

Inférence bayésienne adaptative pour la reconstruction de source en dispersion atmosphérique

Harizo Rajaona

Directeurs de thèse: Yves Delignon, François Septier

Lille

21 novembre 2016

- ① Contexte et problématique
- ② Méthodologie adaptative pour l'inférence bayésienne
- ③ Application au cas expérimental FFT07

① Contexte et problématique

② Méthodologie adaptative pour l'inférence bayésienne

③ Application au cas expérimental FFT07

Les rejets **NRBC**¹ dans l'atmosphère peuvent être d'origine :

- accidentelle (fuite ou explosion sur un site industriel),
- malveillante (actes terroristes)



Fukushima (2011)



Igualada (2015)



Los Angeles (2015)

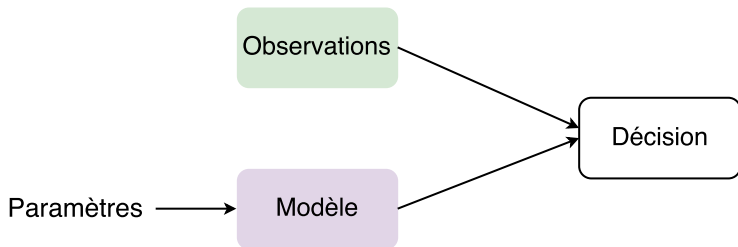
Priorités :

- informer et protéger les populations,
- atténuer/neutraliser le risque.

1. Nucléaires, Radiologiques, Biologiques, Chimiques

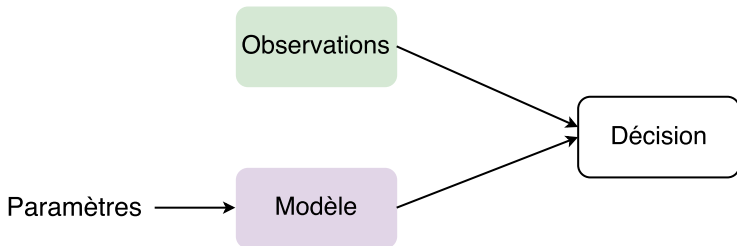
Outils de détection et d'évaluation du risque :

- données d'observation (capteurs mesurant la concentration de polluant)



Outils de détection et d'évaluation du risque :

- données d'observation (capteurs mesurant la concentration de polluant)
- outils de modélisation des phénomènes atmosphériques



Modèle de dispersion

Outil de calcul numérique permettant de simuler la propagation dans l'atmosphère d'un rejet de polluant.

Typologie des modèles selon :

- l'échelle (locale, régionale, synoptique),
- le degré de simplification des équations de la mécanique des fluides

Paramètres d'entrée :

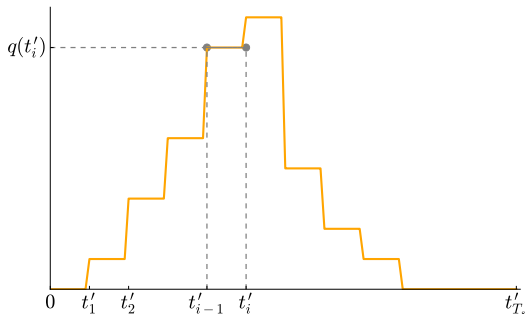
- données météorologiques : vent (direction + vitesse), température, humidité, nébulosité, flux de rayonnement...
- terme source : position, quantités émises, durée, substance émise...

Terme source : définitions

Hypothèses sur la nature de la source :

- localisée (représentée par un point géographique $\mathbf{x}_s \in \mathbb{R}^3$),
- unique (un seul point d'émission),
- non-instantanée, avec un profil temporel d'émission :

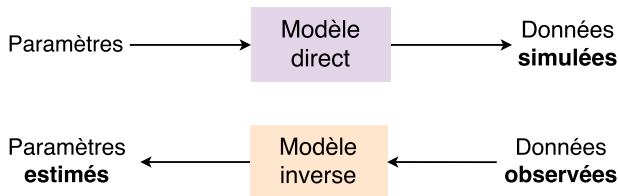
$$\mathbf{q} = (q(t'_1), q(t'_2), \dots, q(t'_{T_s}))$$



⇒ émission
constante sur le
palier $[t'_{i-1}, t'_i]$

Terme source : estimation

Reconstruire les paramètres d'un terme source (STE^2) à partir des observations est un problème inverse.



