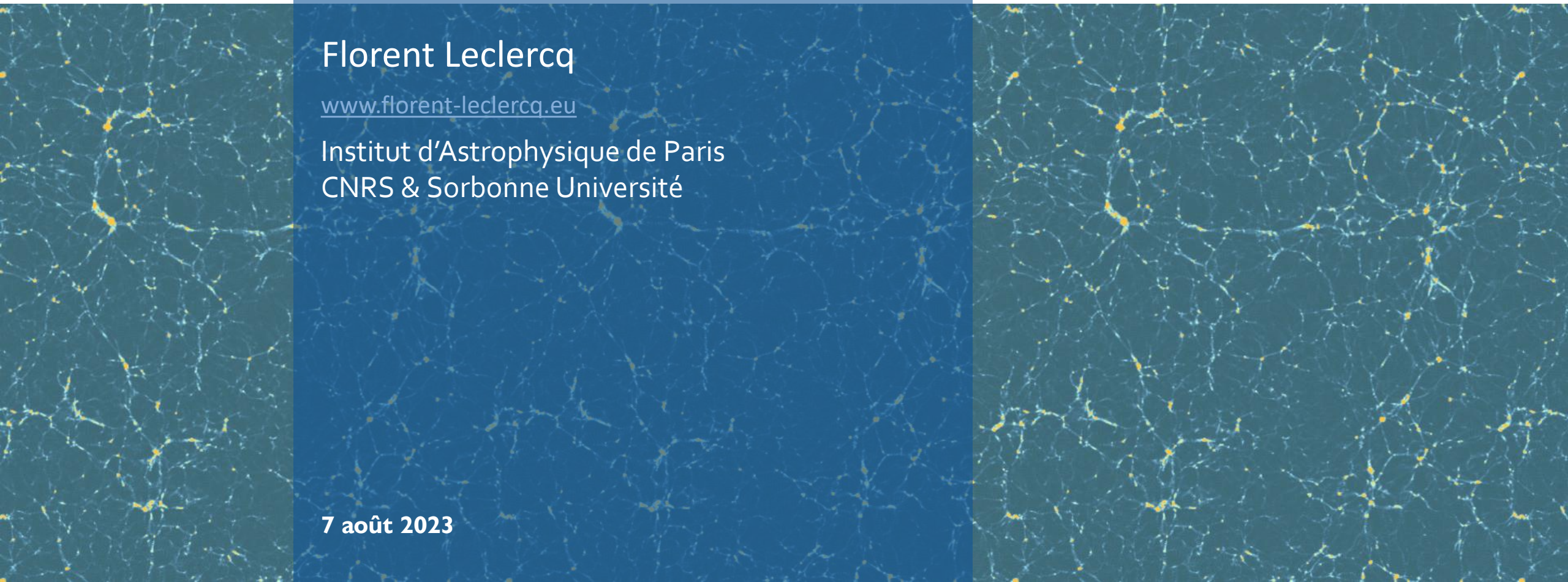
La théorie des probabilités : la logique de la découverte scientifique

33ème Festival d'Astronomie de Fleurance,
Cours Fil Noir

Florent Leclercq

www.florent-leclercq.eu

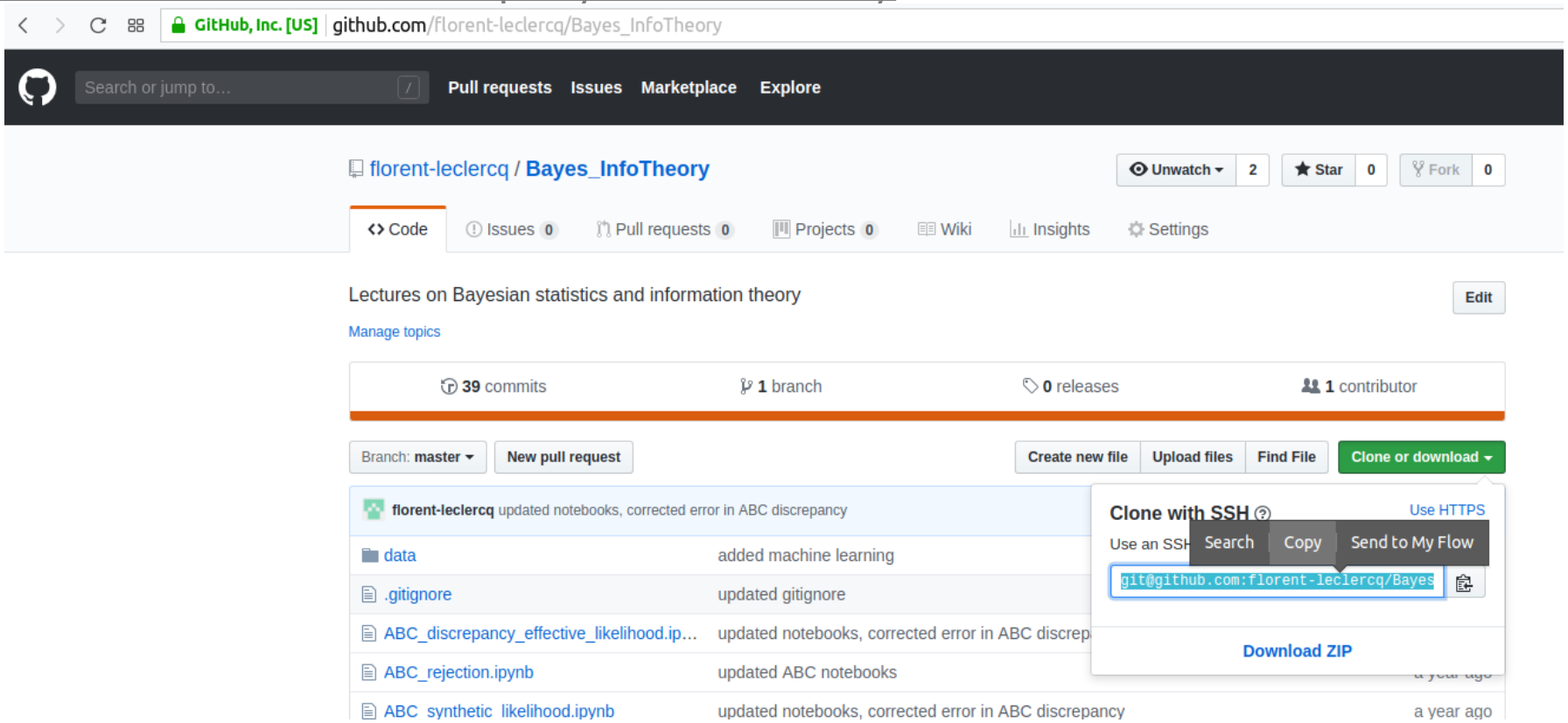
Institut d'Astrophysique de Paris
CNRS & Sorbonne Université



7 août 2023

Le dépôt Github

- https://github.com/florent-leclercq/Bayes_InfoTheory

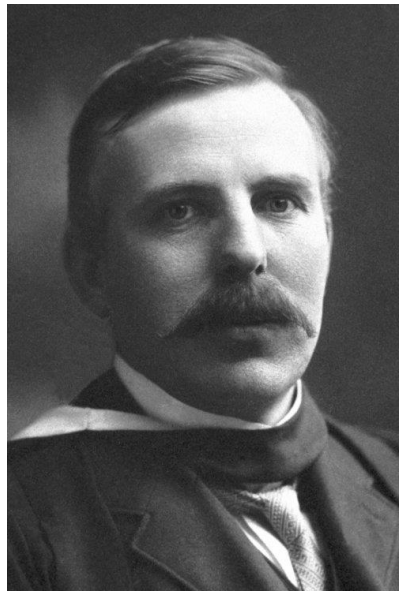


git clone https://github.com/florent-leclercq/Bayes_InfoTheory.git (or with SSH)

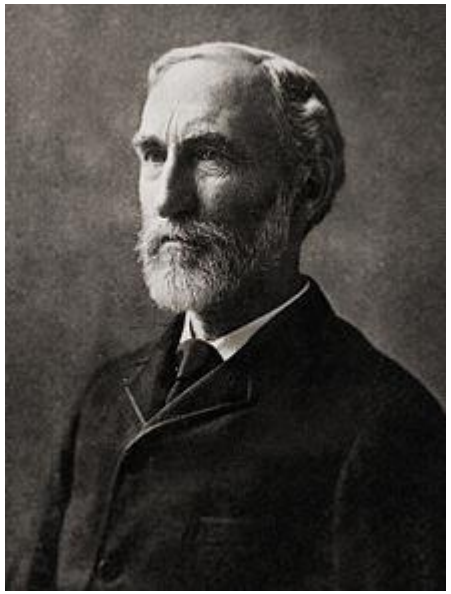
- Page du cours : http://florent-leclercq.eu/teaching.php
- Slides : http://florent-leclercq.eu/documents/teaching/Cours_Fil_Noir_Fleurance_2023.pdf

Introduction : pourquoi faire des statistiques correctement est important
Un exemple historique : le paradoxe de Gibbs

- « Si votre expérience nécessite des statistiques, vous auriez dû réaliser une meilleure expérience. »



Ernest Rutherford
(1871-1937)



J. Willard Gibbs
(1839-1903)

- L'ensemble canonique et l'ensemble grand-canonique de Gibbs, dérivés du principe de l'entropie maximale, échouent à prédire correctement les propriétés thermodynamiques des systèmes physiques réels.
- Les entropies prédites sont toujours plus grandes que celles observées… il doit donc exister des contraintes microphysiques supplémentaires :
  - Discrétisation des niveaux d'énergie : rayonnement : Planck (1900), solides : Einstein (1907), Debye (1912), Ising (1925), atomes individuels : Bohr (1913)…
  - … Mécanique quantique : Heisenberg, Schrödinger (1927).

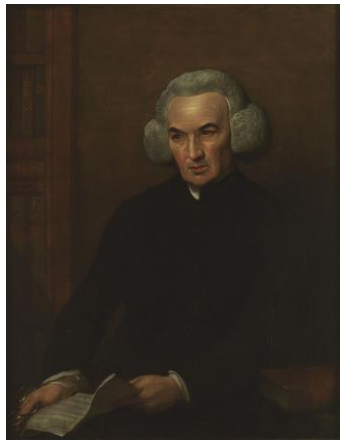
Les premiers indices indiquant le besoin d'une physique *quantique* ont été découverts grâce à des applications apparemment « infructueuses » des statistiques.

Comment raisonner rationnellement en présence d'incertitudes : La formule de Bayes

- Comment mesurer une grandeur ? Comment vérifier une théorie ? Plus généralement, comment la connaissance progresse-t-elle ?
- La formule de Bayes (1763): une identité mathématique sur la façon dont nous **analysons des observations** et **changeons d'avis** lorsque nous obtenons de nouvelles informations.
- Mais pourquoi l'utiliser ?
  - La formule de Bayes est triviale et dépassée.
  - Elle mesure la **croyance**. Elle dit que nous pouvons apprendre même à partir de données manquantes ou incomplètes, à partir d'approximations, à partir de l'ignorance. Elle va à l'encontre de la conviction que la science exige objectivité et précision.
  - Après la mort de Laplace, elle a été déclarée morte et enterrée.

Ce n'est (probablement !) pas la bonne personne.

Thomas Bayes
(1701-1761)



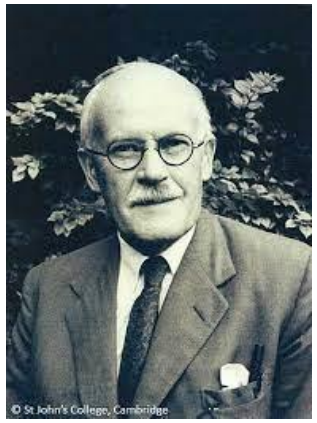
Richard Price
(1723-1791)

Pierre-Simon de Laplace
(1749-1827)

Photos prises à Bunhill Fields Burial Ground, City of London, en 2021

Controverse : fréquentisme versus bayésianisme

- Deux conceptions de la nature des probabilités et des questions scientifiques :
  - Probabilités « objectives » liées à la fréquence de phénomènes aléatoires répétitifs. Questions liées à des expériences déterminées et reproductibles.
  - Probabilités « subjectives » liées à la certitude accordée à une mesure ou à une théorie. Questions liées à des phénomènes et des choix n'impliquant pas l'idée d'une répétition.

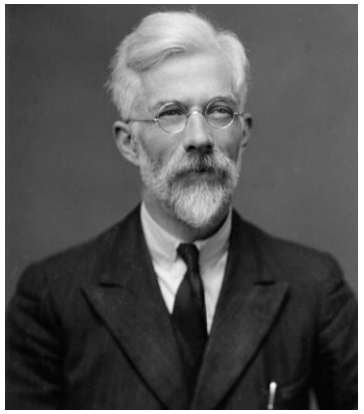


Harold Jeffreys
(1891-18989)


Leonard J. Savage
(1917-1971)

Fisher publiait parfois des insultes que seul un saint pourrait entièrement pardonner.
Savage 1976, Annals of Statistics


Karl Pearson
(1857-1936)

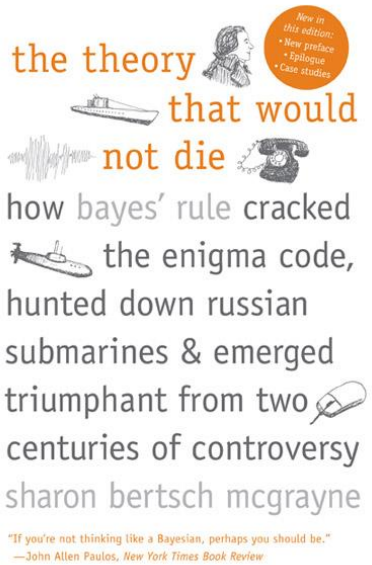

Ronald Aylmer Fisher
(1890-1962)


Jerzy Neyman
(1894-1981)

- Techniques fréquentistes et bayésiennes donnent les mêmes résultats dès qu'on travaille sur des grands échantillons. Ce n'est que sur des petits nombres et des faibles occurrences que l'estimation fréquentiste et l'induction bayésienne diffèrent.

La théorie qui ne voulait pas mourir

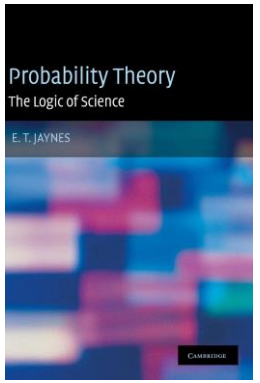
- Et pourtant, la formule de Bayes a aidé à résoudre beaucoup de problèmes pratiques :
  - Prouver l'innocence d'Alfred Dreyfus (Henri Poincaré, 1899-1906),
  - Sauver le système téléphonique Bell de la panique financière (Edward C. Molina, 1907),
  - Prédire les tremblements de terre et les tsunamis (Harold Jeffreys, 1930-1940),
  - Casser le code Enigma de la marine allemande (Alan Turing, 1940-1944),
  - Prouver que la cigarette cause le cancer du poumon (Jerome Cornfield, 1951)…
  - Rechercher un bombe H puis un sous-marin perdus en mer (John P. Craven, 1966-1968)



Sharon Bertsch McGrayne (2012)

- La bataille scientifique a duré 150 ans, jusqu'à l'arrivée des ordinateurs.

> La supériorité des méthodes bayésiennes est aujourd'hui un fait largement démontré dans une centaine de domaines différents. On peut argumenter avec une philosophie ; il n'est pas si facile d'argumenter avec une sortie d'ordinateur, qui nous dit : « Indépendamment de toute votre philosophie, voici les faits de performance réelle ».
> Jaynes 2002, Probability Theory – The logic of science



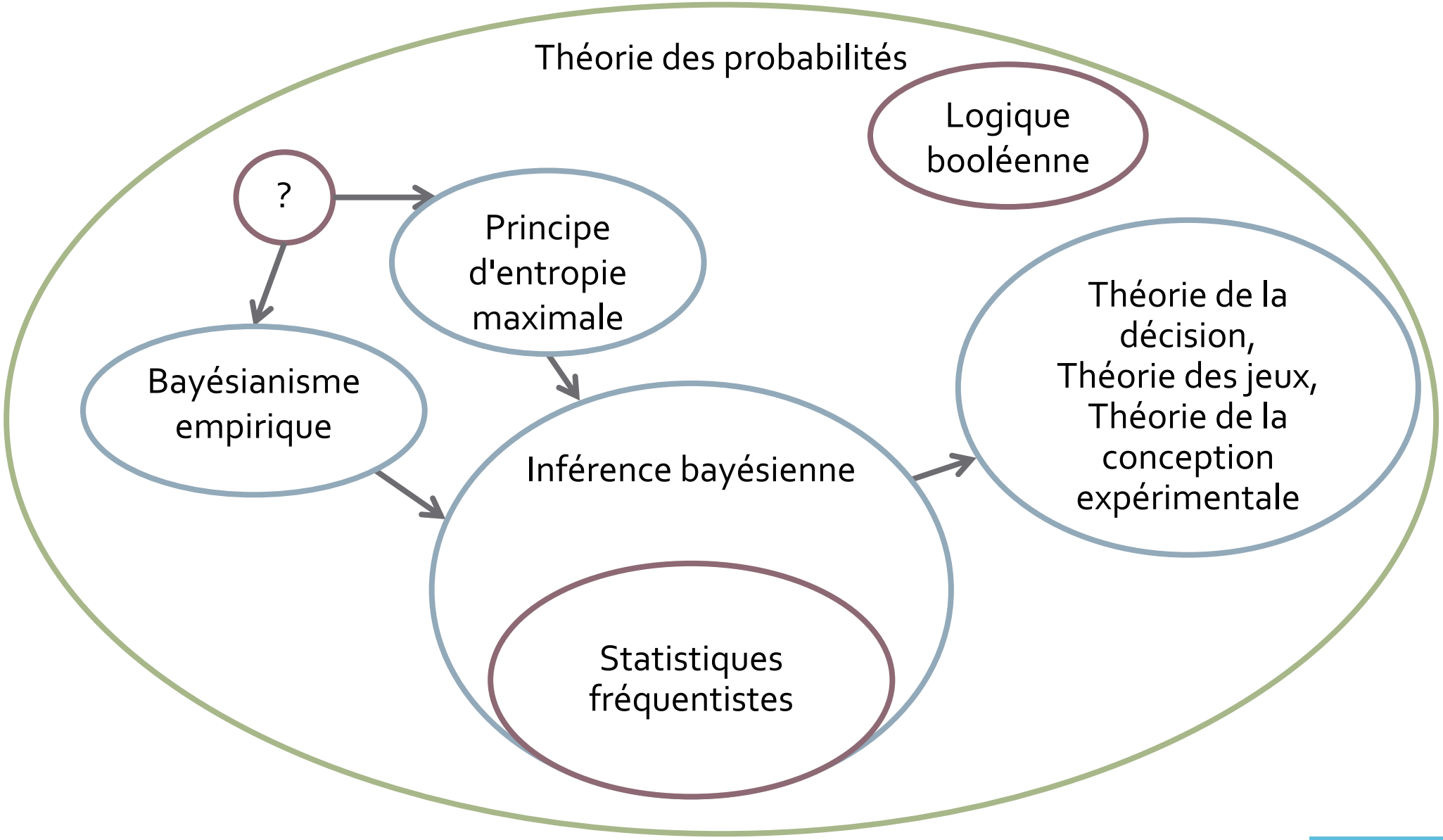
Jaynes (2002)

Richard Threlkeld Cox (1898-1991)

Edwin Thompson Jaynes (1922-1998)

- Théorème de Cox-Jaynes (1946) : la théorie (bayésienne) des probabilités est la seule « logique de la science » possible.

La « théorie des probabilités » d'après Jaynes : une extension de la logique ordinaire



Plan – vous pouvez voter pour les sujets que vous souhaitez !

- Principes de base de la théorie des probabilités
- Champs gaussiens aléatoires
- Traitement bayésien du signal et filtrage de Wiener
  - Débruitage bayésien
  - Séparation bayésienne des composantes
- Méthodes Monte Carlo et chaînes de Markov : l'algorithme de Metropolis-Hastings
- Exemple d'inférence de paramètres (cosmologie avec les données des supernovæ)
  - Chaînes de Markov hamiltoniennes
  - Émulateurs
- Exemple de comparaison de modèles (l'éclipse de 1919)
- Priors d'ignorance et le principe de l'entropie maximale
- Théorie de l'information et apprentissage automatique supervisé
- Théorie bayésienne de la décision et théorie bayésienne de la conception expérimentale
- Réseaux bayésiens, modèles bayésiens hiérarchiques et Bayésianisme empirique

Les principes de base de la théorie des probabilités

- Les probabilités jointes et les probabilités conditionnelles : $p(A, B) = p(A|B)p(B)$
- Les probabilités marginales : $p(A) = \int p(A, B)\,\mathrm{d}B$

- La règle du produit (« et » logique) : $p(AB|C) = p(A|BC)\,p(B|C)$
- La règle de la somme (« ou » logique) : $p(A + B|C) = p(A|C) + p(B|C) - p(AB|C)$

- La formule de Bayes (inférence des paramètres d'un modèle) :

$$p(s|d) = \frac{p(d|s)\,p(s)}{p(d)}$$

posterior ← $p(s|d)$; vraisemblance ← $p(d|s)$; prior ← $p(s)$; évidence ← $p(d)$

- Le facteur de Bayes (comparaison bayésienne de modèles) : $\mathcal{B}_{12} = \frac{p(d|\mathcal{M}_1)}{p(d|\mathcal{M}_2)}$

où $p(d|\mathcal{M}_i) = \int p(d|s_i, \mathcal{M}_i)p(s_i|\mathcal{M}_i)\,\mathrm{d}s_i$

Le ratio des évidences (et non des vraisemblances) prend en compte l'effet « rasoir d'Occam ».

