Statistical Machine Learning, Part I

Regression 2

mcuturi@i.kyoto-u.ac.jp

Last Week

Regression: highlight a functional relationship between a **predicted variable** and **predictors**

Last Week

Regression: highlight a functional relationship between a **predicted variable** and **predictors**

find a function $\boldsymbol{f}$ such that

$$\forall(\mathbf{x}, \boldsymbol{y}) \text{ that can appear }, \boldsymbol{f}(\mathbf{x}) \approx \boldsymbol{y}$$

Last Week

Regression: highlight a functional relationship between a **predicted variable** and **predictors**

to find an accurate function f such that

$\forall(\mathbf{x}, y)$ that can appear , $f(\mathbf{x}) \approx y$

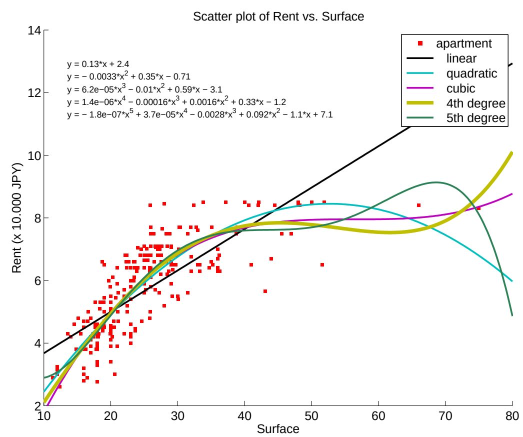
use a data set & the least-squares criterion:

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{j=1}^{N} (y_j - f(x_j))^2$$

Last Week

Regression: highlight a functional relationship between a **predicted variable** and **predictors**

- **when regressing a real number vs a real number :**



- ◦ Least-Squares Criterion $L(b, a_1, \cdots, a_p)$ to fit **lines**, polynomials.
- ◦ results in solving a linear system.

$$\frac{\partial \mathbf{2}^{\text{nd}}\text{order}(b, a_1, \cdots, a_p)}{\partial a_p} = \textbf{linear} \text{ in } (b, a_1, \cdots, a_p)$$

- ◦ When setting $\partial L / \partial a_p = 0$ we get $p + 1$ **linear** equations for $p + 1$ variables.

Last Week

Regression: highlight a functional relationship between a **predicted variable** and **predictors**

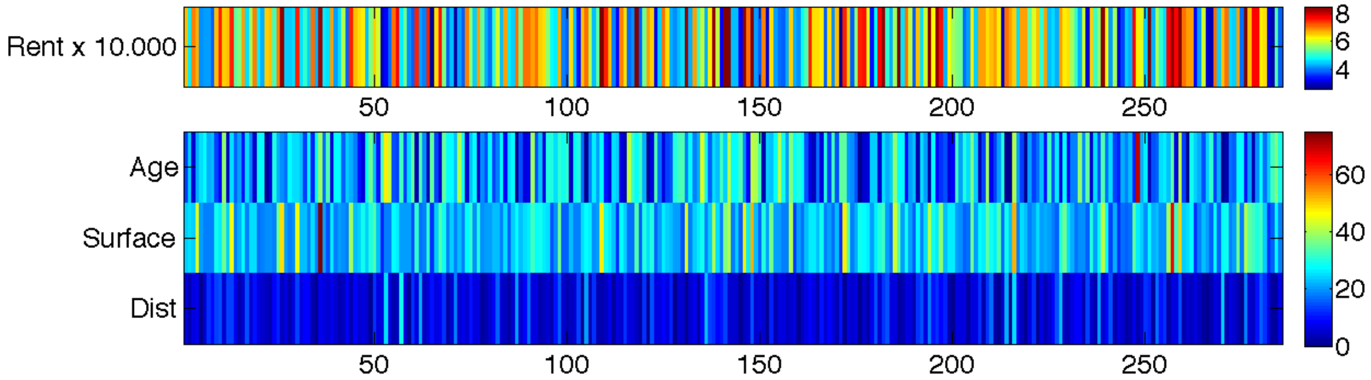
- **when regressing a real number vs d real numbers (vector in $\mathbb{R}^d$)**,
  - find best fit $\alpha \in \mathbb{R}^d$ such that $(\alpha^T \mathbf{x} + \alpha_0) \approx y$.
  - Add to $d \times N$ data matrix, a row of 1's to get the predictors $\boldsymbol{X}$.
  - The row $\boldsymbol{Y}$ of **predicted** values
  - The Least-Squares criterion also applies:

$$L(\alpha) = \|\boldsymbol{Y} - \alpha^T \boldsymbol{X}\|^2 = \left(\alpha^T \boldsymbol{X}\boldsymbol{X}^T \alpha - 2\boldsymbol{Y}\boldsymbol{X}^T \alpha + \|\boldsymbol{Y}\|^2\right).$$

$$\nabla_\alpha L = 0 \quad \Rightarrow \quad \alpha^\star = \boldsymbol{(X X^T)^{-1} X Y^T}$$

- This works if $\boldsymbol{X}\boldsymbol{X}^T \in \mathbb{R}^{d+1}$ is **invertible**.

Last Week




```
>> (X*X')\(X*Y')
ans =
   -0.049332605603095     x age
    0.163122792160298     x surface
   -0.004411580036614     x distance
    2.731204399433800     + 27.300 JPY
```

Today

- A **statistical / probabilistic** perspective on LS-regression
- A few words on **polynomials** in higher dimensions
- A **geometric** perspective
- **Variable co-linearity and Overfitting** problem
- Some solutions: **advanced regression techniques**
  - Subset selection
  - Ridge Regression
  - Lasso

A (very few) words on the statistical/probabilistic interpretation of LS

The Statistical Perspective on Regression

- **Assume that** the values of y are stochastically linked to observations $\mathbf{x}$ as

$$\boldsymbol{y} - (\alpha^T \mathbf{x} + \beta) \sim \mathcal{N}(0, \sigma).$$

- This difference is a random variable called ε and is called a **residue**.

The Statistical Perspective on Regression

- This can be rewritten as,

$$\boldsymbol{y} = (\alpha^T \mathbf{x} + \beta) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma),$$

- We **assume** that the difference between y and $(\alpha^T \mathbf{x} + b)$ behaves like a **Gaussian** (normally distributed) random variable.