Plusieurs approches de résolution possibles :

- rétro-transport,
- résolution d'un système linéaire,
- algorithmes évolutionnaires,
- méthodes bayésiennes et simulation stochastique.

Problématique de recherche

On se concentre sur les méthodes bayésiennes :

- formalisme rigoureux pour estimation et quantification de l'incertitude,
- exploitation d'un nombre limité de mesures (régularisation),
- temps de calcul potentiellement élevés,
- estimation disjointe de la position et du profil d'émission.

Problématique

- ▶ Développer une méthode bayésienne pour estimer la localisation **et** le profil d'émission d'une source.
- ▶ Coupler cette méthode avec un modèle de dispersion atmosphérique dans une chaîne de calcul opérationnelle.

- 1 Contexte et problématique
- 2 Méthodologie adaptative pour l'inférence bayésienne
- 3 Application au cas expérimental FFT07

Principe : estimation probabiliste des paramètres θ d'un système ayant généré un ensemble d'observations η .

- $\theta \Rightarrow$ paramètres du terme source
- $\eta \Rightarrow$ mesures de concentration observées

Règle de Bayes

$$\pi(\theta) = p(\theta|\eta) = \frac{p(\theta)p(\eta|\theta)}{p(\eta)} \propto p(\theta)p(\eta|\theta)$$

- ▶ loi a posteriori $\pi(\theta)$: information sur θ connaissant η ,
- ▶ loi a priori $p(\theta)$: information préalable sur θ ,
- ▶ vraisemblance $p(\eta|\theta)$: probabilité d'observer η pour θ fixé.

- **Problème** : $p(\eta|\theta)$ trop coûteuse (ou impossible) à calculer
⇒ pas d'expression analytique pour $\pi(\theta)$!
⇒ recours à des méthodes d'approximation numérique

Méthodes de Monte-Carlo

Permettent d'approximer l'espérance de toute fonction d'une variable aléatoire de loi π en échantillonnant depuis cette loi :

$$\mathbb{E}_{\pi}[f(\theta)] = \int f(\theta)\pi(\theta)d\theta \simeq \frac{1}{N} \sum_{i=1}^N f(\theta^{(i)}), \quad \theta^{(i)} \sim \pi$$

- Obtention d'estimateurs bayésiens par simulation (ex : poser $f(\theta) = \theta$ pour le MMSE³).

Méthodes d'échantillonnage

- Algorithmes **MCMC**⁴ : π est la distribution stationnaire d'une chaîne de Markov construite par itérations successives.
 - Metropolis-Hastings
 - échantillonneur de Gibbs

Bons résultats obtenus dans la littérature STE :

- en milieu urbain : Keats (2007), Chow (2008)
- en multi-source : Yee (2008)

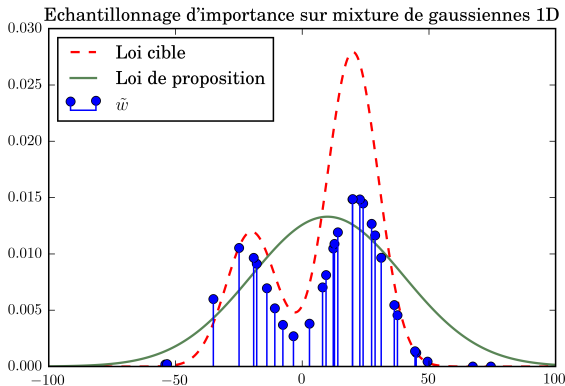
Inconvénients :

- perte d'une partie des échantillons générés (*burn-in*)
 - états corrélés (non-parallélisable)
 - MH : performances liées au choix du noyau et de l'initialisation
 - Gibbs : requiert les lois conditionnelles de θ
- Algorithmes **d'échantillonnage d'importance (IS)**⁵ : tirage d'une population d'échantillons pondérés (ou particules) à partir d'une loi de proposition.