Champs gaussiens aléatoires

Notebook : https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/GRF_and_fNL.ipynb

Champs gaussiens aléatoires

- Définition : tout vecteur aléatoire x de distribution de probabilité (pdf) :

$$p(x|\mu, C) = \frac{1}{\sqrt{|2\pi C|}} \exp\left[-\frac{1}{2}(x-\mu)^\intercal C^{-1}(x-\mu)\right]$$

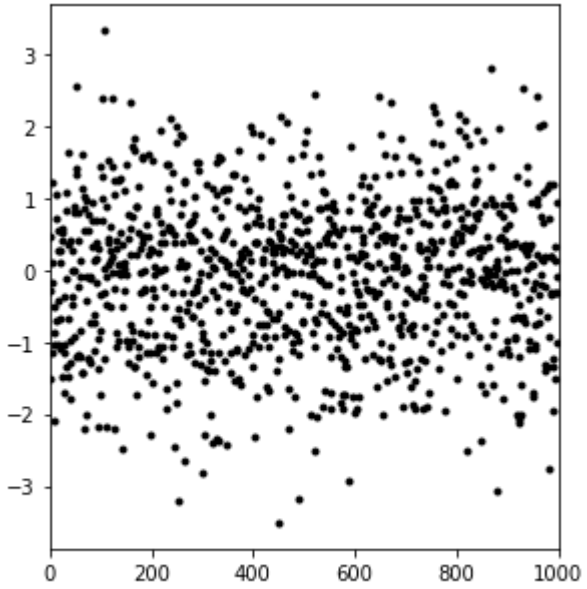
$$-2\ln p(x|\mu, C) = (x-\mu)^\intercal C^{-1}(x-\mu) + \ln|2\pi C|$$

… c'est tout !

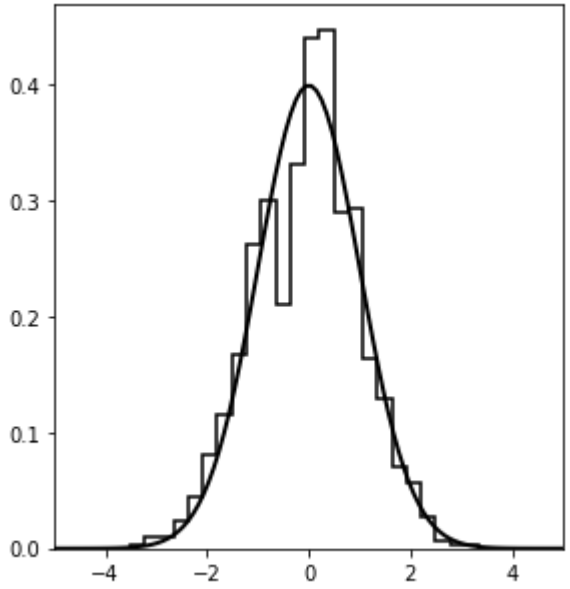
- Génération de champs gaussiens aléatoires :
  - Tirer un vecteur de bruit blanc ξ (variables gaussiennes non corrélées et de moyenne nulle et variance unitaire)
  - Trouver la racine carrée de la matrice C : $\sqrt{C}$ (n'importe quelle matrice satisfaisant cette condition convient)
  - Calculer : $x = \sqrt{C}\xi + \mu$

Champs gaussiens aléatoires : exemples

bruit blanc



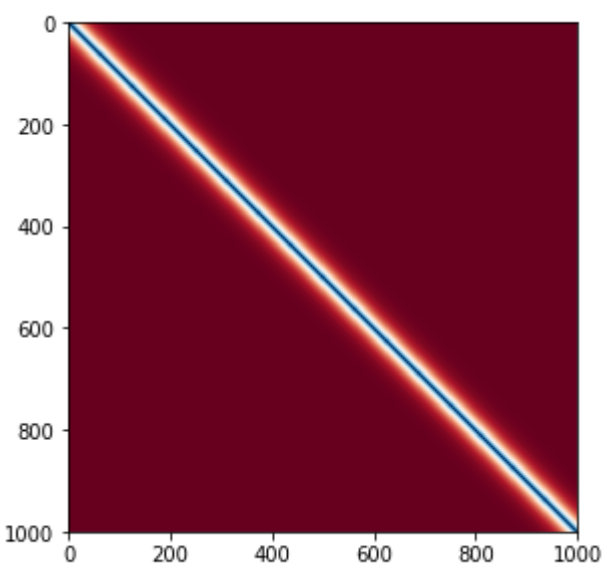
histogramme

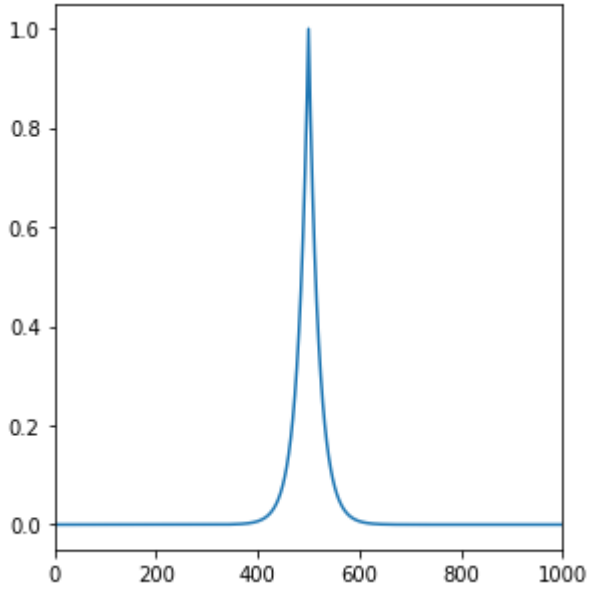


Champs gaussiens aléatoires : exemples

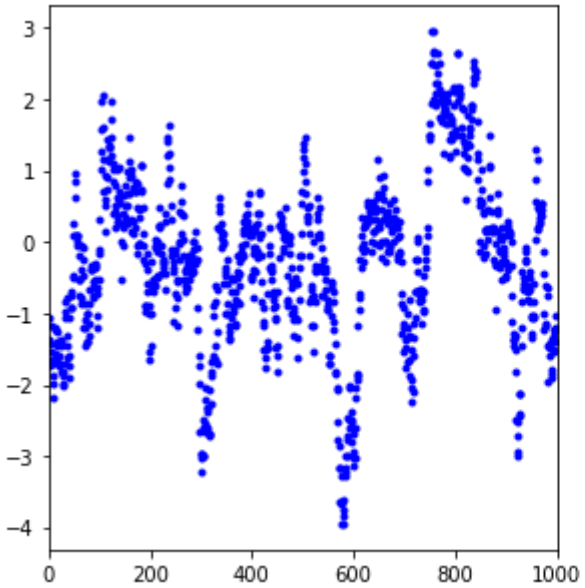
$$C_{ij} = \exp\left(-\frac{|i-j|}{20}\right)$$

matrice de covariance

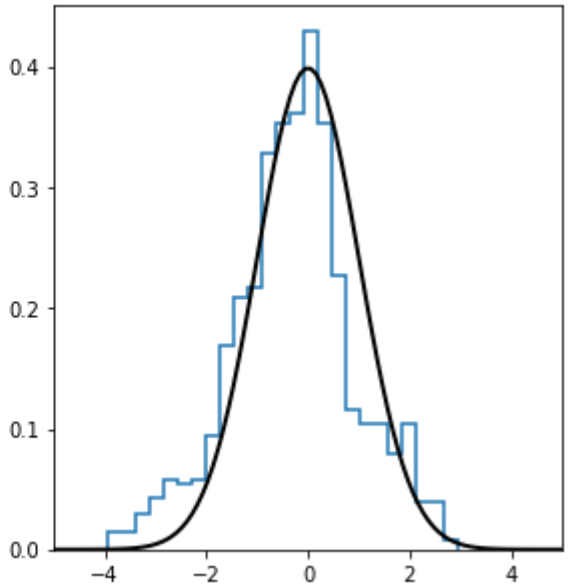




champ gaussien



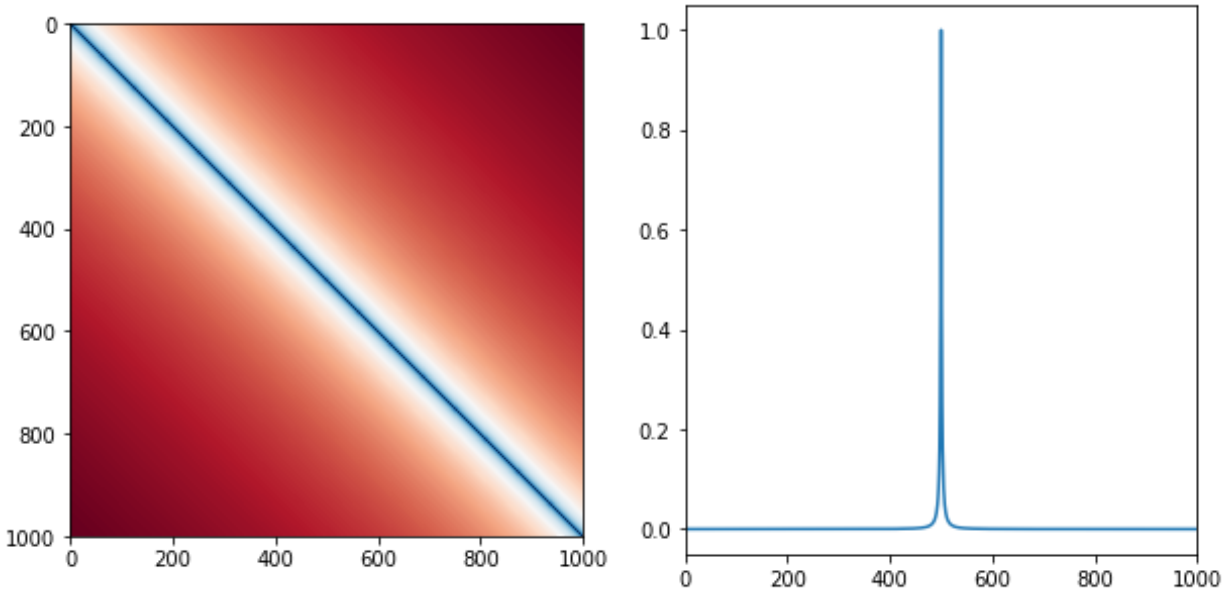
histogramme



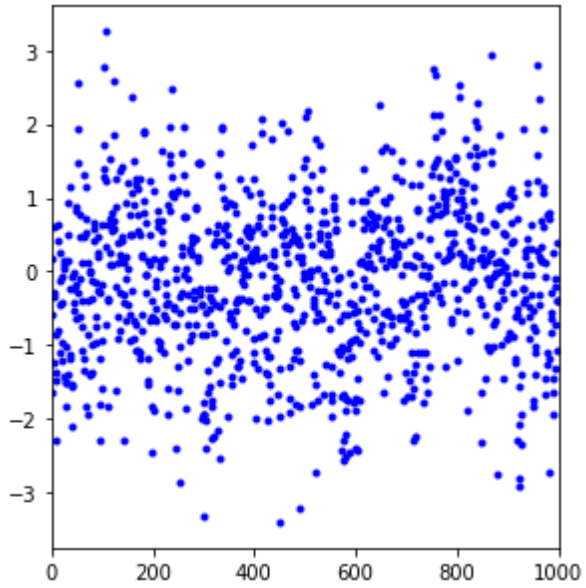
Champs gaussiens aléatoires : exemples

$$C_{ij} = \frac{1}{(1 + |i - j|/2)^2}$$

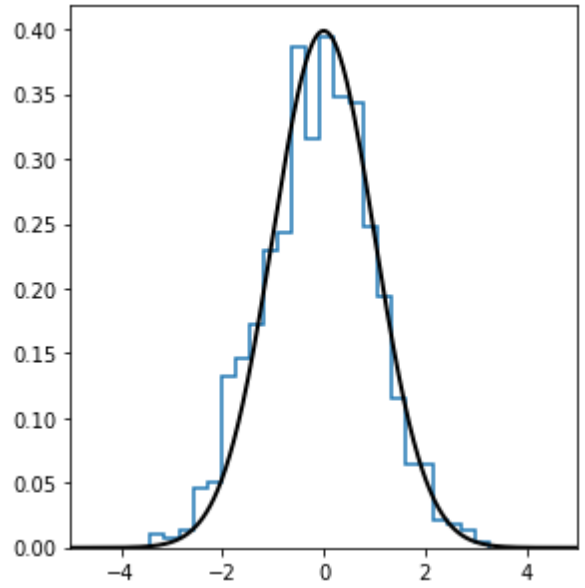
matrice de covariance



champ gaussien



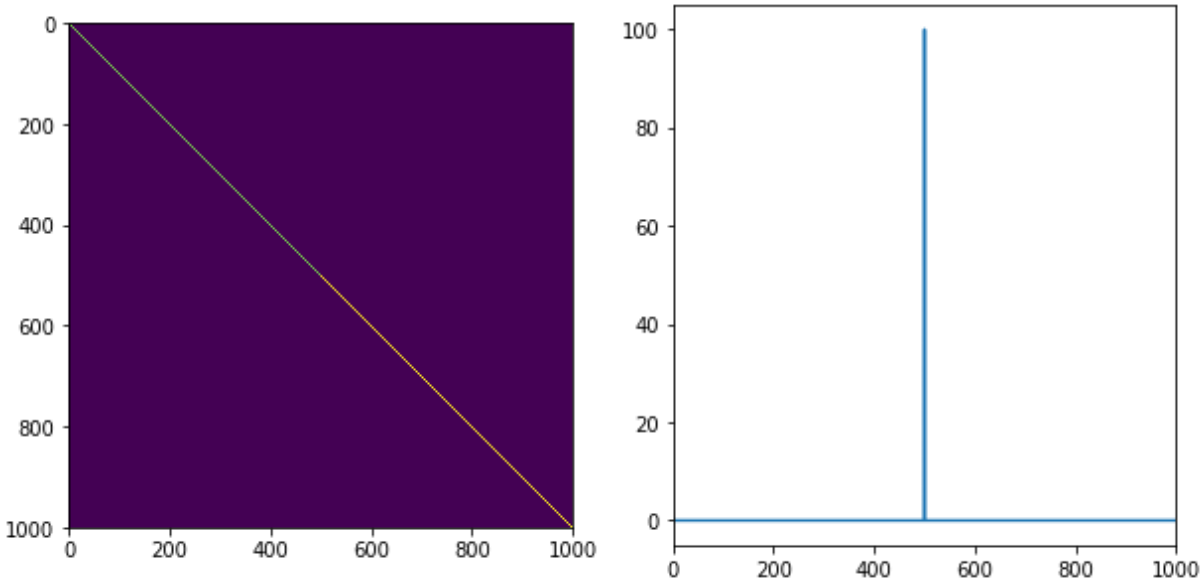
histogramme



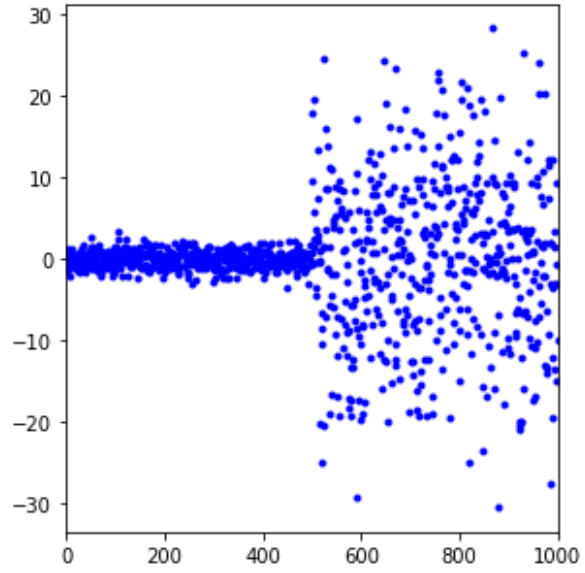
Champs gaussiens aléatoires : exemples

$$C_{ii} = \begin{cases} 1 & \text{if } i < N/2 \\ 100 & \text{otherwise} \end{cases}$$
$$C_{ij} = 0 \quad \text{for} \quad i \neq j$$

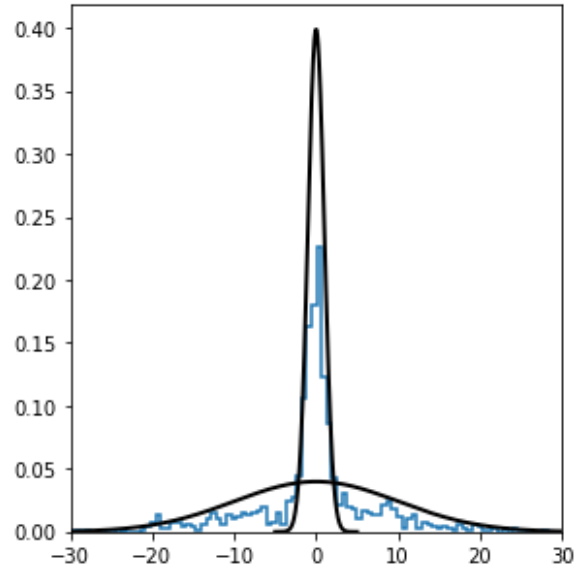
matrice de covariance



champ gaussien



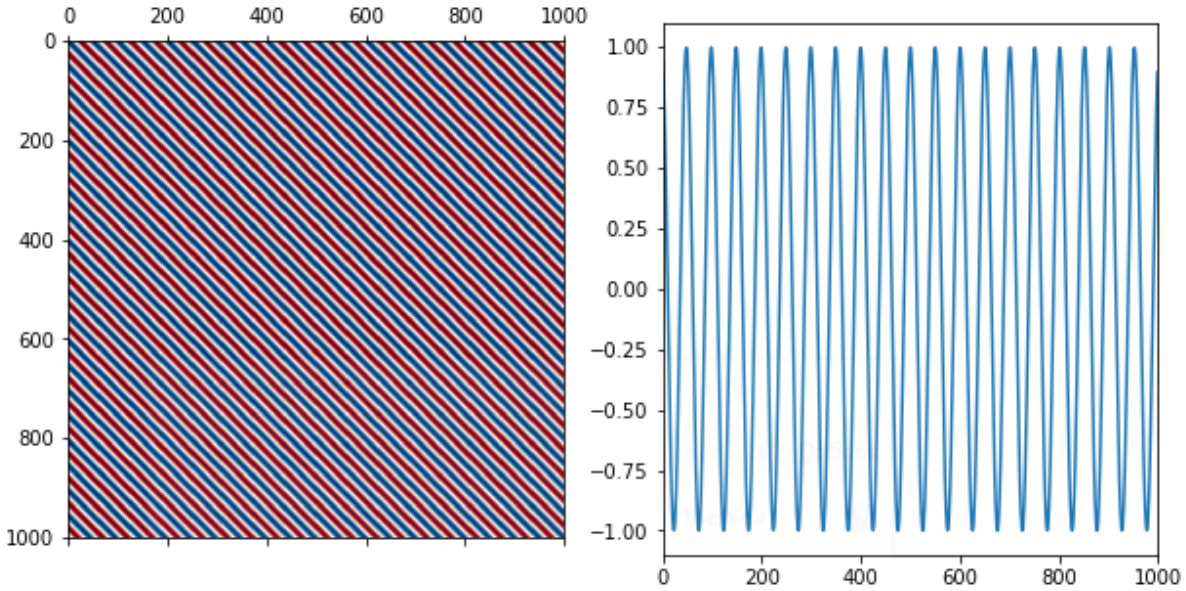
histogramme



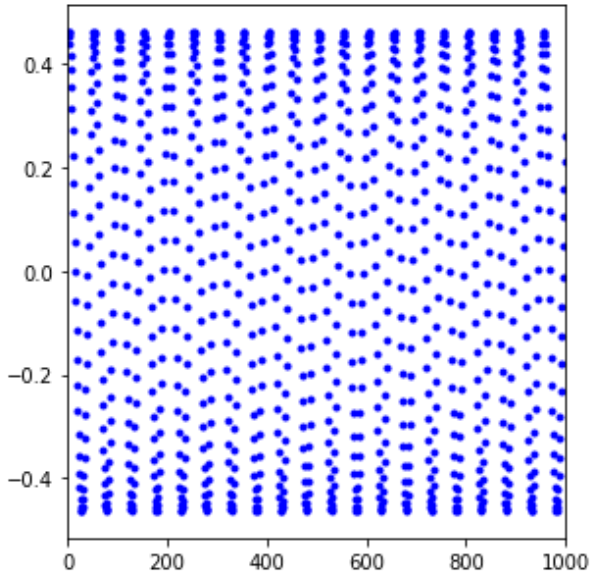
Champs gaussiens aléatoires : exemples

$$C_{ij} = \cos\left(\frac{i-j}{8}\right)$$

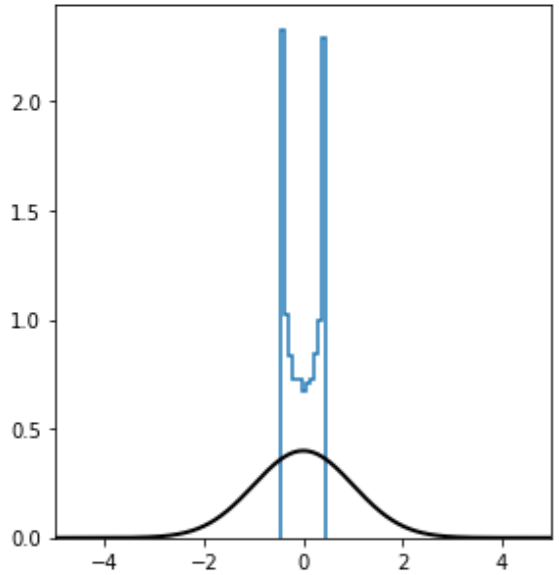
matrice de covariance



champ gaussien



histogramme

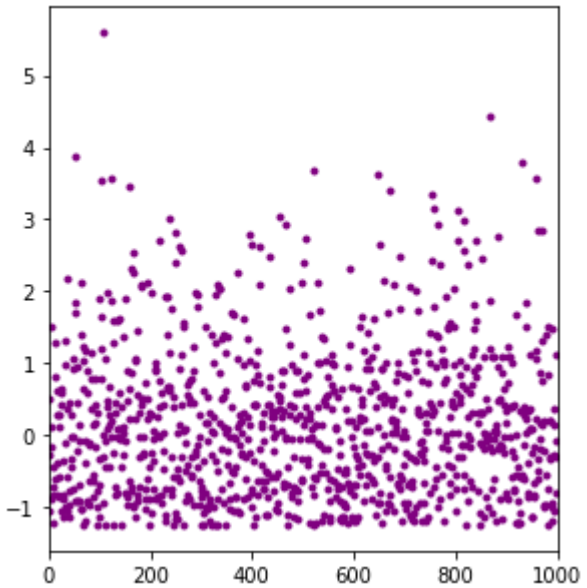


Les histogrammes de champs aléatoires gaussiens ne sont pas toujours gaussiens !