Goal as a statistician: **Estimate** α and β given observations.

Identically Independently Distributed (i.i.d) Observations

- Statistical hypothesis: **assume that the parameters are** $\alpha = \mathbf{a}, \beta = b$

Identically Independently Distributed (i.i.d) Observations

- Statistical hypothesis: **assume that the parameters are $\alpha = \mathbf{a}, \beta = b$**
- In such a case, what would be the **probability** of **each** observation $(\mathbf{x}_j, y_j)$?

Identically Independently Distributed (i.i.d) Observations

- Statistical hypothesis: **assuming that the parameters are $\alpha = \mathbf{a}, \beta = b$**, what would be the **probability** of **each** observation?:
  - For each couple $(\mathbf{x}_j, y_j)$, $j = 1, \cdots, N$,

$$P(\mathbf{x}_j, y_j \,|\, \alpha = \mathbf{a}, \beta = b) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|y_j - (\mathbf{a}^T \mathbf{x}_j + b)\|^2}{2\sigma^2}\right)$$

Identically Independently Distributed (i.i.d) Observations

- Statistical hypothesis: **assuming that the parameters are $\alpha = \mathbf{a}, \beta = b$**, what would be the **probability** of **each** observation?:
  - For each couple $(\mathbf{x}_j, y_j)$, $j = 1, \cdots, N$,

$$P(\mathbf{x}_j, y_j \,|\, \alpha = \mathbf{a}, \beta = b) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|y_j - (\mathbf{a}^T\mathbf{x}_j + b)\|^2}{2\sigma^2}\right)$$

  - Since each measurement $(\mathbf{x}_j, y_j)$ has been **independently sampled**,

$$P\left(\{(\mathbf{x}_j, y_j)\}_{j=1,\cdots,N} \,|\, \alpha = a, \beta = b\right) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|y_j - (\mathbf{a}^T\mathbf{x}_j + b)\|^2}{2\sigma^2}\right)$$

Identically Independently Distributed (i.i.d) Observations

- Statistical hypothesis: **assuming that the parameters are $\alpha = \mathbf{a}, \beta = b$**, what would be the **probability** of **each** observation?:
  - For each couple $(\mathbf{x}_j, y_j)$, $j = 1, \cdots, N$,

$$P(\mathbf{x}_j, y_j \,|\, \alpha = \mathbf{a}, \beta = b) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|y_j - (\mathbf{a}^T\mathbf{x}_j + b)\|^2}{2\sigma^2}\right)$$

  - Since each measurement $(\mathbf{x}_j, y_j)$ has been **independently sampled**,

$$P\left(\{(\mathbf{x}_j, y_j)\}_{j=1,\cdots,N} \,|\, \alpha = a, \beta = b\right) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|y_j - (\mathbf{a}^T\mathbf{x}_j + b)\|^2}{2\sigma^2}\right)$$

  - A.K.A **likelihood** of the dataset $\{(\mathbf{x}_j, y_j)_{j=1,\cdots,N}\}$ as a function of a and b,

$$\mathcal{L}_{\{(\mathbf{x}_j, y_j)\}}(\mathbf{a}, b) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|y_j - (\mathbf{a}^T\mathbf{x}_j + b)\|^2}{2\sigma^2}\right)$$

Maximum Likelihood Estimation (MLE) of Parameters

Hence, for $\mathbf{a}, b$, the **likelihood** function on the dataset $\{(\mathbf{x}_j, y_j)_{j=1,\cdots,N}\}$...

$$\mathcal{L}(\mathbf{a}, b) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|y_j - (\mathbf{a}^T\mathbf{x}_j + b)\|^2}{2\sigma^2}\right)$$

Maximum Likelihood Estimation (MLE) of Parameters

Hence, for $\mathbf{a}, b$, the **likelihood** function on the dataset $\{(\mathbf{x}_j, y_j)_{j=1,\cdots,N}\}$...

$$\mathcal{L}(\mathbf{a}, b) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|y_j - (\mathbf{a}^T\mathbf{x}_j + b)\|^2}{2\sigma^2}\right)$$

Why not use the **likelihood** to **guess** $(\mathbf{a}, b)$ given data?

Maximum Likelihood Estimation (MLE) of Parameters

Hence, for $\mathbf{a}, b$, the **likelihood** function on the dataset $\{(\mathbf{x}_j, y_j)_{j=1,\cdots,N}\}$...

$$\mathcal{L}(\mathbf{a}, b) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|y_j - (\mathbf{a}^T\mathbf{x}_j + b)\|^2}{2\sigma^2}\right)$$

...the **MLE** approach selects the values of $(\mathbf{a}, b)$ which **mazimize** $\mathcal{L}(\mathbf{a}, b)$

Maximum Likelihood Estimation (MLE) of Parameters

Hence, for $\mathbf{a}, b$, the **likelihood** function on the dataset $\{(\mathbf{x}_j, y_j)_{j=1,\cdots,N}\}$...

$$\mathcal{L}(\mathbf{a}, b) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|y_j - (\mathbf{a}^T\mathbf{x}_j + b)\|^2}{2\sigma^2}\right)$$

...the **MLE** approach selects the values of $(\mathbf{a}, b)$ which **mazimize** $\mathcal{L}(\mathbf{a}, b)$

