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

Contexte

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 :

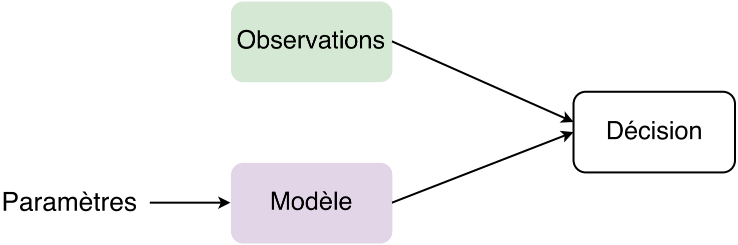
- informer et protéger les populations,
- atténuer/neutraliser le risque.

1. Nucléaires, Radiologiques, Biologiques, Chimiques

Contexte

Outils de détection et d'évaluation du risque :

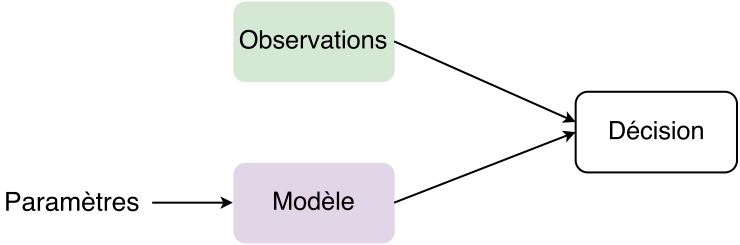
- données d'observation (capteurs mesurant la concentration de polluant)



Contexte

Outils de détection et d'évaluation du risque :

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



Dispersion atmosphérique

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 :

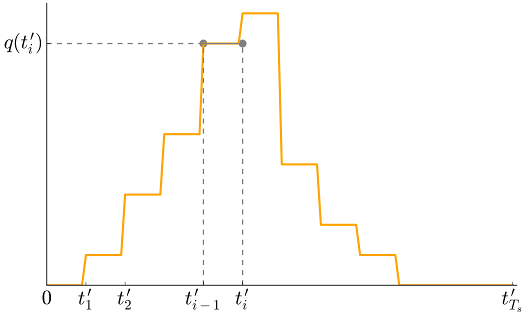
- données météorologiques : vent (direction + vitesse), température, humidité, nébulosité, flux de rayonnement...
- terme source : position, quantités émises, durée, substance émise...

Terme source : définitions

Hypothèses sur la nature de la source :

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

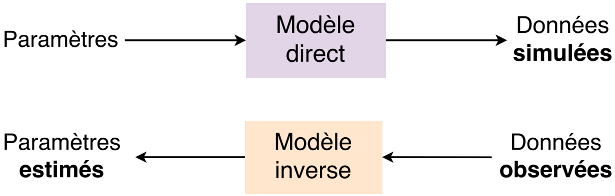
$$\boldsymbol{q} = \big(q(t'_1), q(t'_2), \cdots, q(t'_{T_s})\big)$$



$\Rightarrow$ 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.

2. Source Term Estimation

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.

Inférence bayésienne

Principe : estimation probabiliste des paramètres $\boldsymbol{\theta}$ d'un système ayant généré un ensemble d'observations $\boldsymbol{\eta}$.

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

Règle de Bayes

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

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

- **Problème** : $p(\boldsymbol{\eta}|\boldsymbol{\theta})$ trop coûteuse (ou impossible) à calculer
  $\Rightarrow$ pas d'expression analytique pour $\pi(\boldsymbol{\theta})$!
  $\Rightarrow$ 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 :

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

- Obtention d'estimateurs bayésiens par simulation (ex : poser $f(\boldsymbol{\theta}) = \boldsymbol{\theta}$ pour le MMSE[3]).

3. Minimum Mean Square Estimator

Méthodes d'échantillonnage

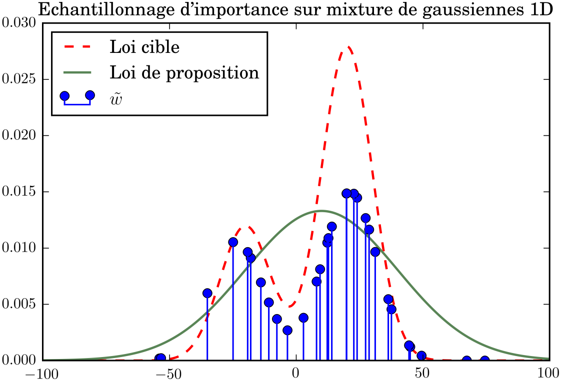
- Algorithmes **MCMC** [4] : π est la distribution stationnaire d'une chaîne de Markov construite par itérations successives.
  - Metropolis-Hastings
  - échantillonneur de Gibbs

  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 $\boldsymbol{\theta}$
- 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