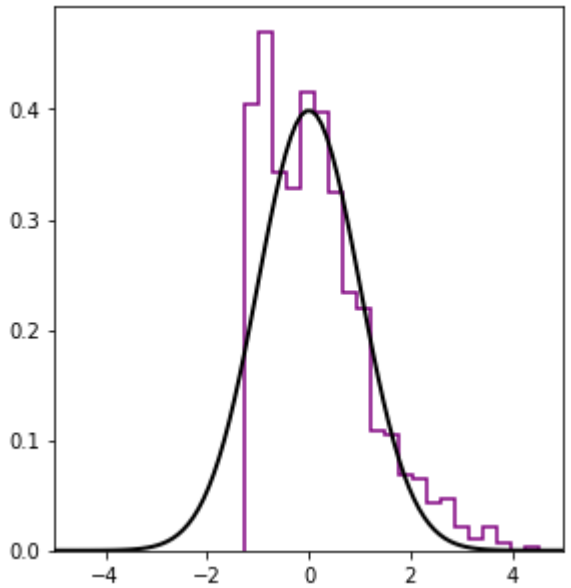
Exemple de signal non-Gaussien

- $s = \Phi + f_{\mathrm{NL}}\Phi^2$ où Φ est un champ gaussien aléatoire.
- En cosmologie, cela s'appelle « non-gaussianité de type local ».
- Les non-gaussianités primordiales renseignent sur la physique de l'inflation cosmologique.

signal



histogramme



- La pdf à un point possède une asymétrie (*skewness*).

Probabilités marginales et conditionnelles de champs gaussiens aléatoires

- On travaille avec deux variables qui constituent **conjointement** un champ gaussien $\begin{pmatrix} x \\ y \end{pmatrix}$.

- Alors x et y sont des champs gaussiens dont les **probabilités marginales** ont pour

$$\text{Moyenne : } \mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \qquad \text{Matrice de covariance : } C = \begin{pmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{pmatrix}$$

(Les moyennes et covariances marginales sont simplement les parties correspondantes de la moyenne et de la covariance jointe.)

- Les **probabilités conditionnelles** sont également gaussiennes avec :

$$\text{Moyenne : } \quad \mu_{x|y} = \mu_x + C_{xy} C_{yy}^{-1} (y - \mu_y)$$

$$\text{Matrice de covariance : } \quad C_{x|y} = C_{xx} - C_{xy} C_{yy}^{-1} C_{yx}$$

Traitement bayésien du signal et filtrage de Wiener

Notebook : https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/WienerFilter_denoising.ipynb

Notebook : https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/WienerFilter_denoising_CMB.ipynb

Notebook : https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/WienerFilter_deblending.ipynb

Le filtrage de Wiener : débruitage bayésien d'un signal

- Modèle de données : $d = s + n$ où $\begin{pmatrix} s \\ d \end{pmatrix}$ est un champ gaussien aléatoire (conjointement).

- Solution :

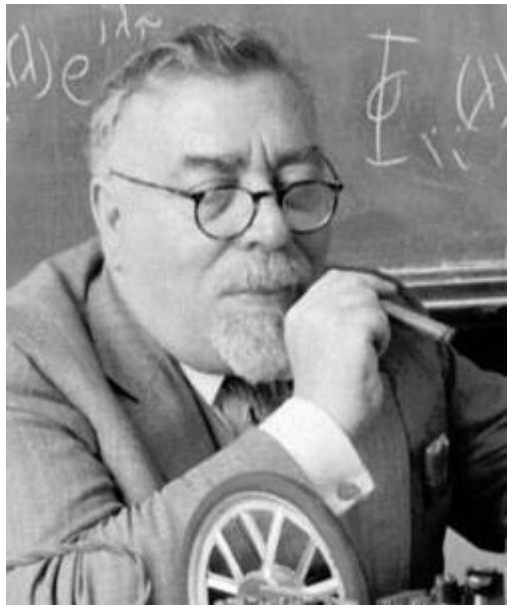
$$\mu_{s|d} = \mu_s + C_{sd} C_{dd}^{-1} (d - \mu_d)$$

$$C_{s|d} = C_{ss} - C_{sd} C_{dd}^{-1} C_{ds}$$

- Notations : $C_{ss} \equiv S$ et $C_{nn} \equiv N$.
- Hypothèses supplémentaires : $C_{sn} = C_{ns} = 0$. Alors :

$$C_{dd} = S + N$$

$$C_{sd} = C_{ss} + C_{sn} = C_{ss} = S$$



Norbert Wiener
(1894-1964)

- Expressions finales :

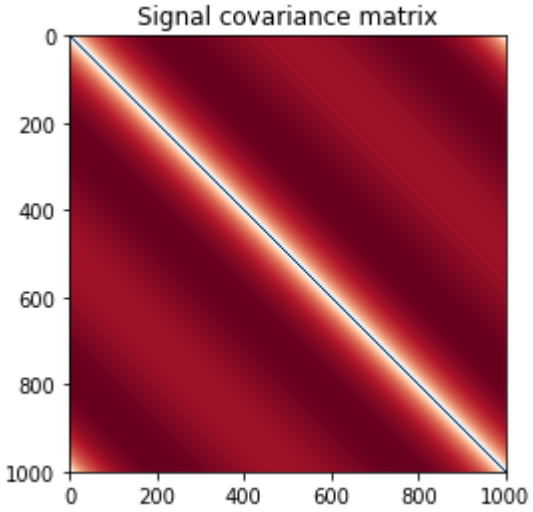
$$\mu_{s|d} = \mu_s + S(S+N)^{-1}(d - \mu_d) = \mu_s + (S^{-1} + N^{-1})^{-1} N^{-1} (d - \mu_d)$$

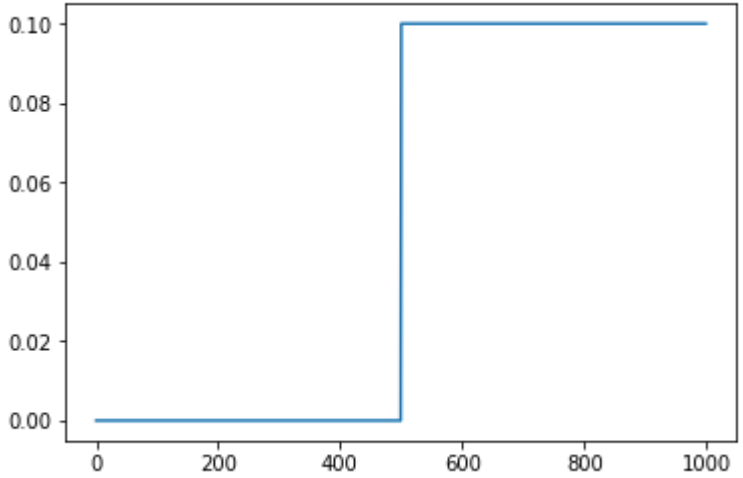
$$C_{s|d} = S - S(S+N)^{-1} S = (S^{-1} + N^{-1})^{-1}$$

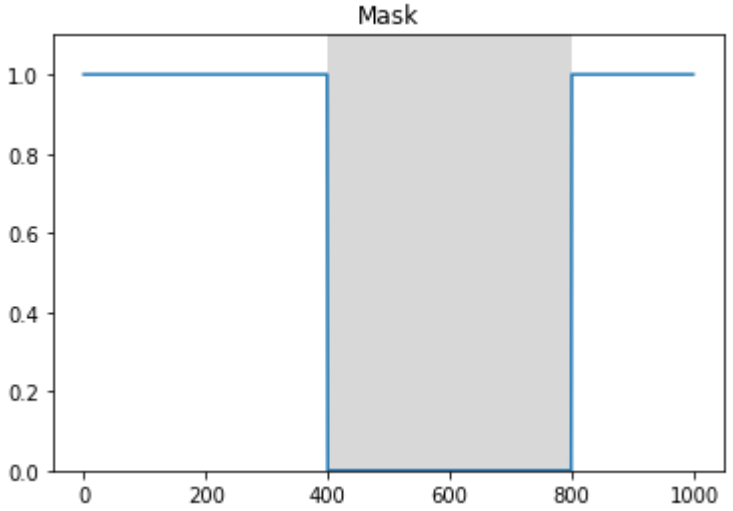
Exemple de débruitage par filtrage de Wiener

- Hypothèses concernant la matrice de covariance du signal et du bruit

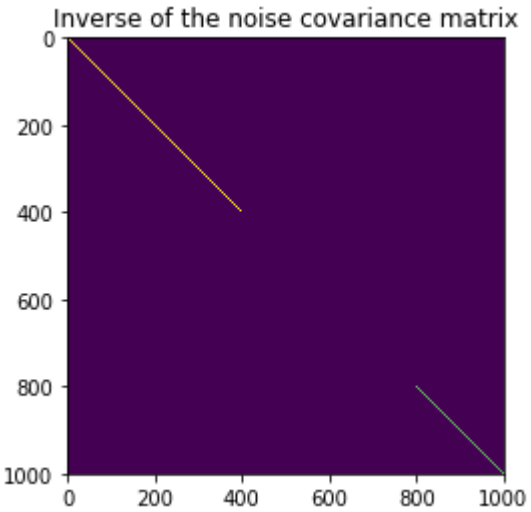
S







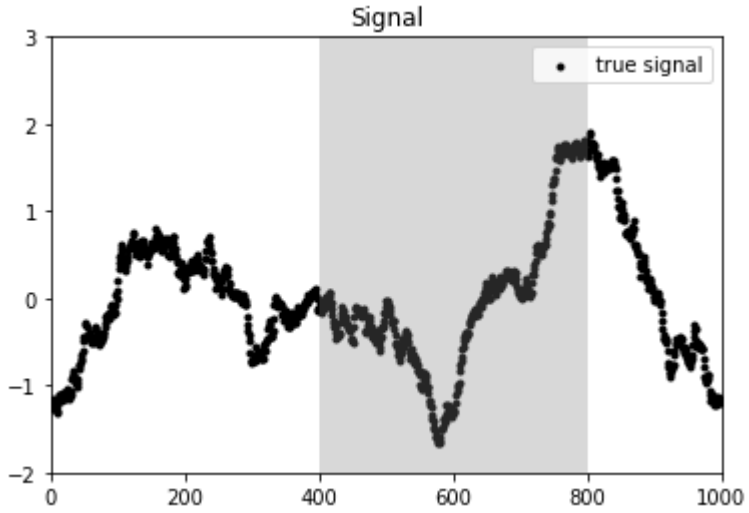
N^{-1}

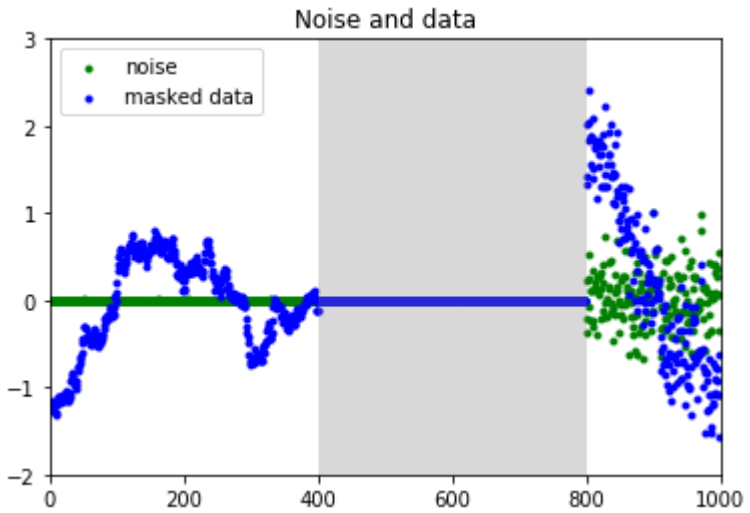


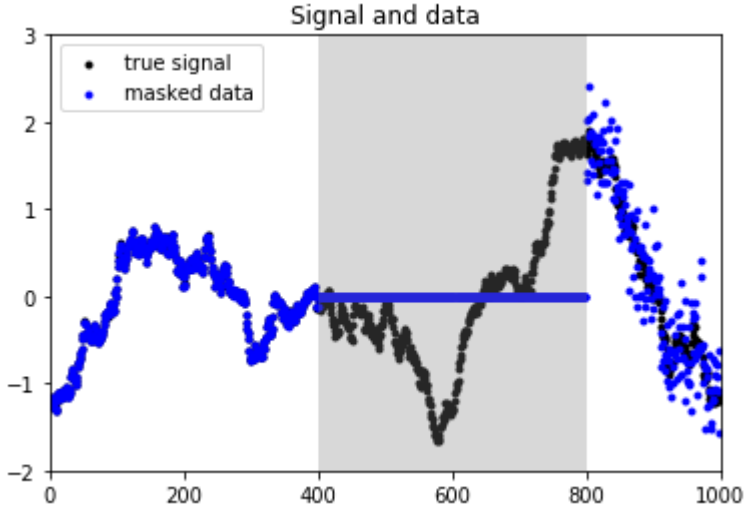
Exemple de débruitage par filtrage de Wiener

- Génération d'un signal et de données synthétiques

$$d = s + n$$



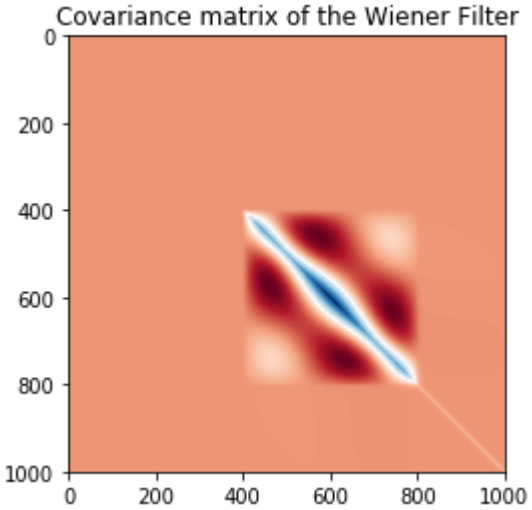




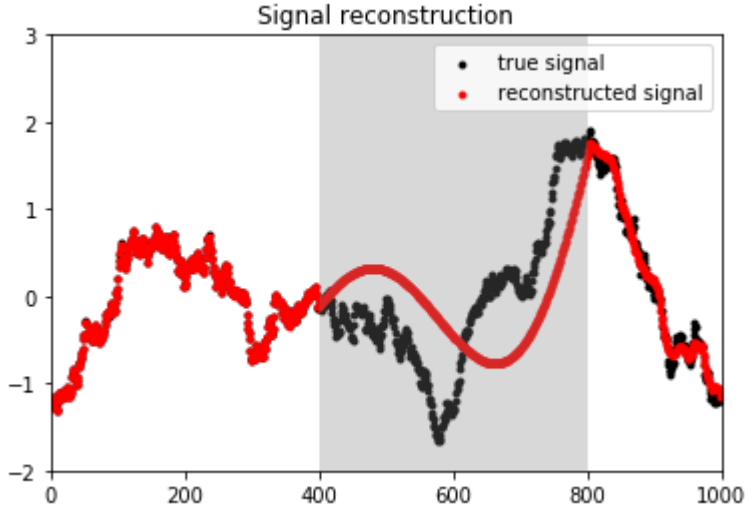
Exemple de débruitage par filtrage de Wiener

- Filtrage de Wiener
- La moyenne de la reconstruction correspond à l'estimateur maximum a posteriori.

$$C_{s|d} = (S^{-1} + N^{-1})^{-1}$$



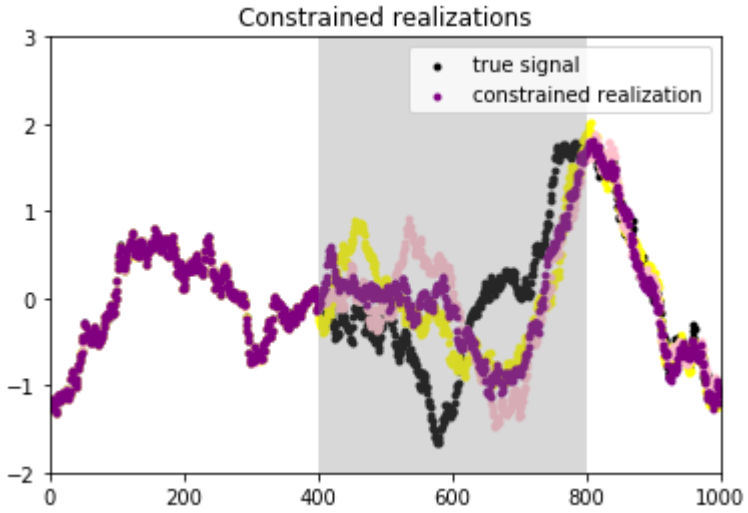
$$\mu_{s|d} = \mu_s + C_{s|d}N^{-1}(d - \mu_d)$$



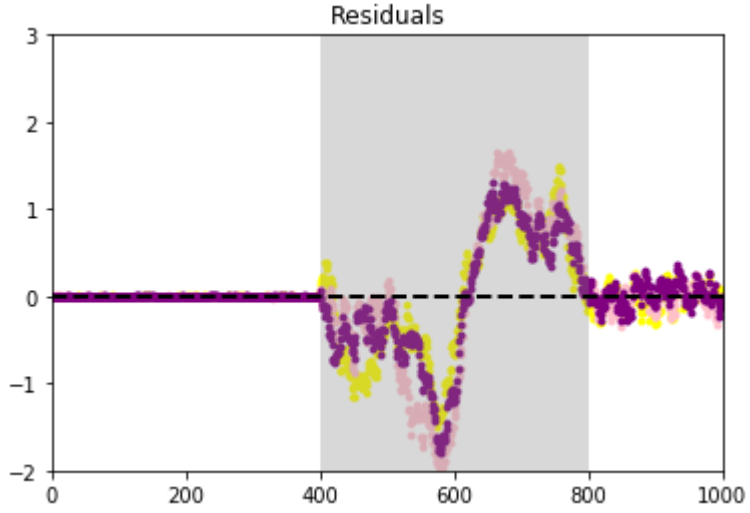
Exemple de débruitage par filtrage de Wiener

- Simulations contraintes du signal débruité

$$s_{\mathrm{sim}} = \mu_{s|d} + \sqrt{C_{s|d}}\,\xi$$



$$s_{\mathrm{sim}} - s_{\mathrm{true}}$$



Le filtrage de Wiener : séparation bayésienne de composantes mélangées et bruitées

- Modèle de données : $d = \begin{pmatrix} x_1 \\ x_2 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} n_1 \\ n_2 \\ n_3 \end{pmatrix}$

- Hypothèses : $C_{x_1 d} = \begin{pmatrix} C_{x_1 x_1} & C_{x_1 x_1} & 0 \end{pmatrix}$
$C_{x_2 d} = \begin{pmatrix} 0 & C_{x_2 x_2} & C_{x_2 x_2} \end{pmatrix}$
$C_{nn} = \begin{pmatrix} C_{n_1 n_1} & 0 & 0 \\ 0 & C_{n_2 n_2} & 0 \\ 0 & 0 & C_{n_3 n_3} \end{pmatrix}$

$$C_{dd} = \begin{pmatrix} C_{x_1 x_1} + C_{n_1 n_1} & C_{x_1 x_1} & 0 \\ C_{x_1 x_1} & C_{x_1 x_1} + C_{x_2 x_2} + C_{n_2 n_2} & C_{x_2 x_2} \\ 0 & C_{x_2 x_2} & C_{x_2 x_2} + C_{n_3 n_3} \end{pmatrix}$$

- Solution :

$$\mu_{x_1|d} = C_{x_1 d} C_{dd}^{-1} d \qquad \mu_{x_2|d} = C_{x_2 d} C_{dd}^{-1} d$$
$$C_{x_1|d} = C_{x_1 x_1} - C_{x_1 d} C_{dd}^{-1} C_{d x_1} \qquad C_{x_2|d} = C_{x_2 x_2} - C_{x_2 d} C_{dd}^{-1} C_{d x_2}$$

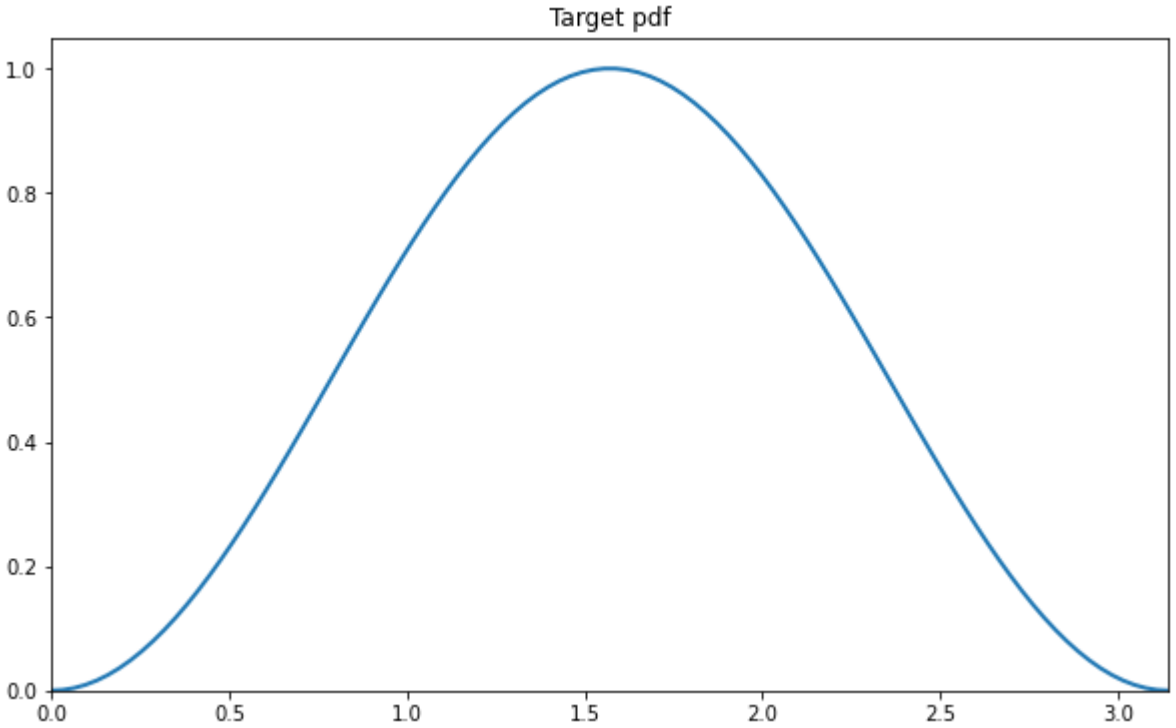
Méthodes Monte Carlo et chaînes de Markov