4. Markov Chain Monte Carlo

5. Importance Sampling

Echantillonnage d'importance



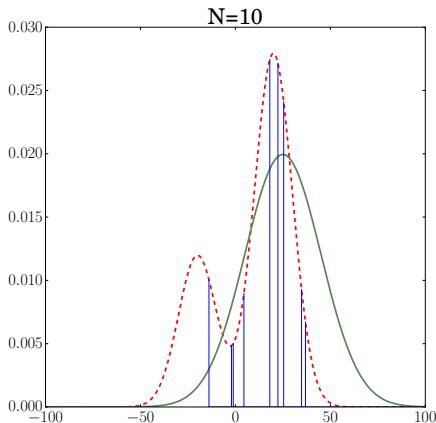
Avantages :

- échantillons i.i.d. : traitement parallélisable
- exploitation de tous les échantillons générés

Echantillonnage d'importance

Inconvénients :

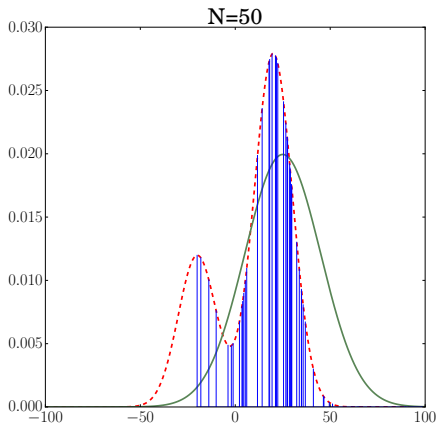
- performances fortement conditionnées par le choix de la loi de proposition !



Echantillonnage d'importance

Inconvénients :

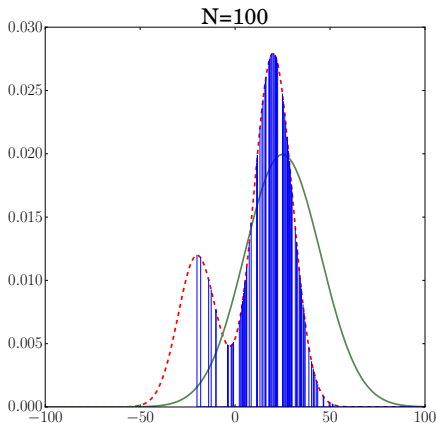
- performances fortement conditionnées par le choix de la loi de proposition !



Echantillonnage d'importance

Inconvénients :

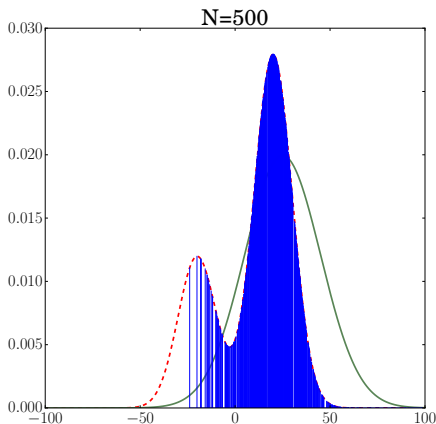
- performances fortement conditionnées par le choix de la loi de proposition !



Echantillonnage d'importance

Inconvénients :

- performances fortement conditionnées par le choix de la loi de proposition !



