Dimension Reduction

David Hofmeyr

Dept. Statistics and Actuarial Science,
Stellenbosch University, South Africa

6-8 December 2021

Outline

What is dimension reduction?

Why do we do dimension reduction?

(Some) Properties of high dimensional random variables

Principal Components Analysis

What is dimension reduction?

- Can (loosely) think of dimension reduction as a map, DR, operating on the data space, $\mathcal{X}$, which creates a summary of its operand, $\mathbf{x} \in \mathcal{X}$.
- What do we want from a summary?
  - Remove irrelevant information
  - Remove redundant information
  - Retain useful information
  - ... It should take less effort to read (process) but still contain salient points

What is dimension reduction?

In other words

- $size(DR(\mathbf{x})) < size(\mathbf{x})$ (less effort to process)
- BUT "useful" dimension reduction should satisfy
  $Useful\ Information(DR(\mathbf{x})) \approx Useful\ Information(\mathbf{x})$
- "Ideally" we have
  $Useful\ Information(DR(\mathbf{x})) = Useful\ Information(\mathbf{x}) \iff$
  "sufficient" dimension reduction
  Example: Regression
  Interested in $F(y|\mathbf{x})$. Sufficient dimension reduction for this problem, DR, satisfies $F(y|DR(\mathbf{x})) = F(y|\mathbf{x})$

What is dimension reduction?

- **Example:** Subset selection
  $DR(\mathbf{x}) = A^\top \mathbf{x}$, where $p \times p'$ matrix A has ith column e_j, and $\hat{\beta}_j$ is the ith included covariate
- Sensible to have $A_{,i} = e_j$ if:
  - Measurements in $\mathbf{x}$ correspond directly to physical quantities
  - We don't only want to make inference on $F(y|DR(\mathbf{x}))$ but also the map DR
- not necessarily sensible in general
  - "modern" data may be derived from features extracted from non-Euclidean objects (network summaries, spectral decompositions of time series, etc.)
  - We only care about $F(y|DR(\mathbf{x}))$

What is dimension reduction?

- We will assume data occupy $\mathbb{R}^p$
- We will only consider linear dimension reduction
- We will not assume (necessarily) that the individual features (dimensions) are informative
- $DR(\mathbf{x}) = V^\top \mathbf{x},\ V \in \mathbb{R}^{p \times p'}$ (or similar)
  - frequently we want $\|A_{,i}\| = 1\ \forall i$ (don't change the scale after "transformation")
  - For some problems it is necessary to consider orthogonal A, i.e. $A_{,i}^\top A_{,j} = 0$ for $i \neq j$
    - Sometimes this arises incidentally from the problem description

Why do we do dimension reduction?

- Computational issues (make the data easier to process)
- Correlated features (remove redundant information)
- Overfitting (don't focus on spurious details)
- The "quirks" of high dimensional data (other things for which I didn't think of an analogy)

Computational Issues

- The most obvious challenge in high dimensions
- Trivially: "Data matrix $X \in \mathbb{R}^{n \times p}$ has more columns than data matrix $Y \in \mathbb{R}^{n \times p'}$ $(p > p')$, and so it takes more memory and computational effort to work with X than with Y"
- When is it worth doing dimension reduction solely for computational benefits?
  - $\Longleftrightarrow$

    $Cost(\tilde{X} \leftarrow DR(X)) + Cost(Process(\tilde{X})) < Cost(Process(X))$

Computational Issues

Example: Linear regression

- Each $\mathbf{x}_i \in \mathbb{R}^p,\ i \in \{1, ..., n\}$
- Computational cost of the basic problem is $\mathcal{O}(p^2(n+p))$
- (Poor) Heuristic dimension reduction: Remove "highly correlated" combinations of variables
  - compute all correlations $\mathcal{O}(p^2)$
  - recover $\tilde{p}$ (maybe unknown) variables
  - new cost $\mathcal{O}(p^2 + \tilde{p}^2(n+\tilde{p}))$
  - for $\tilde{p} \in o_{(P)}(p)$ as $p \to \infty$ (e.g. $\tilde{p} \leq_{(P)} Kp^\delta$ for some $K > 0, 0 < \delta < 1$) this will "always" be beneficial as p grows