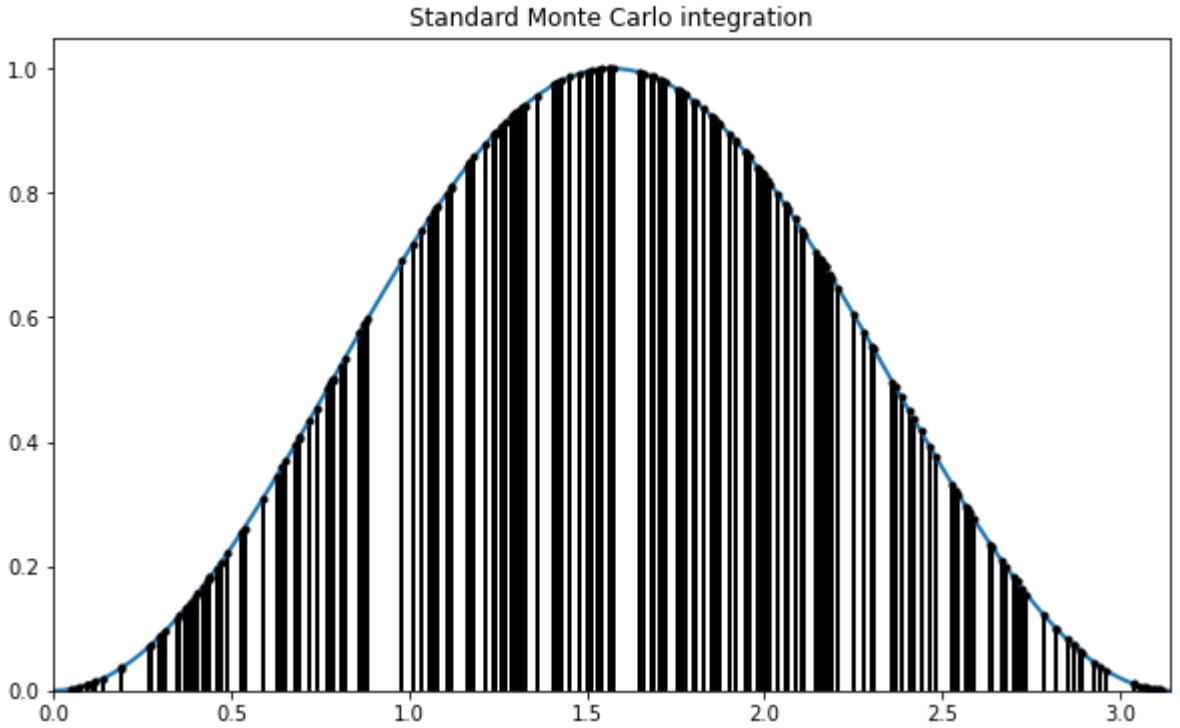
Notebook : https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/Sampling_Importance_Rejection.ipynb

Notebook : https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/MCMC_MH.ipynb

Notebook : https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/Supernova_MCMC_HMC.ipynb

Intégration Monte Carlo : échantillonnage standard





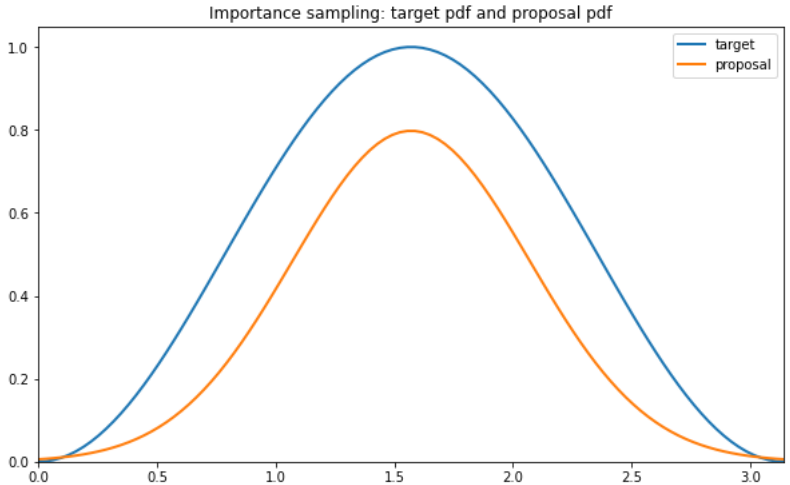
$$I = \int_a^b f(x)\,\mathrm{d}x$$

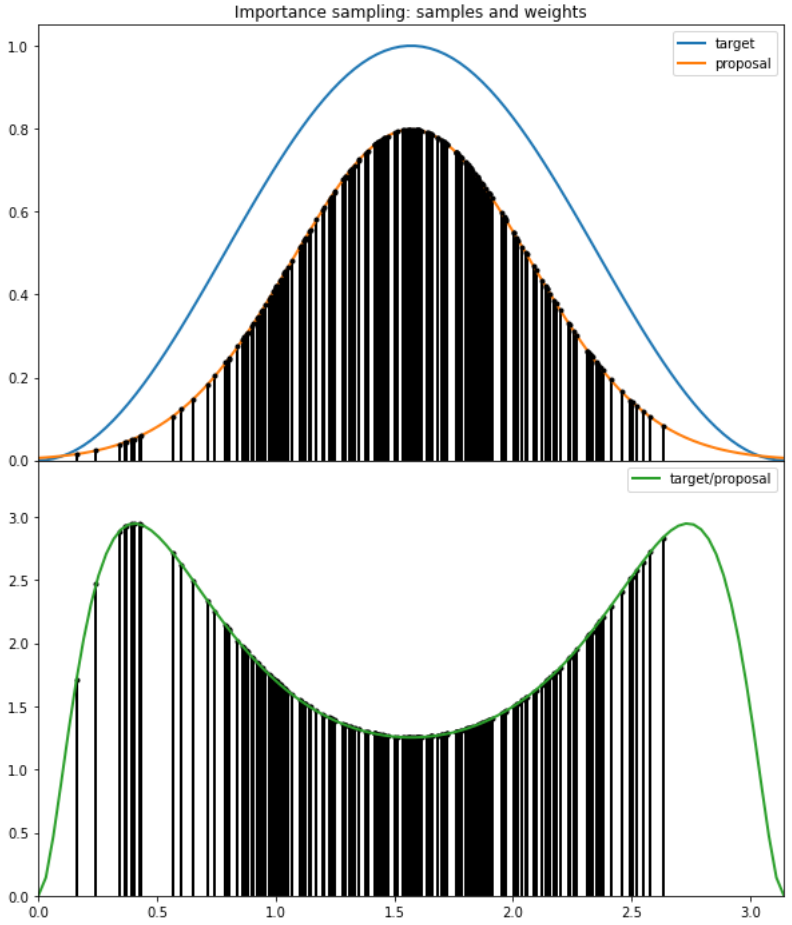
$$I \approx \frac{b-a}{N_{\text{samples}}} \sum_{i=1}^{N_{\text{samples}}} x_i$$

```
In [6]: trueI=quad(target_pdf,a,b)[0]
        trueI

Out[6]: 1.5707963267948966
```

```
In [9]: StandardMonteCarloI=np.sum(samples)*(b-a)/Nsamp
        StandardMonteCarloI

Out[9]: 1.660402945484072
```

Intégration Monte Carlo : échantillonnage préférentiel (*importance sampling*)





$$I \approx \frac{\sum_{i=1}^{N_{\text{samples}}} x_i w_i}{\sum_{i=1}^{N_{\text{samples}}} w_i}$$

```
In [14]: ImportanceI=np.average(samples,weights=weights)
         ImportanceI

Out[14]: 1.5100957559182684
```

Intégration Monte Carlo : ré-échantillonnage préférentiel (*importance resampling*)

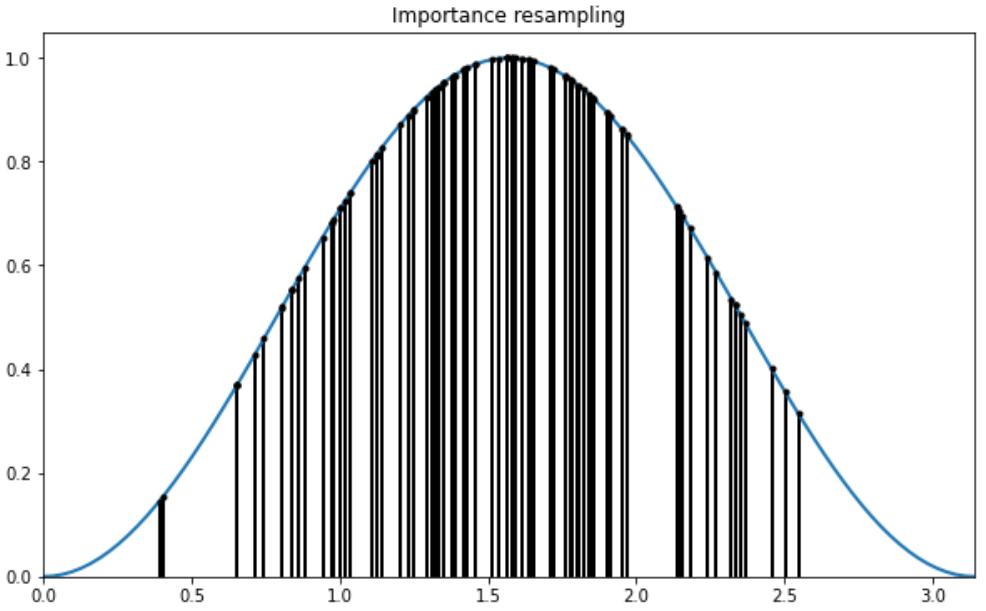
Importance resampling

A problem with importance sampling is the situation in which all but one of the weights are close to zero. To avoid with situation, we can do **importance resampling**. We draw Nresamp new samples from the current sample set with probabilities proportional to their weights. We replace the current samples with this new set, and the the current weights by 1/Nresamp (drawing according to the importance weight replaces likelihoods by frequencies).

In [15]:
```
Nresamp=100
normalizedweights=weights/np.sum(weights)
resamples=np.random.choice(samples, size=Nresamp, replace=True, p=normalizedweights)
reweights=1./Nresamp*np.ones(Nresamp)
```

Weights are then updated given their likelihood, as in the previous importance sampling step.

In [16]:
```
reweights*=target_pdf(resamples)/(1./Nresamp)
```

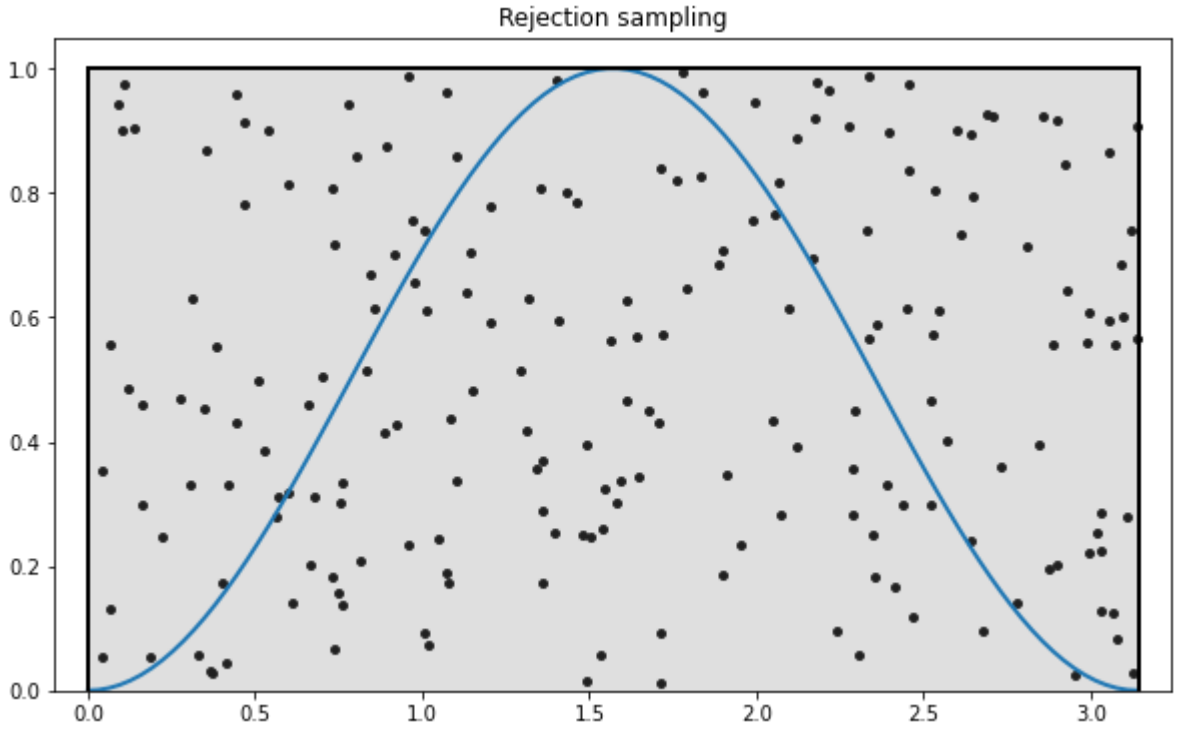


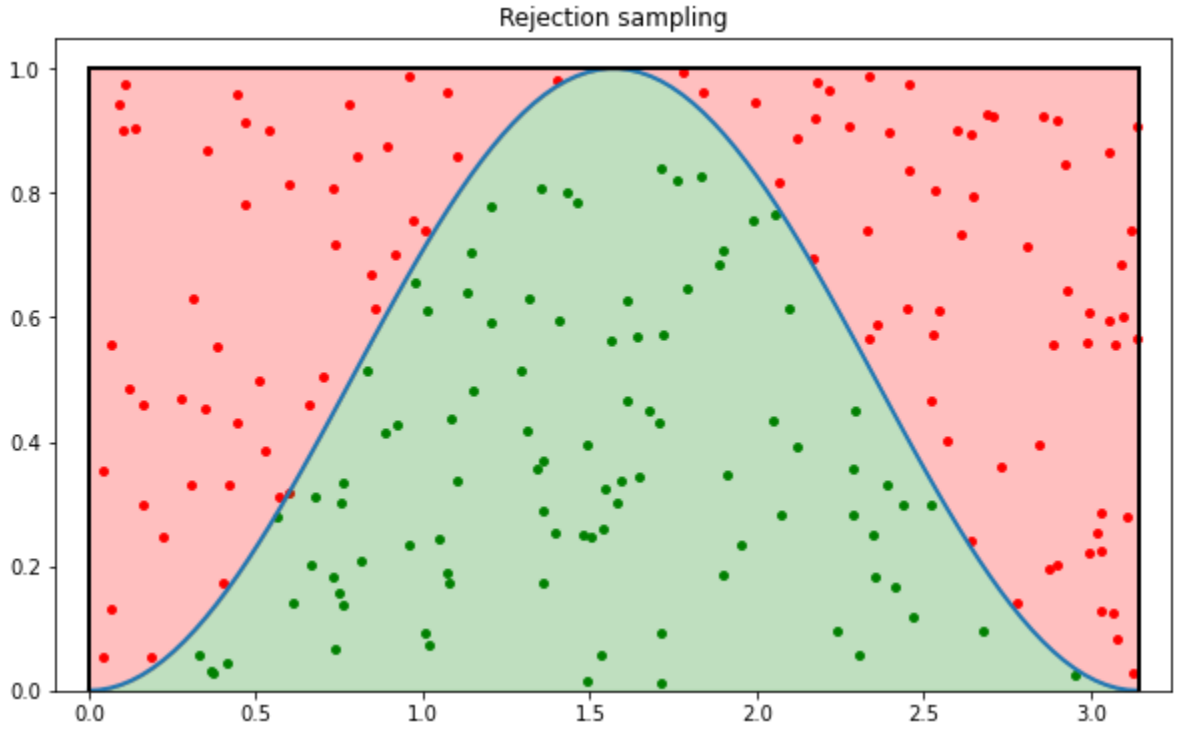
$$I \approx \frac{\sum_{i=1}^{N_{\text{samples}}} x_i w_i}{\sum_{i=1}^{N_{\text{samples}}} w_i}$$

In [18]:
```
ImportanceReI=np.average(resamples,weights=reweights)
ImportanceReI
```

Out[18]: 1.531675529534044

Intégration Monte Carlo : échantillonnage par rejet (*rejection sampling*)






```
In [23]: fraction=float(len(accepted_samples))/Nsamp
         fraction

Out[23]: 0.465

In [24]: fraction=float(len(accepted_samples))/Nsamp
         RejectionI=fraction*upperbound*(b-a)
         RejectionI

Out[24]: 1.460840583919254
```

Chaînes de Markov Monte Carlo (Markov Chain Monte Carlo – MCMC)

- Propriété de Markov :
  - L'information utile pour la prédiction du futur est entièrement contenue dans l'état présent du processus et n'est pas dépendante des états antérieurs (le système n'a pas de « mémoire »).
  - Mathématiquement : la distribution conditionnelle de probabilité des états futurs, étant donnés les états passés et l'état présent, ne dépend que de l'état présent et non pas des états passés.
- Algorithme de Metropolis-Hastings :

begin
  initialise $\boldsymbol{x}_{(0)}$;
  for $i = 1$ *to* n **do**
    $\boldsymbol{x}^* \curvearrowleft q(\boldsymbol{x}^*|\boldsymbol{x})$ (proposal distribution);
    $\alpha \curvearrowleft \mathcal{U}(0,1)$ (uniform distribution);
    if $\alpha < \min[1, r(\boldsymbol{x}, \boldsymbol{x}^*)]$ **then**
      $\boldsymbol{x}_{(i)} = \boldsymbol{x}^*$;
    else
      $\boldsymbol{x}_{(i)} = \boldsymbol{x}_{(i-1)}$;
    end
  end
  return $(\boldsymbol{x}_{(0)}, \ldots, \boldsymbol{x}_{(n)})$;
end

Cas général : $r(\boldsymbol{x}, \boldsymbol{x}^*) \equiv \frac{p(\boldsymbol{x}^*)}{p(\boldsymbol{x})} \frac{q(\boldsymbol{x}|\boldsymbol{x}^*)}{q(\boldsymbol{x}^*|\boldsymbol{x})}$ (ratio de Hastings)

Cas particulier : $r(\boldsymbol{x}, \boldsymbol{x}^*) = \frac{p(\boldsymbol{x}^*)}{p(\boldsymbol{x})}$ (Metropolis update)

dans le cas d'un *proposal* symétrique : $q(\boldsymbol{x}^*|\boldsymbol{x}) = q(\boldsymbol{x}|\boldsymbol{x}^*)$

Algorithme de Metropolis-Hastings : implémentation

begin
  initialise $\boldsymbol{x}_{(0)}$;
  for $i = 1$ *to* n **do**
    $\boldsymbol{x}^* \curvearrowleft q(\boldsymbol{x}^*|\boldsymbol{x})$ (proposal distribution);
    $\alpha \curvearrowleft \mathcal{U}(0,1)$ (uniform distribution);
    if $\alpha < \min[1, r(\boldsymbol{x}, \boldsymbol{x}^*)]$ **then**
      $\boldsymbol{x}_{(i)} = \boldsymbol{x}^*$;
    else
      $\boldsymbol{x}_{(i)} = \boldsymbol{x}_{(i-1)}$;
    end
  end
  return $(\boldsymbol{x}_{(0)}, \ldots, \boldsymbol{x}_{(n)})$;
end

```python
def MH_sampler(Ntries,theta_start,Ntrials,Nsuccesses,lh,prior,proposal_sigma):
    Naccepted=0
    samples=np.zeros(Ntries+1)
    samples[0]=theta_start
    theta=theta_start
    for i in range(Ntries):
        theta_p = theta + proposal_pdf(proposal_sigma).rvs()
        # the Gaussian proposal pdf satisfies the detailed balance equation, so the
        # acceptance ratio simplifies to the Metropolis ratio
        a = min(1, target_pdf(theta_p,Ntrials,Nsuccesses,lh,prior)/target_pdf(theta,Ntrials,Nsuccesses,lh,prior))
        u = np.random.uniform()
        if u < a:
            Naccepted+=1
            theta=theta_p
        samples[i+1] = theta
    return Naccepted, samples
```

Cas général : $r(\boldsymbol{x}, \boldsymbol{x}^*) \equiv \frac{p(\boldsymbol{x}^*)}{p(\boldsymbol{x})} \frac{q(\boldsymbol{x}|\boldsymbol{x}^*)}{q(\boldsymbol{x}^*|\boldsymbol{x})}$ (ratio de Hastings)

Cas particulier : $r(\boldsymbol{x}, \boldsymbol{x}^*) = \frac{p(\boldsymbol{x}^*)}{p(\boldsymbol{x})}$ (Metropolis update)

dans le cas d'un *proposal* symétrique : $q(\boldsymbol{x}^*|\boldsymbol{x}) = q(\boldsymbol{x}|\boldsymbol{x}^*)$

