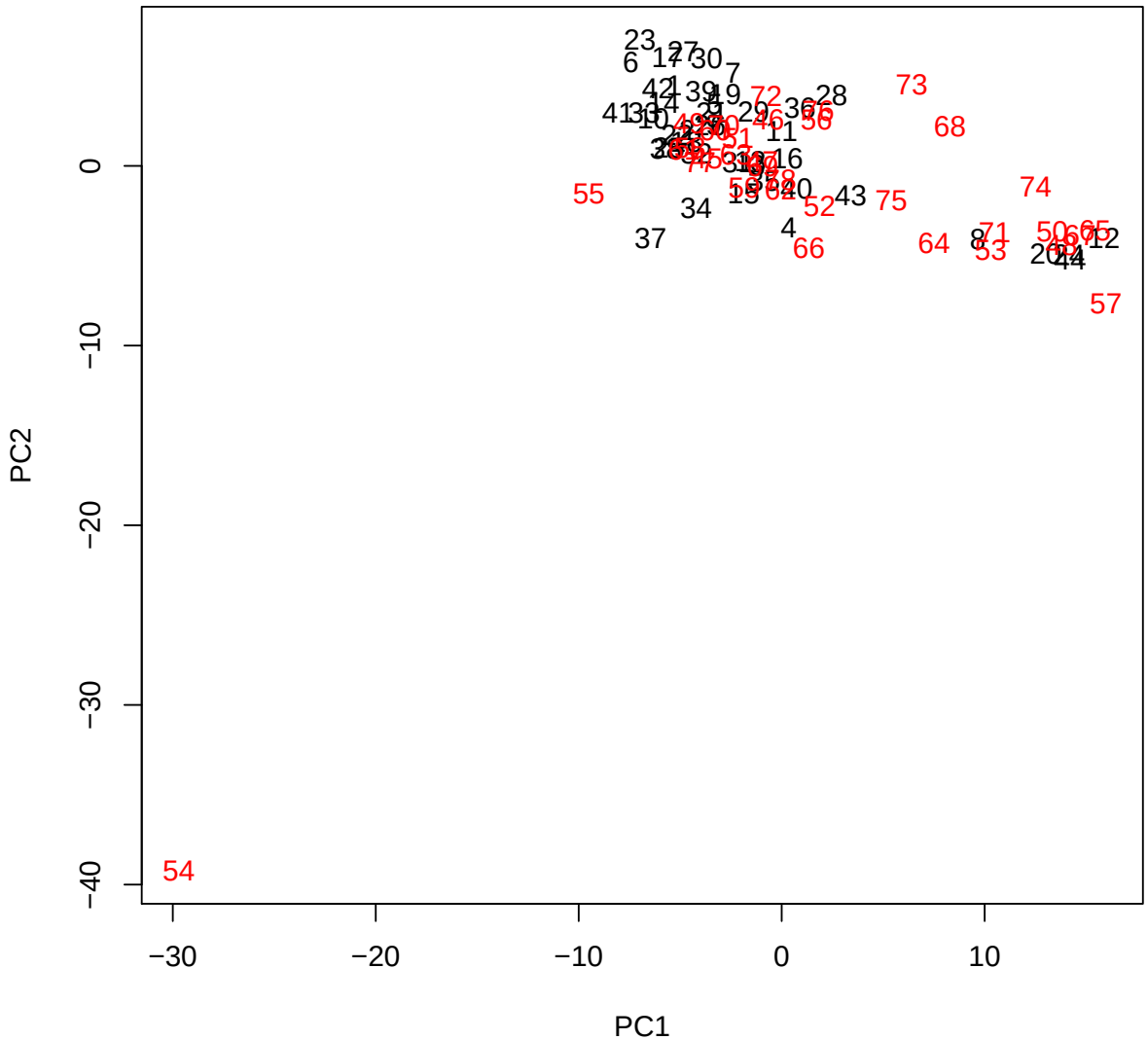
# Breast Cancer Gene Expression

- Data on 78 patients
- 4751 measurements taken on each patient (4751 gene expression levels)
- For each patient we know whether the tumor metastasized or not
- Can we separate these groups using their gene expression measurements?

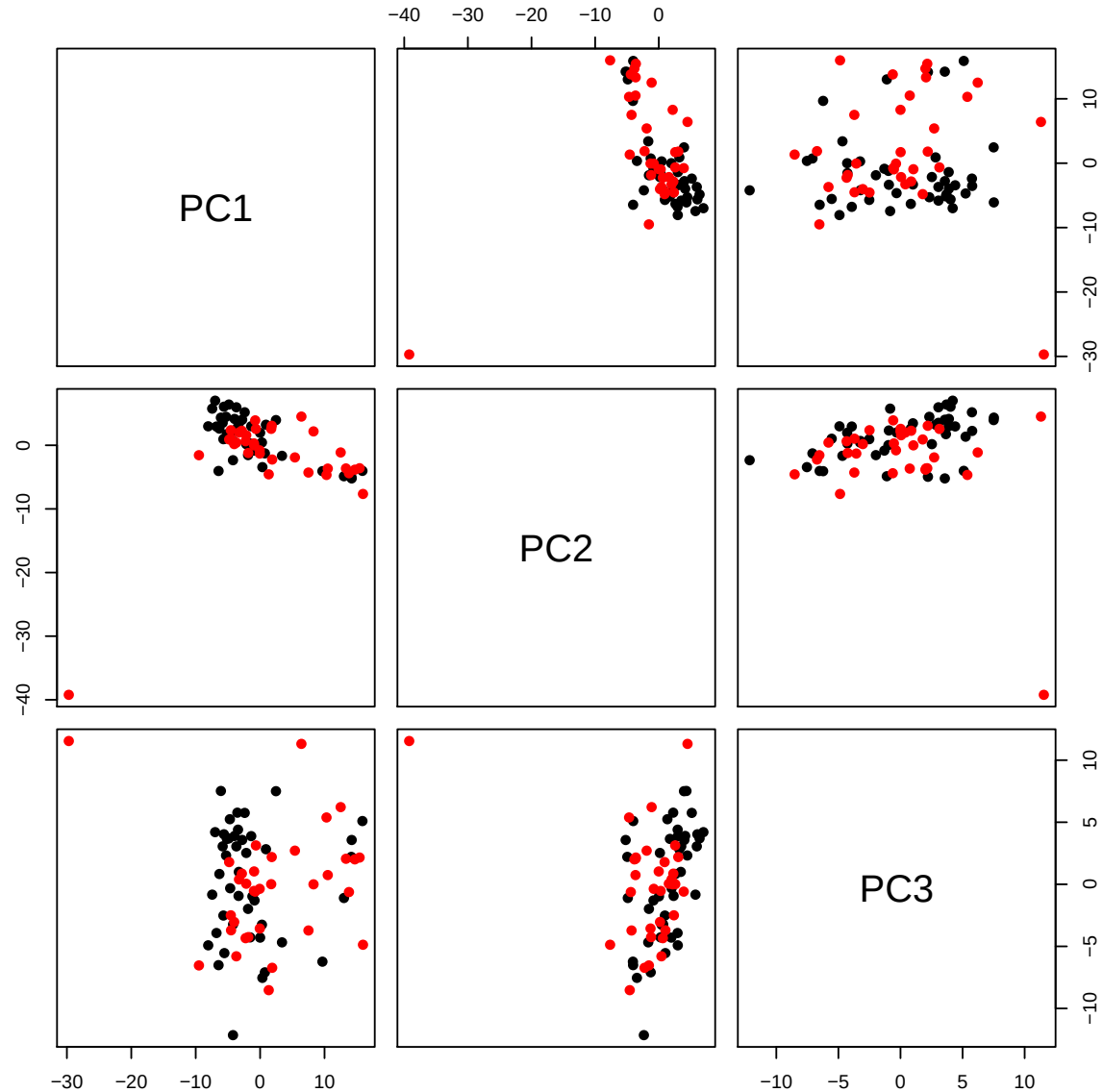
# Breast Cancer Gene Expression

- Can’t plot a 4751-dimensional space
- 78 observations are not too many to reveal structures on such a high-dimensional space
- We try to use fewer measurements, but in such a way that they remain close to the full 4751- dimensional observations

