Autres interprétations

- Quelques mots sur l'utilisation de **polynômes** en plus haute dimension.
- Une perspective **géométrique**
- **co-linéarité des variables et surapprentissage**
- Une perspective **statistique** sur la régression MC.
- Quelques solutions: **techniques de régression avancées**
  - Sélection de variables
  - Régression "Ridge"
  - Lasso

Polynômes en plus haute dimension.

- Chaque observation a d variables, chaque observation $\mathbf{x} \in \mathbb{R}^d$,
  - L'espace des polynômes de degré p est généré par

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

    oú chaque monome $\mathbf{x}^{\mathbf{u}}$ est défini comme $x_1^{u_1} x_2^{u_2} \cdots x_d^{u_d}$

  - Récurrence pour la dimension de cet espace: $\dim_{p+1} = \dim_p + \binom{p+1}{d+p}$
- For $d = 20$ et $p = 5$, $1 + 20 + 210 + 1540 + 8855 + 42504 > 50.000$

Problème avec interpolation polynomiale en **haute-dimension** est l'**explosion** de variables (une pour chaque monome)

Géométrie

Fondamentaux

- Problème posé comme

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

et

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

- Nous cherchons un vecteur α tel que $\alpha^T X \approx Y$.

Fondamentaux

- Si cette expression est transpose, nous obtenons $X^T\alpha \approx Y^T$,

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

- Avec la notation $\mathbf{Y} = Y^T, \mathbf{X} = X^T$ et $\mathbf{X}_k$ pour la $(k+1)^{\text{th}}$ colonne de $\mathbf{X}$,

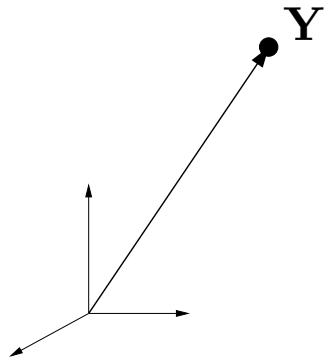
$$\sum_{k=0}^{d} \alpha_k \mathbf{X}_k \approx \mathbf{Y}$$

- $\mathbf{X}_k$ correspond à **toutes** les valeurs prises par la $k^{\text{ème}}$ variable.
- **Problem**: approximer/reconstruire $\mathbf{Y} \in \mathbb{R}^N$ en utilisant $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d \in \mathbb{R}^N$?

Système Linéaire

Approximer $\mathbf{Y} \in \mathbb{R}^N$ avec des vecteurs $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$ dans $\mathbb{R}^N$.

- La capacité d'approximer $\mathbf{Y}$ depend implicitement de l'espace vectoriel généré par $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$

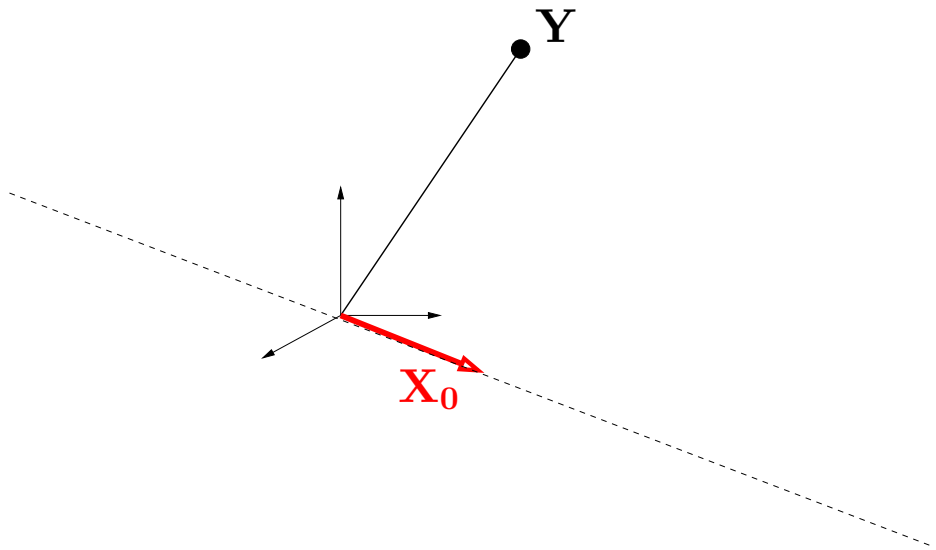


Considérons le vecteur $\mathbf{Y}$ de toutes les valeurs dépendantes.