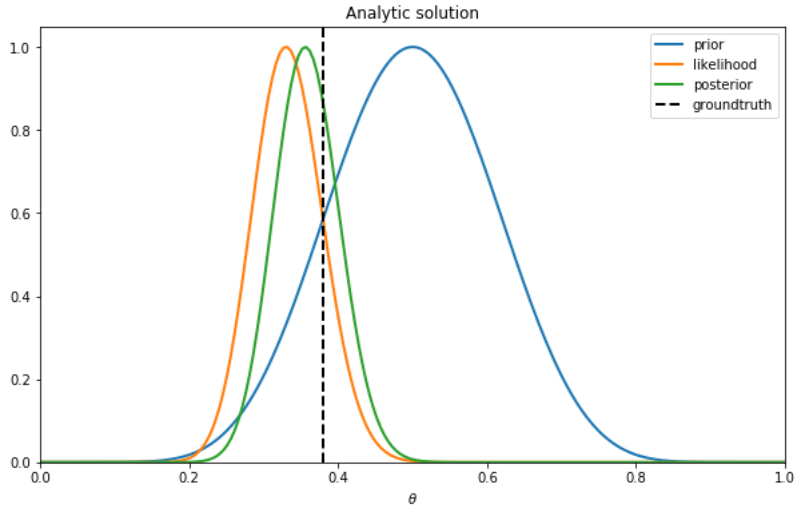
Exemple de MCMC

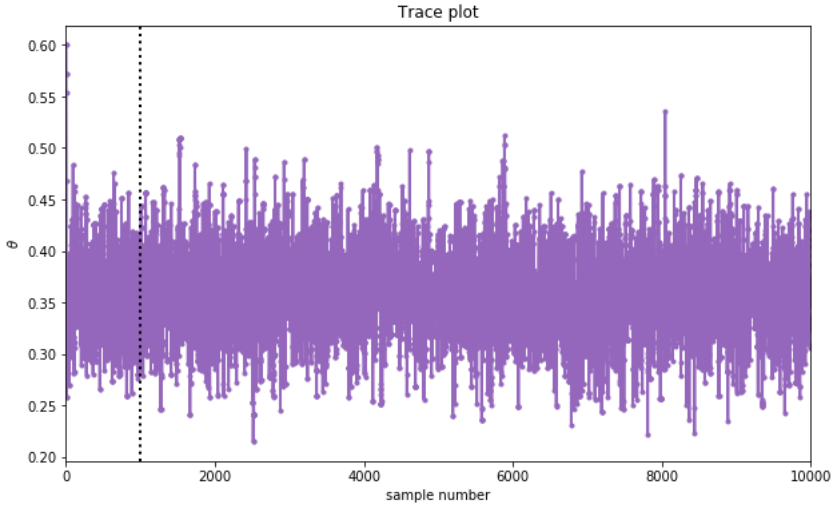
- Problème considéré : une expérience de Bernouilli (N_{trials} expériences indépendantes avec une probabilité de succès θ).
- La vraisemblance pour ce problème est la distribution binomiale :

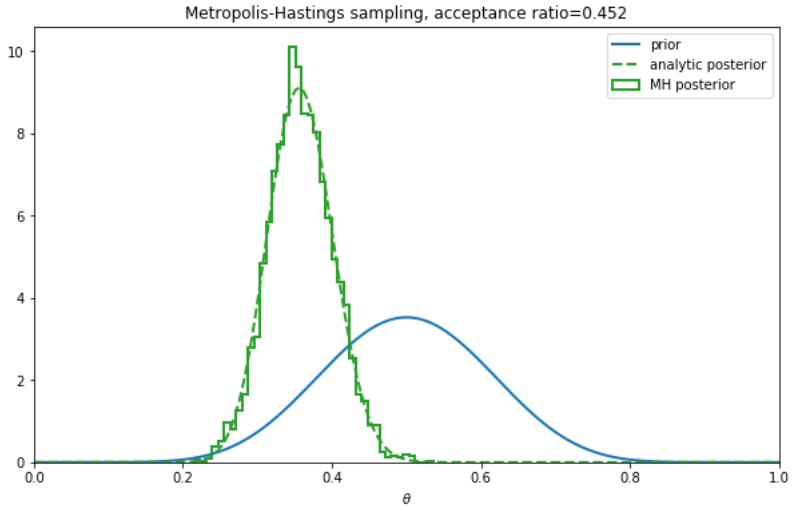
$$p(N_{\text{successes}}, N_{\text{trials}}, \theta) = \binom{N_{\text{trials}}}{N_{\text{successes}}} \theta^{N_{\text{successes}}} (1-\theta)^{N_{\text{trials}} - N_{\text{successes}}}$$

- Résultat analytique : la distribution beta est une famille de **priors conjugués**, c'est-à-dire : si le prior est $\mathcal{B}(\alpha, \beta)$, alors le posterior est $\mathcal{B}(\alpha', \beta')$ avec $\alpha' = \alpha + N_{\text{sucesses}}$

$$\beta' = \beta + N_{\text{trials}} - N_{\text{sucesses}}$$

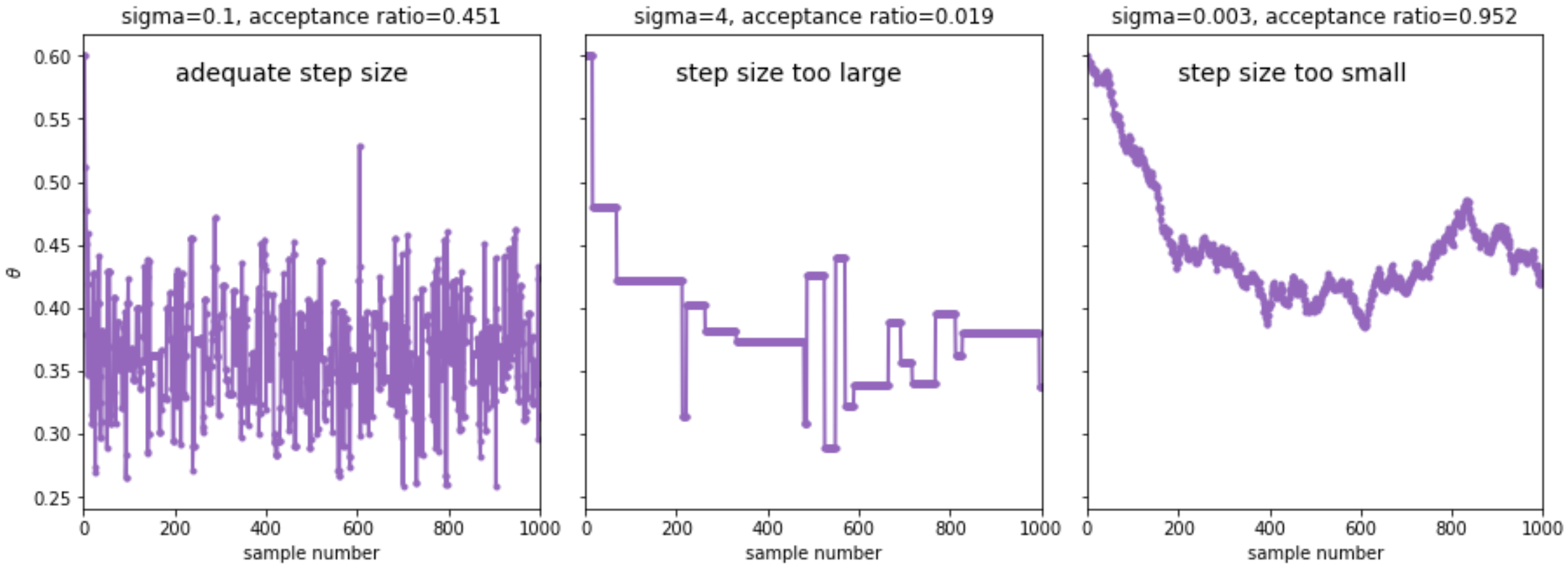






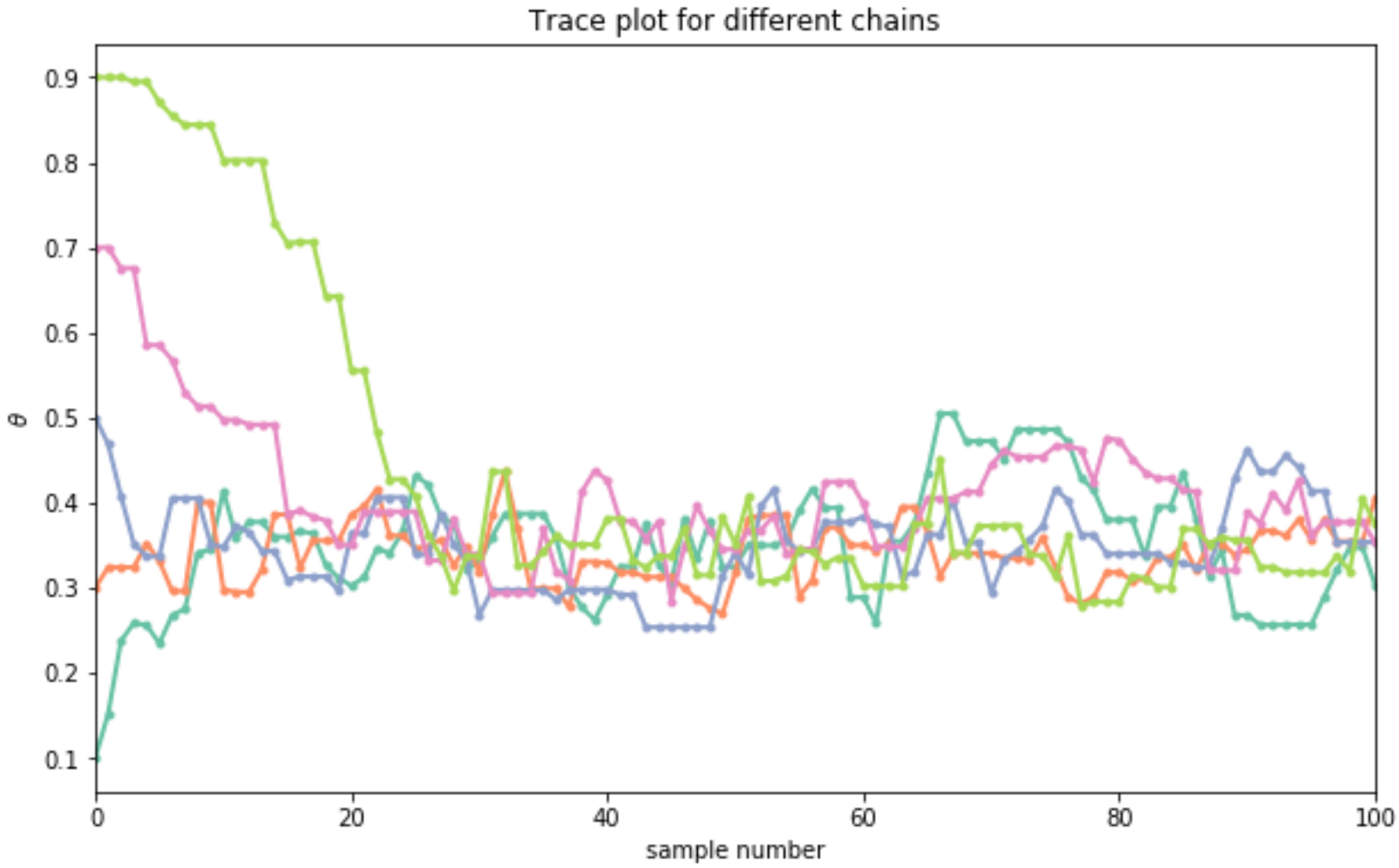
Diagnostics des chaînes de Markov : les *trace plots*

- L'ajustement du *proposal* :



Diagnostics des chaînes de Markov : les *trace plots*

- Plusieurs chaînes indépendantes, différents points de départ :



Diagnostics des chaînes de Markov : convergence – le test de Gelman-Rubin

- Paramètres:
  - m : nombre de chaînes
  - n : longueur des chaînes

$$\bar{\psi}_{.j} = \frac{1}{n}\sum_{i=1}^{n}\psi_{ij}$$

- Définitions :
  - Variance inter-chaînes (*"between" chains variance*) : $B \equiv \frac{n}{m-1}\sum_{j=1}^{m}\left(\bar{\psi}_{.j} - \bar{\psi}_{..}\right)^2$

$$\bar{\psi}_{..} = \frac{1}{m}\sum_{j=1}^{m}\bar{\psi}_{.j}$$

  - Variance intra-chaînes (*"within" chains variance*) : $W \equiv \frac{1}{m}\sum_{j=1}^{m}s_j^2$

$$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\psi_{ij} - \bar{\psi}_{.j}\right)^2$$

- Estimateurs (de la variance du posterior marginal de chacun des paramètres) :
  - $\widehat{\mathrm{var}}^- \equiv W$ sous-estime la variance
  - $\widehat{\mathrm{var}}^+ \equiv \frac{n}{n-1}W + \frac{1}{n}B$ sur-estime la variance
- Test de Gelman-Rubin :
  - Facteur de réduction d'échelle potentiel (*Potential scale reduction factor* – PSRF) : $\widehat{R} \equiv \sqrt{\frac{\widehat{\mathrm{var}}^+}{\widehat{\mathrm{var}}^-}}$
  - Test : $\widehat{R} \to 1$ quand $n \to \infty$. On vise typiquement $\widehat{R} - 1 \lesssim 10^{-2}$.

Priors d'ignorance et le principe de l'entropie maximale

Notebook : https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/LighthouseProblem.ipynb

Notebook : https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/MaximumEntropy.ipynb

- **Priors d'ignorance** : l'idée est d'imposer l'invariance de l'état de connaissance selon une transformation :

$$p(T(x))\mathrm{d}T(x) = p(x)\mathrm{d}x$$

- Cas le plus simple : symétrie sous l'échange de deux modèles $\mathcal{M}_1$ et $\mathcal{M}_2$:

$$\begin{array}{l} p(\mathcal{M}_1) = p(\mathcal{M}_2) \\ p(\mathcal{M}_1) + p(\mathcal{M}_2) = 1 \end{array} \Rightarrow p(\mathcal{M}_1) = p(\mathcal{M}_2) = \frac{1}{2}$$

Symétrie $\mathbb{Z}_2$

- Incertitude maximale sur un « paramètre de localisation » de la distribution : $T(x) = x + s \quad \forall s$

$$\begin{array}{l} \mathrm{d}T = \mathrm{d}x \\ p(x) = p(x+s) \quad \forall s \end{array} \Rightarrow p(x) = C$$

Prior plat

Symétrie $U(1)$

- Incertitude maximale sur un « paramètre d'échelle » de la distribution : $T(x) = ax \quad \forall a$

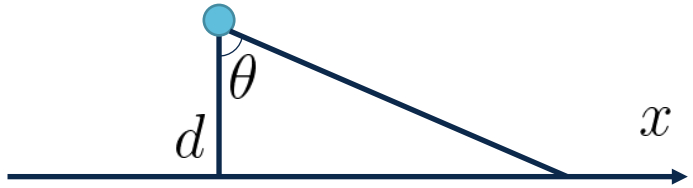
$$\begin{array}{l} \mathrm{d}T = a\,\mathrm{d}x \\ p(x) = a\,p(ax) \quad \forall a \end{array} \Rightarrow p(x) = C/a$$

Prior de Jeffreys

Symétrie $U(1)$

- Cas général : spécifier un **groupe de transformations** et résoudre une **équation fonctionnelle**.

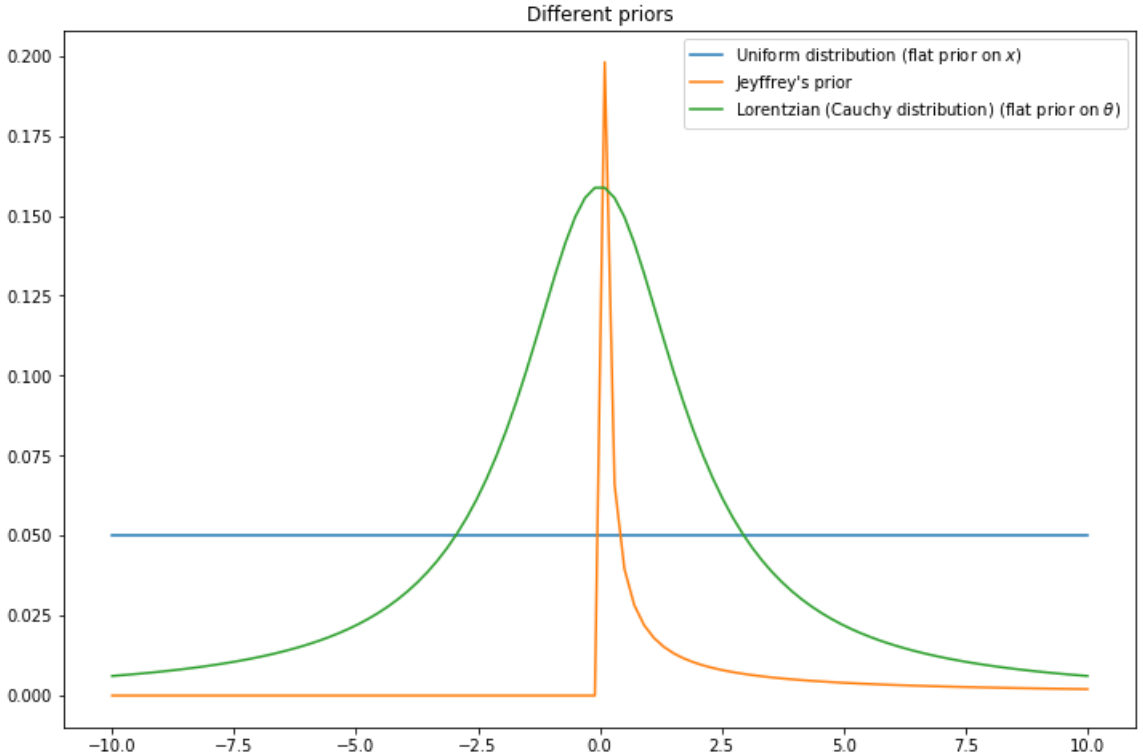
Le problème du phare (*the lighthouse problem*)

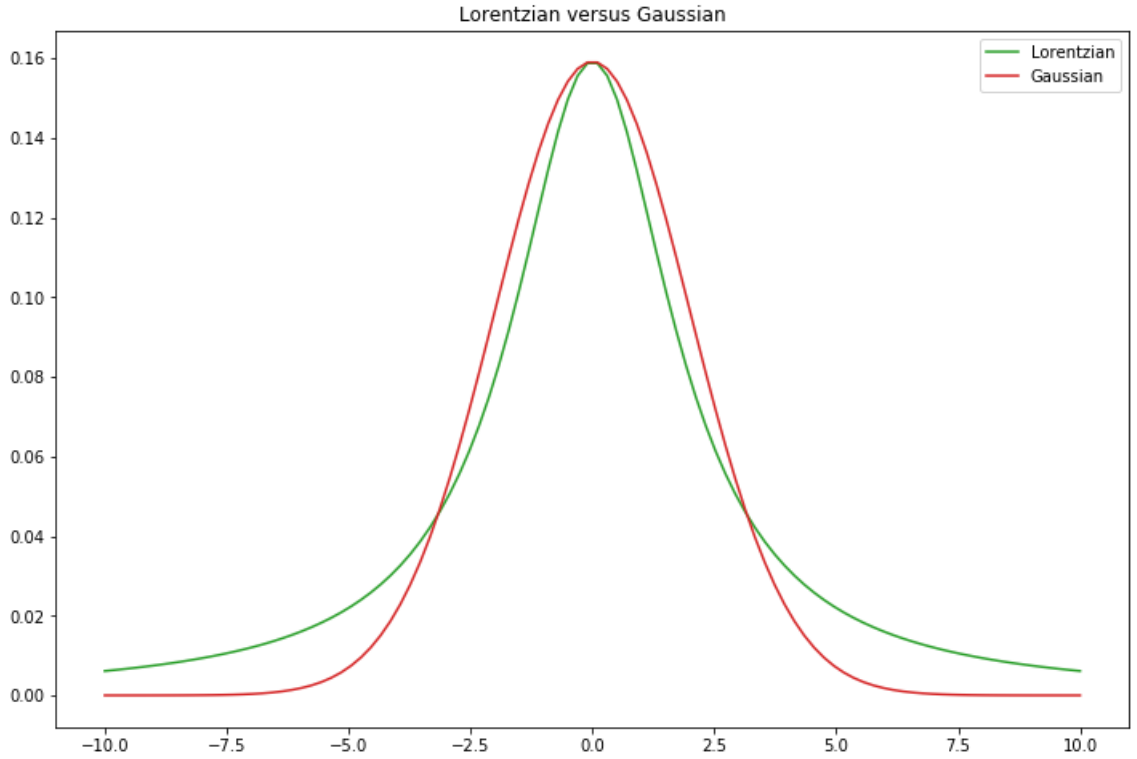


$$x = d\tan\theta$$
$$\mathrm{d}x = d(1+\tan^2\theta)\mathrm{d}\theta$$
$$p(x)\mathrm{d}x = p(\theta)\mathrm{d}\theta$$

Si $p(\theta) = C$ ➡ $p(x) \propto \dfrac{d}{d^2 + x^2}$

Lorentzienne/Distribution de Cauchy





- Un état d'ignorance maximale pour une variable n'est généralement pas la même chose qu'un état d'ignorance maximale pour une fonction non-linéaire de cette variable.