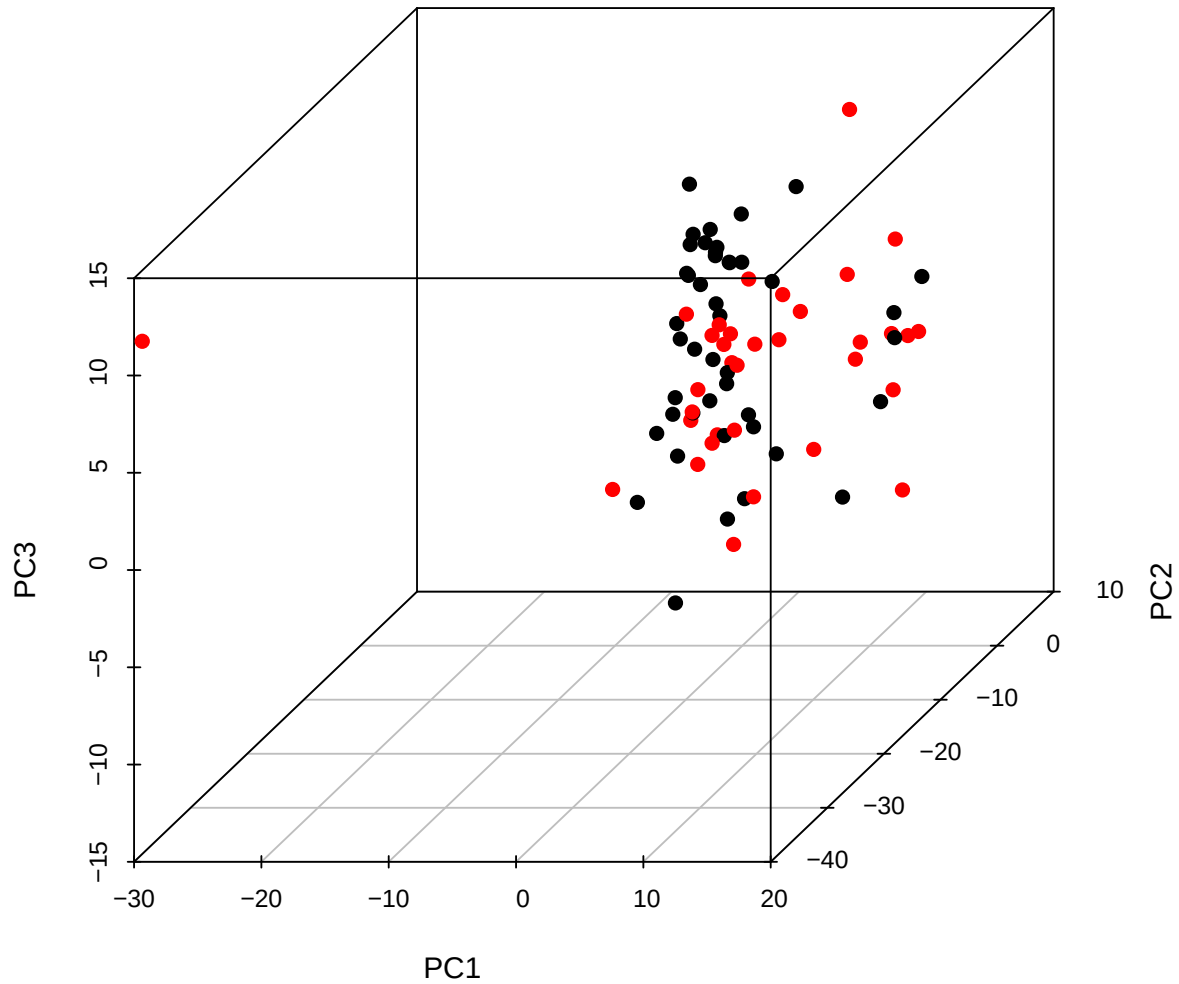
# Gene Expression - Best 2D



# Gene Expression - Best 3D



# Gene Expression - Best 3D



# Principal components analysis

How good is the **best k**-dimensional approximation?

$$\min_{\mathbf{L}} \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{X}_i - P(\mathbf{X}_i, \mathbf{L})\|^2 =$$

$$\min_{\mathbf{L}} \frac{1}{n}\sum_{i=1}^{n} \left[ \|\mathbf{X}_i\|^2 - \|P(\mathbf{X}_i, \mathbf{L})\|^2 \right] =$$

$$\frac{1}{n}\sum_{i=1}^{n} \|\mathbf{X}_i\|^2 - \max_{\mathbf{L}} \frac{1}{n}\sum_{i=1}^{n} \|P(\mathbf{X}_i, \mathbf{L})\|^2 =$$

$$\text{trace}(Q_n) - \sum_{j=1}^{k} \lambda_j = \sum_{j=\mathbf{k}+1}^{p} \lambda_j$$

# Principal components analysis

- Important question: how many PC's should we use?
- The fewer we use the more we will have reduced the dimension
- The more we use the better the approximation to the full data
- There is a trade-off between goodness of the fit and number of principal components

# Principal components analysis

- We know that for the best subspace **L** of dimension **k** we have

$$\min_{\mathbf{L}} \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{X}_i - P(\mathbf{X}_i, \mathbf{L})\|^2 =$$

$$\text{loss}(\mathbf{k}) = \sum_{j=\mathbf{k}+1}^{p} \lambda_j$$

# Principal components analysis

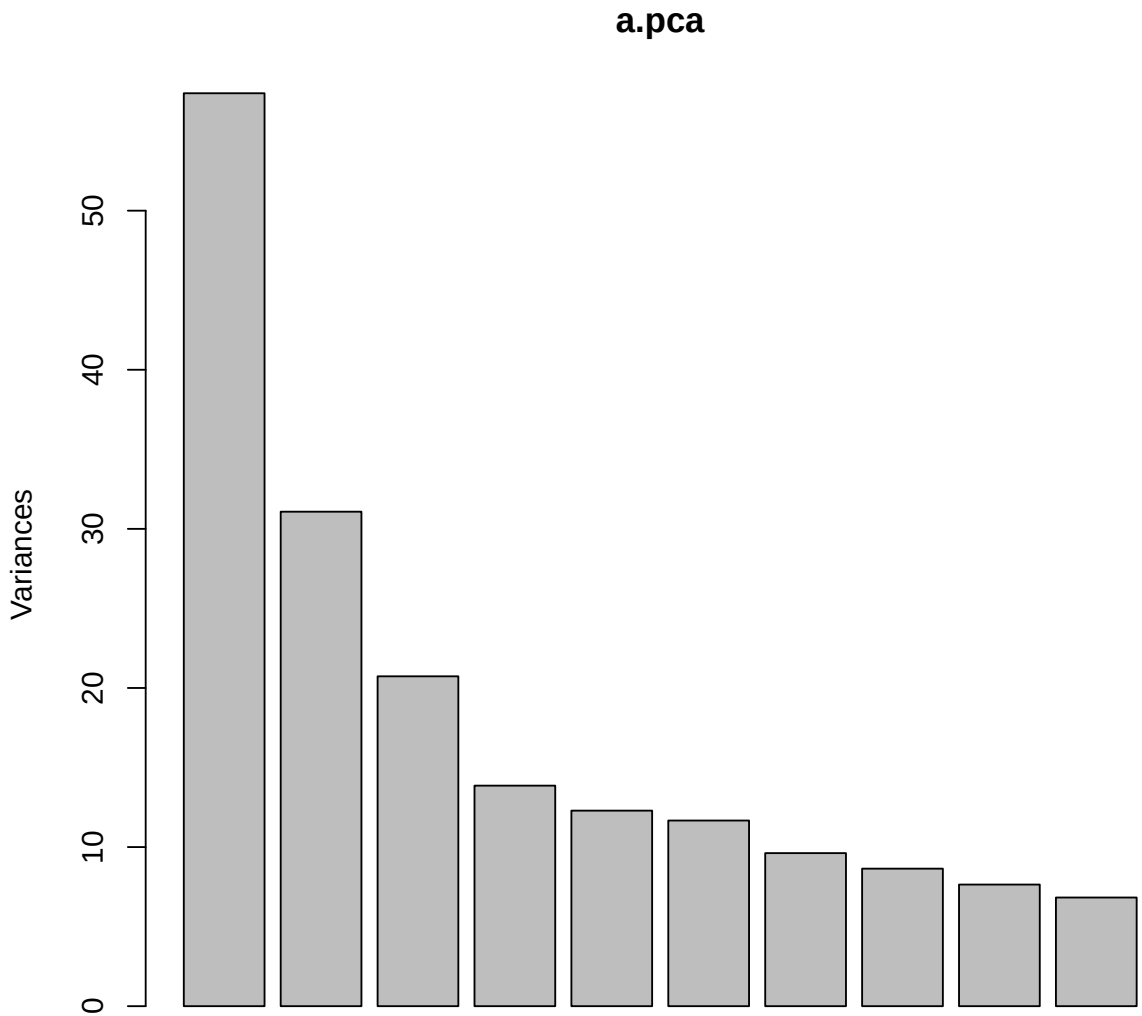
- hence, the gain obtained by using $\mathbf{k+1}$ PC's instead of using only $\mathbf{k}$ is

