Advanced Statistics I : Gaussian Linear Model (and beyond)

Aurélien Garivier

CNRS / Telecom ParisTech

Centrale

Outline

Discrete and continuous distributions

- Discrete distribution $P = \sum_{i=1}^{n} p_i \delta_{x_i}$
  Ex: Binomial, Poisson distributions
- Continuous distribution $Q(dx) = f(x)dx$.
  Ex: Exponential distribution
- A distribution can be neither purely discrete, nor purely continuous !
  Ex: $Z = \min\{X, 1\}$, where $X \sim \mathcal{E}(\lambda)$

Descriptive properties

$$
\begin{aligned}
\text{Expectation} \quad & \mu = \mathbb{E}[X] \\
\text{Variance} \quad & \sigma^2 = \mathbb{E}\left[(X-\mu)^2\right] = \mathbb{E}\left[X^2\right] - \mu^2 \\
\text{Skewness} \quad & \gamma = \frac{\mathbb{E}[(X-\mu)^3]}{\sigma^3} \\
\text{Kurtosis} \quad & \kappa = \frac{\mathbb{E}[(X-\mu)^4]}{\sigma^4} - 3
\end{aligned}
$$

Some remarkable distributions

- Normal: scale- and shift- stable family
- Chi-2: if $X_1, \ldots, X_n$ is a $\mathcal{N}(0,1)$-sample, then

$$Z \sim X_1^2 + \cdots + X_n^2 \sim \chi^2(n)$$

- Student: if $X \sim \mathcal{N}(0,1)$ is independent of $Z \sim \chi^2(n)$, then

$$T = \frac{X}{\sqrt{Z/n}} \sim \mathcal{T}(n)$$

- Fischer: if $X \sim \chi^2(n)$ is independent of $Y \sim \chi^2(m)$, then

$$F = \frac{X/n}{Y/m} \sim \mathcal{F}(n,m)$$

Empirical Distribution and statistics

- Let $X_1, \ldots, X_n$ be a P-sample
- Empirical mean: $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$
- Empirical variance:

$$\Sigma_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - (\bar{X}_n)^2$$

- Empirical distribution $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$
- Unbiased variance estimator

$$\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n-1}\left(\sum_{i=1}^n X_i^2 - n(\bar{X}_n)^2\right)$$

- If $P = \mathcal{N}(0, \sigma^2)$, using Cochran's Theorem we get

$$\bar{X}_n \sim \mathcal{N}(0, \sigma^2/n) \quad \perp\!\!\!\perp \quad \sum_{i=1}^n (X_i - \bar{X}_n)^2 \sim \sigma^2\chi^2(n-1)$$

Convergence properties

Theorem (LLN)

If $\mathbb{E}[|X_i|] < \infty$, *then (in probability, almost surely)*

$$\bar{X}_n \longrightarrow \mu$$

- Application to S_n^2 and $\hat{\sigma}_n^2$, etc. . .
- Convergence of the empirical distribution $P_n \rightharpoonup P$ under appropriate hypotheses

Central Limit Theorem

Theorem (CLT)

If $\mathbb{E}[X_i^2] < \infty$,

$$\frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma} \rightharpoonup \mathcal{N}(0,1)$$

- By Slutsky's Lemma,

$$\frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\hat{\sigma}_n} \rightharpoonup \mathcal{N}(0,1)$$

- Student statistic: if $X_i \sim \mathcal{N}(\mu, \sigma^2)$,

$$\frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\hat{\sigma}_n} \sim \mathcal{T}(n-1)$$

Confidence interval for the mean

- if σ is known,

$$I_\alpha(\mu) = \left[\bar{X}_n \pm \frac{\sigma \phi_{1-\alpha/2}}{\sqrt{n}}\right]$$

- if σ is unknown,

$$I_\alpha(\mu) = \left[\bar{X}_n \pm \frac{\hat{\sigma}_n t^{n-1}_{1-\alpha/2}}{\sqrt{n}}\right]$$

Confidence interval for the variance

- if μ is known, as $\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2(n)$

$$\Longrightarrow \; I_\alpha(\sigma^2) = \left[\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^n_{1-\alpha/2}}, \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^n_{\alpha/2}}\right]$$

- if μ is unknown, as $\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}_n}{\sigma}\right)^2 \sim \chi^2(n-1)$

$$\Longrightarrow \; I_\alpha(\sigma^2) = \left[\frac{\hat{\sigma}_n^2}{\chi^{n-1}_{1-\alpha/2}/(n-1)}, \frac{\hat{\sigma}_n^2}{\chi^{n-1}_{\alpha/2}/(n-1)}\right]$$