Correlated Features

- When features (dimensions/columns in the data matrix) are highly correlated, "standard inference" can be misleading
- In the extreme as $\rho \to \pm 1$, $X_{,i} \approx aX_{,j} + b$ for $a, b \in \mathbb{R}$
- Mathematically $[\mathbf{1}, X]$ is close to a matrix with non-full rank
- Why is this a problem?
  - $[\mathbf{1}, X]'[\mathbf{1}, X]$ has high condition number (hard to invert)
  - The SS objective in regression looks like a "half-pipe", some gradient based methods fail to converge in reasonable time
- Affects inference in the effect it has on variance of regression coefficients, for example.
- At a higher, more generic level, if columns in X are close to linearly dependent, I can (almost) recreate some using others, so their presence is redundant (they don't add much information)

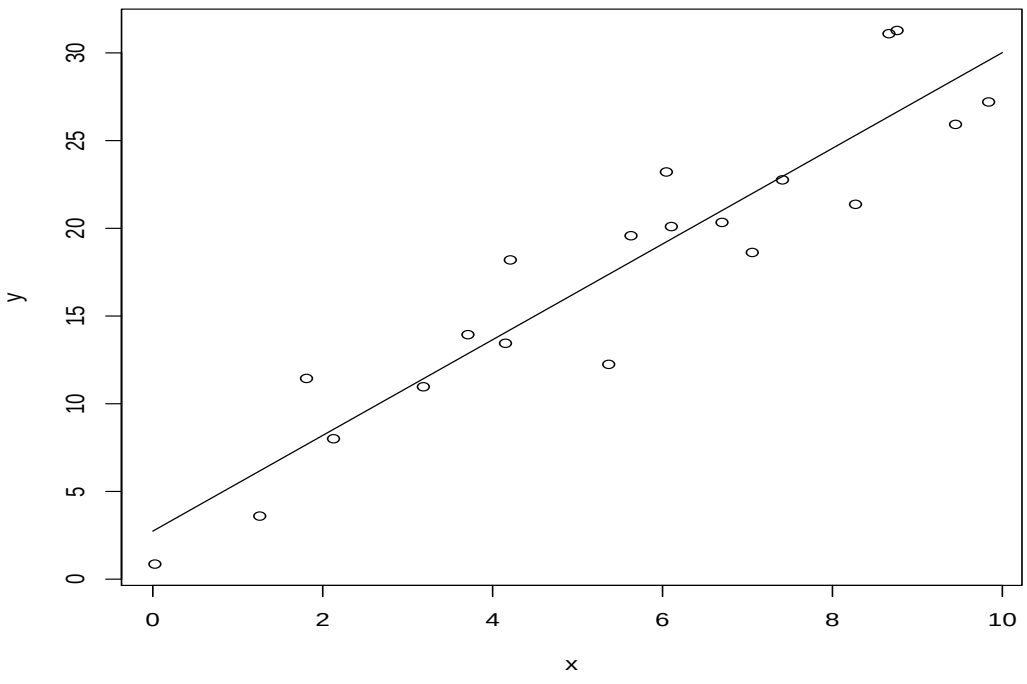
Overfitting

- Standard (supervised) methods fit a model by minimising training error
  - e.g. SS in regression
- Complex models (with too many dfs/decision variables/parameters) can fit data "too well".
- will fit the "noise" as well as the "signal"
- No obvious way to decompose a posteriori
- (Can also be a problem for unsupervised learning... "overlearning")

Overfitting

- If S/N ratio is high a more parsimonious model should first pick up the signal
- If S/N ratio is low... not for us to consider here
- The change from "reasonable" to "terrible" can occur suddenly
- **Example:**
  $X \sim U[0, 10]$
  $Y_i \sim N(1 + 3x_i, 3^2)$
  consider $\hat{y} = polynomial(x, d), d = 1, ..., 10$