$$\text{loss}\,(k) - \text{loss}\,(k+1) \;=\; \lambda_{\mathbf{k}+1}$$

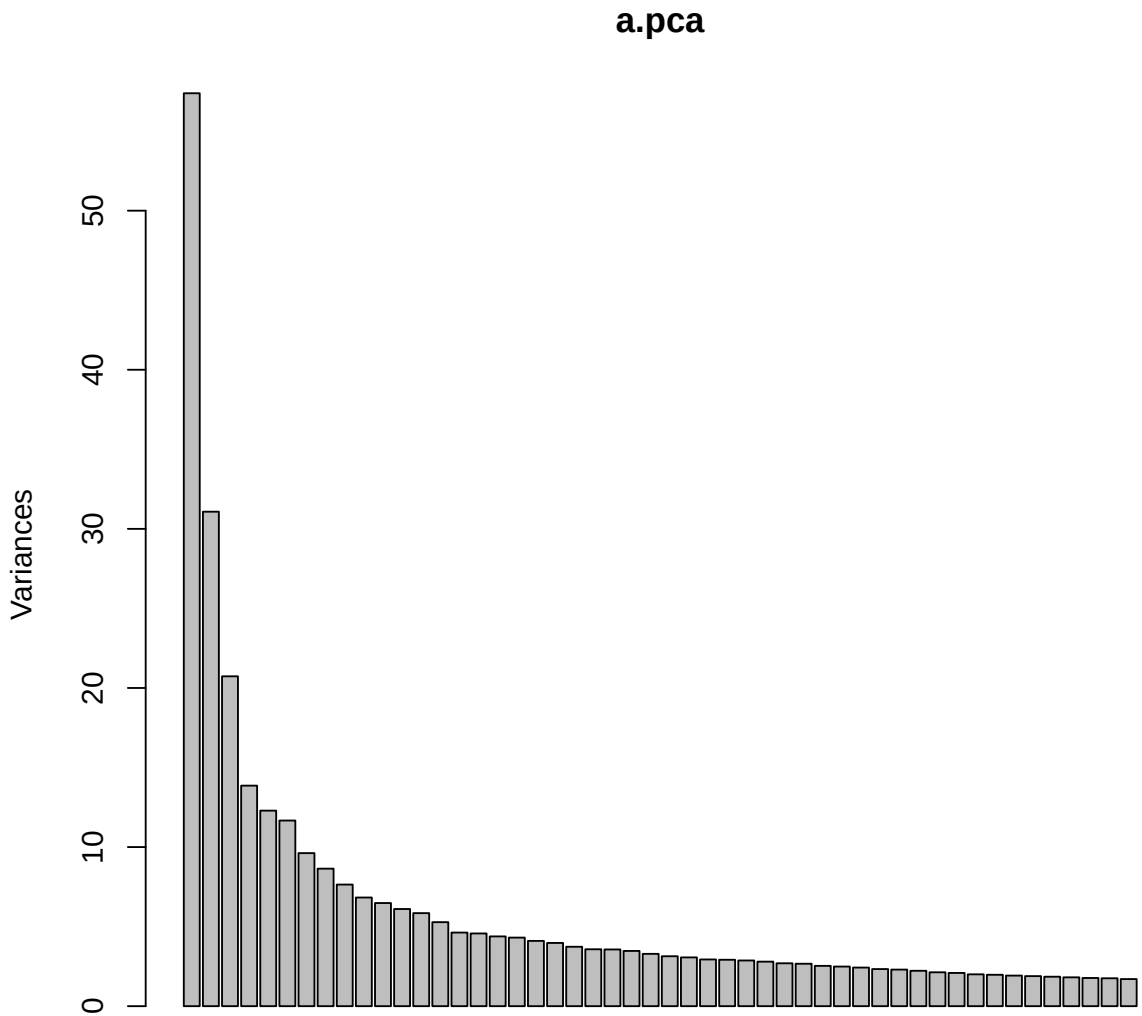
  for $k = 0, 1, 2, \ldots, p$

- the gain is decreasing...

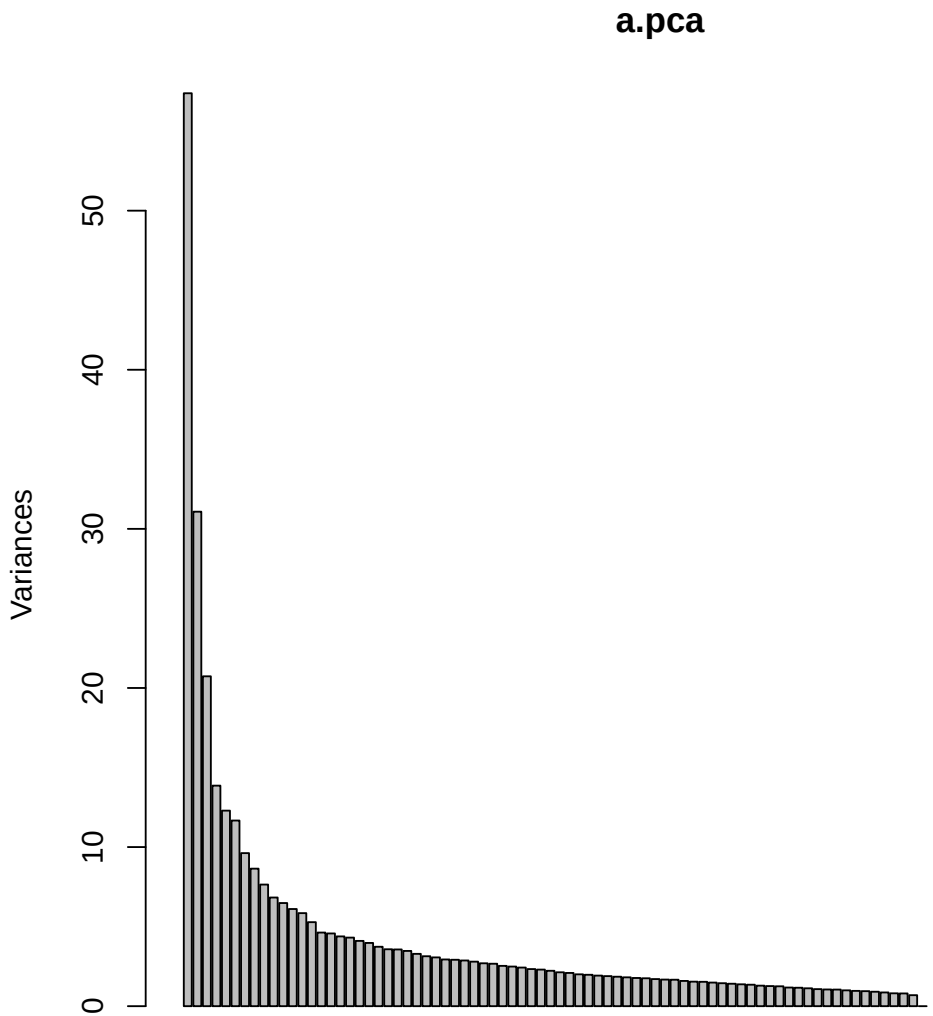
# Breast cancer example - scree plot - first 10



# Breast cancer example - scree plot - first 50



# Breast cancer example - scree plot - first 100



# PCA - summary

- Principal components (PCs) provide “best” lower-dimensional approximations
- The more PCs we use the better the approximation
- The gain in quality of the approximation decreases as we add more PCs
- These gains are the eigenvalues of the covariance matrix

# Principal components analysis

- So, what about Statistics in all this?
- If $\mathbf{X}$ is a random vector with $E[\mathbf{X}] = \mathbf{0}$ and $\text{cov}(\mathbf{X}) = \text{cov}(\mathbf{X}, \mathbf{X}) = \mathbf{\Sigma}$

$$\min_{\mathbf{L}} E \| \mathbf{X} - P(\mathbf{X}, \mathbf{L}) \|^2 =$$

$$\min_{\mathbf{L}} E \left[ \| \mathbf{X} \|^2 - \| P(\mathbf{X}, \mathbf{L}) \|^2 \right]$$

Hence, we need to find

$$\max_{\mathbf{L}} E \| P(\mathbf{X}, \mathbf{L}) \|^2$$

# Principal components analysis

For the case $k = 1$ We have

$$P(\mathbf{X}, \mathbf{L}) = \mathbf{a}\,(\mathbf{a}'\mathbf{X})$$

for some vector $\mathbf{a} \in \mathbb{R}^p$ with $\|\mathbf{a}\| = 1$ Hence

$$E\,\|\,P(\mathbf{X}, \mathbf{L})\,\|^2 = E\,\|\mathbf{a}\,(\mathbf{a}'\mathbf{X})\,\|^2$$