Comparison of two variances

- let $X_{1,1}, \ldots, X_{1,n_1}$ be a sample $\mathcal{N}(\mu_1, \sigma_1^2)$, and $X_{2,1}, \ldots, X_{2,n_2}$ be an independant sample $\mathcal{N}(\mu_2, \sigma_2^2)$,
- in order to test H_0 : "$\sigma_1 = \sigma_2$" versus H_1 : "$\sigma_1 \neq \sigma_2$", use statistic

$$F = \frac{\hat{\sigma_1}^2}{\hat{\sigma_2}^2} \sim_{H_0} F(n_1 - 1, n_2 - 1)$$

Comparison of two means

Theorem

- *let* $X_{1,1}, \dots, X_{1,n_1}$ *be a sample* $\mathcal{N}(\mu_1, \sigma^2)$, *and* $X_{2,1}, \dots, X_{2,n_2}$ *be an independent sample* $\mathcal{N}(\mu_2, \sigma^2)$,
- *To estimate the common variance, use*

$$\hat{\sigma}_{12} = \frac{1}{n_1 + n_2 - 2} \left(\sum_{i=1}^{n_1} (X_{1,i} - \bar{X}_1)^2 + \sum_{i=1}^{n_2} (X_{2,i} - \bar{X}_2)^2 \right)$$
$$\sim \frac{\sigma^2}{n_1 + n_2 - 2} \chi^2(n_1 + n_2 - 2)$$

- *in order to test* H_0 : "$\mu_1 = \mu_2$" *versus* H_1 : "$\mu_1 \neq \mu_2$", *use statistic*

$$T = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{(\bar{X}_1 - \bar{X}_2)}{\hat{\sigma}_{12}} \sim_{H_0} T(n_1 + n_2 - 2)$$

Outline

Generic formulation

- $Y_i = \alpha_1 x_i^1 + \cdots + \alpha_p x_i^p + \sigma Z_i,\ Z_i \sim \mathcal{N}(0,1)$
- Matrice form:

$$Y = X\theta + \sigma Z, \quad Z \sim \mathcal{N}(0_n, I_n)$$

- Ex: ANOVA, regression, rupture in time series

Cochran's Theorem

Theorem

- *let* $X = (X_1, \ldots, X_n)$ *be a standard centered normal sample*
- *let* $E_1, \ldots, E_p$ *be a decomposition of* R^n *by two-by-two orthogonal subspaces of dimensions* $\dim E_j = d_j$
- *for* $1 \leq i \leq p$*, let* $v_1^i, \ldots, v_{j_i}^i$ *be an orhogonal basis of* E_j

Then

- *the components of* X *in base* $(v_1, \ldots, v_n)$ *form another standard centered normal sample*
- *the random vectors* $X_{E_1}, \ldots, X_{E_p}$ *obtained by projecting* X *on* $E_1, \ldots, E_p$ *are independent*
- *so are* $\|X_{E_1}\|, \ldots, \|X_{E_p}\|$*, and they satisfy:*

$$\|X_{E_i}\|^2 \sim \chi^2(d_i)$$

Generic solution

Theorem

- *The Maximum-Likelihood estimator and the least-square estimator are given by:*

$$\hat{\theta} = \left({}^t XX\right)^{-1} {}^t XY \sim \mathcal{N}(\theta, \sigma^2 \left({}^t XX\right)^{-1})$$

- *The variance σ^2 is estimated (without bias) by:*

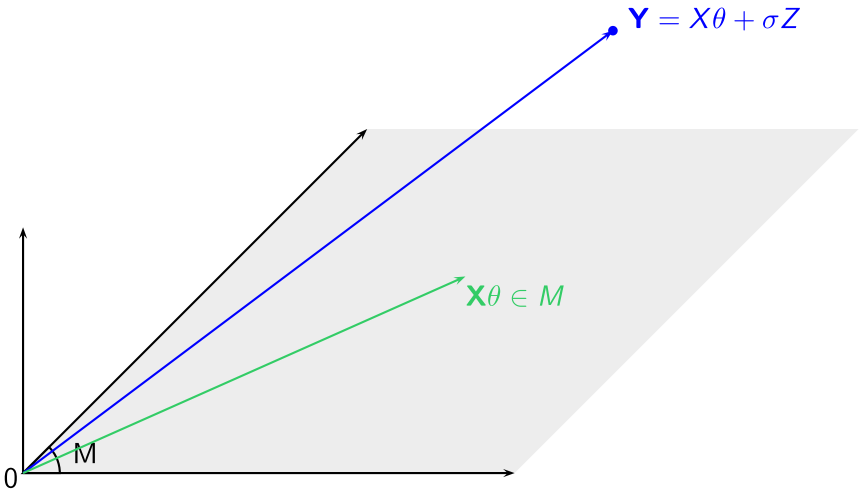
$$\hat{\sigma}^2 = \frac{\|Y - X\hat{\theta}\|^2}{n-p} \sim \frac{\sigma^2}{n-p}\chi^2(n-p)$$