Système Linéaire

Approximer $\mathbf{Y} \in \mathbb{R}^N$ avec des vecteurs $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$ dans $\mathbb{R}^N$.

- La capacité d'approximer $\mathbf{Y}$ depend implicitement de l'espace vectoriel généré par $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$

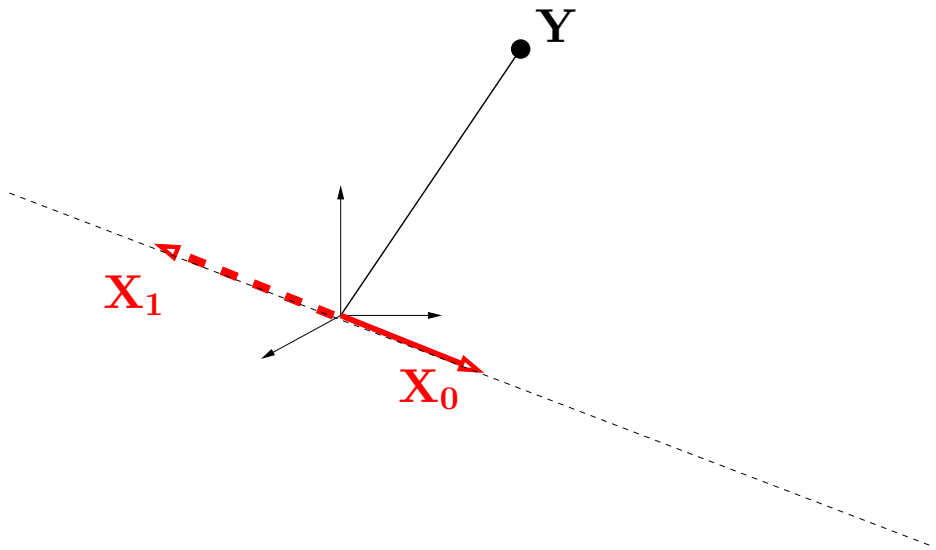


Affichons le premier régresseur $\mathbf{X_0}$...

Système Linéaire

Approximer $\mathbf{Y} \in \mathbb{R}^N$ avec des vecteurs $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$ dans $\mathbb{R}^N$.

- La capacité d'approximer $\mathbf{Y}$ depend implicitement de l'espace vectoriel généré par $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$

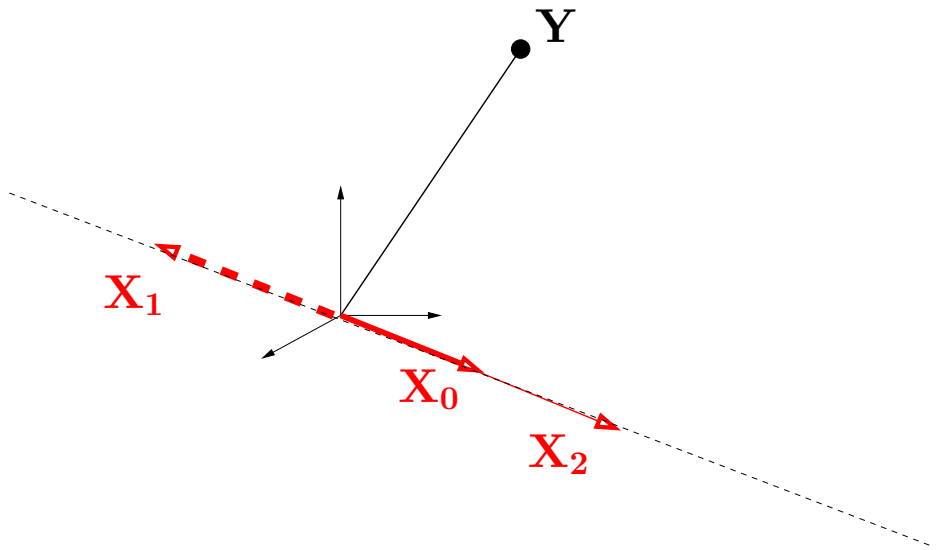


Faisons l'hypothêse que le régresseur suivant $\mathbf{X_1}$ est colinéaire avec $\mathbf{X_0}$...

Système Linéaire

Approximer $\mathbf{Y} \in \mathbb{R}^N$ avec des vecteurs $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$ dans $\mathbb{R}^N$.