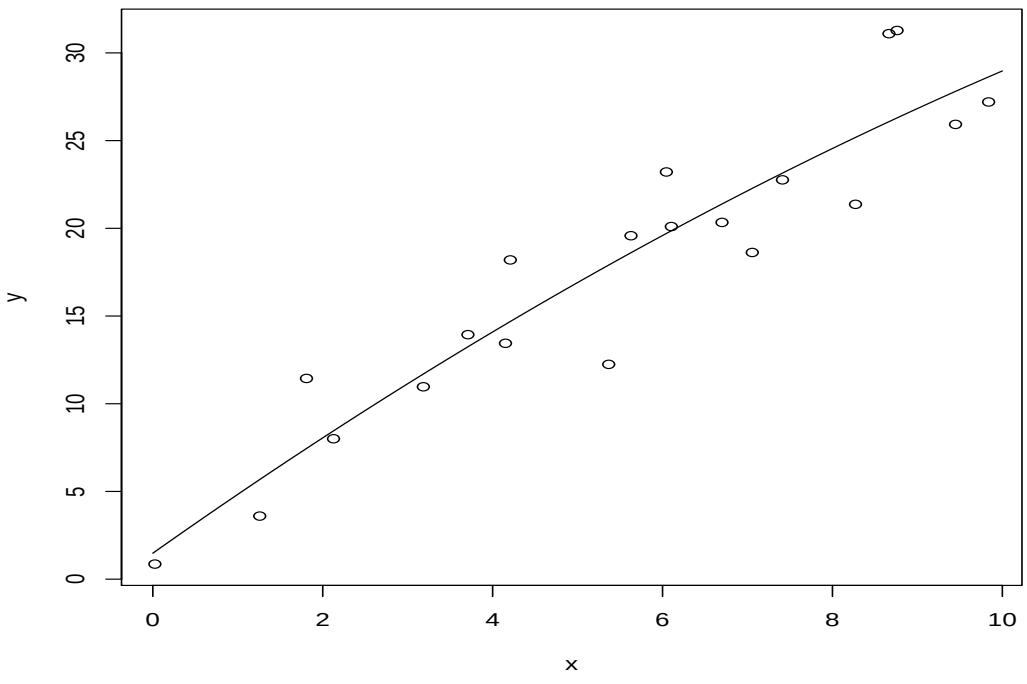
Overfitting regression: d = 1



$R'^2 = 0.865$, ANOVAp-value $\mathcal{O}(10^{-9})$

$\mathbb{E}[SSerror(Model, X_{n+1})] \approx 9.8$

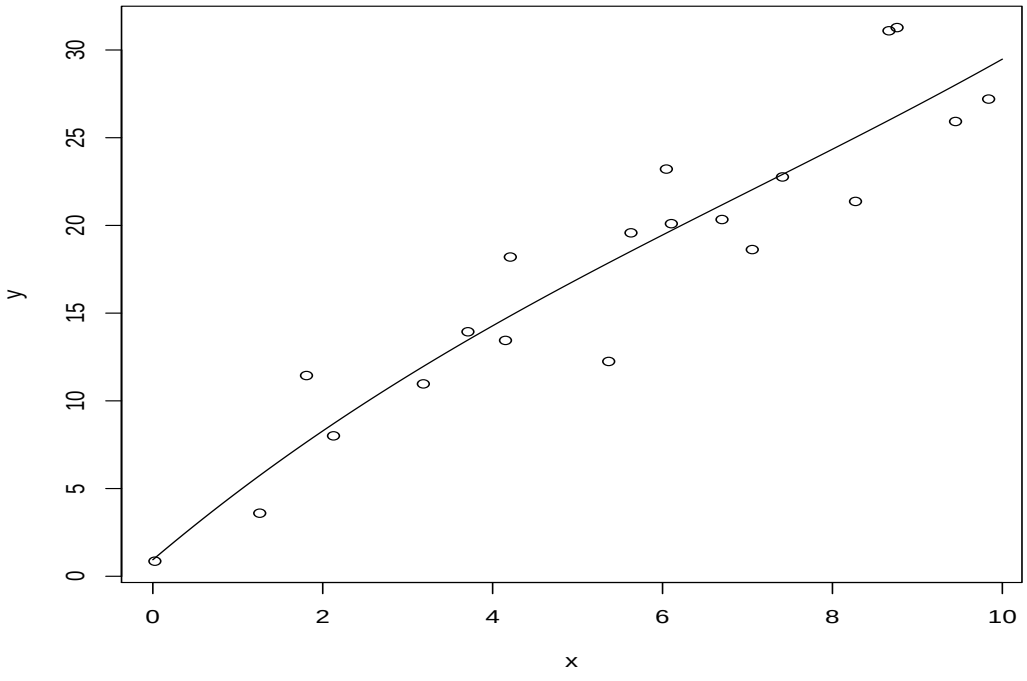
Overfitting regression: d = 2