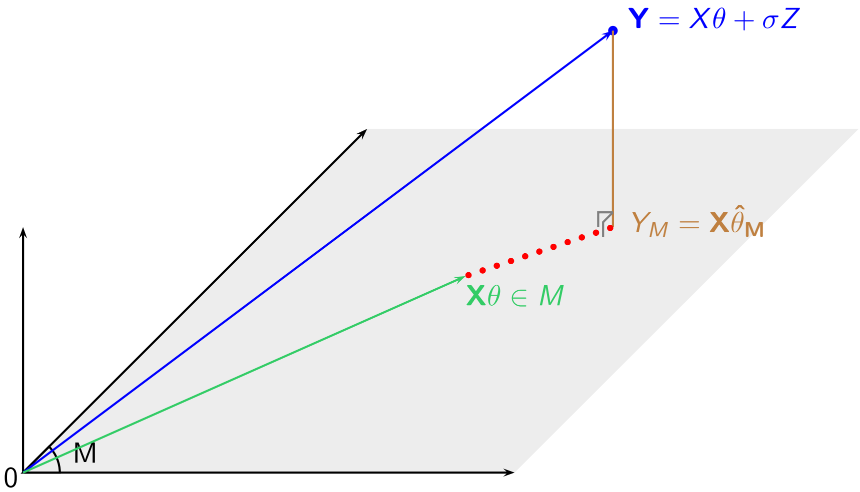
- *$\hat{\theta}$ and σ^2 are independent*

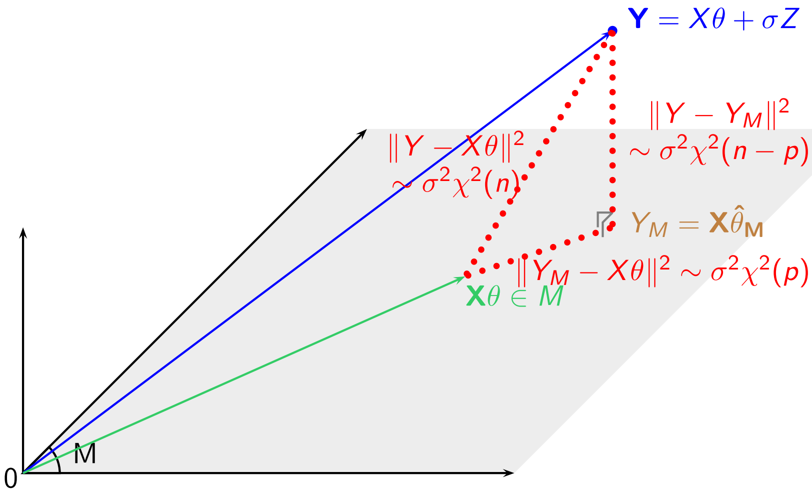
Incremental Gramm-Schmidt procedure

Theorem (Gauss-Markov)

$\hat{\theta}$ has minimal variance among all linear unbiased estimators of θ


$\mathbf{Y} = X\theta + \sigma Z$
$\mathbf{X}\theta \in M$
M
0


$\mathbf{Y} = X\theta + \sigma Z$
$Y_M = \mathbf{X}\hat{\theta}_{\mathbf{M}}$
$\mathbf{X}\theta \in M$
M
0


$\mathbf{Y} = X\theta + \sigma Z$
$\|Y - Y_M\|^2$
$\sim \sigma^2\chi^2(n-p)$
$\|Y - X\theta\|^2$
$\sim \sigma^2\chi^2(n)$
$Y_M = \mathbf{X}\hat{\theta}_{\mathbf{M}}$
$\|Y_M - X\theta\|^2 \sim \sigma^2\chi^2(p)$
$\mathbf{X}\theta \in M$
M
0

Simple regression: $Y_i = \alpha + \beta x_i + \sigma Z_i$

Theorem

- *The ML-estimators are given by:*

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x} \sim \mathcal{N}\left(\alpha, \frac{\sigma^2 \mathbb{E}[x^2]}{n\mathrm{Var}(x)}\right)$$

$$\hat{\beta} = \frac{\mathrm{Cov}(x, Y)}{\mathrm{Var}(x)} \sim \mathcal{N}\left(\beta, \frac{\sigma^2}{n\mathrm{Var}(x)}\right)$$

- *They are correlated:* $Cov\left(\hat{\alpha}, \hat{\beta}\right) = -\frac{\sigma^2 \bar{x}}{n\mathrm{Var}[x]}$
- *The variance can be estimated by:*

$$\hat{\sigma}_n^2 = \frac{1}{n-2} \sum (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2 \sim \frac{\sigma^2}{n-2} \chi^2(n-2)$$

- *Smart reparameterization* $Y_i = \delta + \beta(x_i - \bar{x}) + \sigma Z_i$

Polynomial regression

$$\begin{pmatrix} Y \end{pmatrix} = \begin{pmatrix} 1 & x & x^2 & \dots & x^p \end{pmatrix} \times \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{pmatrix}$$

Can also be used for exponential growth models
$y_i = \exp(ax_i + b_i + \epsilon_i)$ to determine β such that $\mathbb{E}[Y] = \alpha X^\beta, \dots$

Outline

Student test on a regressor

Theorem

In order to test $H_0 = ``\theta_k = a''$ *versus* $H_1 = ``\theta_k \neq a''$

- *estimate the variance of* $\hat{\theta}_k$ *by*

$$\hat{\sigma}^2\left(\hat{\beta}_k\right) = \hat{\sigma}^2\left\{({}^tXX)^{-1}\right\}_{k,k}$$