Echantillonnage d'importance adaptatif

Solution : adapter itérativement la loi de proposition φ

Population Monte Carlo [Cappé et al., 2004]

Introduction du concept d'adaptation par minimisation de :

$$KL(\pi, \varphi) = \int \log \left(\frac{\pi(\boldsymbol{\theta})}{\varphi(\boldsymbol{\theta})} \right) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \quad (\text{divergence KL})$$

D-kernel PMC [Douc et al., 2007]

- Loi de proposition $\varphi_{\alpha} \Rightarrow$ mélange de noyaux fixes pondérés $\{(\alpha_d, \varphi_d)\}_{1 \leq d \leq D}$:

$$\varphi_{\alpha}(\boldsymbol{\theta}) = \sum_{d=1}^D \alpha_d \varphi_d(\boldsymbol{\theta})$$

- Optimisation des α_d par minimisation KL.

M-PMC [Cappé et al., 2008]

- Loi de proposition $\varphi_{(\alpha, \nu)} \Rightarrow$ mélange de noyaux paramétriques pondérés $\left\{ \left(\alpha_d, \varphi_d(\cdot | \nu_d) \right) \right\}_{1 \leq d \leq D}$:

$$\varphi_{(\alpha, \nu)}(\theta) = \sum_{d=1}^D \alpha_d \varphi_d(\theta | \nu_d)$$

- Optimisation des α_d et ν_d par minimisation KL (algorithme EM).

Jusqu'ici : optimisation itérative seulement en fonction de l'itération précédente !

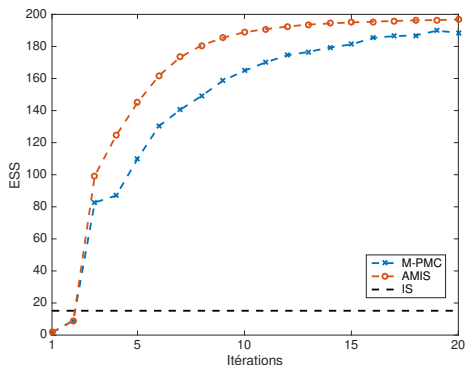
Echantillonnage d'importance adaptatif

Adaptive Multiple Importance Sampling (AMIS) [Cornuet et al., 2012]

- ▶ Loi de proposition identique à celle du M-PMC
- ▶ Ré-utilisation des particules de toutes les itérations pour :
 - le calcul et recyclage de tous les poids d'importance,
 - l'optimisation des α_d et ν_d .

Avantages :

- utilisation efficace de tous les échantillons disponibles
- convergence plus rapide vers la loi cible
- variance d'erreur d'estimation réduite

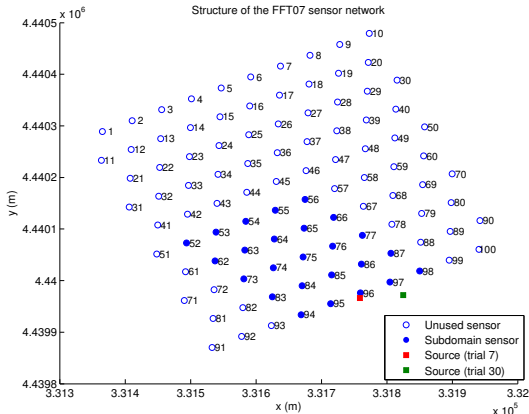


- ① Contexte et problématique
- ② Méthodologie adaptative pour l'inférence bayésienne
- ③ Application au cas expérimental FFT07**

L'expérience FFT07

Campagne expérimentale :

- rejets de gaz traceur sur terrain instrumenté dans diverses configurations (période, météo, nombre de sources...)
- création de données de référence pour validation d'algorithmes STE



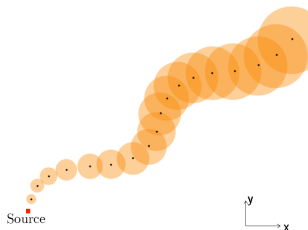
L'expérience FFT07

Caractéristiques des cas étudiés :

- restriction à $N_C = 25$ capteurs proches de la source
- T_C instants d'observations moyennées sur fenêtres de 10s
- capteurs et source à même altitude : $\mathbf{x}_s \in \mathbb{R}^2$
- rejet non-instantané, conditions atmosphériques stables
- étude avec données simulées et observations réelles

Modèle de dispersion gaussien à bouffées :

- implémentation simple
- temps de calcul faibles
- émissions non-instantanées
- variabilité météorologique



Formalisation du problème STE

Objectif : estimer les paramètres de position \mathbf{x}_s et d'émission \mathbf{q} de la source pour une configuration donnée (*trial*) de l'expérience FFT07.