- La capacité d'approximer $\mathbf{Y}$ depend implicitement de l'espace vectoriel généré par $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$

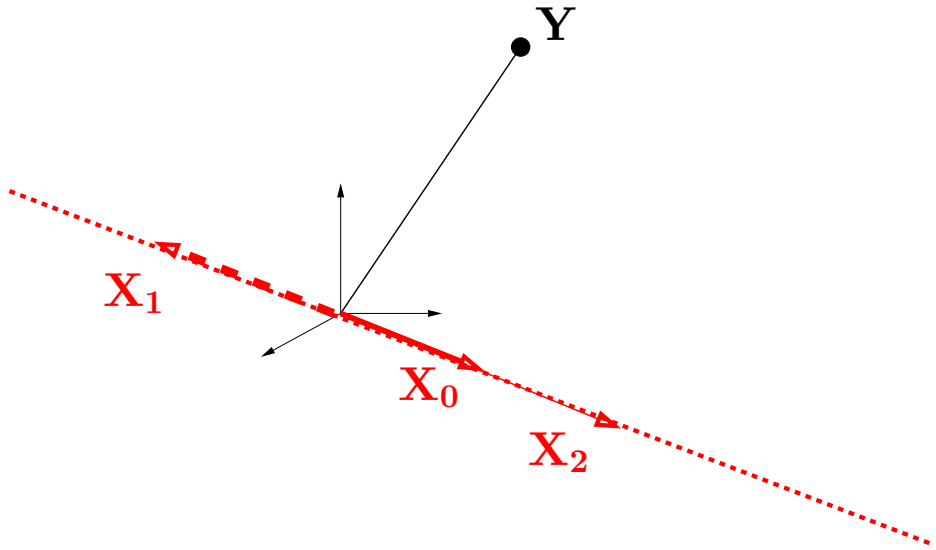


... ainsi que $\mathbf{X_2}$...

Système Linéaire

Approximer $\mathbf{Y} \in \mathbb{R}^N$ avec des vecteurs $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$ dans $\mathbb{R}^N$.

- La capacité d'approximer $\mathbf{Y}$ depend implicitement de l'espace vectoriel généré par $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$

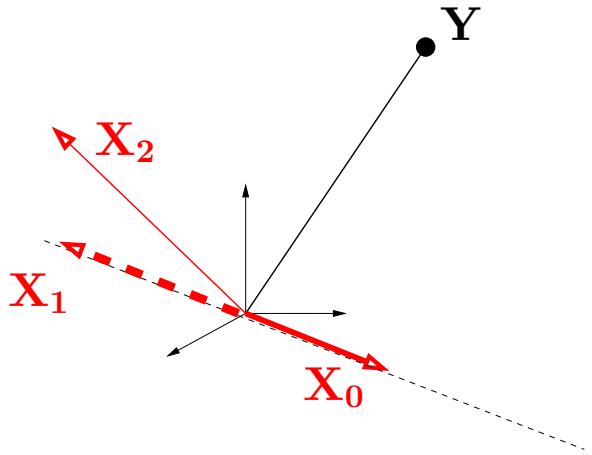


Peu d'options pour approximer $\mathbf{Y}$...

Système Linéaire

Approximer $\mathbf{Y} \in \mathbb{R}^N$ avec des vecteurs $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$ dans $\mathbb{R}^N$.

- La capacité d'approximer $\mathbf{Y}$ depend implicitement de l'espace vectoriel généré par $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$

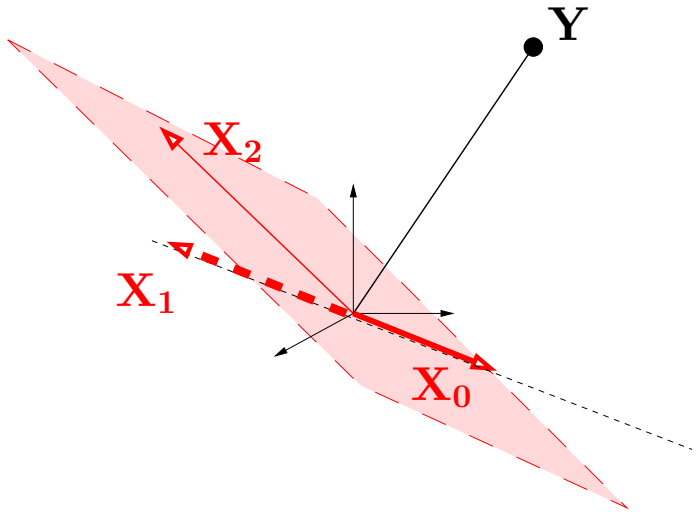


Supposons que $\mathbf{X_2}$ ne soit pas colinéaire avec $\mathbf{X_0}$.