# Principal components analysis

$$
\begin{aligned}
\max_{\mathbf{L}} \; E \, \| P(\mathbf{X}, \mathbf{L}) \|^2 &= \max_{\|\mathbf{a}\|=1} \; E \, \| \mathbf{a} \, (\mathbf{a}' \mathbf{X}) \|^2 \\
&= \max_{\|\mathbf{a}\|=1} \; E \, [\mathbf{X}' \mathbf{a} \, \mathbf{a}' \mathbf{X}] \\
&= \max_{\|\mathbf{a}\|=1} \; E \, [\mathbf{a}' \mathbf{X} \, \mathbf{X}' \mathbf{a}] \\
&= \max_{\|\mathbf{a}\|=1} \; \mathbf{a}' E \, [\mathbf{X} \, \mathbf{X}'] \, \mathbf{a} \\
&= \max_{\|\mathbf{a}\|=1} \; \mathbf{a}' \boldsymbol{\Sigma} \, \mathbf{a}
\end{aligned}
$$

# Principal components analysis

Finding the "best" (in the $L_2$ sense) approximating 1-dimensional subspace is equivalent to finding the direction $\mathbf{a}_0$ such that

$$\mathbf{a}_0{}' \mathbf{\Sigma} \mathbf{a}_0 = \max_{\|\mathbf{a}\|=1} \mathbf{a}' \mathbf{\Sigma} \mathbf{a}$$

Also note that, for any $\mathbf{a} \in \mathbb{R}^p$

$$\mathbf{a}' \mathbf{\Sigma} \mathbf{a} = \text{var}(\mathbf{a}' \mathbf{X})$$

# Principal components analysis

Thus, the problem is equivalent to

- finding the direction $\mathbf{a}_0$ along which $\mathbf{X}$ has largest variability;
- finding the coefficients $\mathbf{a}_0$ that produce the linear combination of components of $\mathbf{X}$ with maximum variance.

# Principal components analysis

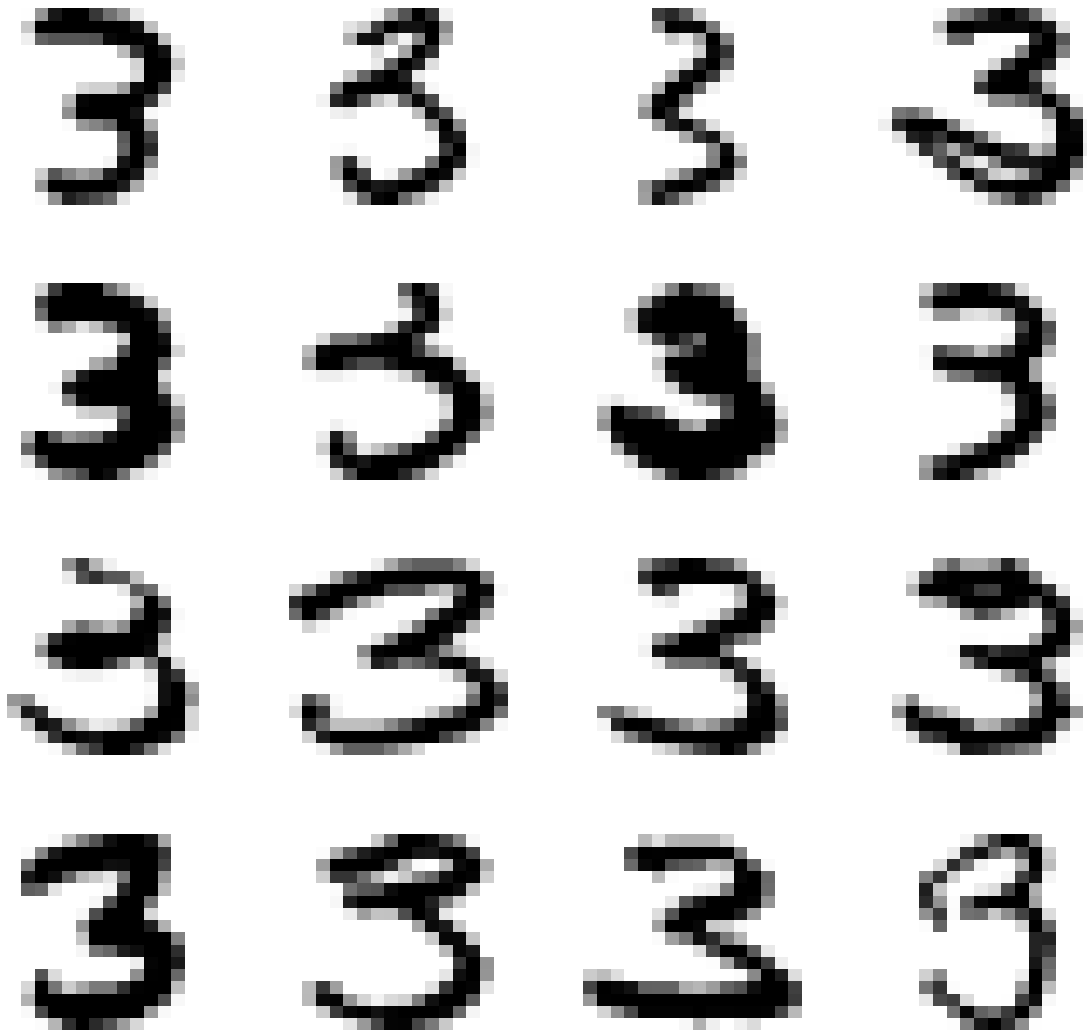
- The same approach we used for a sample applies to the population case. To find the 2nd PC we need to find

$$\max_{\|\mathbf{a}\|=1, \text{cov}(\mathbf{a}_0'\mathbf{X}, \mathbf{a}'\mathbf{X})=0} \text{Var}(\mathbf{a}'\mathbf{X}) = \mathbf{a}_1' \boldsymbol{\Sigma} \mathbf{a}_1$$

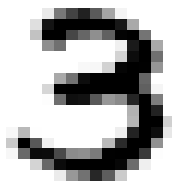
  where $\mathbf{a}_1$ is the eigenvector of $\boldsymbol{\Sigma}$ associated with the second largest eigenvalue.

- etc. etc.

# Interesting PCA example



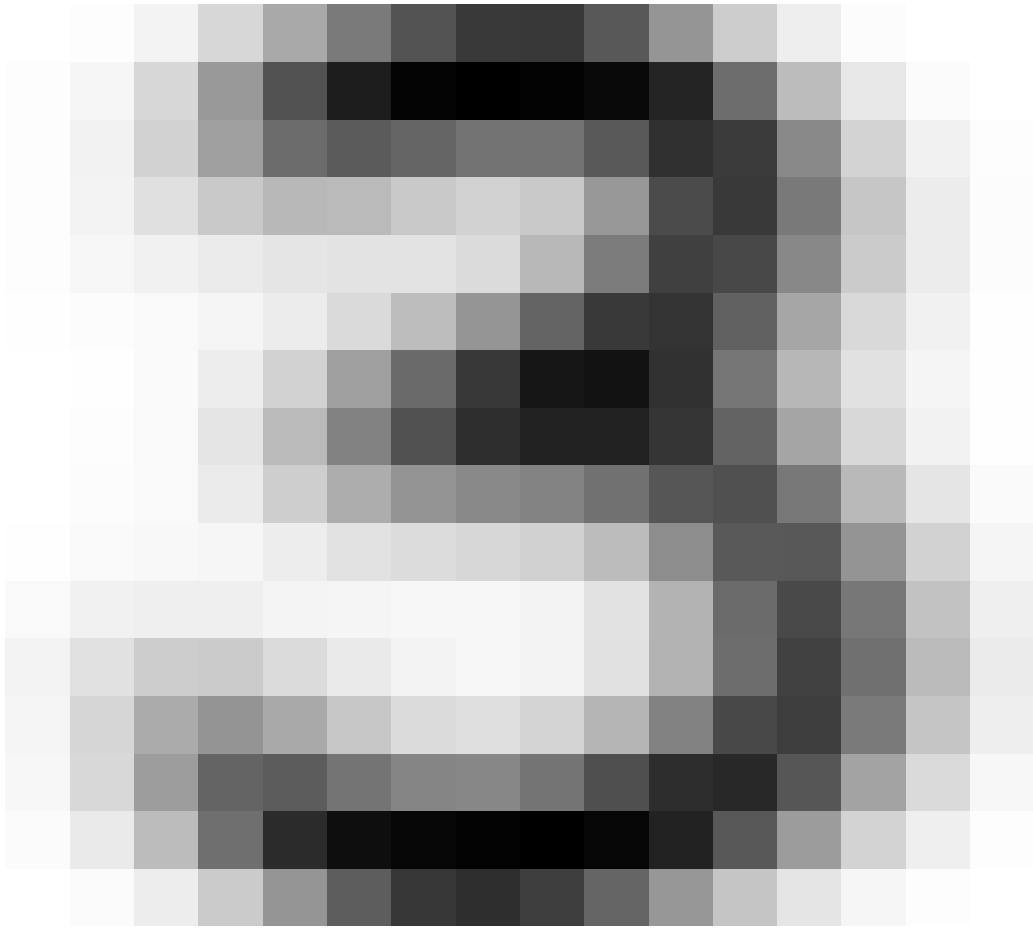
# Interesting PCA example



$$\longrightarrow \quad \mathbf{x}_i = (-1, -1, -0.989, -0.018, \cdots, -0.967, -1)$$