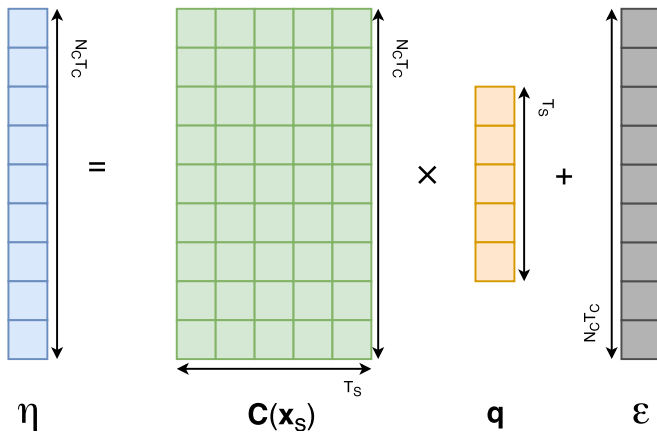
Modèle de données :

$$\boldsymbol{\eta} = \mathbf{C}(\mathbf{x}_s)\mathbf{q} + \boldsymbol{\varepsilon}$$

où :

- $\boldsymbol{\eta} \in \mathbb{R}^{N_C T_C}$: observations concaténées par capteur
- $\mathbf{C}(\mathbf{x}_s) \in \mathbb{R}^{N_C T_C \times T_s}$: matrice source-récepteur construite avec un modèle de dispersion
- $\mathbf{q} \in \mathbb{R}^{T_s}$: profil d'émission
- $\boldsymbol{\varepsilon} \in \mathbb{R}^{N_C T_C}$: erreurs (observation, modèle)

Formalisation du problème STE



Formalisation du problème STE

The diagram illustrates the formalization of the STE problem, showing the relationship between several variables:

- η : A vertical vector (blue) with 10 cells. The top cell is labeled R_1 .
- $=$: Equality symbol.
- $C(x_s)$: A 10x5 grid (green). The top row is labeled $1_{t'_1}$.
- \times : Multiplication symbol.
- q : A vertical vector (orange) with 5 cells.
- $+$: Addition symbol.
- ε : A vertical vector (gray) with 10 cells.

Formalisation du problème STE

The diagram illustrates the formalization of the STE problem as a matrix equation. It shows the following components:

- R_1 : A vertical column vector of 10 light blue squares, with the top square highlighted in a darker blue. Labeled R_1 in blue text to its left.
- η : A label in black text centered below the R_1 vector.
- $=$: An equals sign indicating the equation.
- $1_{t'_2}$: A label in green text centered above the first row of the $C(x_s)$ matrix.
- $C(x_s)$: A 10x5 grid of light green squares. Labeled $C(x_s)$ in black text centered below the grid.
- \times : A multiplication symbol indicating the matrix product.
- q : A vertical column vector of 5 orange squares. Labeled q in black text centered below the vector.
- $+$: A plus sign indicating the addition.
- ε : A vertical column vector of 10 squares. The top square is dark gray, and the remaining 9 squares are light gray. Labeled ε in black text centered below the vector.

The overall equation represented is:

$$\eta = C(x_s) \times q + \varepsilon$$

Formalisation du problème STE

The diagram illustrates the formalization of the STE problem as a matrix equation. It shows the relationship between the observed data η , the model matrix $C(x_s)$, the parameters q , and the noise term ε .