Système Linéaire

Approximer $\mathbf{Y} \in \mathbb{R}^N$ avec des vecteurs $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$ dans $\mathbb{R}^N$.

- La capacité d'approximer $\mathbf{Y}$ depend implicitement de l'espace vectoriel généré par $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$

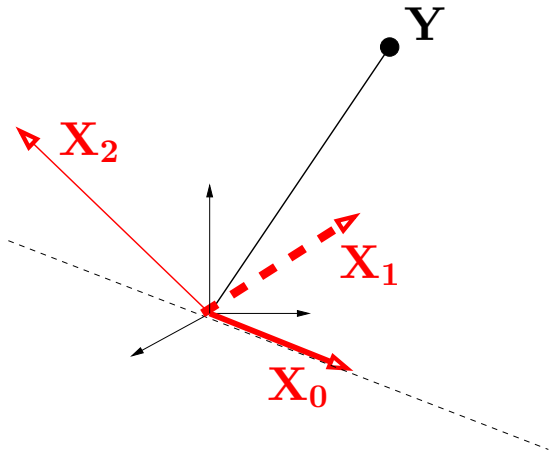


Ceci nous donne de nouvelles possibilités pour reconstruire $\mathbf{Y}$.

Système Linéaire

Approximer $\mathbf{Y} \in \mathbb{R}^N$ avec des vecteurs $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$ dans $\mathbb{R}^N$.

- La capacité d'approximer $\mathbf{Y}$ depend implicitement de l'espace vectoriel généré par $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$

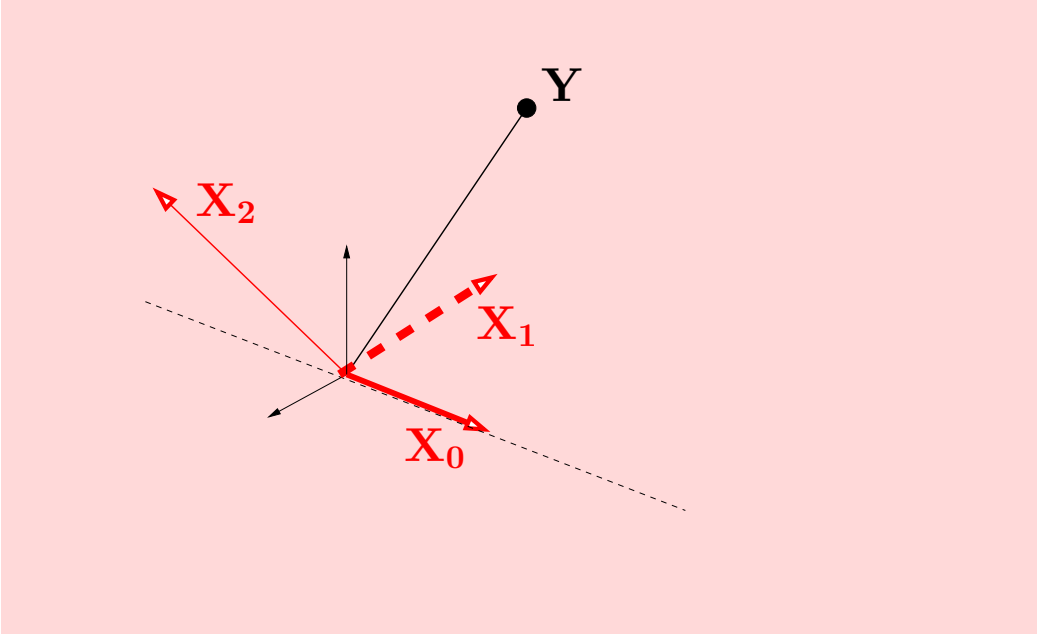


Quet $\mathbf{X_0}, \mathbf{X_1}, \mathbf{X_2}$ sont linéairement independants,

Système Linéaire

Approximer $\mathbf{Y} \in \mathbb{R}^N$ avec des vecteurs $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$ dans $\mathbb{R}^N$.

- La capacité d'approximer $\mathbf{Y}$ depend implicitement de l'espace vectoriel généré par $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$



$\mathbf{Y}$ est dans l'espace engendré

Système Linéaire

Approximer $\mathbf{Y} \in \mathbb{R}^N$ avec des vecteurs $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$ dans $\mathbb{R}^N$.