Le principe de l'entropie maximale

- Maximiser l'entropie : une méthode générale pour sélectionner un prior en tenant compte de :
  - l'indifférence à l'égard des états de connaissance égale
  - d'informations préalables pertinentes

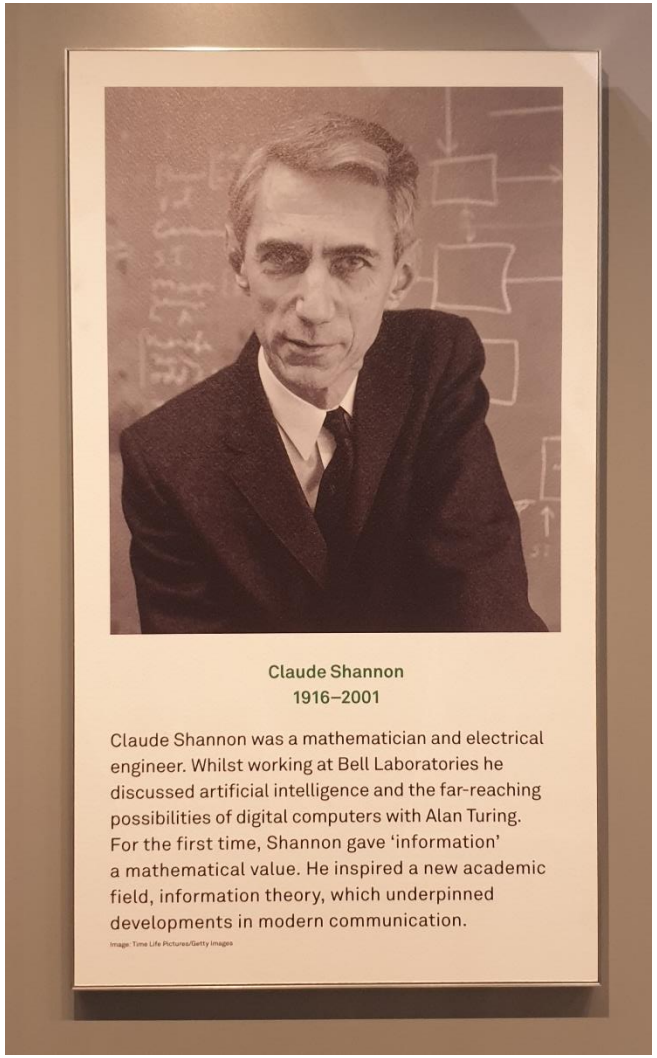
- Quelle forme doit avoir $H[p]$ pour une source d'informations produisant N « mots » finis avec des probabilités respectives p_n ?

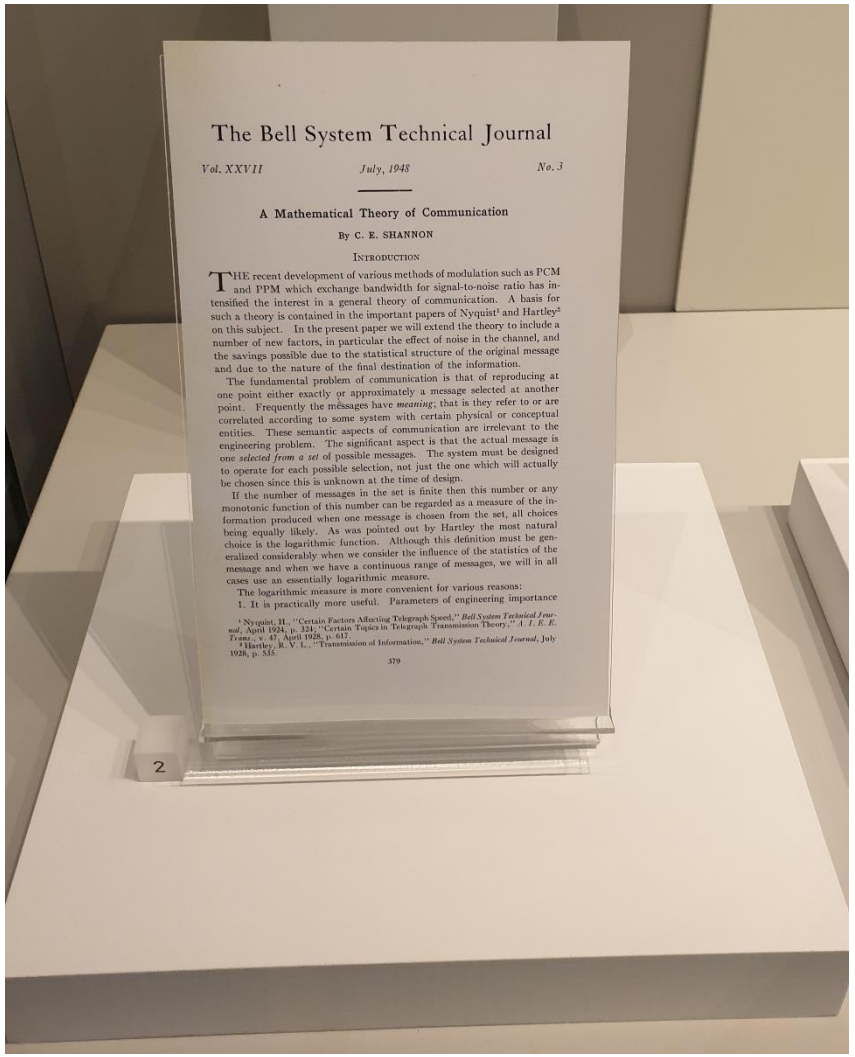
- Désidératas :
  - Si tous les mots sont équiprobables ($\forall n,\ p_n = \frac{1}{N}$), alors $H[p]$ doit croître avec N.
  - Si les mots sont générés en deux étapes (1 : choisir un sous-ensemble de mots ; 2 : choisir un mot dans ce sous-ensemble), alors l'entropie doit être la somme des entropies associées à chaque étape.

➡ Théorème (Shannon): $H[p] \propto -\sum_n p_n \log_2 p_n$

La naissance de la théorie de l'information

- Photos prises au *Science Museum* (South Kensington, Londres) en 2021 :





Pourquoi ne pas appeler ça de l'entropie ? Premièrement, un développement mathématique très semblable au vôtre existe déjà dans la mécanique statistique de Boltzmann, et deuxièmement, personne ne comprend très bien l'entropie, donc dans n'importe quelle discussion, vous serez dans une position avantageuse.

von Neumann à Shannon, à propos d'un nom pour « l'information manquante »

Le problème du dé pipé

- Pour un dé non pipé, $p_n = \frac{1}{6} \quad \forall n \in [\![1, 6]\!]$: le principe d'indifférence suffit.
- Maintenant supposons que la valeur moyenne après de nombreux lancés n'est pas 3,5 mais 4. Quelle est la loi de probabilité dans ce cas ?
- Il nous faut maximiser $H[p]$ avec deux contraintes :

$$< n >_p = \sum_{n=1}^{6} n\, p_n = 4 \qquad \text{et} \qquad \sum_{n=1}^{6} p_n = 1$$

- Méthode (1) : par force brute !
  - Obtenir p_5 et p_6 en fonction de p_1, p_2, p_3, p_4.
  - Exprimer $H[p] = \sum_{n=1}^{6} p_n \ln p_n$ en fonction de p_1, p_2, p_3, p_4.
  - Dériver et résoudre $\frac{\partial H}{\partial p_n} = 0$ pour tout $n \in [\![1, 4]\!]$.

Le problème du dé pipé

- Méthode (2) : une solution plus élégante qui respecte la symétrie des variables : la méthode des **multiplicateurs de Lagrange**.
- On écrit le Lagrangien : $\mathcal{L}[\{p_n\}, \lambda, \mu] = -\sum_{n=1}^{6} p_n \ln p_n - \lambda \left(\sum_{n=1}^{6} n p_n - 4 \right) - \mu \left(\sum_{n=1}^{6} p_n - 1 \right)$
- Les deux contraintes sont : $\dfrac{\partial \mathcal{L}}{\partial \lambda} = 0$ et $\dfrac{\partial \mathcal{L}}{\partial \mu} = 0$

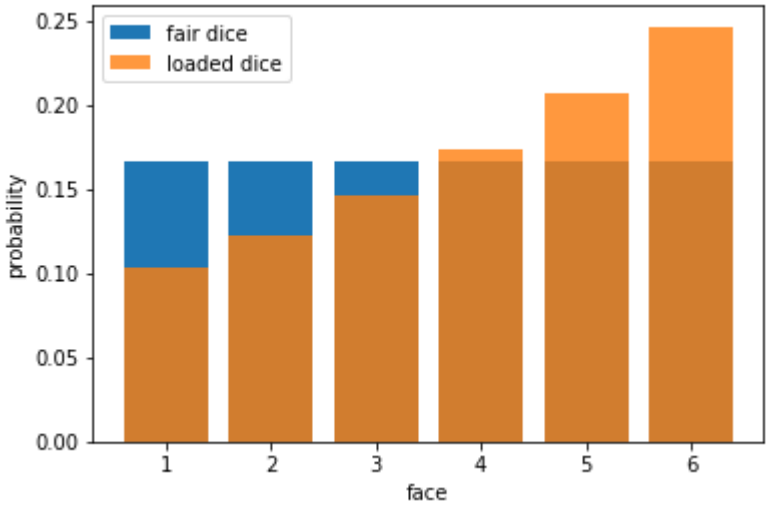
  $\dfrac{\partial \mathcal{L}}{\partial p_n} = 0$ donne $-1 - \ln p_n - \lambda n - \mu = 0$

  $p_n = \dfrac{\mathrm{e}^{-\lambda n}}{Z}$ avec $\ln Z \equiv 1 + \mu$
- La contrainte de normalisation impose : $Z = \sum_{n=1}^{6} \mathrm{e}^{-\lambda n} = \dfrac{1 - \mathrm{e}^{-6\lambda}}{\mathrm{e}^{\lambda} - 1}$
- La contrainte sur la moyenne est obtenue en remarquant que :

$$-\frac{\mathrm{d} \ln Z}{\mathrm{d}\lambda} = -\frac{1}{Z}\frac{\mathrm{d}Z}{\mathrm{d}\lambda} = \sum_{n=1}^{6} n \frac{\mathrm{e}^{-\lambda n}}{Z} = \sum_{n=1}^{6} n\, p_n = 4$$

- Cela donne une équation à résoudre pour e^{λ} : $\mathrm{e}^{\lambda}/(\mathrm{e}^{\lambda} - 1) - 6/(\mathrm{e}^{6\lambda} - 1) = 4$

Le problème du dé pipé



- Ceci est un exemple de théorie des probabilités au-delà des statistiques bayésiennes : nous avons obtenu numériquement les valeurs d'une distribution de probabilité, conditionnée à certaines observations, sans utiliser le théorème de Bayes.
- Analogie thermodynamique :
  - Dé non pipé = ensemble micro-canonique $p_n = \frac{1}{N}$
  - Dé pipé (avec moyenne connue) = ensemble canonique

$$p_n = \frac{\mathrm{e}^{-\beta E_n}}{Z} \qquad \beta \equiv \frac{1}{k_\mathrm{B} T}$$

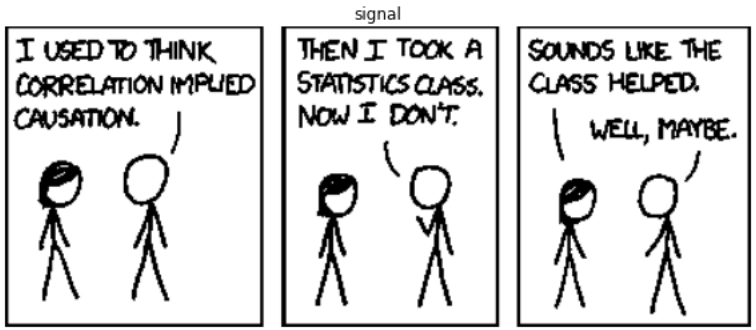
E_n = énergie des différents états

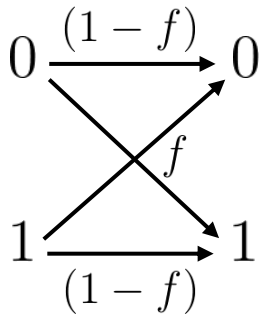
Z = fonction de partition = évidence en statistiques bayésiennes

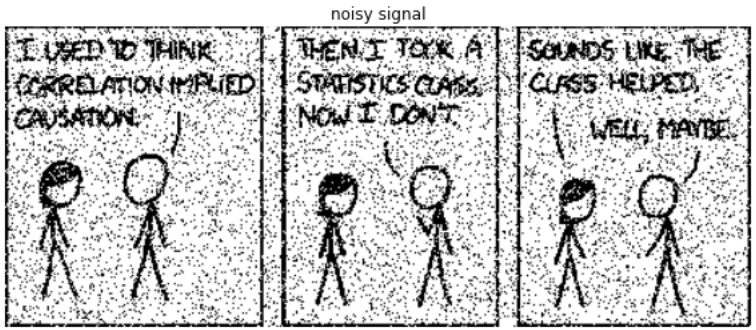
Théorie de l'information

Notebook : https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/IT_noisy_binary_channel.ipynb

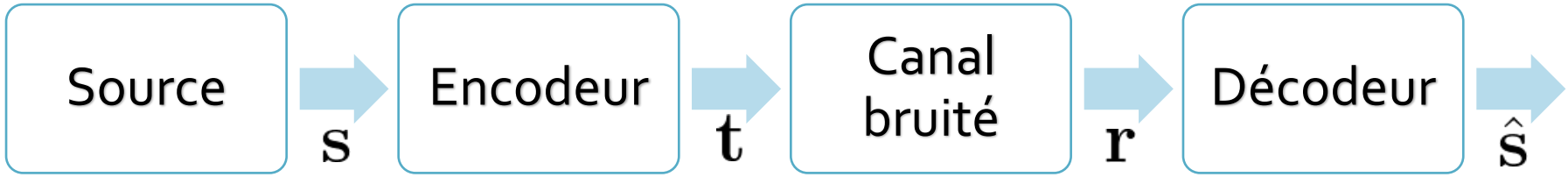
Le canal symétrique binaire bruité







https://xkcd.com/552/



Taux de transfert d'informations : $R = \frac{\#\mathbf{s}}{\#\mathbf{t}} = \frac{K}{N}$

Le code R3

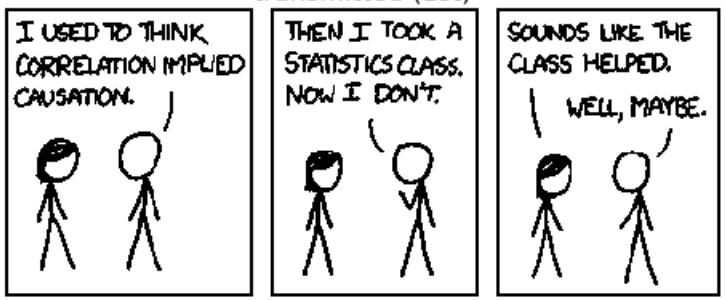
s	0	0	1	0	1	1	0
t	000	000	111	000	111	111	000
n	000	001	000	000	101	000	000
r	000	001	111	000	010	111	000
$\hat{\mathbf{s}}$	0	0	1	0	0	1	0
corrected errors		$\star$					
undetected errors					$\star$		

Taux de transfert d'informations : $R[R_3] = \frac{1}{3}$

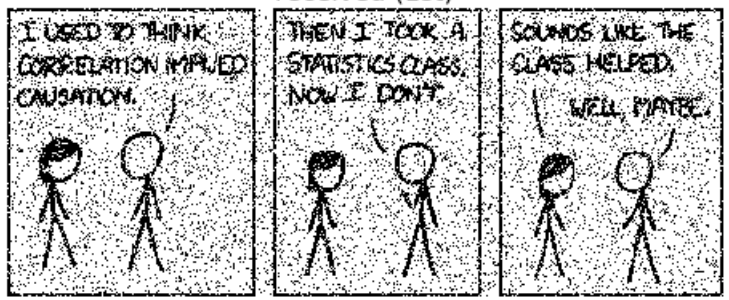
MacKay 2003, chap. 1

Le code R3 : exemple

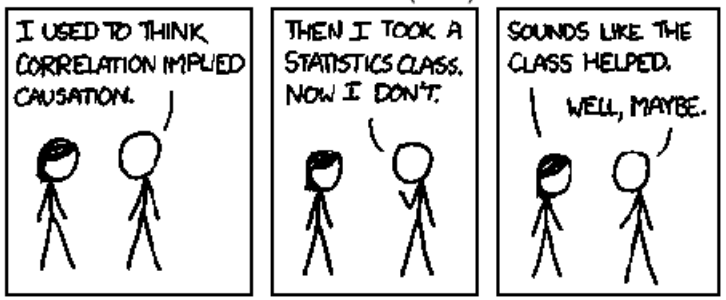
transmitted (1st)



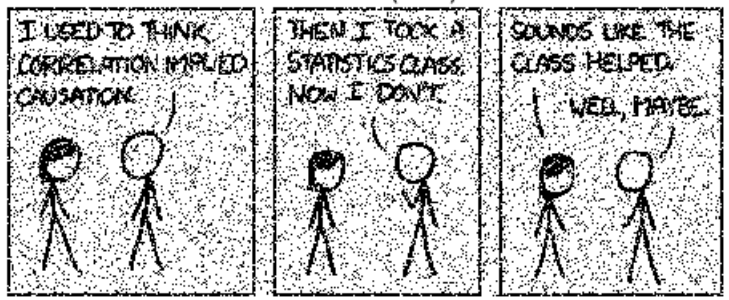
received (1st)



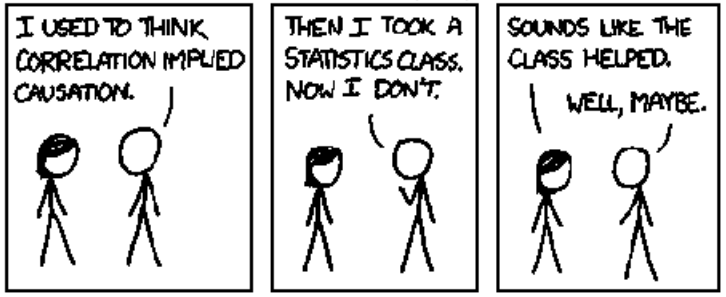
transmitted (2nd)



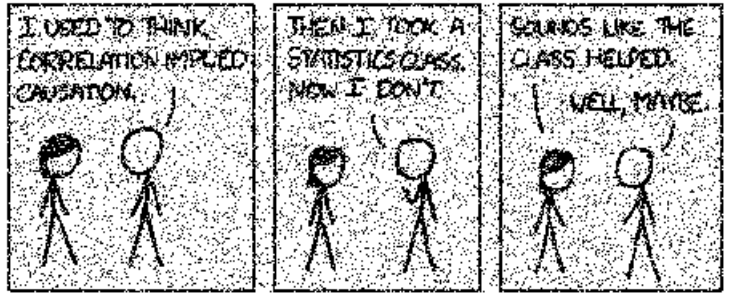
received (2nd)



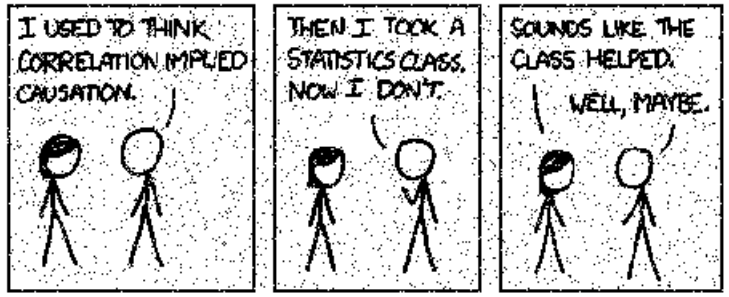
transmitted (3rd)



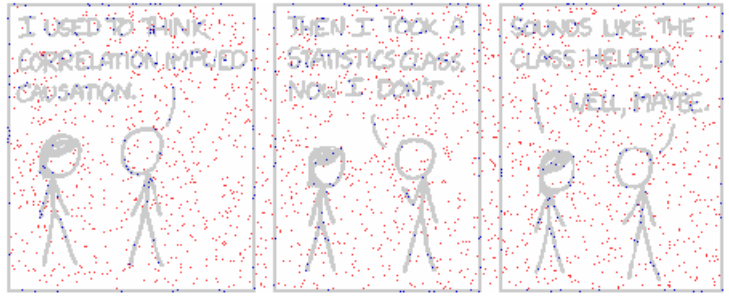
received (3rd)



decoded

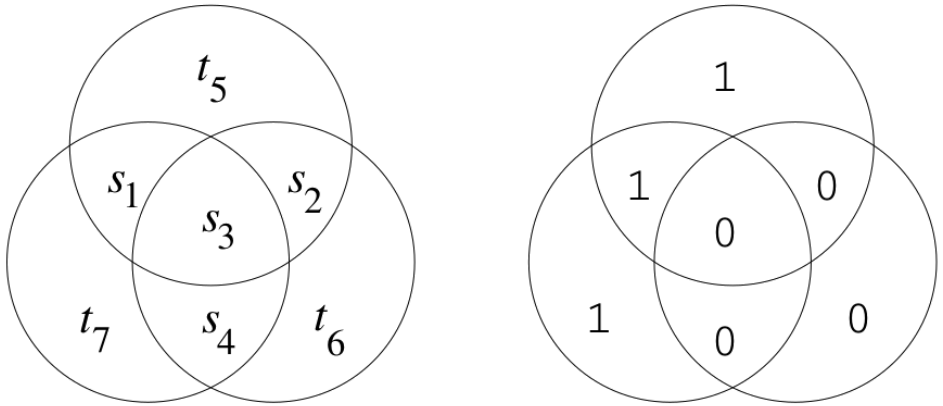


uncorrected errors



Le code de Hamming (7,4) : encodeur

- On introduit le concept de **contrôle de parité** :

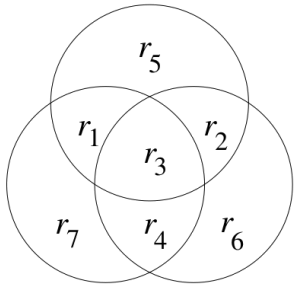


$$\mathbf{t} = \mathbf{G}\mathbf{s} \qquad \mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix}$$