$$\mathbf{x}_i \in \mathbb{R}^{256} \qquad i = 1, \ldots, 658$$

# Interesting PCA example



Picture of $\bar{\mathbf{X}}_n$ – the average "3" in the data

# Interesting PCA example

- The basic idea is to use PCA to obtain a (much) lower dimensional approximation to these
- Let $L \subset \mathbb{R}^{256}$ be a subspace of dimension $k$, generated by the vectors $\mathbf{b}_1, \ldots, \mathbf{b}_k \in \mathbb{R}^{256}$, i.e.

$$L = < \mathbf{b}_1, \ldots, \mathbf{b}_k >$$

# Interesting PCA example

- If

$$\mathbf{B} = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ \mathbf{b}_1 & \mathbf{b}_2 & \vdots & \mathbf{b}_k \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \in \mathbb{R}^{256 \times k}$$

then the orthogonal projection of $\mathbf{X}_i \in \mathbb{R}^{256}$ onto $L$ is

$$P(\mathbf{X}_i, L) = \mathbf{B}\ (\mathbf{B}'\mathbf{X}_i)$$

# Interesting PCA example

- Let $S_n \in \mathbb{R}^{256\times 256}$ be the sample covariance matrix of $\mathbf{X}_1, \ldots, \mathbf{X}_n$, and let $\mathbf{u}_j$, $j = 1, \ldots, 256$, be its eigenvectors, associated with decreasing eigenvalues:

$$S_n \, \mathbf{u}_j = \lambda_j \, \mathbf{u}_j \, , \qquad j = 1, \ldots, 256 \, ,$$

with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{256} \geq 0$.

# Interesting PCA example

- Then, we have seen that the choice

$$\mathbf{b}_j = \mathbf{u}_j\,, \qquad j = 1, \ldots, k$$

spans the subspace $L$ that minimizes

$$\sum_{i=1}^{n} \| \mathbf{X}_i - P(\mathbf{X}_i, L) \|^2$$

over all possible subspaces $L \subset \mathbb{R}^{256}$ of dimension $k$.

# Interesting PCA example

- We first pick $k = 2$. Let

$$\mathbf{U}_2 = \begin{bmatrix} \vdots & \vdots \\ \mathbf{b}_1 & \mathbf{b}_2 \\ \vdots & \vdots \end{bmatrix} \in \mathbb{R}^{256\times 2}$$

and let

$$\mathbf{V}_i = \mathbf{U}_2'\mathbf{X}_i \in \mathbb{R}^2, \quad i = 1, \ldots, n$$

be the first 2 principal components. Note that

$$P(\mathbf{X}_i, L) = \mathbf{U}_2\mathbf{V}_i.$$

# Interesting PCA example

- Since it may be hard to interpret the 256 coefficients in each of the 2 principal components, we will instead find which images $\mathbf{X}_i$ are being approximated as we move along a grid on the 2-dimensional $\mathbf{V}_i$-space.
- We illustrate this idea in the following slides. Here $\mathbf{X}_i \in \mathbb{R}^2$ and we use a single principal component ($k = 1$), represented by the **grey** line.

# Interesting PCA example

- We select a grid of points (indicated in **blue**) along this 1-dimensional subspace of $\mathbb{R}^2$, and find which points $\mathbf{X}_i$ are being approximated by them. These are the points in the sample closest to

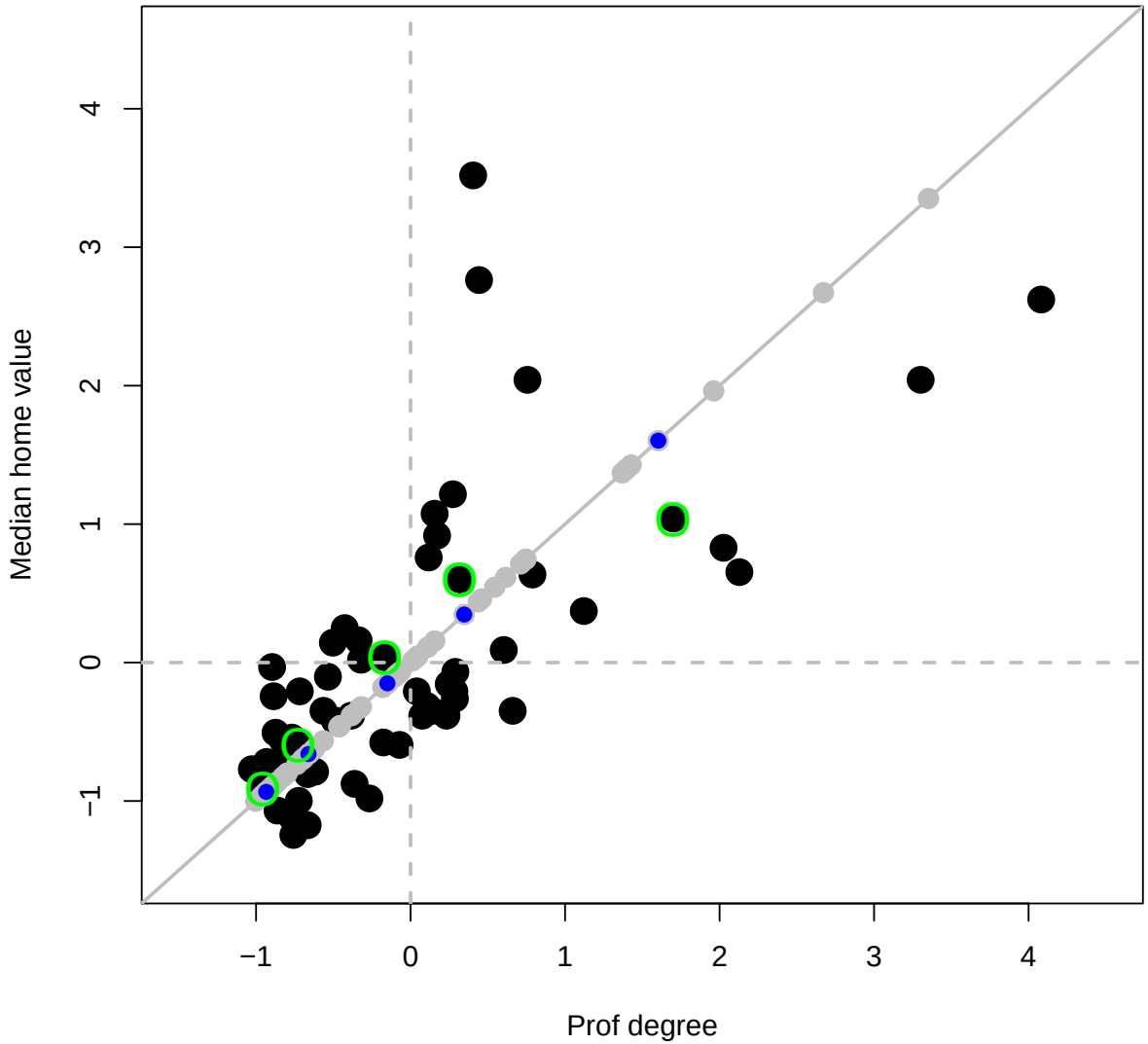
$$P(\mathbf{X}_i, L) = \mathbf{U}_1 \mathbf{V}_i$$

for each $\mathbf{V}_i \in \mathbb{R}$ in the grid.

# Interesting PCA example

- The grid of points along the 1-dimensional principal component subspace are represented with **blue** dots. The corresponding closest points in the original sample are indicated by **green** circles.
- In our “images” application, we will see how these **green** points change as we move from one **blue** point to the next, along the **grey** line.