- La capacité d'approximer $\mathbf{Y}$ depend implicitement de l'espace vectoriel généré par $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$

La dimension de l'espace engendré est $\mathbf{Rank(X)}$, le rang de $\mathbf{X}$

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

Système Linéaire

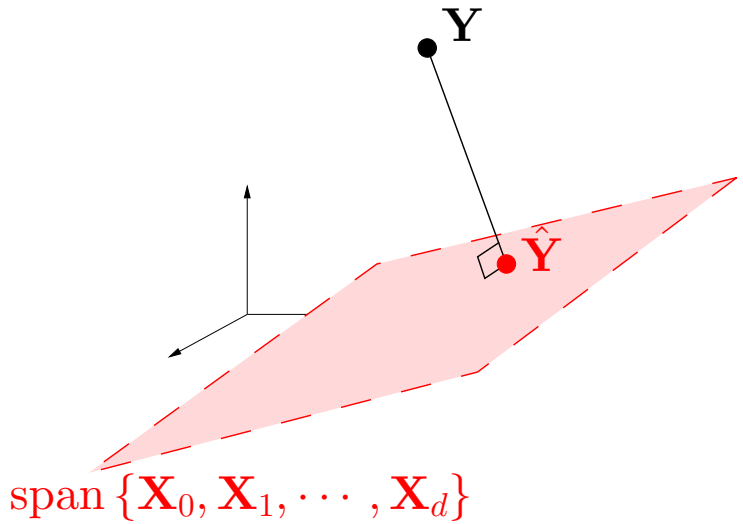
Trois cas, selon les valeurs de $\mathbf{Rank}\,\mathbf{X}$ et d, N

1. $\mathbf{Rank}\,\mathbf{X} < N$. $d+1$ **les vecteurs colonnes ne génèrent pas** $\mathbb{R}^N$
   - Pour un vecteur Y arbitraire, **pas de solution** à $\alpha^T X = Y$
2. $\mathbf{Rank}\,\mathbf{X} = N$ et $d+1 > N$, **bcp de variables, elles engendrent** $\mathbb{R}^N$
   - Nombre **infini** de solutions valables $\alpha^T X = Y$.
3. $\mathbf{Rank}\,\mathbf{X} = N$ et $d+1 = N$, **# variables = # observations**
   - Solution unique: $\alpha = \mathbf{X}^{-1}\mathbf{Y}$ we have $\alpha^T X = Y$

Dans toutes les applications, $d+1 \neq N$ et donc les cas 1 ou 2 sont pertinents.

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

- **Pas de solution** à $\alpha^T X = Y$ (équivalent à $\mathbf{X}\alpha = \mathbf{Y}$) dans le cas général.
- Quid de la **projection orthogonale** de $\mathbf{Y}$ sur l'**espace généré** par les $\mathbf{X}$?

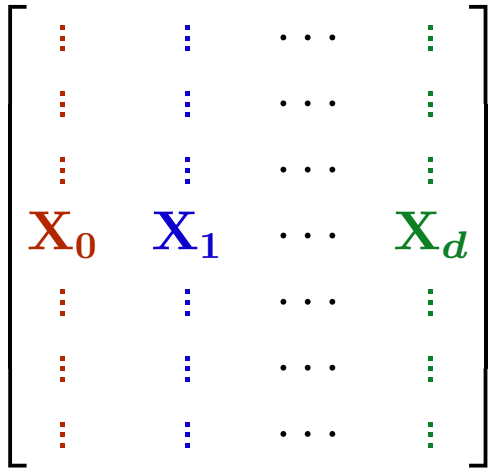


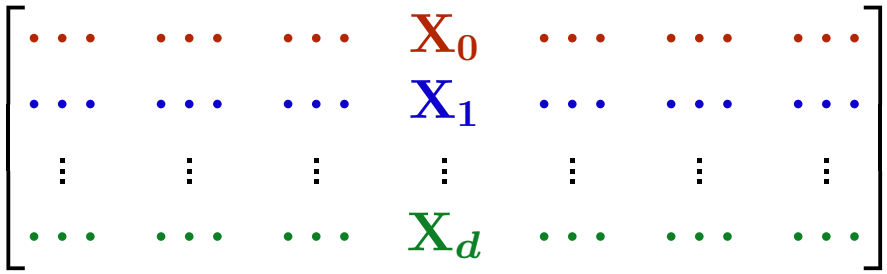
- Plus exactement, le point $\hat{\mathbf{Y}}$ tel que

$$\hat{\mathbf{Y}} = \underset{\mathbf{u} \in \mathbf{span}\, \mathbf{X_0}, \mathbf{X_1}, \cdots, \mathbf{X_d}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{u}\|.$$

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

Lemma 1. $\{\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d\}$ *est une famille* ***l.i.*** $\Leftrightarrow \mathbf{X}^T\mathbf{X}$ *est inversible*





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

- Facile de calculer la **projection** $\hat{\omega}$ d'un point ω sûr un **sous-espace** V.
- Si $(\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d)$ est une **base** de $\mathrm{span}\{\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d\}$...