$R'^2 = 0.861$, ANOVAp-value $\mathcal{O}(10^{-8})$

$\mathbb{E}[SSerror(Model, X_{n+1})] \approx 9.9$

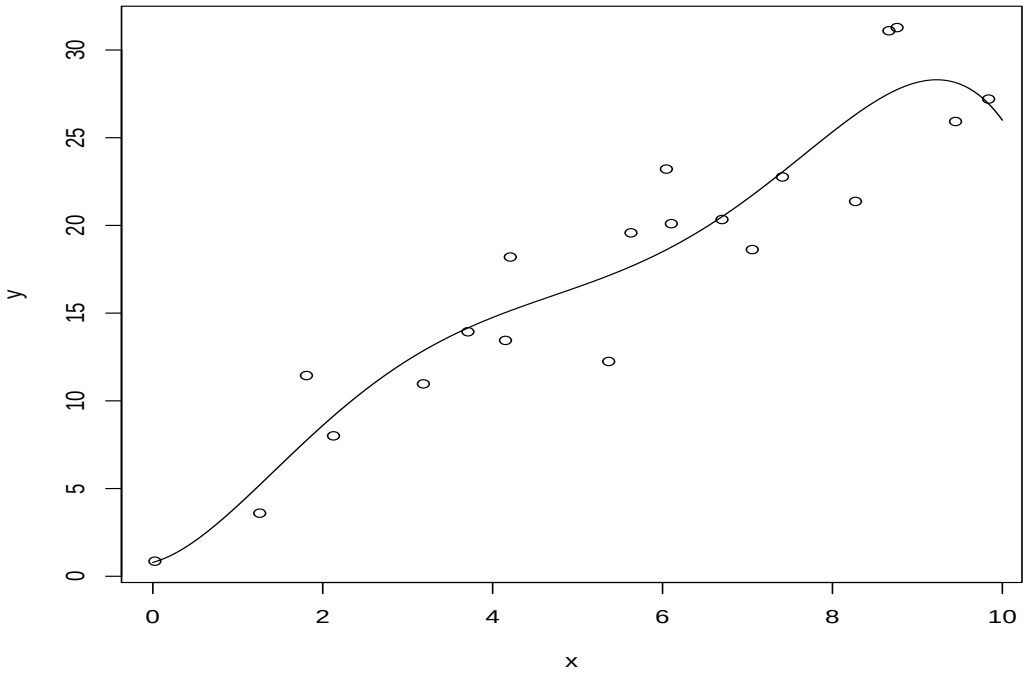
Overfitting regression: d = 3



$R'^2 = 0.877$, ANOVA p-value $\mathcal{O}(10^{-7})$, no individual "significant" $\hat{\beta}$'s