The equation is represented as:

$$\eta = C(x_s) \times q + \varepsilon$$

The components are visualized as follows:

- η : A vertical column vector of 10 light blue squares. The top square is labeled R_1 in blue.
- $=$: An equals sign.
- $C(x_s)$: A 10x5 grid of light green squares. The top row is highlighted in a darker green and labeled $1_{t'_3}$ in green.
- \times : A multiplication symbol.
- q : A vertical column vector of 5 orange squares.
- $+$: A plus sign.
- ε : A vertical column vector of 10 squares. The top square is dark gray, and the remaining 9 squares are light gray.

Formalisation du problème STE

The diagram illustrates the formalization of the STE problem as a matrix equation. It shows a vertical vector η (blue) on the left, followed by an equals sign. To the right of the equals sign is a 10x5 grid of light green squares representing the matrix $C(x_s)$. Above the top row of this grid is the label $1_{t'_s}$. To the right of the grid is a multiplication symbol \times , followed by a vertical vector q (orange) with 5 elements. To the right of q is a plus sign $+$, followed by a vertical vector ε (gray) with 10 elements. The top element of ε is a darker gray, while the others are light gray.

$$\eta = C(x_s) \times q + \varepsilon$$

η $C(x_s)$ q ε

Formalisation du problème STE

The diagram illustrates the formalization of the STE problem as a matrix equation. It shows the relationship between a vector η , a matrix $C(x_s)$, a vector q , and a vector ε .

On the left, a vertical column of 10 light blue squares represents the vector η . The second square from the top is highlighted with a darker blue border and labeled R_2 to its left. Below this column is the symbol η .

In the center, a 10x5 grid of light green squares represents the matrix $C(x_s)$. The second row from the top is highlighted with a darker green border. Below this grid is the symbol $C(x_s)$.

To the right of the matrix is a multiplication symbol \times .

Next is a vertical column of 6 light orange squares representing the vector q . Below this column is the symbol q .

To the right of the vector q is an addition symbol $+$.

On the far right, a vertical column of 10 light gray squares represents the vector ε . The second square from the top is highlighted with a darker gray border. Below this column is the symbol ε .

The entire equation is represented as: $\eta = C(x_s) \times q + \varepsilon$.

Formalisation du problème STE

The diagram illustrates the formalization of the STE problem as a matrix equation. It shows four vertical structures representing vectors or matrices, connected by mathematical operators.

- η** : A vertical vector of 10 light blue squares. The third square from the top is highlighted with a darker blue border and labeled **R3** to its left.
- =**: An equals sign.
- $C(x_s)$** : A 10x5 grid of light green squares. The third row from the top is highlighted with a darker green border.
- \times** : A multiplication symbol.
- q** : A vertical vector of 6 light orange squares.
- +**: A plus sign.
- ε** : A vertical vector of 10 light gray squares. The third square from the top is highlighted with a darker gray border.

The equation is represented as:

$$\eta = C(x_s) \times q + \varepsilon$$

Formalisation du problème STE

The diagram illustrates the formalization of the STE problem as a matrix equation. It shows four vertical structures representing matrices or vectors, connected by mathematical operators.

- η** : A vertical column of 10 light blue squares. The bottom square is highlighted in a darker blue and labeled $R_{N_C T_C}$ in blue text.
- $=$** : An equals sign.
- $C(x_s)$** : A 10x5 grid of light green squares. The bottom row of 5 squares is highlighted in a darker green.
- \times** : A multiplication symbol.
- q** : A vertical column of 6 orange squares.
- $+$** : A plus sign.
- ε** : A vertical column of 10 light gray squares. The bottom square is highlighted in a darker gray.

The equation is represented as:

$$\eta = C(x_s) \times q + \varepsilon$$

Démarche de résolution

- Objectif : calculer la loi a posteriori $p(\mathbf{x}_s, \mathbf{q}|\boldsymbol{\eta})$
- Problème : source non-instantanée
 - dimension $T_s + 2$ potentiellement élevée du vecteur de paramètres,
 - calcul coûteux pour une simulation Monte-Carlo.