(i.e. $\{\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d\}$ est une **famille linéairement indépendante**)

... alors $(\mathbf{X}^T\mathbf{X})$ est inversible et ...

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

- Cette projection nous permet d'obtenir α:

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

- Numériquement?

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

- Si $\mathbf{X}^T\mathbf{X}$ est inversible,

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

- Si $\mathbf{X}^T\mathbf{X}$ n'est pas inversible... problème.
- Si le conditionnement de $\mathbf{X}^T\mathbf{X}$,

$$\frac{\lambda_{\max}(\mathbf{X}^T\mathbf{X})}{\lambda_{\min}(\mathbf{X}^T\mathbf{X})},$$

  est très grand, une petit variation dans $\mathbf{Y}$ peut entraîner une grande variation dans α.
- Dans ce cas, le système linéaire est dit **mal conditionné**...
- La formule

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

  peut se révéler instable, comme décrit dans l'example numérique suivant.

Cas 2: $\mathbf{Rank\,X} = N$ et $d + 1 > N$

dimension élevé, échantillon petit

- **Problème inverse mal posé**: l'ensemble

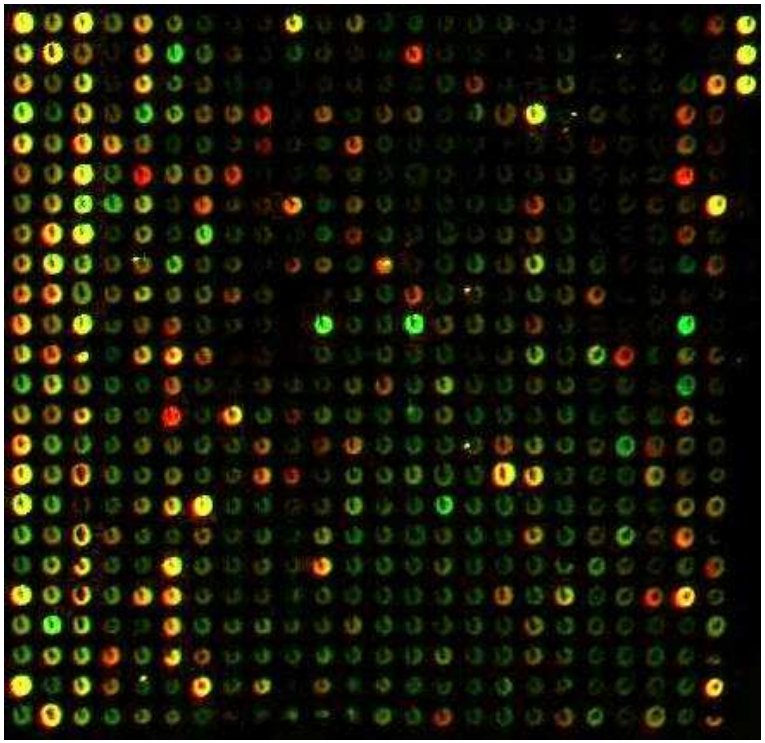
$$\{\alpha \in \mathbb{R}^d \quad | \quad \mathbf{X}\alpha = \mathbf{Y}\}$$

  est un **espace vectoriel**. Nous avons besoin de choisir **une** parmi **plusieurs** solutions **admissibles**.

- Quet cela se produit-il?
  - Senseurs très détaillés (puce ADN, scanners, $etc.$)
- Comment résoudre ce problème?
  - Antidote d'un problème mal posé = régularisation.