Taux de transfert d'informations : $R[H(7,4)] = \frac{4}{7}$

MacKay 2003, chap. 1

Le code de Hamming (7,4) : décodeur



syndrome

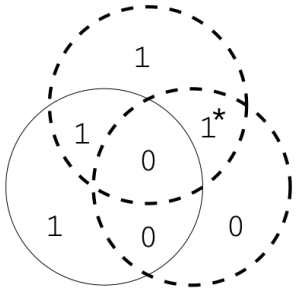
$$\mathbf{z} = \mathbf{Hr}$$

$$\mathbf{H} = \left[\begin{array}{cccc|ccc} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right]$$

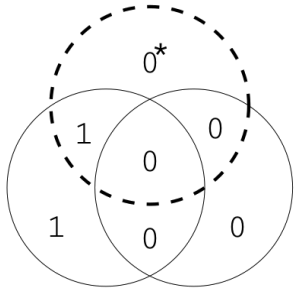
contrôles de parité identité

- On introduit le concept de **syndrome** :

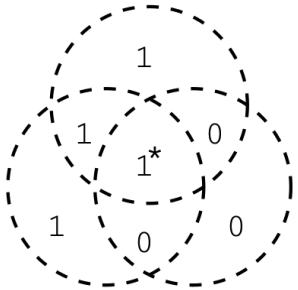
- Algorithme pour la correction d'erreurs :



$\mathbf{z} = 110$



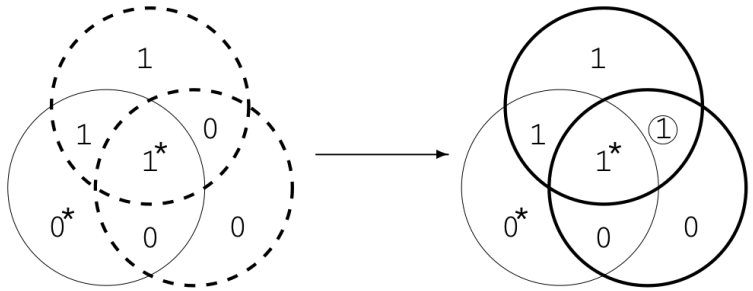
$\mathbf{z} = 100$



$\mathbf{z} = 111$

Syndrome $\mathbf{z}$	000	001	010	011	100	101	110	111
Unflip this bit	None	r_7	r_6	r_4	r_5	r_1	r_2	r_3

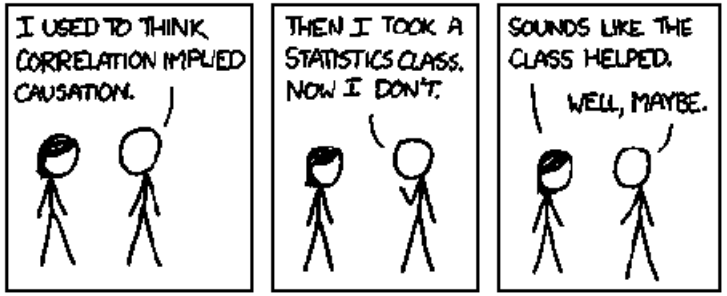
- Exemple de correction d'erreur infructueuse :



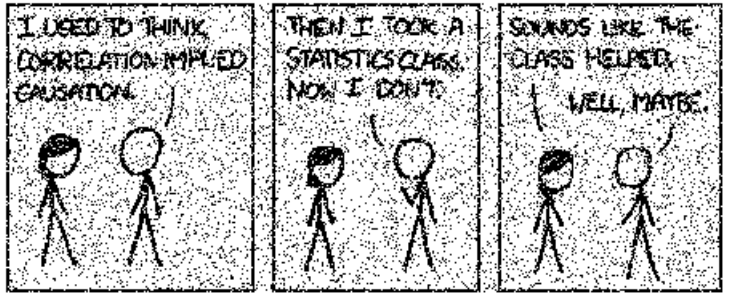
MacKay 2003, chap. 1

Le code de Hamming (7,4) : exemple

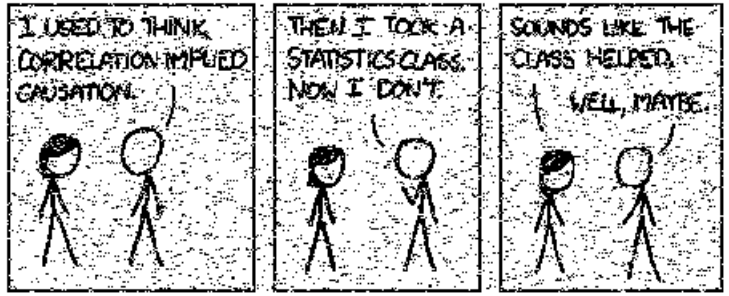
transmitted image



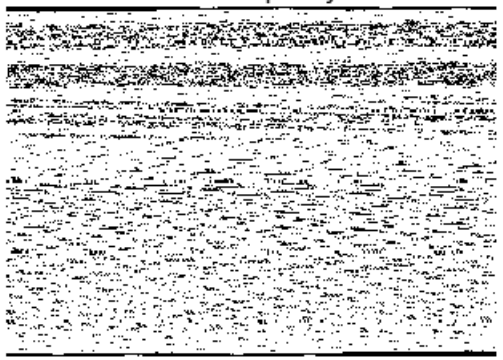
received noisy image



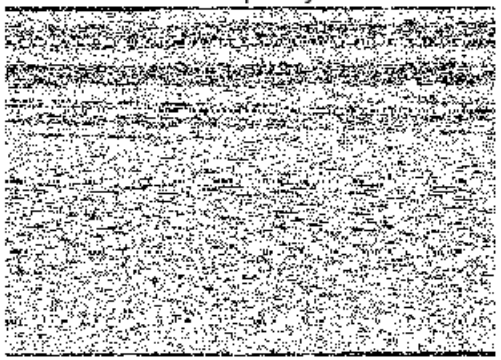
decoded



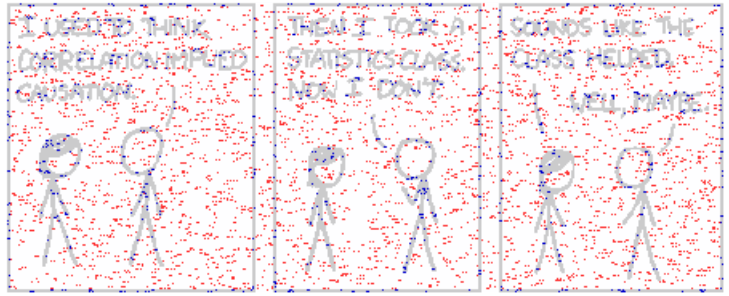
transmitted parity bits



received parity bits

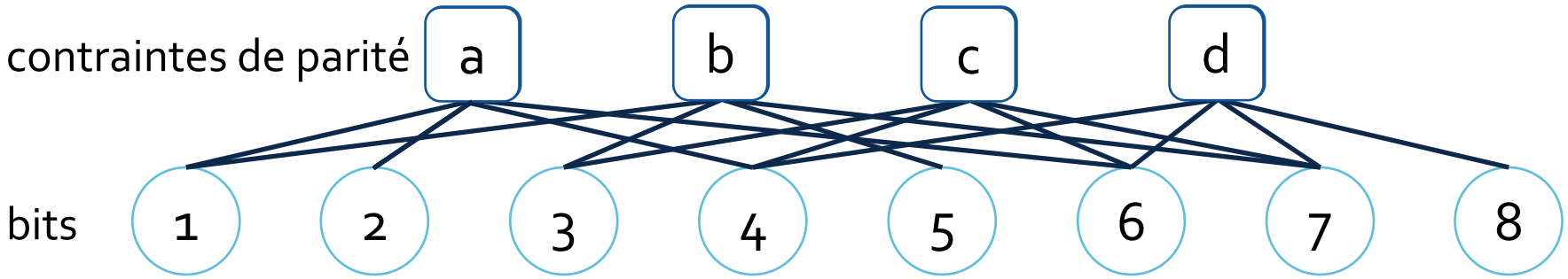


uncorrected errors



Les codes de parité à faible densité (low-density parity-check codes – LDPC)

- Graphe de Tanner :



- *N* bits, *M=(1-R)N* contraintes de parité à contrôler
  - 2^{RN} « mots » possibles le dictionnaire
  - une matrice de contrôle de parité creuse (*sparse matrix*)
- Décodage des codes LDPC : la théorie générale emprunte à la physique statistique : les spins d'Ising en interaction et l'approximation du champ moyen BP (Bethe-Peierls – Belief Propagation).

Le théorème de Shannon pour le canal symétrique binaire bruité

- Taux de transfert d'informations $R[R_N]$ et probabilité d'erreur p_b pour les codes à répétition (pour N impair) :

$$R[R_N] = \frac{1}{N} \qquad p_\mathrm{b} = \sum_{(N+1)/2}^{N} \binom{N}{n} f^n (1-f)^{N-n}$$

MacKay 2003, equation (1.24)

- Définitions :
  - L'entropie (en base 2) :

$$H_2(x) \equiv x \log_2 \frac{1}{x} + (1-x) \log_2 \frac{1}{1-x}$$

  - La capacité du canal bruité :

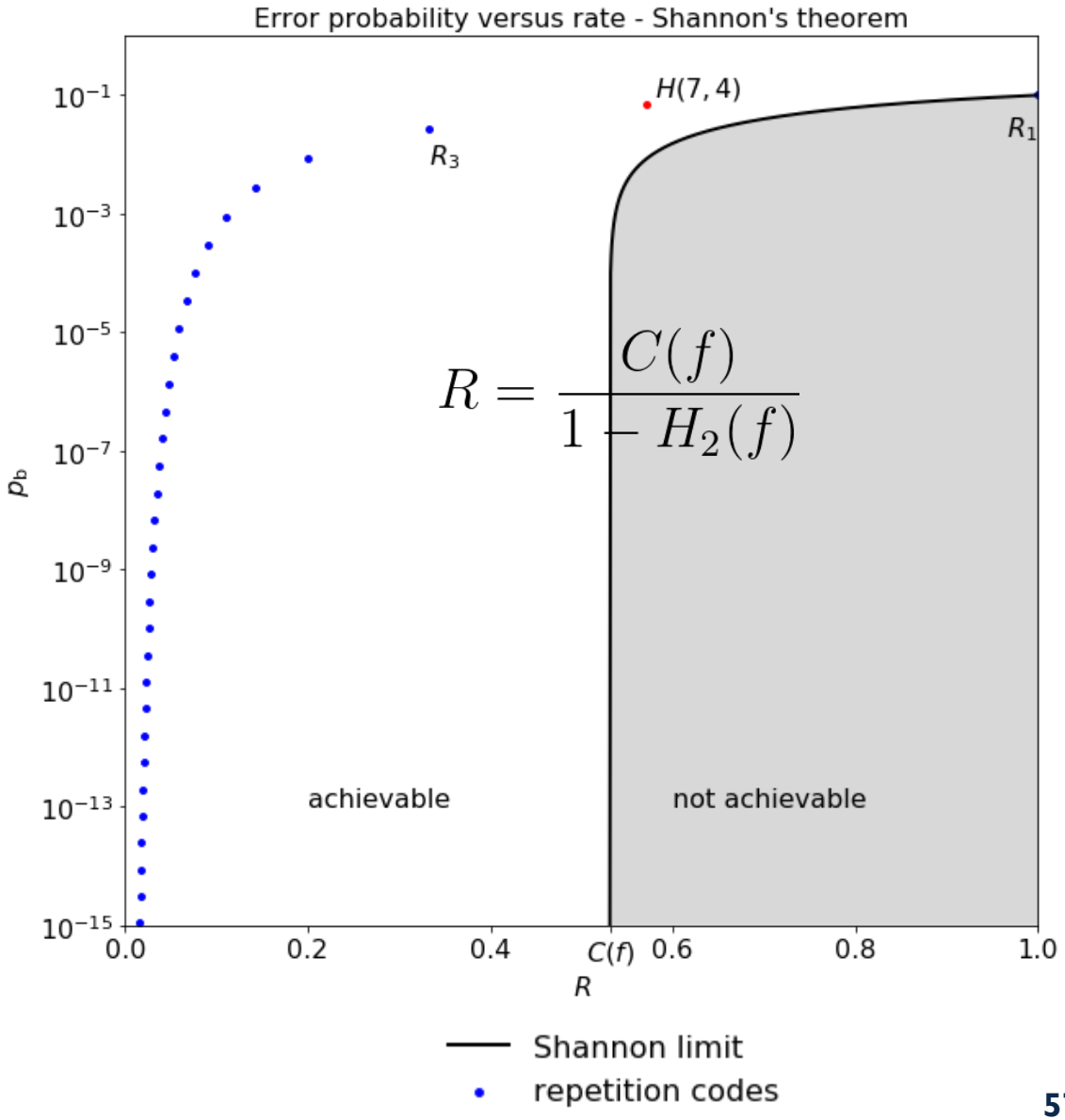
$$C(f) \equiv 1 - H_2(f)$$

  - La limite de Shannon :

$$R = \frac{C(f)}{1 - H_2(f)}$$

Le théorème de Shannon pour le canal symétrique binaire bruité

- Théorème de Shannon (1948) : les taux de transfert d'informations supérieurs à la limite de Shannon sont réalisables, les taux inférieurs à la limite de Shannon ne sont pas réalisables.
- La frontière entre régions réalisables et non-réalisables rencontre l'axe des R à une valeur non-nulle $R = C(f)$. On n'a pas nécessairement $R \to 0$ si $p_\mathrm{b} \to 0$.
- Les taux de transfert d'informations jusqu'à la capacité du canal $R = C(f)$ sont réalisables avec une probabilité d'erreur p_b arbitrairement faible.



Théorie bayésienne de la décision

Notebook : https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/DecisionTheory.ipynb

Théorie bayésienne de la décision

- La théorie bayésienne de la décision est un cadre pour la **prise de décision optimale**, étant donné un ensemble d'actions possibles et un état de connaissance incertain, représenté par une pdf $p(\theta|I)$ (habituellement le posterior d'une inférence bayésienne préalable à la prise de décision).

- Notations :
  - $\{\theta\}$ = ensemble de paramètres (variables observées)
  - $\{a\}$ = ensemble d'actions possibles

- Hypothèse de l'utilité attendue : étant donné un ensemble de fonctions de gain $G(a|\theta)$, la règle de décision optimale consiste à effectuer l'action qui maximise l'utilité attendue $U(a|I)$, définie par

$$U(a|I) \equiv <G(a|\theta)>_{p(\theta|I)} = \int G(a|\theta)p(\theta|I)\,\mathrm{d}\theta$$

- Il faut donc effectuer l'action $a^* = \mathrm{argmax}_a\, U(a|I)$.

Exemple : alertes bayésiennes

- On recherche un évènement E. On a accès à $p(E|I)$ et $p(\bar{E}|I) = 1 - p(E|I)$.
- Il y a deux actions possibles :
  - a_1 = donner l'alerte
  - a_2 = ne rien faire
- Les fonctions d'utilité sont :

détection correcte (un « hit »)

faux positif (une « fausse alarme »)

$$U(a_1|I) = G(a_1|E)p(E|I) + G(a_1|\bar{E})\left[1 - p(\bar{E}|I)\right]$$
$$U(a_2|I) = G(a_2|E)p(E|I) + G(a_2|\bar{E})\left[1 - p(\bar{E}|I)\right]$$

faux négatif (un « miss »)

rejet correct

- Un choix typique de fonctions de gain :

$$G(a_1|E) = G - C \qquad G(a_1|\bar{E}) = -C$$
$$G(a_2|E) = 0 \qquad G(a_2|\bar{E}) = 0$$

le gain attendu pour une détection

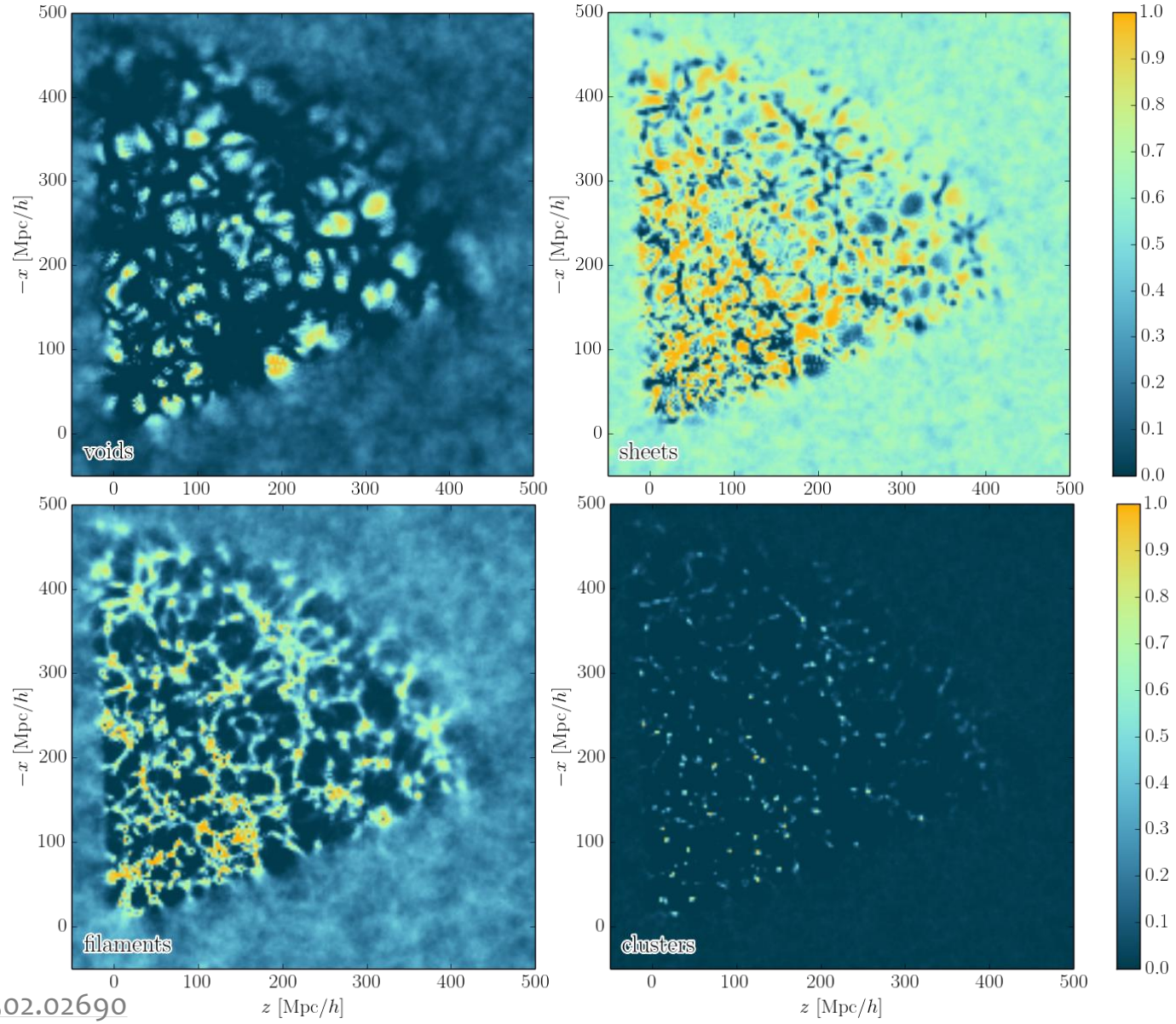
le coût de donner une alerte

- On a donc $U(a_1|I) = p(E|I)(G - C) + [1 - p(E|I)](-C)$

$$U(a_2|I) = 0$$

⇒ il faut donner l'alerte si et seulement si $p(E|I) \geq \frac{G}{C}$

Classification des structures dans la toile cosmique



FL, Jasche & Wandelt 2015a, 1502.02690

Une règle décisionnelle pour la classification des structures dans la toile cosmique

- Une espace de « paramètres d'entrée » :

$$\{\mathrm{T}_0 = \mathrm{void}, \mathrm{T}_1 = \mathrm{sheet}, \mathrm{T}_2 = \mathrm{filament}, \mathrm{T}_3 = \mathrm{cluster}\}$$

- Un espace de « décisions possibles » :

$$\{a_0 = \text{“decide void”}, a_1 = \text{“decide sheet”}, a_2 = \text{“decide filament”},$$
$$a_3 = \text{“decide cluster”}, a_{-1} = \text{“do not decide”}\}$$

- C'est donc un problème de théorie bayésienne de la décision : il faut effectuer l'action qui maximise l'utilité attendue

$$U(a_j(\vec{x}_k)|d) = \sum_{i=0}^{3} G(a_j|\mathrm{T}_i)\, \mathcal{P}(\mathrm{T}_i(\vec{x}_k)|d)$$

- Comment choisir les fonctions de gain ?