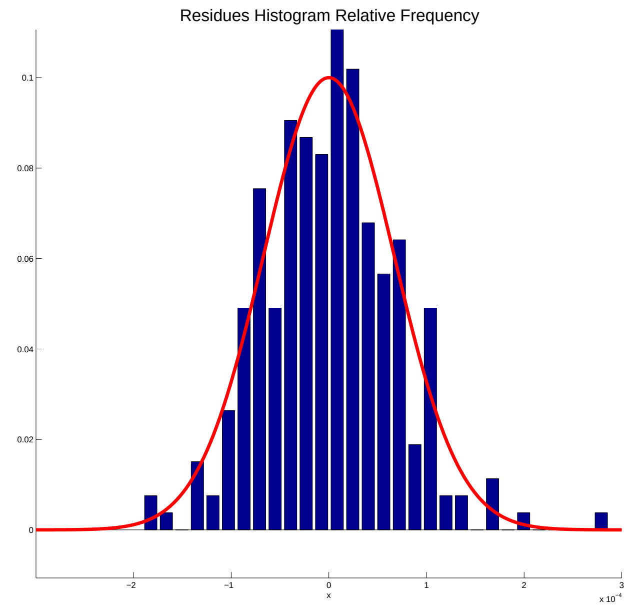
- **Since** $\max_{(\mathbf{a},b)} \mathcal{L}(\mathbf{a}, b) \Leftrightarrow \max_{(\mathbf{a},b)} \log \mathcal{L}(\mathbf{a}, b)$

$$\log L(\mathbf{a}, b) = C - \frac{1}{2\sigma^2} \sum_{j=1}^{N} \|y_j - (\mathbf{a}^T\mathbf{x}_j + b)\|^2.$$

- **Hence** $\max_{(\mathbf{a},b)} \mathcal{L}(\mathbf{a}, b) \Leftrightarrow \min_{(\mathbf{a},b)} \sum_{j=1}^{N} \|y_j - (\mathbf{a}^T\mathbf{x}_j + b)\|^2$...

Statistical Approach to Linear Regression

- Properties of the MLE estimator: convergence of $\|\alpha - \mathbf{a}\|$?
- Confidence intervals for coefficients,
- Tests procedures to assess if model "fits" the data,



- Bayesian approaches: instead of looking for **one** optimal fit $(\mathbf{a}, b)$ juggle with a whole density on $(\mathbf{a}, b)$ to make decisions
- *etc.*

A few words on polynomials in higher dimensions

A few words on polynomials in higher dimensions

- For d variables, that is for points $\mathbf{x} \in \mathbb{R}^d$,
  - the space of polynomials on these variables up to degree p is generated by

$$\{\mathbf{x}^{\mathbf{u}} \,|\, \mathbf{u} \in \mathbb{N}^d, \mathbf{u} = (u_1, \cdots, u_d), \sum_{i=1}^{d} u_i \leq p\}$$

    where the monomial $\mathbf{x}^{\mathbf{u}}$ is defined as $x_1^{u_1} x_2^{u_2} \cdots x_d^{u_d}$

  - Recurrence for dimension of that space: $\dim_{p+1} = \dim_p + \binom{p+1}{d+p}$
- For $d = 20$ and $p = 5$, $1 + 20 + 210 + 1540 + 8855 + 42504 > 50.000$

Problem with polynomial interpolation in **high-dimensions** is the **explosion** of relevant variables (one for each monomial)

Geometric Perspective

Back to Basics

- Recall the problem:

$$X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_N \\ \vdots & \vdots & \cdots & \vdots \end{bmatrix} \in \mathbb{R}^{d+1 \times N}$$

and

$$Y = \begin{bmatrix} y_1 & \cdots & y_N \end{bmatrix} \in \mathbb{R}^N.$$

- We look for α such that $\alpha^T X \approx Y$.

Back to Basics

- If we transpose this expression we get $X^T\alpha \approx Y^T$,

$$\begin{bmatrix} 1 & x_{1,1} & \cdots & x_{d,1} \\ 1 & x_{1,2} & \cdots & x_{d,2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1,k} & \cdots & x_{d,k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1,N} & \cdots & x_{d,N} \end{bmatrix} \times \begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_d \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_2 \\ \vdots \\ y_\cdot \\ \vdots \\ y_N \end{bmatrix}$$

- Using the notation $\mathbf{Y} = Y^T, \mathbf{X} = X^T$ and $\mathbf{X}_k$ for the $(k+1)^{\text{th}}$ column of $\mathbf{X}$,

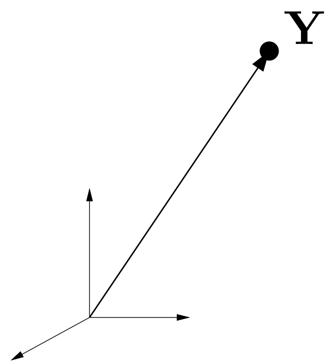
$$\sum_{k=0}^{d} \alpha_k \mathbf{X}_k \approx \mathbf{Y}$$

- Note how the $\mathbf{X}_k$ corresponds to **all** values taken by the k^{th} variable.
- **Problem**: approximate/reconstruct Reconstructing $\mathbf{Y} \in \mathbb{R}^N$ using $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d \in \mathbb{R}^N$?

Linear System

Reconstructing $\mathbf{Y} \in \mathbb{R}^N$ using $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$ vectors of $\mathbb{R}^N$.

- Our ability to approximate $\mathbf{Y}$ depends implicitly on the space spanned by $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$

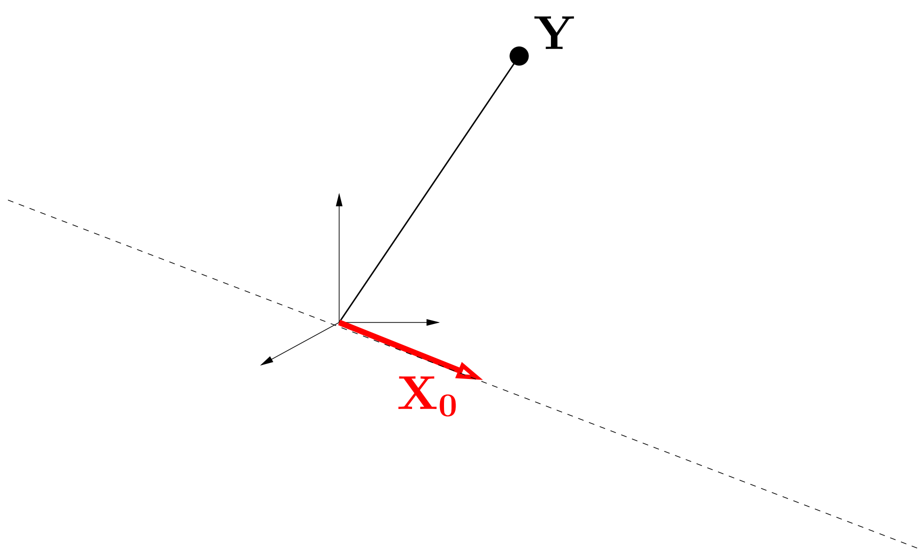


Consider the observed vector in $\mathbb{R}^N$ of predicted values

Linear System

Reconstructing $\mathbf{Y} \in \mathbb{R}^N$ using $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$ vectors of $\mathbb{R}^N$.