Quelques examples en haute dimension / peu d'échantillons

- Les puces ADN produisent des vecteurs mesurant l'activité de chaque gène à un instant donné

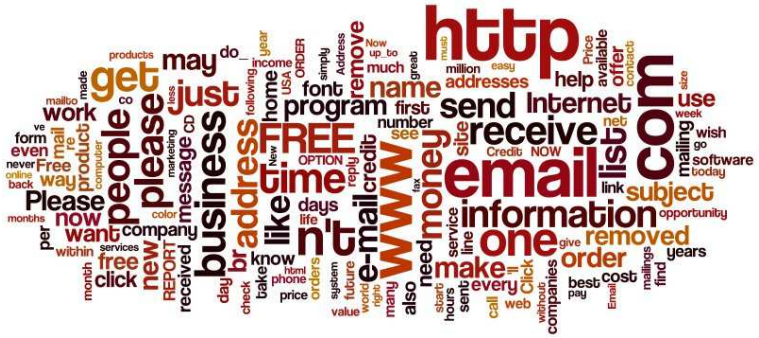


- Tâche: régression d'une variable liée à la santé du patient avec les niveaux d'expression.

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

Quelques examples en haute dimension / peu d'échantillons

- Emails représentés comme des sacs-de-mots $\text{email}_j = \begin{bmatrix} \vdots \\ \text{please} = 2 \\ \vdots \\ \text{send} = 1 \\ \vdots \\ \text{money} = 2 \\ \vdots \\ \text{assignment} = 0 \\ \vdots \end{bmatrix} \in \mathbb{R}^d$



- Tâche: prédire la proba que ce mail soit un spam en fonction du sac-de-mots

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

Variables corrélées

- Example: analyse du marché immobilier.

- Pour chaque apartement: plusieurs **centaines** de prédicteurs, $e.g.$
  - distance supermarché, pharmacie, parking, pressing, écoles $etc.$
  - charactéristiques socio-économiques du voisinage
  - charactéristiques de l'apartement
- Certaines seront **corrélées** (corrélées = "presque" colinéaires)
  - distance à la poste / distance à un distributeur
- Problèmes apparaissent dans ce cas.

Overfitting - Sur-apprentissage

- Etant données d variables (constante comprise), critère des moindres carrées:

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

- Ajoutons **n'importe quelle** variable $\boldsymbol{x_{d+1,j}}, j = 1, \cdots, N$, pour définir

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

Overfitting - Sur-apprentissage

- Etant données d variables (constante comprise), critère des moindres carrées:

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

- Ajoutons **n'importe quelle** variable $\boldsymbol{x_{d+1,j}}, j = 1, \cdots, N$, pour définir

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

- **Alors**

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

Overfitting - Sur-apprentissage

- Etant données d variables (constante comprise), critère des moindres carrées:

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

- Ajoutons **n'importe quelle** variable $\boldsymbol{x_{d+1,j}}, j = 1, \cdots, N$, pour définir

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

- **Alors**

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

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

Overfitting - Sur-apprentissage

- Etant données d variables (constante comprise), critère des moindres carrées:

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

- Ajoutons **n'importe quelle** variable $\boldsymbol{x_{d+1,j}}, j = 1, \cdots, N$, pour définir

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

- **Alors**

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

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

Somme des Résidus Quadratiques décroit... est-ce **pertinent** ?

Rasoir d'Occam et surapprentissage

Minimiser les moindres carrés (RSS) n'est pas **suffisant**.
Nous avons besoin d'une **autre idée** pour éviter le **surapprentissage**.

- **Rasoir d'Occam** :*lex parsimoniae*

- **loi de la parsimonie**: choisir l'hypothèse qui fait le minimum d'hypothèse parmi toutes celles qui expliquent correctement un phénomène.

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

Wikipedia: William de Ockham, born 1287- died 1347

Statistique & régression par MC

Statistique & régression par MC

- **Assumons** que la distribution d'une v.a. Y pour une observation **x** est déterminée ŧravers ω, β comme

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

- On peut réécrire cette relation comme

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

Statistique & régression par MC

- On peut réécrire cette relation comme

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

- Soit encore

$$\boldsymbol{y} \sim \mathcal{N}(\omega^T \mathbf{x} + \beta, \sigma),$$

But du statisticien: **Estimer** (ω, β) au vu d'observations aléatoires y, couplées aux observations non-aléatoires $\mathbf{x}$

Observations (i.i.d)

- Supposons que les vrais paramètres sont $\omega = \mathbf{w}, \beta = b$

Observations (i.i.d)

- Supposons que les vrais paramètres sont $\omega = \mathbf{w}, \beta = b$
- Dans ce cas, quelle serait la **probabilité** de **chaque** observation y_j sachant qu'elle est associée un $\mathbf{x}_j$?