# Interesting PCA example

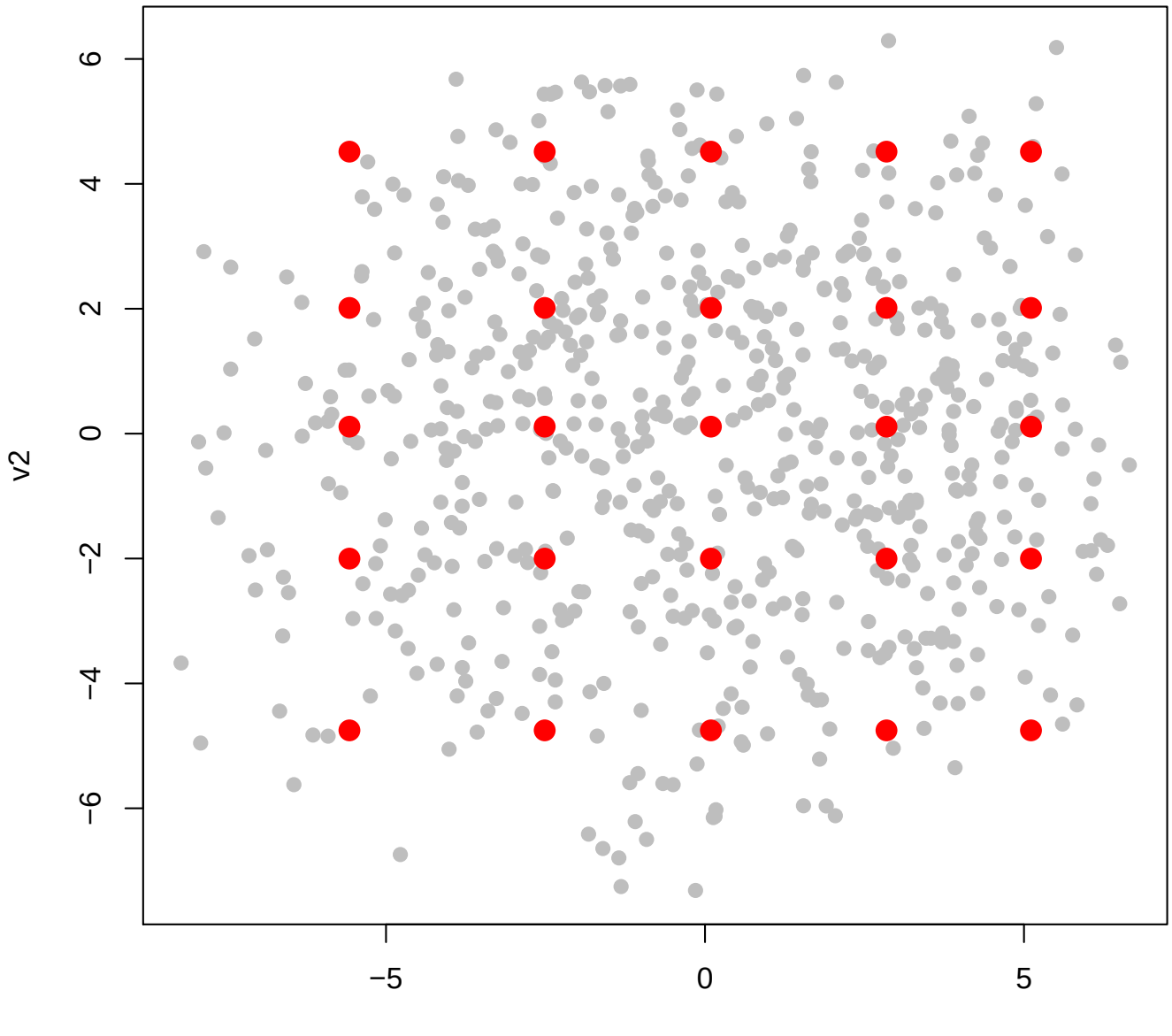



A simpler data set

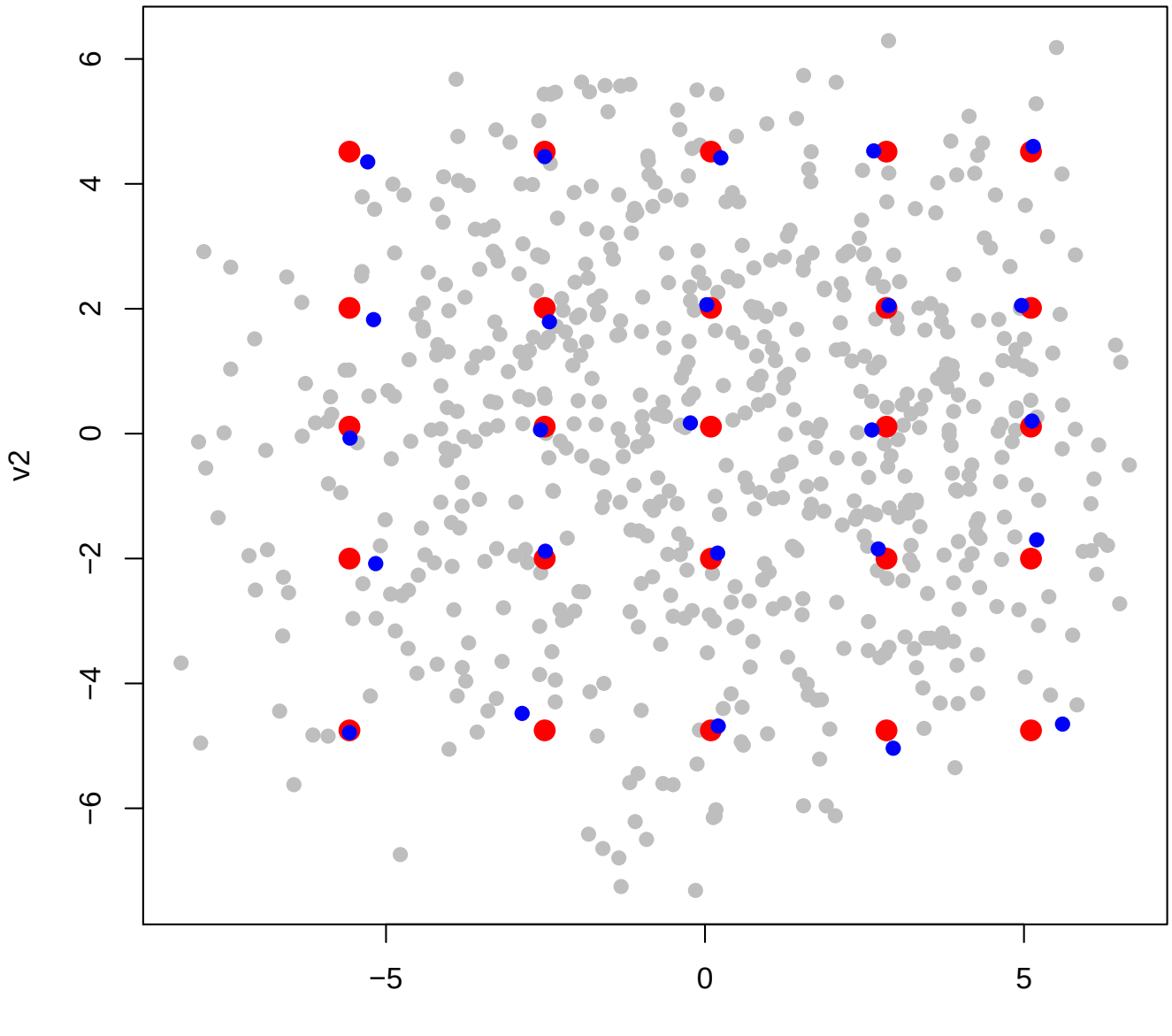
# Interesting PCA example

- Now, we calculate the first 2 principal components of our 256-dimensional images. These are vectors $\mathbf{V}_i \in \mathbb{R}^2$, $i = 1, \ldots, n$, pictured as **grey** points in the next 2 slides.
- Next, we select 25 of these vectors by overimposing a $5 \times 5$ grid over the $\mathbf{V}_i$'s, and selecting the closest $\mathbf{V}_j$ to each of these **red** points. These $\mathbf{V}_j$'s are indicated by **blue** points.
- The corresponding figures are in the next 2 slides.