$\mathbb{E}[SSerror(Model, X_{n+1})] \approx 9.9$

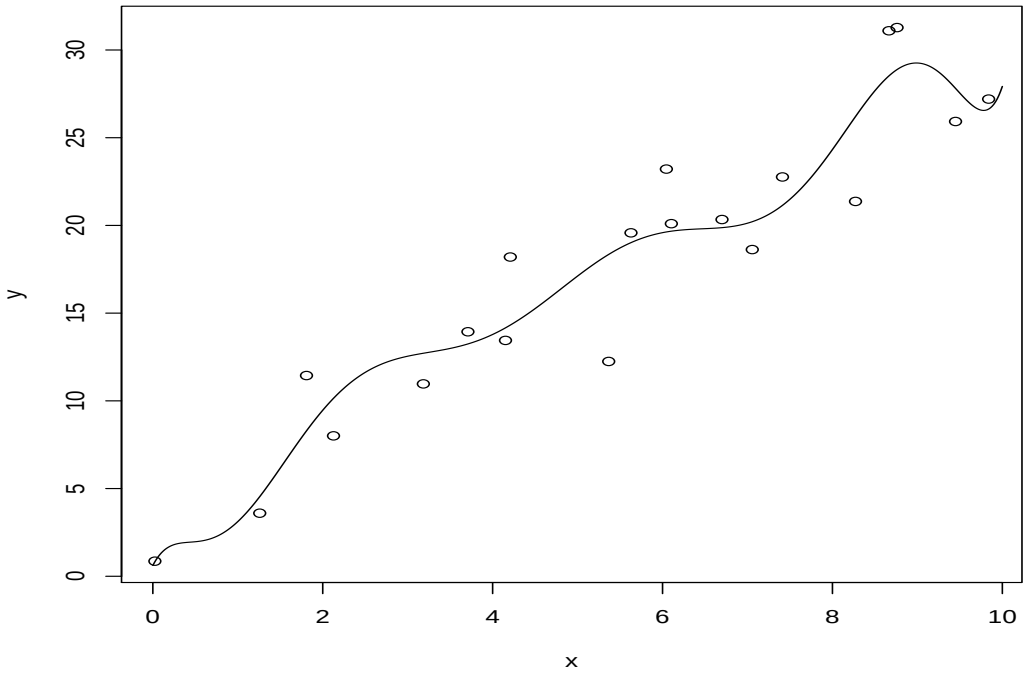
Overfitting regression: d = 5



$R'^2 = 0.889$, ANOVAp-value $\mathcal{O}(10^{-6})$, no individual "significant" $\hat{\beta}$'s

$\mathbb{E}[SSerror(Model, X_{n+1})] \approx 10.9$

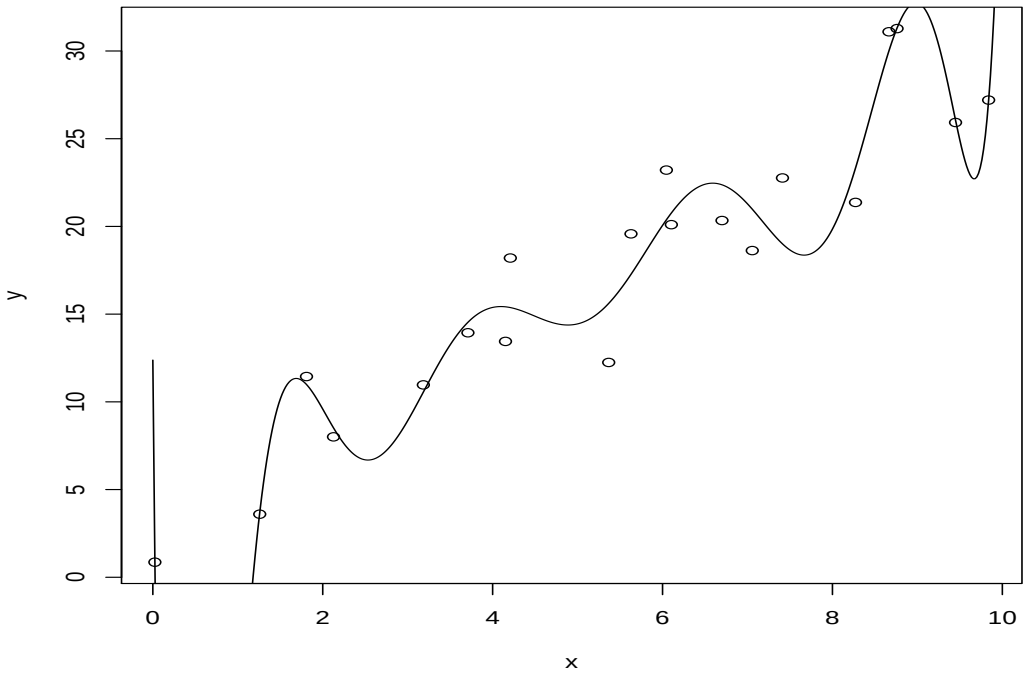
Overfitting regression: d = 9



$R'^2 = 0.902$, ANOVAp-value $\mathcal{O}(10^{-4})$, no individual "significant" $\hat{\beta}$'s

$\mathbb{E}[SSerror(Model, X_{n+1})] \approx 11.7$

Overfitting regression: d = 10



$R'^2 = 0.950$, ANOVAp-value $\mathcal{O}(10^{-5})$, ALL individual "significant" $\hat{\beta}$'s

$\mathbb{E}[SSerror(Model, X_{n+1})] \approx 240.9$

Overfitting

- With enough degrees of freedom we can obtain 0 training error
- **Example:** Linear regression
  $X \in \mathbb{R}^{p \times n}, p > n$
  $y = \beta X$ is underdetermined $\Rightarrow \exists$ solution β for ANY y
  $\therefore \hat{y}_i = y_i, \ \forall i$
  But there may be no (true) relationship between any elements of X and y
- In the last example we had the benefit of visualisation, but not always possible

The quirks of high dimensional data (random variables)