- *use the statistic*

$$T = \frac{\hat{\beta}_k - a}{\hat{\sigma}\left(\hat{\beta}_k\right)} \sim_{H_0} T(n-p)$$

- *Generalization: to test* $H_0 = ``{}^tb\theta = a''$ *versus* $H_1 = ``{}^tb\theta \neq a''$, *use*

$$T = \frac{{}^tb\hat{\beta} - a}{\hat{\sigma}\sqrt{{}^tb({}^tXX)^{-1}b}} \sim_{H_0} T(n-p)$$

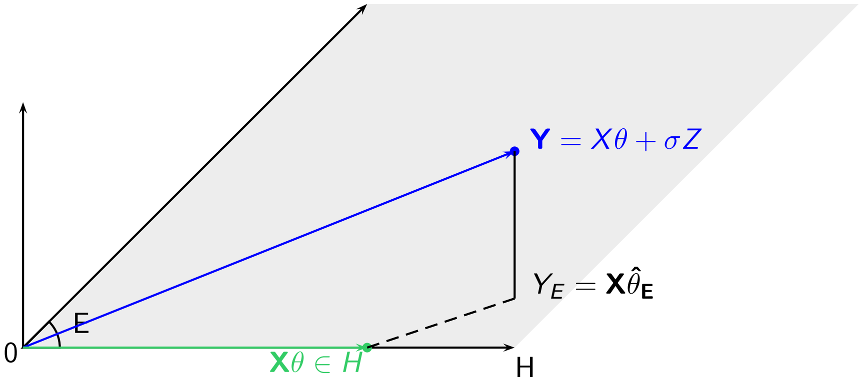
Fischer Test "model vs submodel"

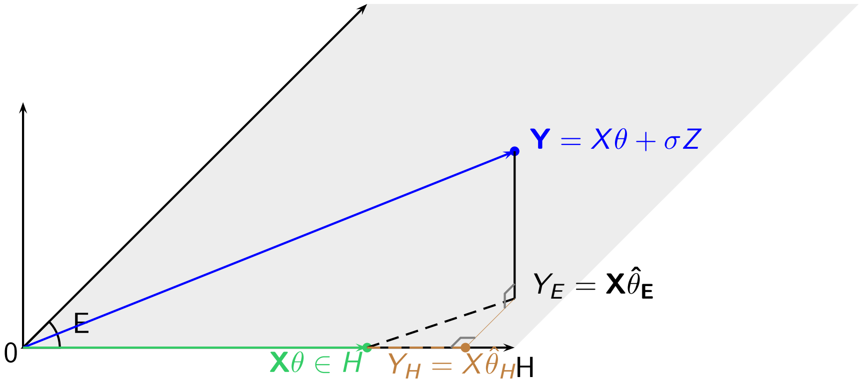
Theorem

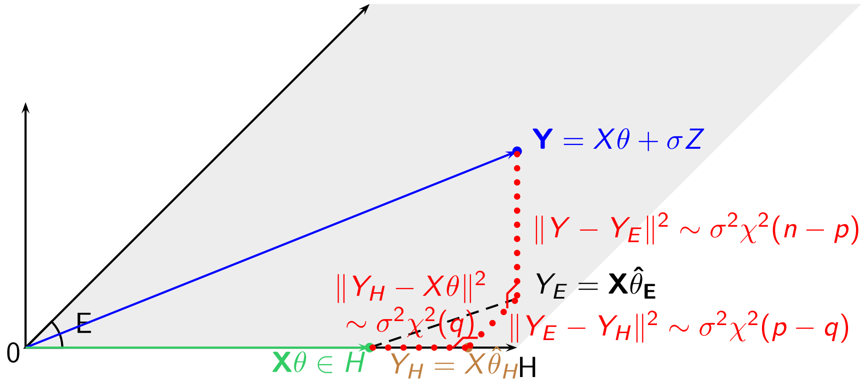
- *let* $H \subset E \subset \mathbb{R}^n$, $\dim H = q, \dim E = p$
- *to test* $H_0 =$ "$\theta \in H$" *versus* $H_1 =$ "$\theta \in E \setminus H$", *use the statistic*

$$F = \frac{\|Y_E - Y_H\|^2/(p-q)}{\|Y - Y_E\|^2/(n-p)} \sim_{H_0} \mathcal{F}(p-q, n-p)$$

- *reject if* $F > \mathcal{F}_{1-\alpha}^{p-q,n-p}$


$\mathbf{Y} = X\theta + \sigma Z$
$Y_E = \mathbf{X}\hat{\theta}_\mathbf{E}$
E
0
$\mathbf{X}\theta \in H$
H


$\mathbf{Y} = X\theta + \sigma Z$
$Y_E = \mathbf{X}\hat{\theta}_{\mathbf{E}}$
E
0
$\mathbf{X}\theta \in H$
$Y_H = X\hat{\theta}_H$
H