Observations (i.i.d)

- Hypothèse: **En supposant que** $\omega = \mathbf{w}, \beta = b$, quelle serait la **probabilité** de **chaque** observation?:
  - Pour chaque couple, $(\mathbf{x}_j, y_j)$, $j = 1, \cdots, N$,

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

Observations (i.i.d)

- Hypothèse: **En supposant que** $\omega = \mathbf{w}, \beta = b$, quelle serait la **probabilité** de **chaque** observation?:
  - Pour chaque couple, $(\mathbf{x}_j, y_j)$, $j = 1, \cdots, N$,

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

  - Chaque mesure y_j est conditionellement **indépendante** des autres

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

Observations (i.i.d)

- Hypothèse: **En supposant que $\omega = \mathbf{w}, \beta = b$**, quelle serait la **probabilité** de **chaque** observation?:
  - For each couple $(\mathbf{x}_j, y_j)$, $j = 1, \cdots, N$,

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

  - Chaque mesure y_j est conditionellement **indépendante** des autres

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

  - A.K.A **vraisemblance** conditionelle des $\{(y_j)_{j=1,\cdots,N}\}$ en fonctions des $\{(\mathbf{x}_j)_{j=1,\cdots,N}\}$, a et b,

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

EMV des Paramètres de Régression

Ainsi, pour $\mathbf{w}, b$ donnés la **vraisemblance** des valeurs $\{y_j\}_{j=1,\cdots,N}$

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

EMV des Paramètres de Régression

Ainsi, pour $\mathbf{w}, b$ donnés la **vraisemblance** des valeurs $\{y_j\}_{j=1,\cdots,N}$

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

Utiliser la **vraisemblance** pour **estimer** $(\mathbf{w}, b)$ avec les données?

EMV des Paramètres de Régression

Ainsi, pour $\mathbf{w}, b$ donnés la **vraisemblance** des valeurs $\{y_j\}_{j=1,\cdots,N}$

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

...l'**EMV** sélectionne les valeurs $(\mathbf{w}, b)$ qui **maximisent** $\mathcal{L}(\mathbf{w}, b)$

EMV des Paramètres de Régression

Ainsi, pour $\mathbf{w}, b$ donnés la **vraisemblance** des valeurs $\{y_j\}_{j=1,\cdots,N}$

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

...l'**EMV** sélectionne les valeurs $(\mathbf{w}, b)$ qui **maximisent** $\mathcal{L}(\mathbf{w}, b)$

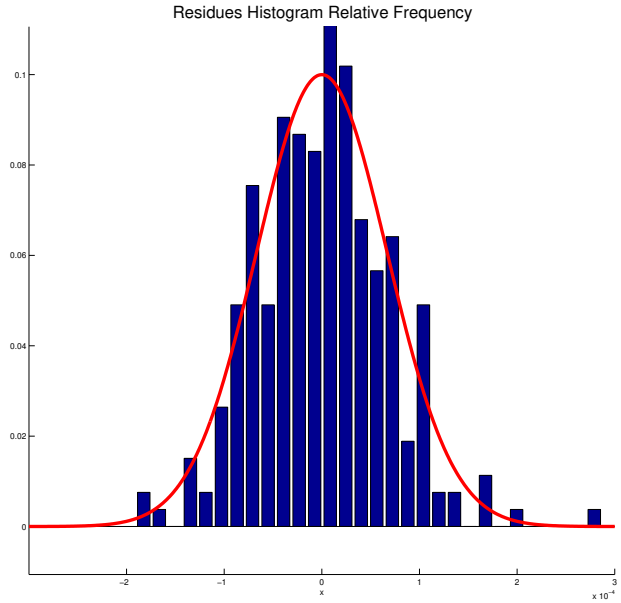
- **Puisque** $\max_{(\mathbf{w},b)} \mathcal{L}(\mathbf{w}, b) \Leftrightarrow \max_{(\mathbf{w},b)} \log \mathcal{L}(\mathbf{w}, b)$

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

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

Approche Statistique pour la Régression Linéaire

- Propriétés de l'EMV: convergence de $\hat{\alpha}_{\text{EMV}}$ vers $\alpha = (\omega, \beta)$?
- Intervalles de Confiance pour les coefficients,
- Tests pour verifier que les hypothèse sur les résidus réalistes



- Approches bayésiennes: au lieu de ne considérer qu'**une** estimation du paramètre optimal $\theta = (\omega, \beta)$, on étudie une densité (a posteriori) des paramètres $\theta = (\mathbf{w}, b)$ à partir d'une distribution a priori et des données.