FL, Jasche & Wandelt 2015b, 1503.00730

Un jeu de hasard avec l'Univers

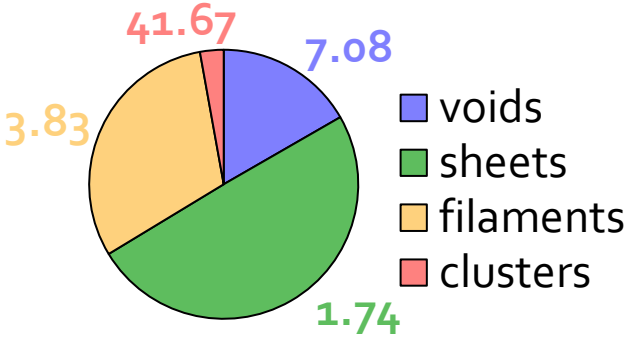
- Une proposition :

$$G(a_j|\mathrm{T}_i) = \begin{cases} \dfrac{1}{\mathcal{P}(\mathrm{T}_i)} - \alpha & \text{if } j \in [\![0,3]\!] \text{ and } i = j \\ -\alpha & \text{if } j \in [\![0,3]\!] \text{ and } i \neq j \\ 0 & \text{if } j = -1. \end{cases}$$

« gagner »

« perdre »

« ne pas jouer »



- En l'absence de données, l'utilité attendue est :

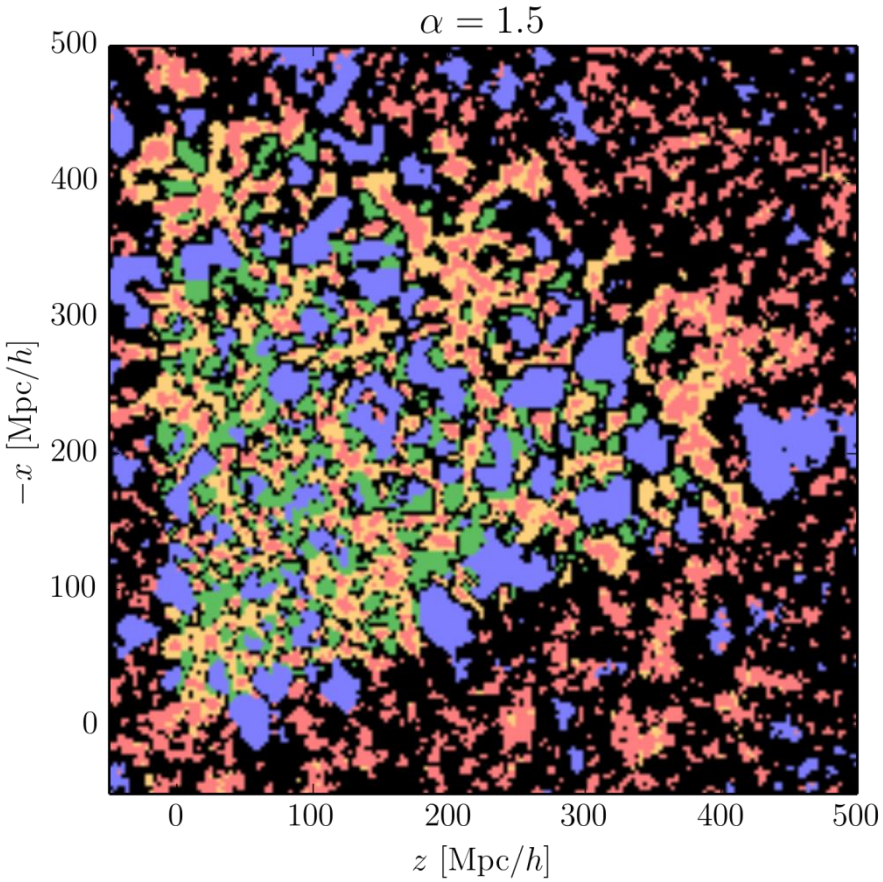
$$U(a_j) = 1 - \alpha \quad \text{if } j \neq 1$$
« jouer au jeu »

$$U(a_{-1}) = 0$$
« ne pas jouer au jeu »

- Pour $\alpha = 1$, c'est un jeu équilibré (*fair game*) ⇒ on va donc toujours jouer ⇒ on construit une « carte spéculative » des structures cosmiques.
- Des valeurs $\alpha > 1$ représentent une aversion pour le risque ⇒ on ne joue que si les données sont informatives ⇒ on construit des « cartes conservatives » des structures cosmiques.

FL, Jasche & Wandelt 2015b, 1503.00730

Jouons au jeu…



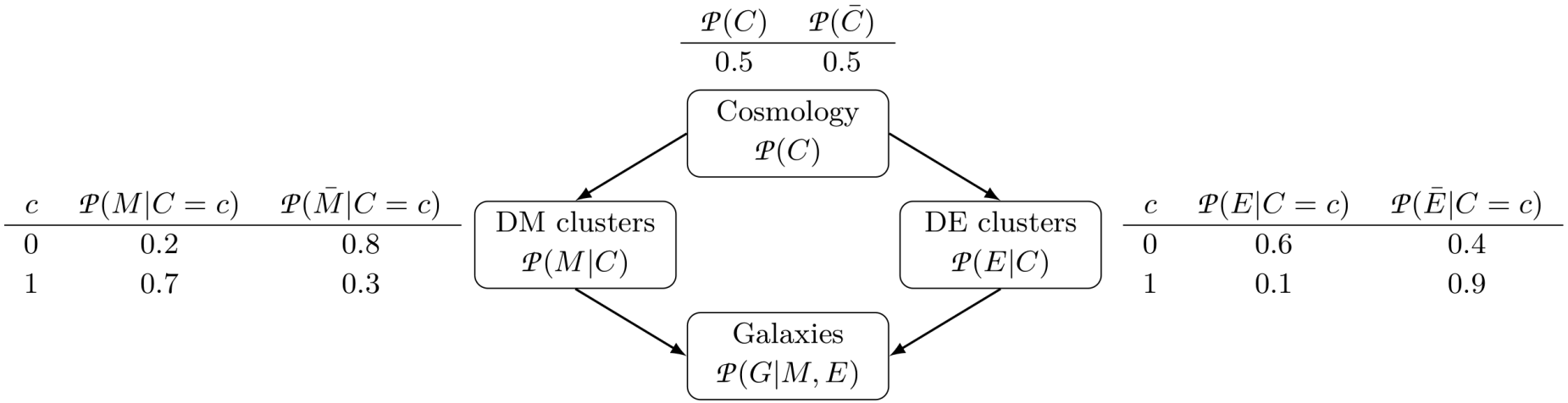
voids filaments undecided

sheets clusters

FL, Jasche & Wandelt 2015b, 1503.00730

Réseaux bayésiens, modèles bayésiens hiérarchiques et Bayésianisme empirique

Réseaux bayésiens



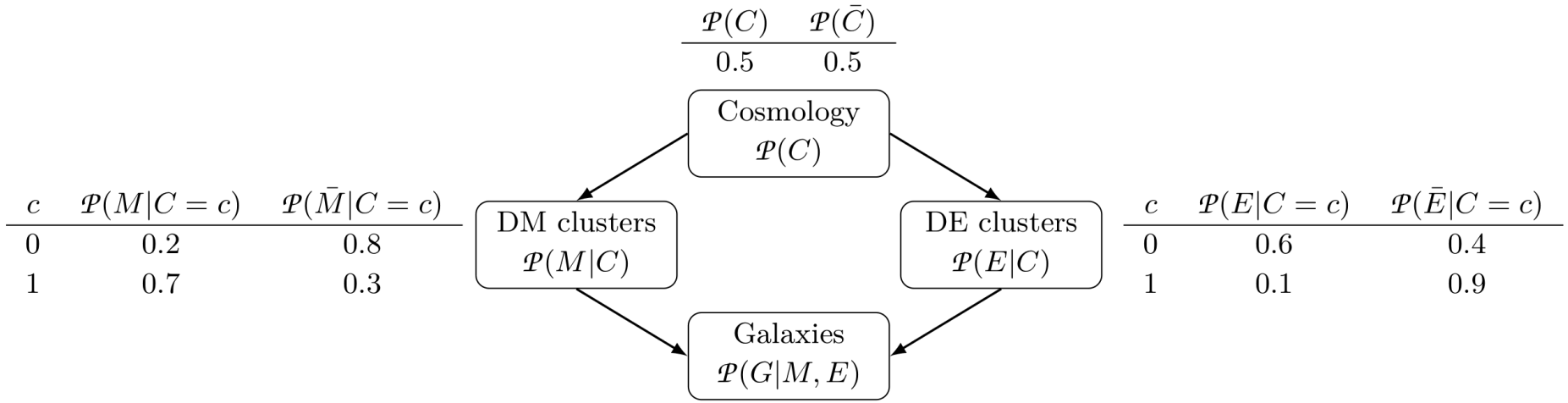
c	$\mathcal{P}(M\|C=c)$	$\mathcal{P}(\bar{M}\|C=c)$
0	0.2	0.8
1	0.7	0.3

c	$\mathcal{P}(E\|C=c)$	$\mathcal{P}(\bar{E}\|C=c)$
0	0.6	0.4
1	0.1	0.9

m	e	$\mathcal{P}(G\|M=m,E=e)$	$\mathcal{P}(\bar{G}\|M=m,E=e)$
0	0	0.10	0.90
1	0	0.95	0.05
0	1	0.95	0.05
1	1	0.99	0.01

- Les réseaux bayésiens sont des modèles probabilistes (avec une représentation graphique), comprenant :
  - Une graphe orienté acyclique (directed acyclic graph – DAG)
  - À chacun des nœuds, les distributions de probabilités conditionnelles

Réseaux bayésiens



c	$\mathcal{P}(M\|C=c)$	$\mathcal{P}(\bar{M}\|C=c)$
0	0.2	0.8
1	0.7	0.3

c	$\mathcal{P}(E\|C=c)$	$\mathcal{P}(\bar{E}\|C=c)$
0	0.6	0.4
1	0.1	0.9

m	e	$\mathcal{P}(G\|M=m,E=e)$	$\mathcal{P}(\bar{G}\|M=m,E=e)$
0	0	0.10	0.90
1	0	0.95	0.05
0	1	0.95	0.05
1	1	0.99	0.01

- Le graphe permet de simplifier l'écriture de la probabilité jointe de toutes les variables :

$$p(C, M, E, G) = p(C)\, p(E|C)\, p(M|C, E)\, p(G|C, M, E)$$

$$p(C, M, E, G) = p(C)\, p(E|C)\, p(M|C)\, p(G|M, E)$$

- Inférence :

$$p(M|G) = \frac{p(M,G)}{p(G)} = \frac{\sum_{c,e} p(C{=}c,M{=}1,E{=}e,G{=}1)}{\sum_{c,m,e} p(C{=}c,M{=}m,E{=}e,G{=}1)} = \frac{0.4313}{0.70305} \approx 0.6135$$

$$p(E|G) = \frac{p(E,G)}{p(G)} = \frac{\sum_{c,m} p(C{=}c,M{=}m,E{=}1,G{=}1)}{\sum_{c,m,e} p(C{=}c,M{=}m,E{=}e,G{=}1)} = \frac{0.3363}{0.70305} \approx 0.4783$$

$$p(\bar{M},\bar{E}|G) = \frac{p(\bar{M},\bar{E},G)}{p(G)} = \frac{\sum_{c} p(C{=}c,M{=}0,E{=}0,G{=}1)}{\sum_{c,m,e} p(C{=}c,M{=}m,E{=}e,G{=}1)} = \frac{0.0295}{0.70305} \approx 0.0420$$

- Prédiction :

$$p(G|C) = \frac{p(G,C)}{p(C)} = \frac{\sum_{m,e} p(C{=}1,M{=}m,E{=}e,G{=}1)}{p(C{=}1)} = 0.7233$$

$$p(E|M,G) = \frac{p(E,M,G)}{p(M,G)} = \frac{\sum_c p(C{=}c,M{=}1,E{=}1,G{=}1)}{\sum_{c,e} p(C{=}c,M{=}1,E{=}e,G{=}1)} = \frac{0.09405}{0.4313} \approx 0.2181$$

$$p(E|G) = \frac{p(E,G)}{p(G)} = \frac{\sum_{c,m} p(C{=}c,M{=}m,E{=}1,G{=}1)}{\sum_{c,m,e} p(C{=}c,M{=}m,E{=}e,G{=}1)} = \frac{0.3363}{0.70305} \approx 0.4783$$

- On a donc à la fois :

$$p(E|M) = p(E)$$
$$p(E|M,G) < p(E|G)$$

- G est appelé un collisionneur. Ce phénomène est appelé « biais de collision » ou « explaining away » : deux causes rentrent en compétition pour expliquer le même effet.
- Cas particulier : le « biais de sélection » ou « paradoxe de Berkson » :

$$0 < p(A) < 1; \quad 0 < p(B) < 1; \quad p(A|B) = p(A)$$

$$C = A + B \quad \Rightarrow \quad \begin{array}{l} p(A|B,C) < p(A|C) \\ p(A|\bar{B},C) = 1 > p(A|C) \end{array}$$

Le biais de Malmquist

- Le biais de Malmquist (1925) : dans les relevés astronomiques limités par la magnitude des objets, les objets éloignés sont préférentiellement détectés s'ils sont intrinsèquement brillants.

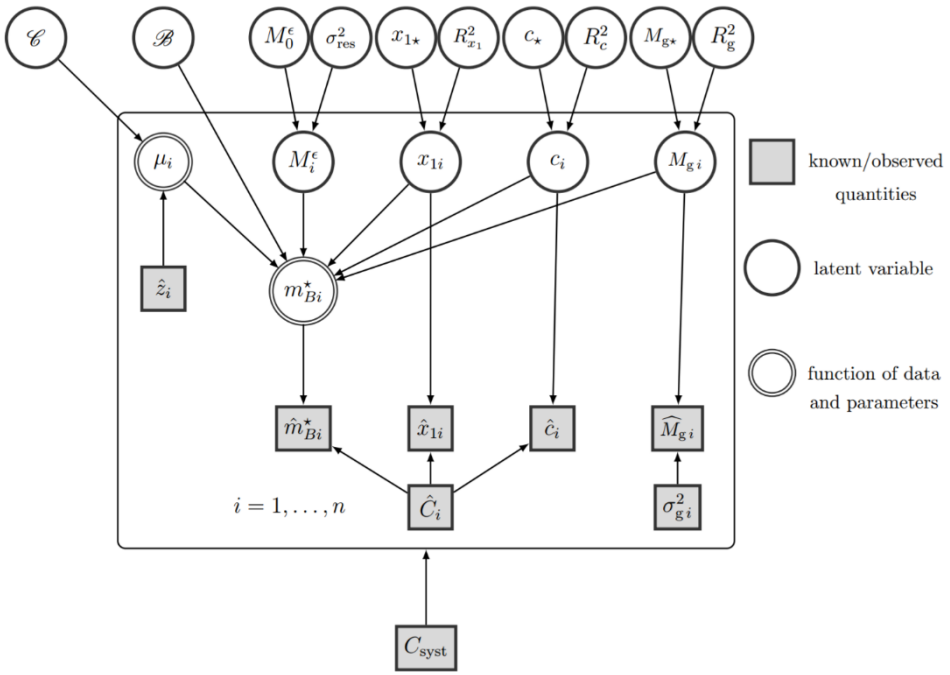
Gunnar Malmquist
(1893-1982)

$$0 < p(A) < 1; \quad 0 < p(B) < 1; \quad p(A|B) = p(A)$$

$$C = A + B \quad \Rightarrow \quad p(A|\bar{B}, C) = 1 > p(A|C)$$

détecté lumineux proche

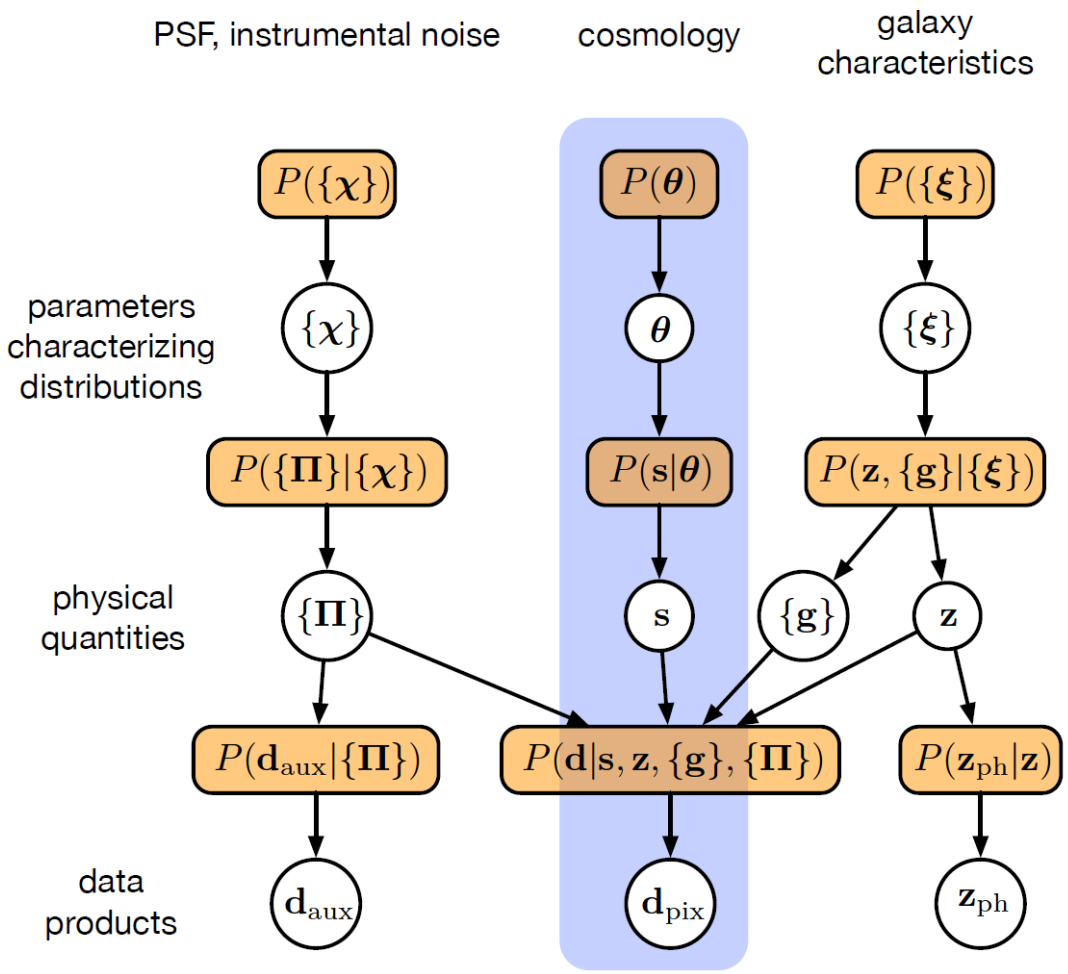
Exemple de réseau bayésien : supernovæ (BAHAMAS)



Parameter	Notation and Prior Distribution
Cosmological parameters	
Matter density parameter	$\Omega_{\rm m} \sim \text{UNIFORM}(0, 2)$
Cosmological constant density parameter	$\Omega_\Lambda \sim \text{UNIFORM}(0, 2)$
Dark energy EOS	$w \sim \text{UNIFORM}(-2, 0)$
Hubble parameter	$H_0/\text{km/s/Mpc} = 67.3$
Covariates	
Coefficient of stretch covariate	$\alpha \sim \text{UNIFORM}(0, 1)$
Coefficient of color covariate	β (or β_0) $\sim \text{UNIFORM}(0, 4)$
Coefficient of interaction of color correction and z	$\beta_1 \sim \text{UNIFORM}(-4, 4)$
Jump in coefficient of color covariate	$\Delta\beta \sim \text{UNIFORM}(-1.5, 1.5)$
Redshift of jump in color covariate	$z_t \sim \text{UNIFORM}(0.2, 1)$
Coefficient of host galaxy mass covariate	$\gamma \sim \text{UNIFORM}(-4, 4)$
Population-level distributions	
Mean of absolute magnitude	$M_0^\epsilon \sim \mathcal{N}(-19.3, 2^2)$
Residual scatter after corrections	$\sigma_{\rm res}^2 \sim \text{INVGAMMA}(0.003, 0.003)$
Mean of absolute magnitude, low galaxy mass	$M_0^{\rm lo} \sim \mathcal{N}(-19.3, 2^2)$
SD of absolute magnitude, low galaxy mass	${\sigma_{\rm res}^{\rm lo}}^2 \sim \text{INVGAMMA}(0.003, 0.003)$
Mean of absolute magnitude, high galaxy mass	$M_0^{\rm hi} \sim \mathcal{N}(-19.3, 2^2)$
SD of absolute magnitude, high galaxy mass	${\sigma_{\rm res}^{\rm hi}}^2 \sim \text{INVGAMMA}(0.003, 0.003)$
Mean of stretch	$x_{1\star} \sim \mathcal{N}(0, 10^2)$
SD of stretch	$R_{x_1} \sim \text{LOGUNIFORM}(-5, 2)$
Mean of color	$c_\star \sim \mathcal{N}(0, 1^2)$
SD of color	$R_c \sim \text{LOGUNIFORM}(-5, 2)$
Mean of host galaxy mass	$M_{\rm g\star} \sim \mathcal{N}(10, 100^2)$
SD of host galaxy mass	$R_{\rm g} \sim \text{LOGUNIFORM}(-5, 2)$

Shariff *et al.* 2015, 1510.05954

Exemple de réseau bayésien : lentillage gravitationnel faible



Can include:

Mask
Intrinsic alignments
Baryon feedback
Shape measurement
Photometric redshifts

Alsing *et al.* 2015, 1505.07840

Le bayésianisme empirique : une alternative au principe d'entropie maximale pour choisir un prior

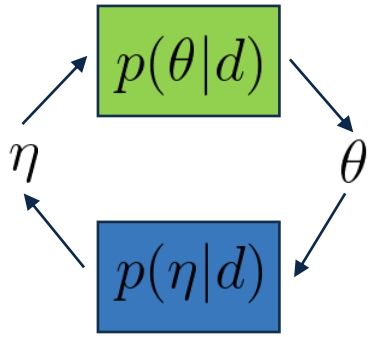
prior hyperprior

$$p(\theta|d) \propto p(d|\theta)\, p(\theta|\eta)\, p(\eta)$$

$$\underline{p(\theta|d)} = \int p(\theta|\eta, d)\, p(\eta|d)\, \mathrm{d}\eta = \int \frac{p(d|\theta)\, p(\theta|\eta)}{p(d|\eta)}\, \underline{p(\eta|d)}\, \mathrm{d}\eta$$

$$\underline{p(\eta|d)} = \int p(\eta|\theta)\, \underline{p(\theta|d)}\, \mathrm{d}\theta$$

➡ Schéma itératif (analogue à un échantillonneur de Gibbs)



- Le bayésianisme empirique (Empirical Bayes) est une troncature de ce schéma après quelques étapes (souvent une seule).
- Cas particulier : $p(\eta|d) \approx \delta_{\mathrm{D}}(\eta - \eta^\star(d))$ ➡ $\underline{p(\theta|d)} \approx \frac{p(d|\theta)\, p(\theta|\eta^\star)}{p(d|\eta^\star)}$

Cet algorithme s'appelle l'algorithme Espérance-Maximisation (Expectation-Maximisation – EM) en machine learning, data mining.