$\mathbf{Y} = X\theta + \sigma Z$
$\|Y - Y_E\|^2 \sim \sigma^2\chi^2(n-p)$
$Y_E = \mathbf{X}\hat{\theta}_{\mathbf{E}}$
$\|Y_H - X\theta\|^2$
$\sim \sigma^2\chi^2(q)$
$\|Y_E - Y_H\|^2 \sim \sigma^2\chi^2(p-q)$
E
0
$\mathbf{X}\theta \in H$
$Y_H = X\hat{\theta}_H$
H

SSS-notations and R^2

- For a model M (relative to a matrix X), define

$$\textit{total} \text{ variance} \quad SSY = \|Y - \bar{Y}1_n\|^2 = SSE(1_n)$$
$$\textit{residual} \text{ variance} \quad SSE(M) = \|Y - X\hat{\theta}\|^2$$
$$\textit{explained} \text{ variance} \quad SSR(M) = \|X\hat{\theta} - \bar{Y}1_n\|^2$$

$$SSY = SSE(M) + SSR(M)$$

- The quality of fit is quantified by

$$R^2(M) = \frac{SSR(M)}{SSY}$$

- The Fischer statistic can be written:

$$F = \frac{(SSE(H) - SSE(E))/(p-q)}{SSE(E)/(n-p)} = \frac{n - \dim(E)}{\dim(E) - \dim(H)} \times \frac{R^2(E) - R^2(H)}{1 - R^2(E)}$$

ANOVA

- The model can be written:

$$Y_{i,k} = \theta_i + \sigma\epsilon_{i,k}, 1 \leq i \leq p, 1 \leq k \leq n_i$$

- Let $Y_{i,\bullet} = \frac{1}{n_i}\sum_k Y_{i,k}$ and $Y_{\bullet,\bullet} = \frac{1}{n}\sum_{i,k} Y_{i,k}$
- The variance can be decomposed as:

$$\begin{aligned} SSY &= SSR(M) + SSE(M) \\ &= \sum_i n_i \left(Y_{i,\bullet} - Y_{\bullet,\bullet}\right) + \sum_{i,k} \left(Y_{i,k} - Y_{i,\bullet}\right) \end{aligned}$$

- To test $H_0 =$ "$\theta_1 = \cdots = \theta_p$" versus $H_1 = \bar{H}_0$, the Fischer statistic is:

$$F = \frac{n-p}{p-1} \frac{\sum_i n_i (Y_{i,\bullet} - Y_{\bullet,\bullet})^2}{\sum_{i,k} (Y_{i,k} - Y_{i,\bullet})^2} \sim F(p-1, n-p)$$

Exhaustive, Forward, Backward and Stepwise selection

- Exhaustive search: for all sizes $1 \leq k \leq p$, find the combination of directions with highest R^2.
- Forward selection: at each step, add the direction most correlated with Y. Stop when the Fischer test for this direction is not rejected
- Backward selection: start with full model, and remove the direction with smallest t-statistic. Stop when all remaining t-statistics are significant
- Stepwise selection: like Forward selection, but after each inclusion remove all directions with unsignificant F-statistic
- Note: unless specified, 1_n is always included into the models.

Quadratic Risk: Bias-Variance decomposition

- To simplify the discussion, we consider the model

$$Y = \theta + \sigma Z$$

  where θ is arbitrary but aims at be understood by the family of models $\mathcal{M}$
- The quadratic risk of model $M \in \mathcal{M}$ is defined as

$$r(M) = \mathbb{E}\left[\left\|\theta - \hat{\theta}_M\right\|^2\right]$$

- It can be decomposed as:

$$r(M) = \|\theta - \theta_M\|^2 + \sigma^2 \dim(M)$$

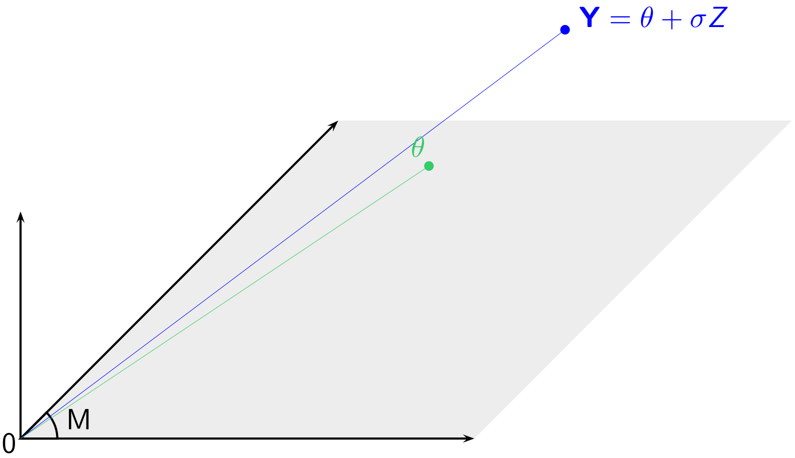
Risk Estimation and Mallow's criterion

- Goal: choose model $M \in \mathcal{M}$ with minimal quadratic risk $r(M)$.
- Problem: the bias $\|\theta - \theta_M\|^2$ is unknown
- Idea: penalize complexity $\dim(M)$
- Mallow's criterion: choose model M minimizing