# Interesting PCA example



# Interesting PCA example

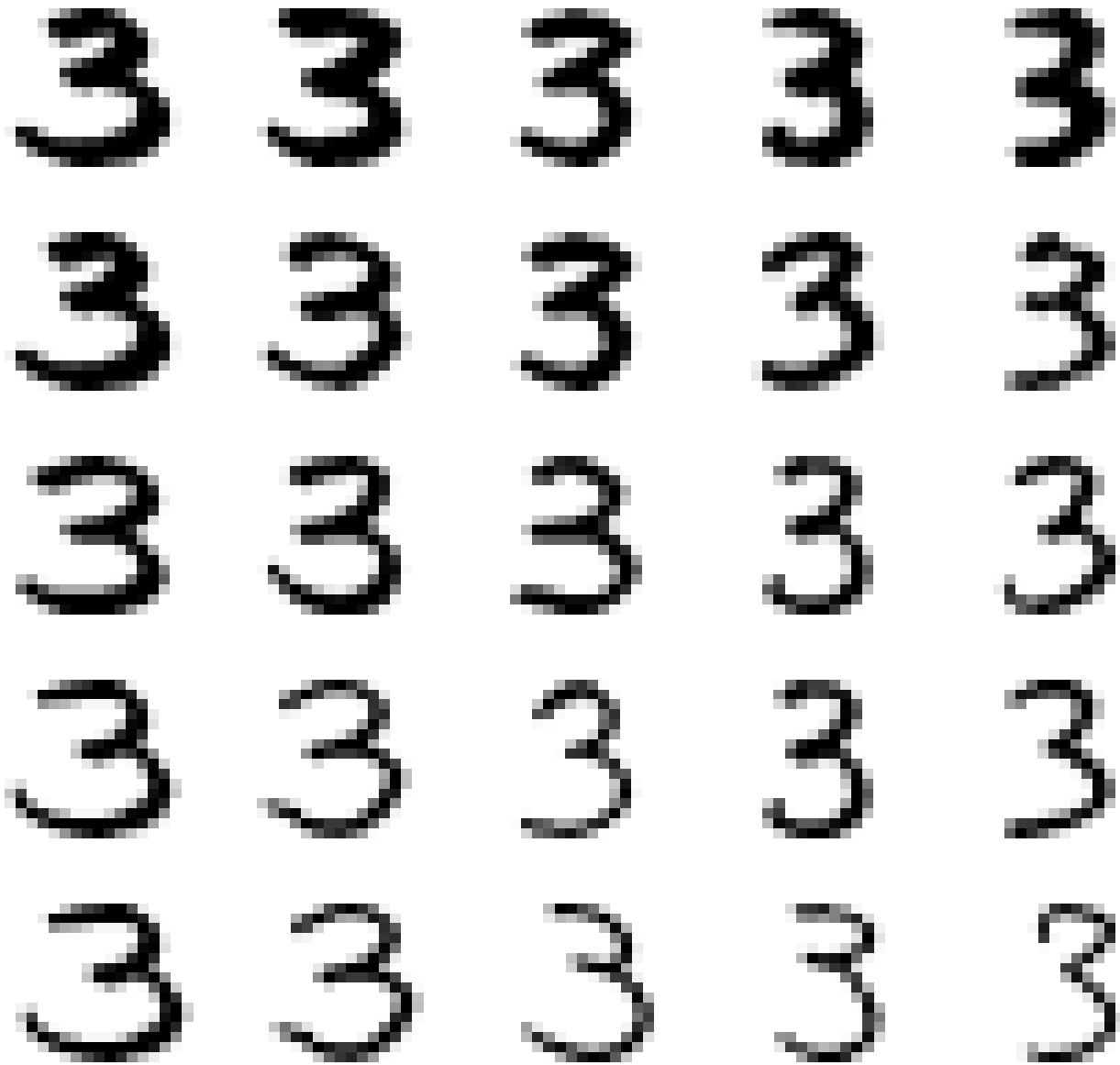


# Interesting PCA example

- In the next slide we display the images closest to the 256-dimensional vector associated with each bivariate **blue** point in the previous plot.

- The images are arranged in a grid, and can be “read” left-right and top-bottom, corresponding to the same trajectories across the selected **blue** bivariate principal components in the previous plot.

# 2 Principal Components



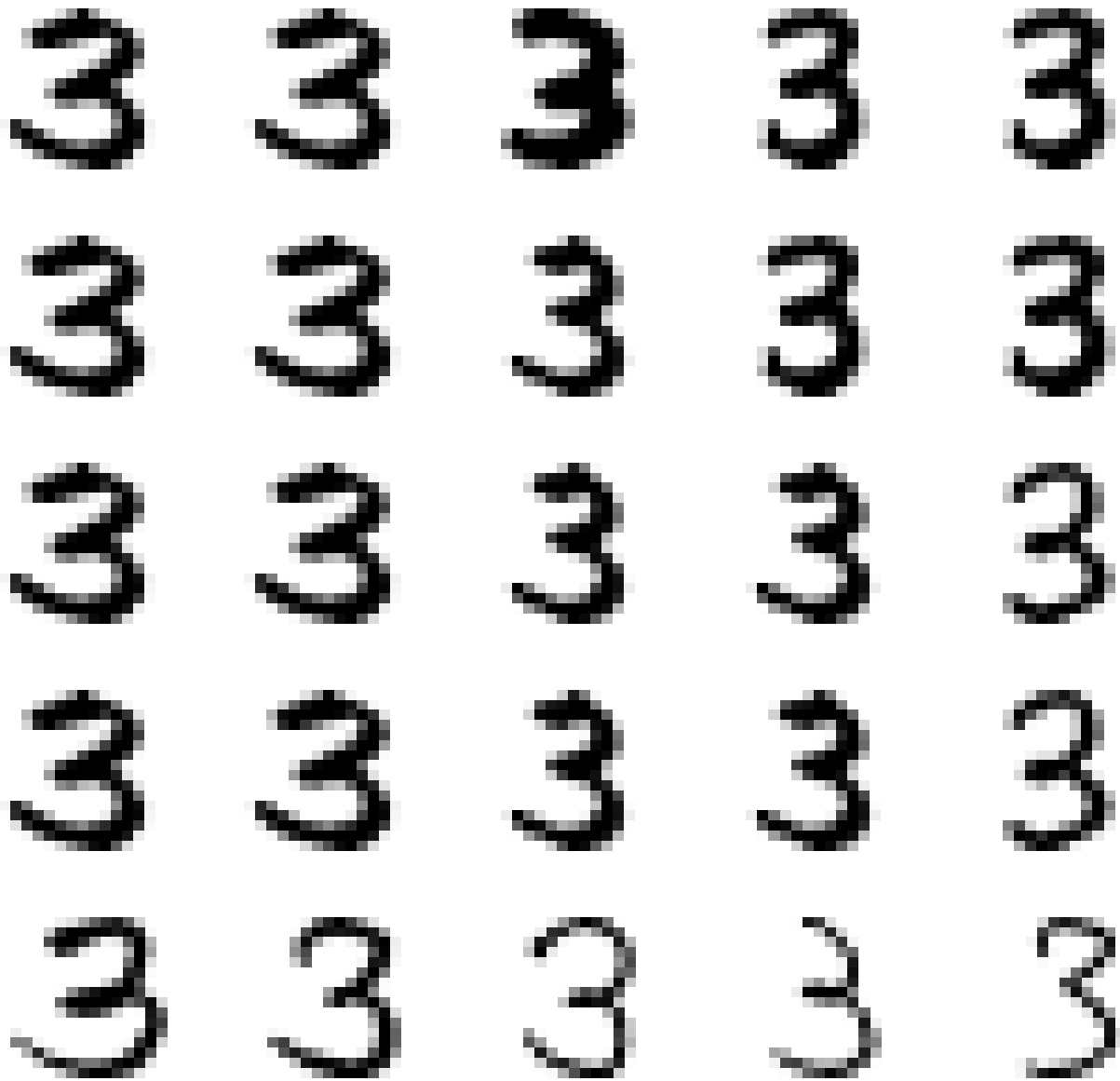
# Interesting PCA example

- Now we repeat the experiment but with 3 principal components.
- We have $\mathbf{V}_i \in \mathbb{R}^3$, and select a 3-dimensional grid of dimension $75 = 5 \times 5 \times 3$; that is: 5 values for each of the first and second PCs, and 3 for the third one.
- As before, we select the $\mathbf{V}_j$ closest to each of the 75 points in the 3-dimensional grid.