Marginalisation du profil d'émission

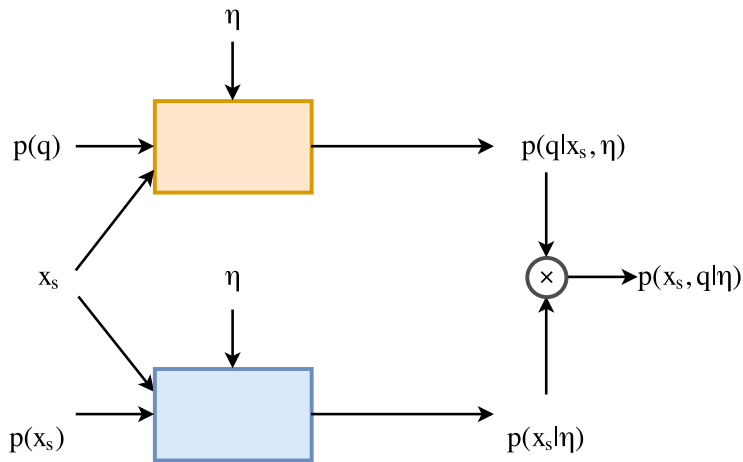
La loi a posteriori des paramètres de la source peut s'écrire comme :

$$p(\mathbf{x}_s, \mathbf{q}|\boldsymbol{\eta}) = p(\mathbf{q}|\mathbf{x}_s, \boldsymbol{\eta})p(\mathbf{x}_s|\boldsymbol{\eta})$$

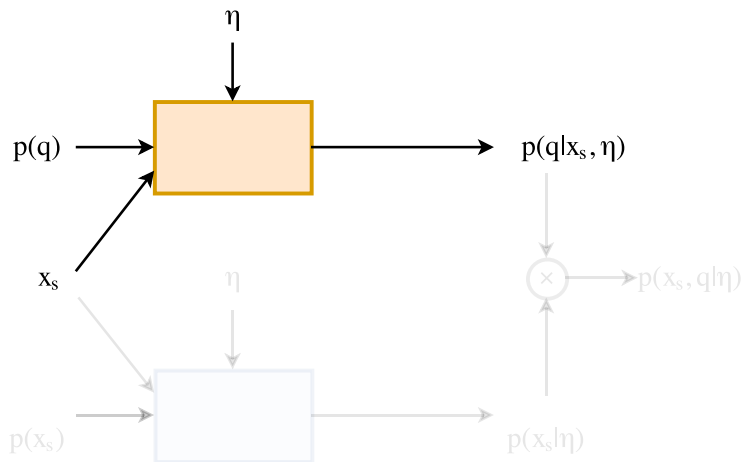
- ▶ $p(\mathbf{q}|\mathbf{x}_s, \boldsymbol{\eta})$: loi a posteriori conditionnelle de \mathbf{q} ,
- ▶ $p(\mathbf{x}_s|\boldsymbol{\eta})$: loi a posteriori marginale de \mathbf{x}_s .

Démarche de résolution

La marginalisation permet de n'échantillonner que les x_s :



Loi conditionnelle de q



Loi conditionnelle de \mathbf{q}

L'erreur sur $\boldsymbol{\eta}$ est supposée gaussienne : $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma_{obs}^2 \mathbf{I})$

\Rightarrow la vraisemblance est gaussienne :

$$p(\boldsymbol{\eta} | \mathbf{x}_s, \mathbf{q}) = \prod_{i=1}^{N_c} \prod_{j=1}^{T_c} \mathcal{N}(\eta_{i,j} | \mathbf{C}_{i,j}(\mathbf{x}_s) \mathbf{q}, \sigma_{obs}^2)$$

A priori gaussien sur \mathbf{q}

Dans ces conditions, si $p(\mathbf{q}) = \mathcal{N}(\mathbf{q} | \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$ alors :

$$p(\mathbf{q} | \mathbf{x}_s, \boldsymbol{\eta}) = \mathcal{N}(\mathbf{q} | \tilde{\boldsymbol{\mu}}_q, \tilde{\boldsymbol{\Sigma}}_q)$$

avec $\tilde{\boldsymbol{\mu}}_q$ et $\tilde{\boldsymbol{\Sigma}}_q$ obtenus analytiquement par résolution d'un système linéaire gaussien.