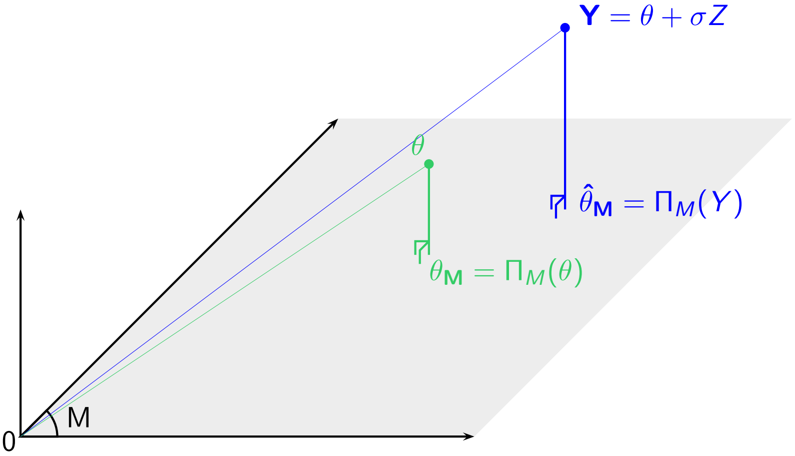
$$C_p(M) = SSE(M) + 2\sigma^2 \dim(M)$$

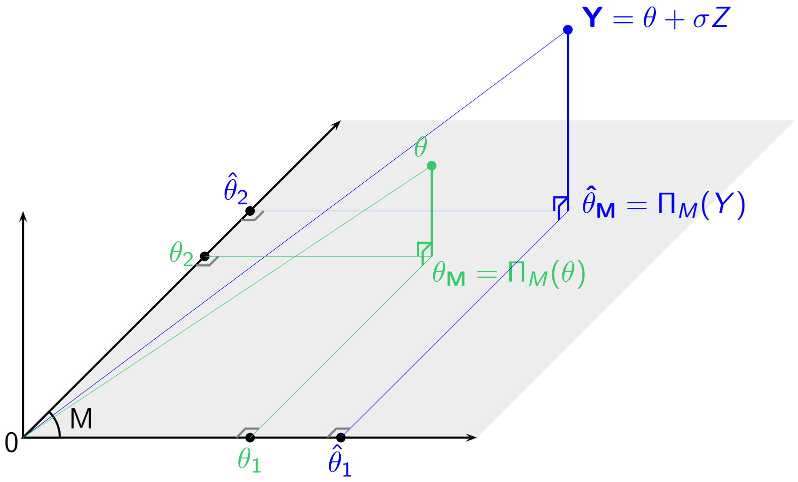
Heuristic: $r(M) = \|\theta\|^2 - \|\theta_M\|^2 + \sigma^2 \dim(M)$, but $\mathbb{E}\left[\|\hat{\theta}_M\|^2\right] = \|\theta_M\|^2 + \sigma^2 \dim(M)$, hence $\tilde{r}(M) = \|\theta\|^2 - \left(\|\hat{\theta}_M\|^2 - \sigma^2 \dim(M)\right) + \sigma^2 \dim(M)$ has expectation $r(M)$, but maximizing $\tilde{r}(M)$ over M is equivalent to maximizing

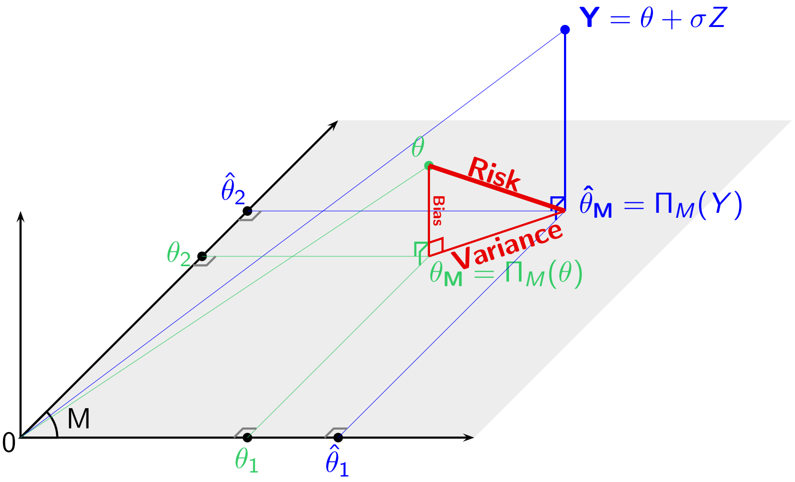
$$\tilde{r}(M) - \|\theta\|^2 + \|Y\|^2 = \|Y - \hat{\theta}_M\|^2 + 2\sigma^2 \dim(M)$$