- Our ability to approximate $\mathbf{Y}$ depends implicitly on the space spanned by $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$

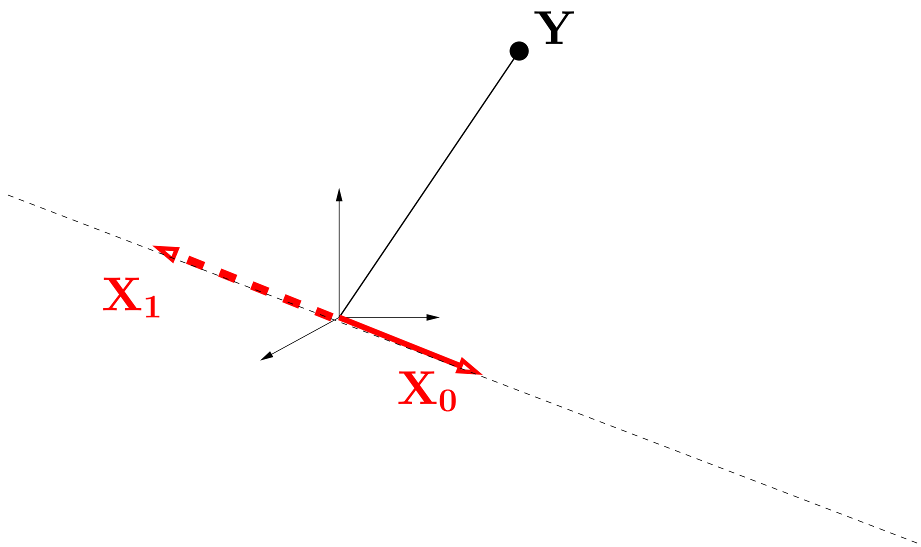


Plot the first regressor $\mathbf{X_0}$...

Linear System

Reconstructing $\mathbf{Y} \in \mathbb{R}^N$ using $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$ vectors of $\mathbb{R}^N$.

- Our ability to approximate $\mathbf{Y}$ depends implicitly on the space spanned by $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$

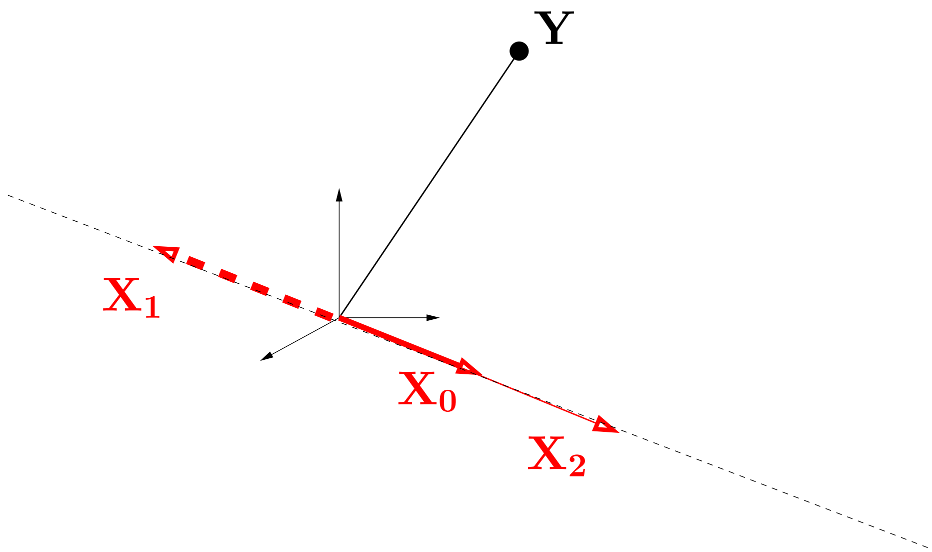


Assume the next regressor $\mathbf{X_1}$ is colinear to $\mathbf{X_0}$...

Linear System

Reconstructing $\mathbf{Y} \in \mathbb{R}^N$ using $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$ vectors of $\mathbb{R}^N$.

- Our ability to approximate $\mathbf{Y}$ depends implicitly on the space spanned by $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$

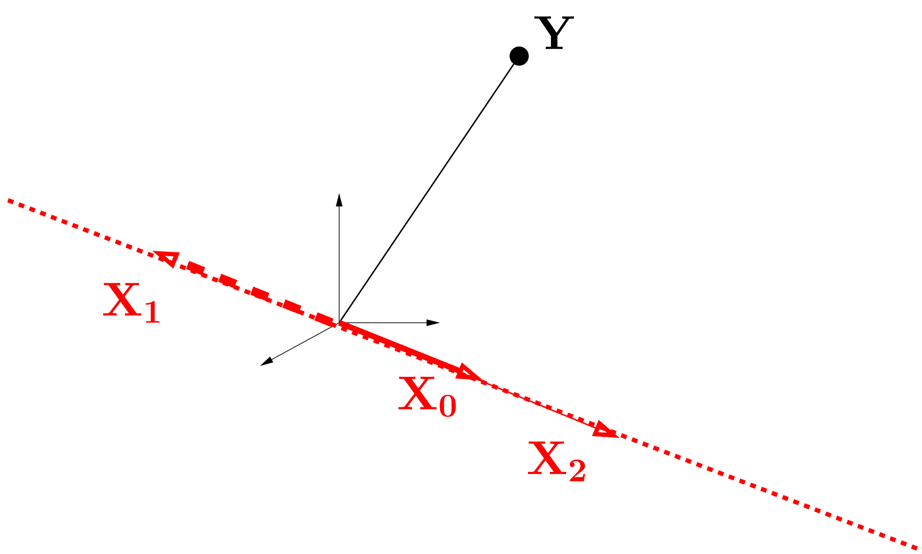


and so is $\mathbf{X_2}$...