- High dimensional random variables display some surprising "quirks"
- Some may be at first surprising, but even with minimal consideration are easy to explain
  - **Example** "All the mass of a uniform random variable over the hyper-cube lies on the surface"

The quirks of high dimensional data (random variables)

- High dimensional random variables display some surprising (?) "quirks"
- Some may be at first surprising, but even with minimal consideration are easy to explain
- Others are just plain elusive (at least to me)
  - **Example** "All the mass of a uniform random variable over the hyper-cube lies on the surface"
    Let $\mathbf{U}^p \sim U[0,1]^p$, then for any $0 < \epsilon < 1$

$$\lim_{p \to \infty} \mathbb{P}\left(d(\mathbf{U}^p, \text{boundary}[0,1]^p) < \frac{1}{p^\epsilon} \right) = 1$$

The quirks of high dimensional data (random variables)

- High dimensional random variables display some surprising (?) "quirks"
- Some may be at first surprising, but even with minimal consideration are easy to explain
- Others are just plain elusive (at least to me)
  - **Example** "All the mass of a uniform random variable over the hyper-cube lies on the surface"
    Let $\mathbf{U}^p \sim U[0,1]^p$, then for any $0 < \epsilon < 1$

$$\lim_{p \to \infty} \mathbb{P}\left(d(\mathbf{U}^p, \text{boundary}[0,1]^p) < \frac{1}{p^\epsilon} \right) = 1$$

    Or the weaker but maybe easier to interpret
    For any $\delta > 0$

$$\lim_{p \to \infty} \mathbb{P}\left(d(\mathbf{U}^p, \text{boundary}[0,1]^p) < \delta \right) = 1$$

The quirks of high dimensional data (random variables)

- Why?

The quirks of high dimensional data (random variables)

- Why?

$$
\begin{aligned}
& d(\mathbf{u}^p, \text{boundary}[0,1]^p) \geq 1/p^\epsilon \\
\Rightarrow\ & \min \mathbf{u}^p \geq 1/p^\epsilon \\
\iff\ & \mathbf{u}_i^p \geq 1/p^\epsilon,\ \forall i \in \{1, ..., p\}
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow \mathbb{P}\left(d(\mathbf{U}^p, \text{boundary}[0,1]^p) < \frac{1}{p^\epsilon} \right) &\geq 1 - \mathbb{P}(\mathbf{U}_i^p \geq 1/p^\epsilon,\ \forall i) \\
&= 1 - \left(1 - \frac{1}{p^\epsilon}\right)^p \\
&\xrightarrow{as\ \ p \to \infty} 1
\end{aligned}
$$

The quirks of high dimensional data (random variables)

- The same is true of many other solids, including the sphere
  Example: "All the mass of a uniform random variable over the unit sphere lies on the surface"
- Slightly less obvious results also hold for the sphere
  Example: "All the mass of a uniform random variable over the unit sphere lies at the equator"
  That is, if $\mathbf{U} \sim U(\mathcal{S}^{p-1})$ then for any $\epsilon > 0$

$$\lim_{p \to \infty} \mathbb{P}\left(-\epsilon < \mathbf{U}_1 < \epsilon\right) = 1$$

The quirks of high dimensional data (random variables)

- Some appear surprising, and remain so even after consideration
- **Example:** "As dimension increases the relative distance between random points becomes more uniform"

The quirks of high dimensional data (random variables)

- Some appear surprising, and remain so even after consideration
- **Example:** "As dimension increases the relative distance between random points becomes more uniform"
  Theorem [**?**] For $p \in \mathbb{N}$ let $X_1^p, ..., X_n^p$ be i.i.d. p dimensional random variables (with each component of X_1^p having the same distribution, F, with finite, non-zero second moment). Define

$$DMIN_p := \min\{d(X_i^p, X^p) | i \in \{1, ..., n\}\}$$
$$DMAX_p := \max\{d(X_i^p, X^p) | i \in \{1, ..., n\}\},$$

  (where X^p has the same distribution as X_1^p). Then

$$\frac{DMAX_p}{DMIN_p} \to_P 1, \text{ as } p \to \infty$$

The quirks of high dimensional data (random variables)

Proof:

The quirks of high dimensional data (random variables)

Proof:

- Consider
$$\frac{1}{p} d(X_1^p, X^p)^2 = \frac{1}{p} \sum_{i=1}^{p} (X_{1,i}^p - X_i^p)^2.$$
- WLLN $\Rightarrow \frac{1}{p} d(X_1^p, X^p)^2 \to_P \mathbb{E}[D^2] \neq 0$, where D is the difference between two independent random variables with distribution function F.
- $\therefore \frac{1}{p} \max\{d(X_1^p, X^p)^2, ..., d(X_n^p, X^p)^2\} \to_P \mathbb{E}[D^2]$. similarly for min.
- $\therefore \frac{\frac{1}{p} \max\{d(X_1^p, X^p)^2, ..., d(X_n^p, X^p)^2\}}{\frac{1}{p} \min\{d(X_1^p, X^p)^2, ..., d(X_n^p, X^p)^2\}} \to_P 1$
- Simplifying and taking square root gives the result.

The quirks of high dimensional data (random variables)

Remarks:

- i.i.d dimensions is not a necessary condition
  - Authors discuss many other situations in which it holds
- "Query point" X^p does not have to have the same distribution as X_1^p, but must be independent.
- Equivalent statement of result: $\forall \epsilon > 0$

$$\lim_{p \to \infty} \mathbb{P}(DMAX_p \leq DMIN_p(1+\epsilon)) = 1.$$

- Especially important if data recorded with noise

Principal Components Analysis

- (Probably) the most popular dimension reduction technique
- Numerous formulations, with the most persuasive (to me) based on "reconstruction error":

$$\min_{V \in \mathbb{R}^{p \times p'}} ||X - XVV'||_F^2,$$

  where X has been centered.
- Here we have $DR(X) = XV^*$ and $XV^*V^{*'}$ can be seen as taking this reduced form of X and putting it "back" into the original input space
  - This is not dissimilar from auto-encoders, where the objective is to "find a reduced form of X which can be almost inverted"
  - If I can more-or-less "undo" the dimension reduction, then I cannot have lost much information

Principal Components Analysis

- A common re-formulation is "find the orthonormal V which maximises the sum of the variances of the columns of XV."
- This re-formulation looks more like a "projection pursuit" formulation:
  - Find a projection of the data which is as "interesting as possible": maximise Interestingness(XV)
  - Projection taken to mean different things in different contexts, is XV or XVV' the projection? In projection pursuit it is XV, whereas in (general) linear algebra it is XVV'
- Why is/might the projection pursuit formulation be preferable?
  - Faster computationally
  - Invariant to rotations which are not permutations
- Why is/might the first formulation be preferable?
  - Convexity

Robust Principal Components Analysis

- Disclaimer: There are much fancier alternatives in other contexts. I will look only at simple variations from the standard model
- We can replace the squared loss with any "loss" function

$$||X - XVV'||_2^F = \sum_{i,j}(X_{ij} - X_{i:}VV'_{j:})^2$$
$$= \sum_{i,j} L(X_{ij}, X_{i:}VV'_{j:})$$

- The projection pursuit alternative (now not equivalent) is to maximise $\sum_{i,k} L(c(XV_{:k}), X_{i:}V_{:k})$ where $c(\cdot)$ is some measure of the center of the (projected) sample.

Let's play around in R

- We'll just use standard R optimisation for which we'll need objective functions and their gradients
- We have

$$D_V \left(\sum_{i,j} \ell(X_{ij} - X_{i:} V V'_{j:}) \right) = -\ell'(X - XVV')'XV - X'\ell'(X - XVV')V$$

Let's play around in R

- There are more elegant ways of enforcing orthonormality, but we will use a "deflation" scheme for the projection pursuit formulation:
  - First find $V_{:1}$
  - Then repeatedly find the subsequent columns by applying the same procedure applied to the data projected into the null space of the columns found so far
- To enforce $||V_{:k}|| = 1$ we simply include in the evaluation of the objective the projection onto the unit sphere, i.e., we maximise over $v \in \mathbb{R}^p$

$$\sum_{i=1}^{n} \ell(c(X\vec{v}) - X_{i:}\vec{v}),$$

where $\vec{v} = v/||v||$, and then set $V_{:k} = \vec{v^*}$

- If we set $c(\cdot)$ to be the mean and first center the observations then we have

$$\nabla_v \sum_{i=1}^{n} \ell(c(X\vec{v}) - X_{i:}\vec{v}) = \frac{1}{||v||}(I - \vec{v}\vec{v}')X'\ell'(c(X\vec{v}) - X\vec{v})$$