- hypothèse simplifiant la résolution du problème
- souvent employée dans la littérature STE
- perte potentielle de la positivité sur l'estimation de \mathbf{q} !

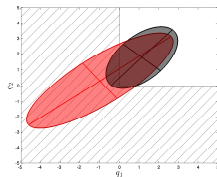
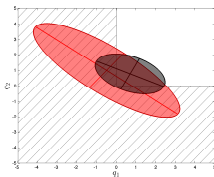
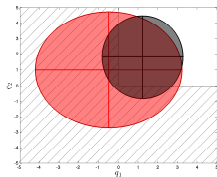
Loi conditionnelle de q

Contrainte de positivité par troncature de la densité

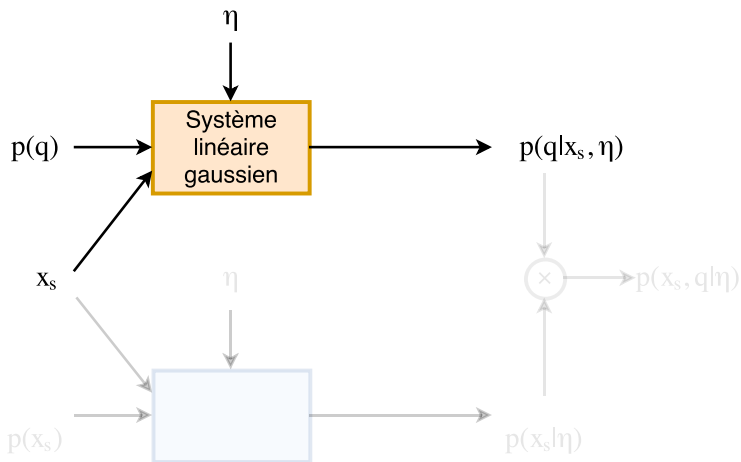
Objectif : restreindre $p(q|x_s, \eta)$ à des valeurs positives en conservant la nature gaussienne de la densité d'origine.

- ▶ assure la cohérence physique de la solution
- ▶ rallonge le temps de calcul
- ▶ modifie potentiellement les valeurs initiales

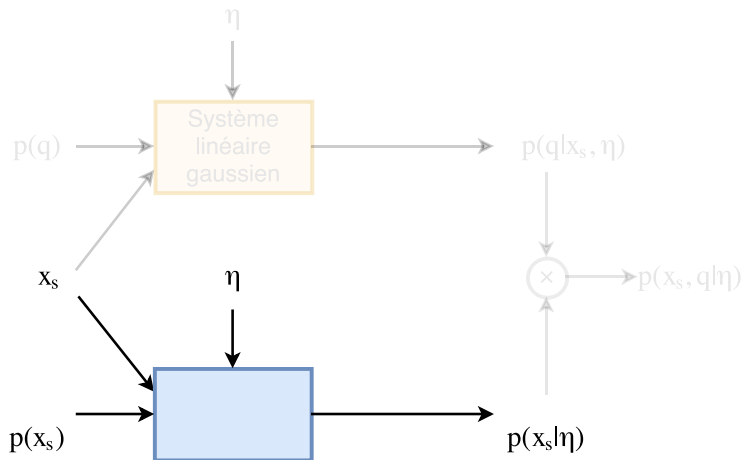
Exemples sur lois gaussiennes bivariées :



Loi conditionnelle de q



Loi conditionnelle de q



Loi a posteriori marginale de \mathbf{x}_s

On utilise l'AMIS pour calculer $p(\mathbf{x}_s|\boldsymbol{\eta})$:

- génération d'un échantillon de KN_p particules sur K itérations,
- loi de proposition : mélange de $D = 4$ noyaux gaussiens, paramètres $(\alpha_d, \boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d)$
- initialisation "uniforme" sur le domaine