Linear System

Reconstructing $\mathbf{Y} \in \mathbb{R}^N$ using $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$ vectors of $\mathbb{R}^N$.

- Our ability to approximate $\mathbf{Y}$ depends implicitly on the space spanned by $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$

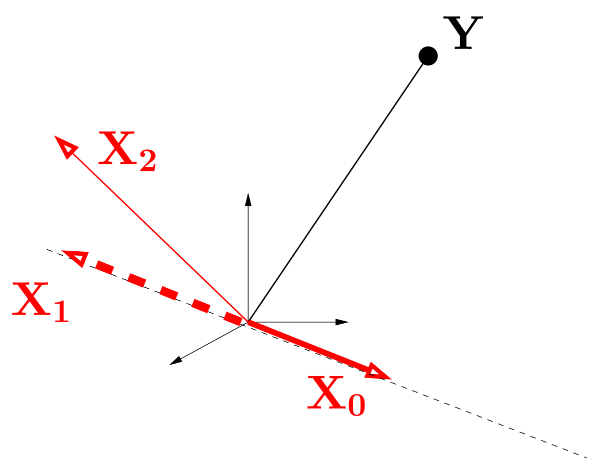


Very little choices to approximate $\mathbf{Y}$...

Linear System

Reconstructing $\mathbf{Y} \in \mathbb{R}^N$ using $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$ vectors of $\mathbb{R}^N$.

- Our ability to approximate $\mathbf{Y}$ depends implicitly on the space spanned by $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$

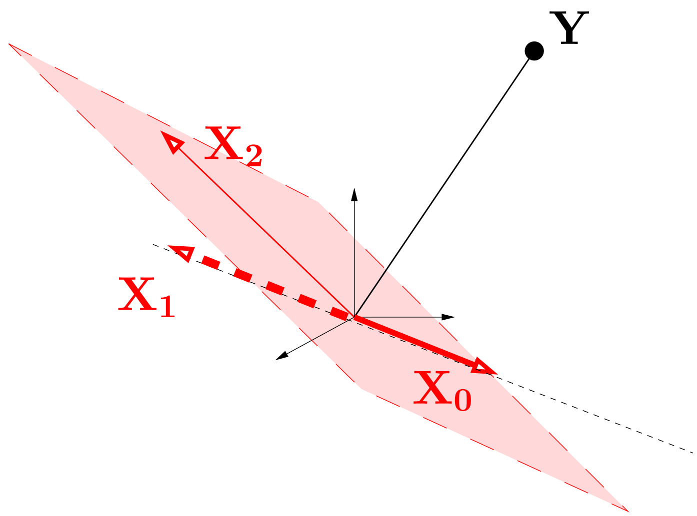


Suppose $\mathbf{X_2}$ is actually not colinear to $\mathbf{X_0}$.

Linear System

Reconstructing $\mathbf{Y} \in \mathbb{R}^N$ using $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$ vectors of $\mathbb{R}^N$.

- Our ability to approximate $\mathbf{Y}$ depends implicitly on the space spanned by $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$

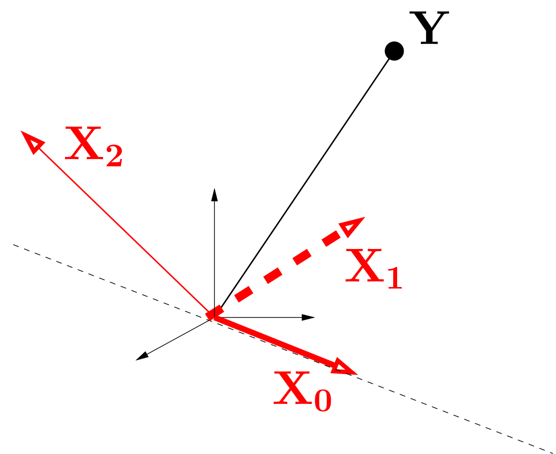


This opens new ways to reconstruct $\mathbf{Y}$.

Linear System

Reconstructing $\mathbf{Y} \in \mathbb{R}^N$ using $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$ vectors of $\mathbb{R}^N$.

- Our ability to approximate $\mathbf{Y}$ depends implicitly on the space spanned by $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$

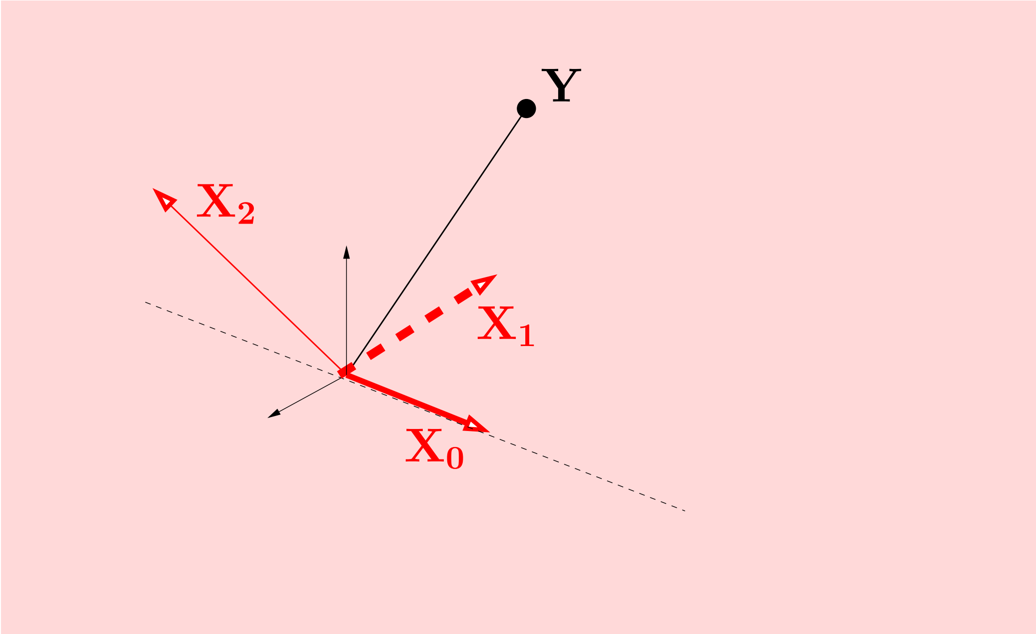


When $\mathbf{X_0}, \mathbf{X_1}, \mathbf{X_2}$ are linearly independent,

Linear System

Reconstructing $\mathbf{Y} \in \mathbb{R}^N$ using $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$ vectors of $\mathbb{R}^N$.