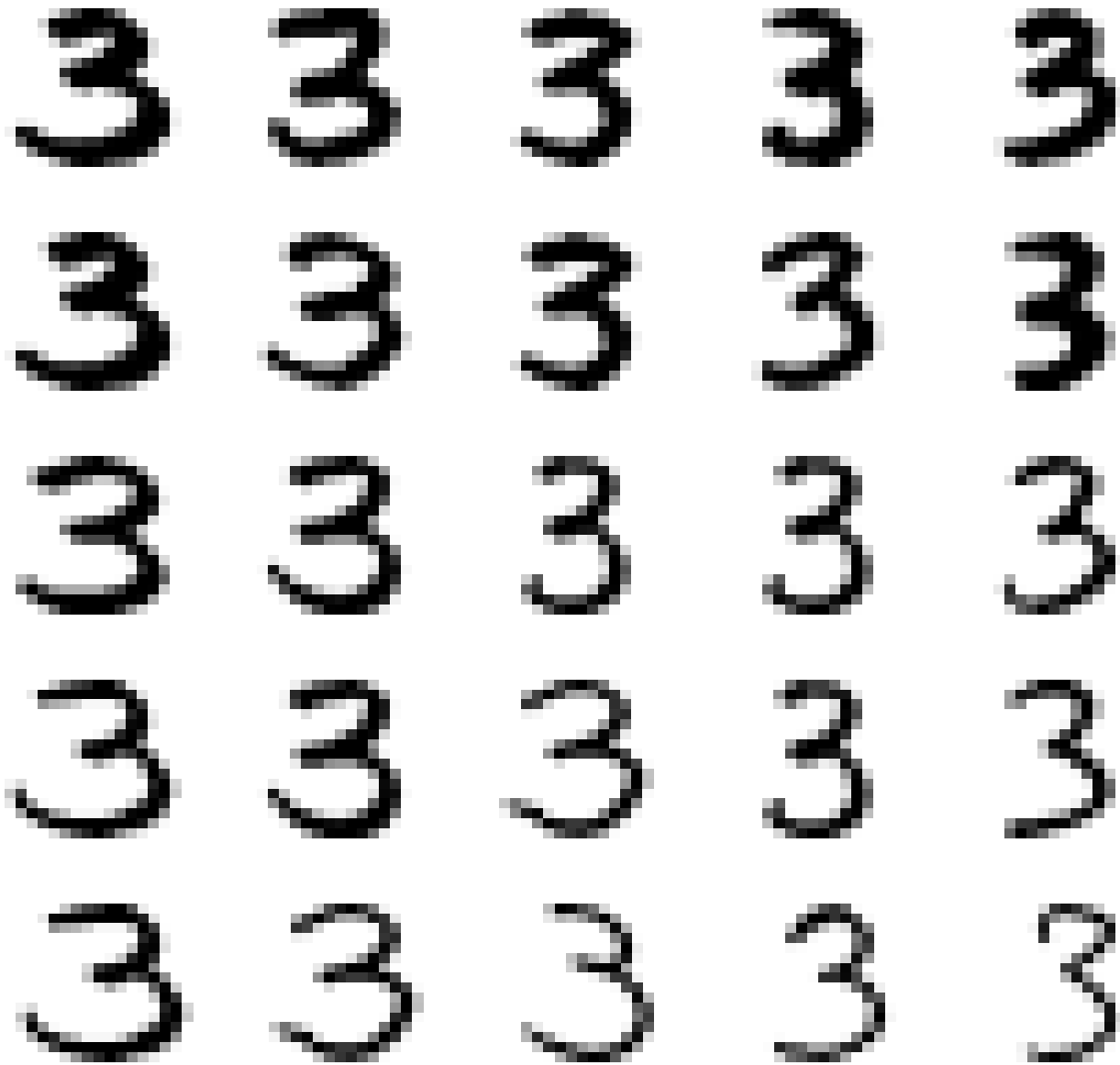
# Interesting PCA example

- We now have 75 **blue** points in $\mathbb{R}^{256}$, and find the images closest to them.
- We represent this 3-dimensional array of images as 3 consecutive "slices" of $5 \times 5$. These are displayed in the next 3 slides.
- Further details can be found in the (somewhat commented) `R` code file

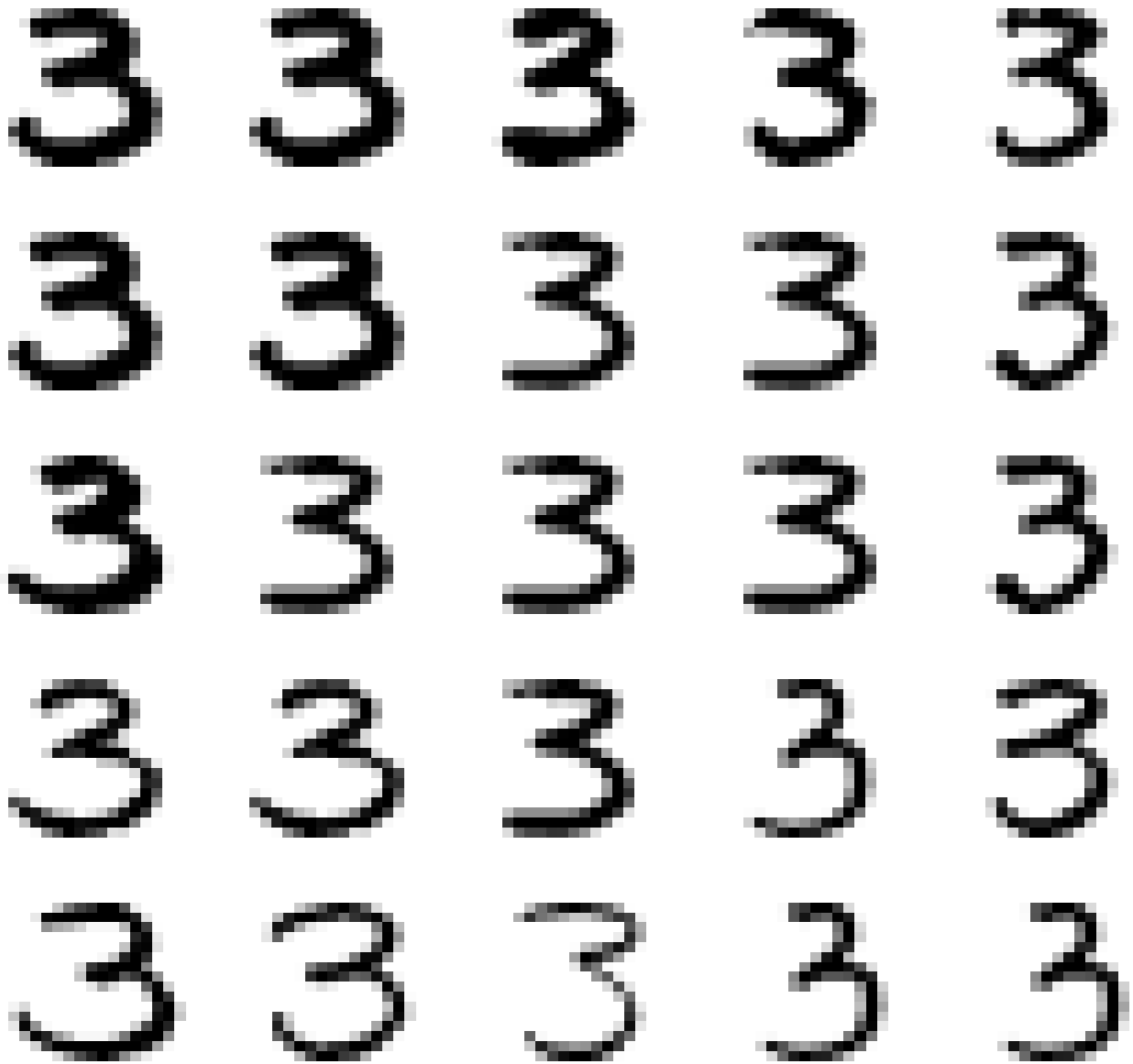
# Low 3rd PC



# Med 3rd PC



# High 3rd PC



# Another interesting PCA example

- PCs can be very sensitive to a few "outliers"
- These can sometimes be alleviated by using a "robust" covariance estimator $\hat{\mathbf{\Sigma}}_n \neq S_n$
- However these robust covariance matrix estimators are very costly to compute (the complexity of the algorithms increases exponentially with the dimension).
- Sometimes,"$n < p$"

# Interesting PCA example

Application: Image processing

- Goal: to detect frames in a clip where "significant" changes occur
- Data: A 3-min scene with $n = 600$ frames. Each frame is represented by a vector $\mathbf{X}_i$, $i = 1, \ldots, 600$.

# Interesting PCA example

Application: Image processing

- The idea is to identify “different” frames, by looking at their Mahalanobis distance to the “center” of the data, i.e.

$$d_i^2 = (\mathbf{X}_i - \hat{\mu})' \hat{\mathbf{\Sigma}}_n^{-1} (\mathbf{X}_i - \hat{\mu})$$

where $\hat{\mu}$ and $\hat{\mathbf{\Sigma}}_n$ are “robust” estimators of the center and covariance matrix of the “good” frames.

# Interesting PCA example

- Each $\mathbf{X}_i \in \mathbb{R}^{640\times 479} = \mathbb{R}^{306560}$
- In this application the frames were reduced to $\mathbf{X}_i \in \mathbb{R}^{49152}$ before starting the analysis.
- **Robust PCs** were computed using projection-pursuit-type algorithms.
- The original clip, and results of the analysis

  `http://www.youtube.com/user/msalibian`