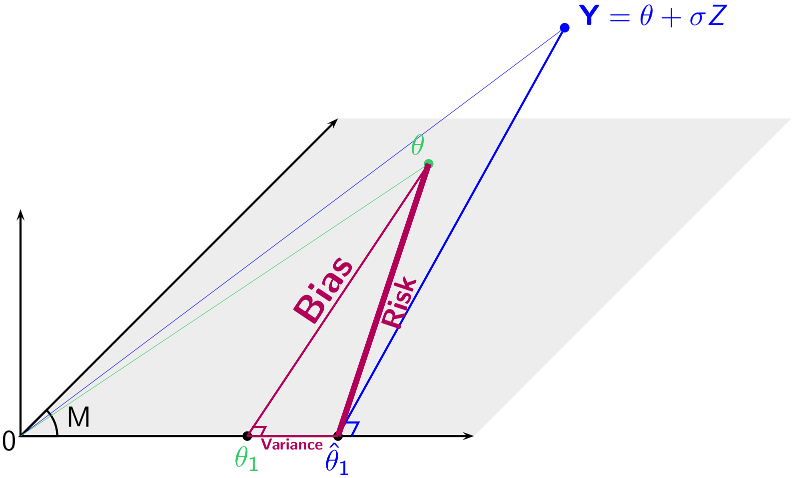


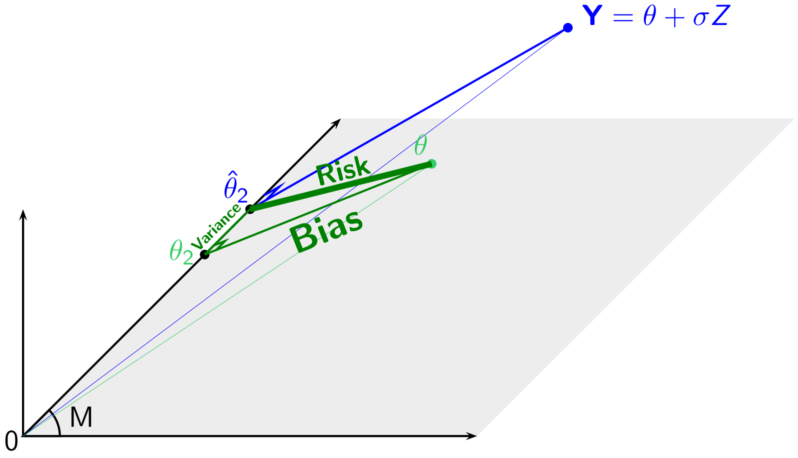
$\mathbf{Y} = \theta + \sigma Z$
θ
M
0

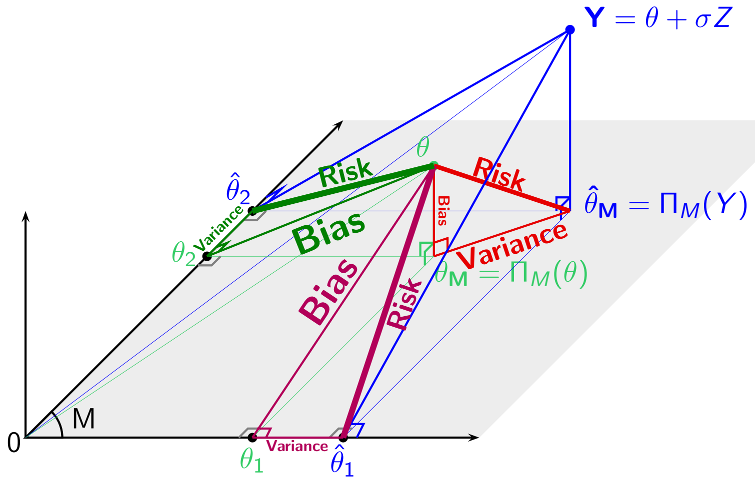

$\mathbf{Y} = \theta + \sigma Z$
θ
$\hat{\theta}_{\mathbf{M}} = \Pi_M(Y)$
$\theta_{\mathbf{M}} = \Pi_M(\theta)$
M
0


$\mathbf{Y} = \theta + \sigma Z$
θ
$\hat{\theta}_2$
$\hat{\theta}_{\mathbf{M}} = \Pi_M(Y)$
θ_2
$\theta_{\mathbf{M}} = \Pi_M(\theta)$
0
M
θ_1
$\hat{\theta}_1$


$\mathbf{Y} = \theta + \sigma Z$
θ
Risk
$\hat{\theta}_2$
Bias
$\hat{\theta}_{\mathbf{M}} = \Pi_M(Y)$
θ_2
Variance
$\theta_{\mathbf{M}} = \Pi_M(\theta)$
M
0
θ_1
$\hat{\theta}_1$


$\mathbf{Y} = \theta + \sigma Z$
θ
Bias
Risk
M
0
θ_1
Variance
$\hat{\theta}_1$


$\mathbf{Y} = \theta + \sigma Z$
θ
Risk
$\hat{\theta}_2$
Variance
θ_2
Bias
M
0


$\mathbf{Y} = \theta + \sigma Z$
θ
Risk
Risk
$\hat{\theta}_2$
Variance
θ_2
Bias
Bias
$\hat{\theta}_{\mathbf{M}} = \Pi_M(Y)$
Variance
$\theta_{\mathbf{M}} = \Pi_M(\theta)$
Bias
Risk
M
0
θ_1
Variance
$\hat{\theta}_1$