- Our ability to approximate $\mathbf{Y}$ depends implicitly on the space spanned by $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$



$\mathbf{Y}$ is in their span since the space is of dimension 3

Linear System

Reconstructing $\mathbf{Y} \in \mathbb{R}^N$ using $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$ vectors of $\mathbb{R}^N$.

- Our ability to approximate $\mathbf{Y}$ depends implicitly on the space spanned by $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$

The dimension of that space is $\mathbf{Rank(X)}$, the rank of $\mathbf{X}$

$$\mathbf{Rank}(\mathbf{X}) \leq \min(d+1, N).$$

Linear System

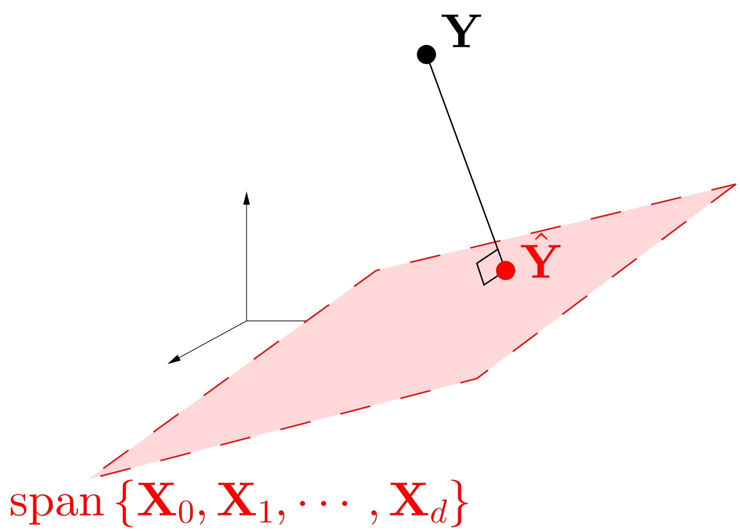
Three cases depending on $\mathbf{Rank}\,\mathbf{X}$ and d, N

1. $\mathbf{Rank}\,\mathbf{X} < N$. $d+1$ **column vectors do not span** $\mathbb{R}^N$
   - For arbitrary Y, there is **no solution** to $\alpha^T X = Y$
2. $\mathbf{Rank}\,\mathbf{X} = N$ and $d+1 > N$, **too many variables span the whole of** $\mathbb{R}^N$
   - **infinite** number of solutions to $\alpha^T X = Y$.
3. $\mathbf{Rank}\,\mathbf{X} = N$ and $d+1 = N$, **# variables = # observations**
   - Exact and unique solution: $\alpha = \mathbf{X}^{-1}\mathbf{Y}$ we have $\alpha^T X = Y$

In most applications, $d+1 \neq N$ so we are either in case 1 or 2

Case 1: $\text{Rank}\,\mathbf{X} < N$

- **no solution** to $\alpha^T X = Y$ (equivalently $\mathbf{X}\alpha = \mathbf{Y}$) in general case.
- What about the **orthogonal projection** of $\mathbf{Y}$ on the **image** of $\mathbf{X}$



- Namely the point $\hat{\mathbf{Y}}$ such that

$$\hat{\mathbf{Y}} = \underset{\mathbf{u} \in \text{span}\, \mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d}{\text{argmin}} \|\mathbf{Y} - \mathbf{u}\|.$$

Case 1: $\mathbf{Rank}\,\mathbf{X} < N$

Lemma 1. $\{\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d\}$ *is a* ***l.i.*** *family* $\Leftrightarrow \mathbf{X}^T\mathbf{X}$ *is invertible*

Case 1: $\text{Rank}\,\mathbf{X} < N$

- Computing the **projection** $\hat{\omega}$ of a point ω on a **subspace** V is well understood.
- In particular, if $(\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d)$ is a **basis** of $\text{span}\{\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d\}$...

(that is $\{\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d\}$ is a **linearly independent** family)

... then $(\mathbf{X}^T\mathbf{X})$ is invertible and ...

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

- This gives us the α vector of weights we are looking for:

$$\hat{\mathbf{Y}} = \mathbf{X}\underbrace{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}}_{\hat{\boldsymbol{\alpha}}} = \mathbf{X}\hat{\boldsymbol{\alpha}} \approx \mathbf{Y} \text{ or } \hat{\boldsymbol{\alpha}}^T X = Y$$

- What can go wrong?

Case 1: Rank $\mathbf{X} < N$

- If $\mathbf{X}^T\mathbf{X}$ is invertible,
$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$
- If $\mathbf{X}^T\mathbf{X}$ is not invertible... we have a problem.
- If $\mathbf{X}^T\mathbf{X}$'s condition number
$$\frac{\lambda_{\max}(\mathbf{X}^T\mathbf{X})}{\lambda_{\min}(\mathbf{X}^T\mathbf{X})},$$
is very large, a small change in $\mathbf{Y}$ can cause dramatic changes in α.
- In this case the linear system is said to be **badly conditioned**...
- Using the formula
$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$
might return garbage as can be seen in the following Matlab example.

Case 2: $\text{Rank}\,\mathbf{X} = N$ and $d + 1 > N$

high-dimensional low-sample setting

- **Ill-posed inverse problem**, the set

$$\{\alpha \in \mathbb{R}^d \quad | \quad \mathbf{X}\alpha = \mathbf{Y}\}$$

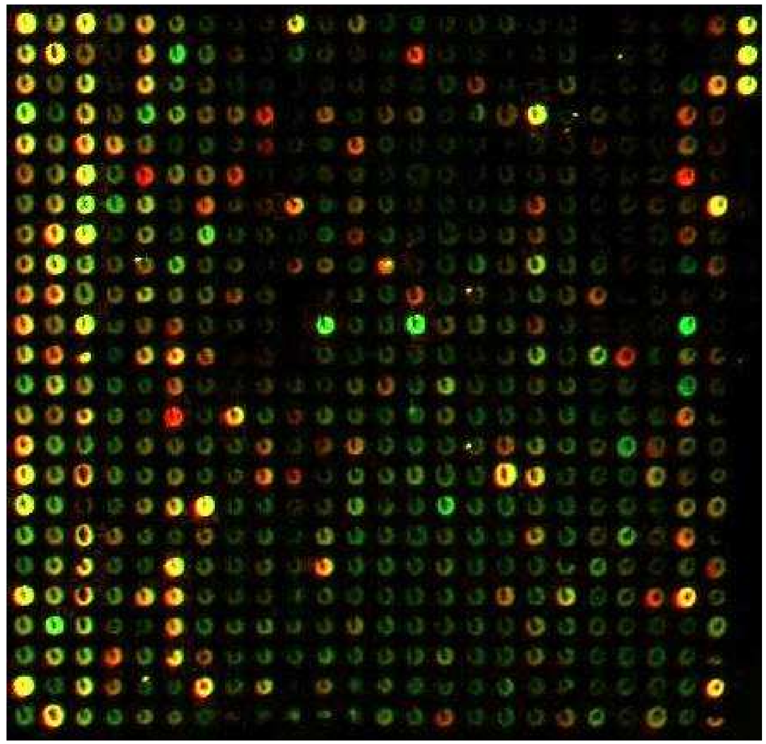
  is a whole **vector space**. We need to choose **one** from **many admissible** points.

- When does this happen?
  - High-dimensional low-sample case (DNA chips, multimedia *etc.*)
- How to solve for this?
  - Use something called regularization.

A practical perspective:
Colinearity and Overfitting

A Few High-dimensions Low sample settings

- DNA chips are very long vectors of measurements, one for each gene



- Task: regress a health-related variable against gene expression levels

Image:http://bioinfo.cs.technion.ac.il/projects/Kahana-Navon/DNA-chips.htm

A Few High-dimensions Low sample settings

- Emails represented as bag-of-words $\text{email}_j = \begin{bmatrix} \vdots \\ \text{please} = 2 \\ \vdots \\ \text{send} = 1 \\ \vdots \\ \text{money} = 2 \\ \vdots \\ \text{assignment} = 0 \\ \vdots \end{bmatrix} \in \mathbb{R}^d$

- Task: regress probability that this is an email against bag-of-words

Image: http://clg.wlv.ac.uk/resources/junk-emails/index.php

Correlated Variables

- Suppose you run a real-estate company.

- For each apartment you have compiled a **few hundred** predictor variables, *e.g.*
  - distances to conv. store, pharmacy, supermarket, parking lot, *etc.*
  - distances to all main locations in Kansai
  - socio-economic variables of the neighboorhood
  - characteristics of the apartment
- Some are obviously **correlated** (correlated= "almost" colinear)
  - distance to Post Office / distance to Post ATM
- In that case, we may have some problems (Matlab example)

Source: http://realestate.yahoo.co.jp/

Overfitting

- Given d variables (including constant variable), consider the least squares criterion

$$L_d\left(\alpha_1, \cdots, \alpha_d\right) = \sum_{i=1}^{j} \left(y_j - \sum_{i=1}^{d} \alpha_i x_{i,j} \right)^2$$

- Add **any** variable vector $\boldsymbol{x_{d+1,j}}, j = 1, \cdots, N$, and define

$$L_{d+1}(\alpha_1, \cdots, \alpha_d, \boldsymbol{\alpha_{d+1}}) = \sum_{i=1}^{j} \left(y_j - \sum_{i=1}^{d} \alpha_i x_{i,j} - \boldsymbol{\alpha_{d+1} x_{d+1,j}} \right)^2$$

Overfitting

- Given d variables (including constant variable), consider the least squares criterion

$$L_d(\alpha_1, \cdots, \alpha_d) = \sum_{i=1}^{j} \left(y_j - \sum_{i=1}^{d} \alpha_i x_{i,j} \right)^2$$

- Add **any** variable vector $\boldsymbol{x_{d+1,j}}, j = 1, \cdots, N$, and define

$$L_{d+1}(\alpha_1, \cdots, \alpha_d, \boldsymbol{\alpha_{d+1}}) = \sum_{i=1}^{j} \left(y_j - \sum_{i=1}^{d} \alpha_i x_{i,j} - \boldsymbol{\alpha_{d+1} x_{d+1,j}} \right)^2$$

THEN $\min_{\alpha \in \mathbb{R}^{d+1}} L_{d+\mathbf{1}(\alpha)} \leq \min_{\alpha \in \mathbb{R}^d} L_d(\alpha)$

Overfitting

- Given d variables (including constant variable), consider the least squares criterion

$$L_d\left(\alpha_1, \cdots, \alpha_d\right) = \sum_{i=1}^{j}\left(y_j - \sum_{i=1}^{d}\alpha_i x_{i,j}\right)^2$$

- Add **any** variable vector $\boldsymbol{x_{d+1,j}}, j = 1, \cdots, N$, and define

$$L_{d+1}(\alpha_1, \cdots, \alpha_d, \boldsymbol{\alpha_{d+1}}) = \sum_{i=1}^{j}\left(y_j - \sum_{i=1}^{d}\alpha_i x_{i,j} - \boldsymbol{\alpha_{d+1} x_{d+1,j}}\right)^2$$

Then $\min_{\alpha\in\mathbb{R}^{d+1}} L_{d+\mathbf{1}(\alpha)} \leq \min_{\alpha\in\mathbb{R}^d} L_d(\alpha)$

why? $L_d\left(\alpha_1, \cdots, \alpha_d\right) = L_{d+\mathbf{1}}\left(\alpha_1, \cdots, \alpha_d, \mathbf{0}\right)$

Overfitting

- Given d variables (including constant variable), consider the least squares criterion

$$L_d(\alpha_1, \cdots, \alpha_d) = \sum_{i=1}^{j} \left(y_j - \sum_{i=1}^{d} \alpha_i x_{i,j} \right)^2$$

- Add **any** variable vector $\boldsymbol{x_{d+1,j}}, j = 1, \cdots, N$, and define

$$L_{d+1}(\alpha_1, \cdots, \alpha_d, \boldsymbol{\alpha_{d+1}}) = \sum_{i=1}^{j} \left(y_j - \sum_{i=1}^{d} \alpha_i x_{i,j} - \boldsymbol{\alpha_{d+1} x_{d+1,j}} \right)^2$$

Then $\min_{\alpha \in \mathbb{R}^{d+1}} L_{d+\mathbf{1}(\alpha)} \leq \min_{\alpha \in \mathbb{R}^d} L_d(\alpha)$

why? $L_d(\alpha_1, \cdots, \alpha_d) = L_{d+\mathbf{1}}(\alpha_1, \cdots, \alpha_d, \mathbf{0})$

Residual-sum-of-squares goes down... but is it **relevant** to add variables?