Other Criteria

- Adjusted R^2 :

$$R_a^2(M) = 1 - \frac{n-1}{n-\dim(M)}(1 - R^2(M))$$
$$= 1 - \frac{n-1}{n-\dim(M)} \times \frac{SSE(M)}{SSY}$$

- Bayesian Information Criterion:

$$\mathrm{BIC}(M) = SSE(M) + \sigma^2 \dim(M) \log n$$

Application: denoising a signal

- discretized and noisy version of $f : [0,1] \to \mathbb{R}$:

$$Y_k = f(k/n) + \sigma Z_k, 0 \leq k \leq n-1$$

- choice of an orthogonal basis of $\mathbb{R}^n$: Fourier

$$\Omega_n = \left\{ \left[\sin\left(\frac{2\pi kl}{n}\right)\right]_{0\leq l\leq N-1}, 1 \leq k \leq \left\lfloor\frac{n-1}{2}\right\rfloor, \left[\cos\left(\frac{2\pi kl}{n}\right)\right]_{0\leq l\leq N-1}, 0 \leq k \leq \left\lfloor\frac{n}{2}\right\rfloor \right\}$$

- nested models with increasing number of non-zero Fourier coefficients

Logistic Regression

- The Gaussian model does obviously not apply everywhere; think e.g. of a regression age/heart disease.
- Logistic model:

$$Y_i \sim \mathcal{B}\left(\mu({}^tX_i\theta)\right),$$

  where $\mu(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)}$ is the *inverse logit function*.
- Maximum likelihood estimation is possible numerically (Newton-Raphson method)

Outline

Discovery of R

- Understand and modify the source codes available on the website.
- The data frame called 'cars' contains two arrays: cars$dist and cars$speed. Its gives the speed of cars and the distances taken to stop (recorded in the 1920s).
  A relation $\mathrm{dist} = A \times \mathrm{speed}^B$ is expected. How to estimate A and B ?
  Test if $B = 0$, and then if $B = 1$.
- Find out how logistic regression can be done with R. Illustrate on some data you choose.

Simple Exercices

- Show that if $X_1, \ldots, X_n$ is a $\mathcal{N}(0,1)$-sample, then

$$\bar{X}_n \sim \mathcal{N}(0, 1/n) \quad \perp\!\!\!\perp \quad \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2 \sim \chi^2(n-1)$$

- Re-compute the formula giving $\hat{\alpha}$ and $\hat{\beta}$ in the simple regression model by analytic minimization of the total squared errors $\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$.
- Compute the squared prediction error $\mathbb{E}\left[(\hat{y}^* - \alpha - \beta x^*)^2\right]$ for a new observation at point x^* in the simple regression model.
- Same exercices for the general gaussian linear model.

Exercice: weighting methods

A two plate weighing machine is called unbiased with precision σ^2 if, an object of true weight m on the left plate is balanced by a random weight y such that $y = m + \sigma\epsilon$ on the right plate, where ϵ is a centered standard normal variable.

Mister M. has three objects of mass a, b and c to weigh with such a machine, and he is allowed to proceed to three measurements. He thinks of three possibilities

- weighting each object separately : (a —), (b —), (c —);
- weighting the objects two at a time : (ab —), (ac —) and (bc —);
- putting each object one time on the right plate alone and two times with another on the right plate (ab — c), (ac — b), (bc — a).

What would you advice him?

Exercice: weighting methods

A two plate weighing machine is called unbiased with precision σ^2 if, an object of true weight m on the left plate is balanced by a random weight y such that $y = m + \sigma\epsilon$ on the right plate, where ϵ is a centered standard normal variable.

Mister M. has three objects of mass a, b and c to weigh with such a machine, and he is allowed to proceed to three measurements.

He thinks of three possibilities

- weighting each object separately : (a —), (b —), (c —);
- weighting the objects two at a time : (ab —), (ac —) and (bc —);
- putting each object one time on the right plate alone and two times with another on the right plate (ab — c), (ac — b), (bc — a).

What would you advice him?

More precisely: compute the individual variance for each possibility and give a first conclusion. Does it hold if one is interested in linear combinations of the weight?

Exercice: multi-intercept regression

Botanists want to quantify the average difference of height between the trees of two forests A and B. In their model, the height of a tree is the sum of three terms:

- a term q depending on the quality of the ground, which is assumed to be constant in each forest: q_A for the trees of forest A, q_B for the trees of forest B;
- an unknown biological constant times the quantity of humus around the tree;
- a random term proper to each tree.

Precisely, they want to estimate the difference $D = q_A - q_B$. For their study, they have collected the height of n_A trees in forest A, n_B trees in forest B, as well as the quantities $(h_i^A)_{1 \leq j \leq n_A}$ and $(h_i^B)_{1 \leq j \leq n_B}$ of humus at the basis of all thoses trees.
Tell them how to do it.