Occam's razor formalization of overfitting

Minimizing least-squares (RSS) is **not clever enough**.
We need **another idea** to avoid **overfitting**.

- **Occam's razor**:*lex parsimoniae*

- **law of parsimony**: principle that recommends selecting the hypothesis that makes the fewest assumptions.

one should always opt for an explanation in terms of the fewest possible causes, factors, or variables.

Wikipedia: William of Ockham, born 1287- died 1347

Advanced Regression Techniques

Quick Reminder on Vector Norms

- For a vector $\mathbf{a} \in \mathbb{R}^d$, the Euclidian norm is the quantity

$$\|\mathbf{a}\|_2 = \sqrt{\sum_{i=1}^{d} a_i^2}.$$

- More generally, the q-norm is for $q > 0$,

$$\|\mathbf{a}\|_q = \left(\sum_{i=1}^{d} |a_i|^q\right)^{\frac{1}{q}}.$$

- In particular for $q = 1$,

$$\|\mathbf{a}\|_1 = \sum_{i=1}^{d} |a_i|$$

- In the limit $q \to \infty$ and $q \to 0$,

$$\|\mathbf{a}\|_\infty = \max_{i=1,\cdots,d} |a_i|. \quad \|\mathbf{a}\|_0 = \#\{i | a_i \neq 0\}.$$

Tikhonov Regularization '43 - Ridge Regression '62

- Tikhonov's motivation : solve **ill-posed inverse problems** by **regularization**
- If $\min_\alpha L(\alpha)$ is achieved on many points... consider

$$\min_\alpha L(\alpha) + \lambda\|\alpha\|_2^2$$

- We can show that this leads to selecting

$$\hat{\alpha} = (\mathbf{X}^T\mathbf{X} + \boldsymbol{\lambda I_{d+1}})^{-1}\mathbf{X}\mathbf{Y}$$

- The condition number has changed to

$$\frac{\lambda_{\max}(\mathbf{X}^T\mathbf{X}) + \lambda}{\lambda_{\min}(\mathbf{X}^T\mathbf{X}) + \lambda}.$$

Subset selection : Exhaustive Search

- Following Ockham's razor, ideally we would like to know for any value p

$$\min_{\alpha, \|\alpha\|_0 = p} L(\alpha)$$

- $\rightarrow$ select the **best** vector α which **only** gives weights to p **variables**.
- $\rightarrow$ Find the **best** combination of p variables.

Practical Implementation

- For $p \leq n$, $\binom{n}{p}$ possible combinations of p variables.
- Brute force approach: generate $\binom{n}{p}$ regression problems and select the one that achieves the best RSS.

Impossible in practice with moderately large n and p... $\binom{30}{5} = 150.000$

Subset selection : Forward Search

Since the **exact** search is **intractable in practice**, consider the **forward** heuristic

- **In Forward search**:
  - define $I_1 = \{0\}$.
  - given a set $I_k \subset \{0, \cdots, d\}$ of k variables, **what is the most informative variable one could** ***add*****?**
    - ▷ Compute for each variable i in $\{0, \cdots, d\} \setminus I_k$

$$t_{\boldsymbol{i}} = \min_{(\alpha_k)_{k \in I_k}, \boldsymbol{\alpha}} \quad \sum_{j=1}^{N} \left(y_j - \left(\sum_{k \in I_k} \alpha_k x_{k,j} + \boldsymbol{\alpha} x_{\boldsymbol{i},j} \right) \right)^2$$

    - ▷ Set $I_{k+1} = I_k \cup \{i^\star\}$ for any $i^\star$ such that $i^\star = \mathbf{min}\, t_i$.
    - ▷ $k = k + 1$ until desired number of variablse

Subset selection : Backward Search

... or the **backward** heuristic

- **In Backward search**:
  - define $I_d = \{0, 1, \cdots, n\}$.
  - given a set $I_k \subset \{0, \cdots, d\}$ of k variables, **what is the least informative variable one could *remove*?**
    - ▷ Compute for each variable i in I_k

$$t_{\boldsymbol{i}} = \min_{(\alpha_k)_{k \in I_k \setminus \{\boldsymbol{i}\}}} \sum_{j=1}^{N} \left\| y_j - \left(\sum_{k \in I_k \setminus \{\boldsymbol{i}\}} \alpha_k x_{k,j} \right) \right) \right\|^2$$

    - ▷ Set $I_{k-1} = I_k \setminus \{i^\star\}$ for any $i^\star$ such that $i^\star = \mathbf{max}\, t_i$.
    - ▷ $k = k - 1$ until desired number of variables

Subset selection : LASSO

Naive Least-squares

$$\min_{\alpha} L(\alpha)$$

Best fit with p variables (Occam!)

$$\min_{\alpha, \|\boldsymbol{\alpha}\|_0 = \boldsymbol{p}} L(\alpha)$$

Tikhonov regularized Least-squares

$$\min_{\alpha} L(\alpha) + \boldsymbol{\lambda}\|\boldsymbol{\alpha}\|_{\mathbf{2}}^{\mathbf{2}}$$

LASSO (least absolute shrinkage and selection operator)

$$\min_{\alpha} L(\alpha) + \boldsymbol{\lambda}\|\boldsymbol{\alpha}\|_{\mathbf{1}}$$