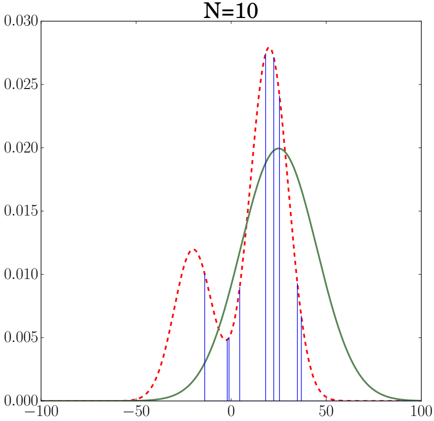
Avantages :

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

Echantillonnage d'importance

Inconvénients :

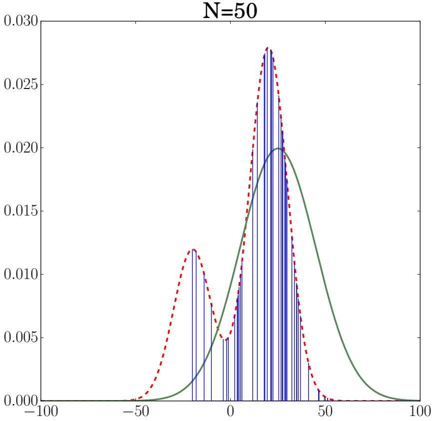
- performances fortement conditionnées par le choix de la loi de proposition !



Echantillonnage d'importance

Inconvénients :

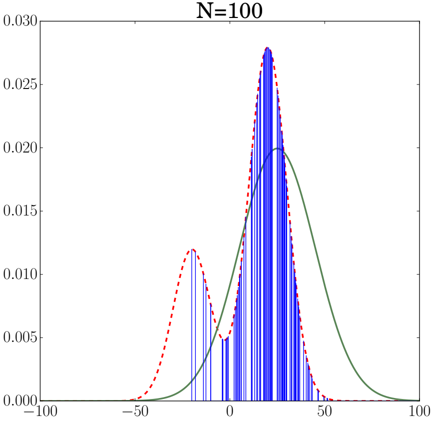
- performances fortement conditionnées par le choix de la loi de proposition !



Echantillonnage d'importance

Inconvénients :

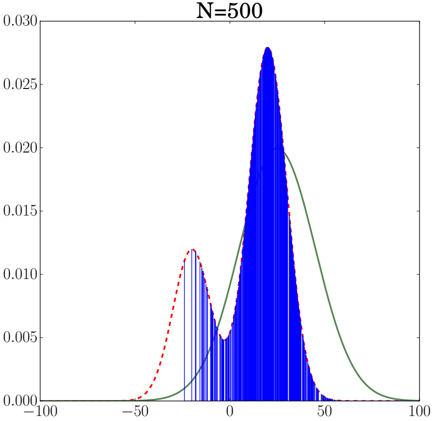
- performances fortement conditionnées par le choix de la loi de proposition !



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

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

- Optimisation des α_d par minimisation KL.

Echantillonnage d'importance adaptatif

M-PMC [Cappé et al., 2008]

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

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

- Optimisation des α_d et $\boldsymbol{\nu}_d$ par minimisation KL (algorithme EM).

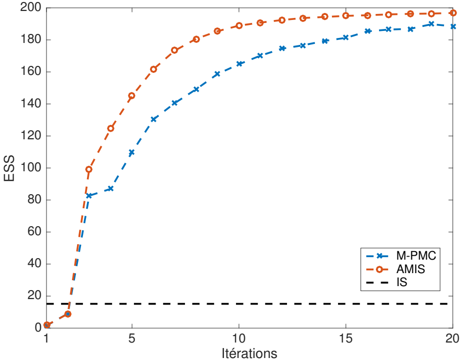
Jusqu'ici : optimisation itérative seulement en fonction de l'itération précédente !

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 $\boldsymbol{\nu}_d$.

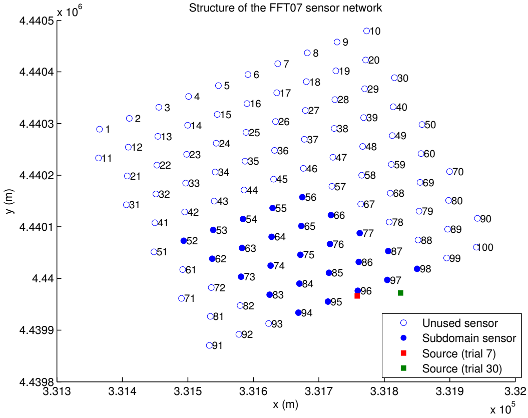
Avantages :

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



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



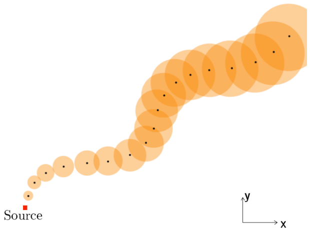
L'expérience FFT07

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 : $\boldsymbol{x_s} \in \mathbb{R}^2$
- 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 $\boldsymbol{x_s}$ et d'émission $\boldsymbol{q}$ de la source pour une configuration donnée (*trial*) de l'expérience FFT07.

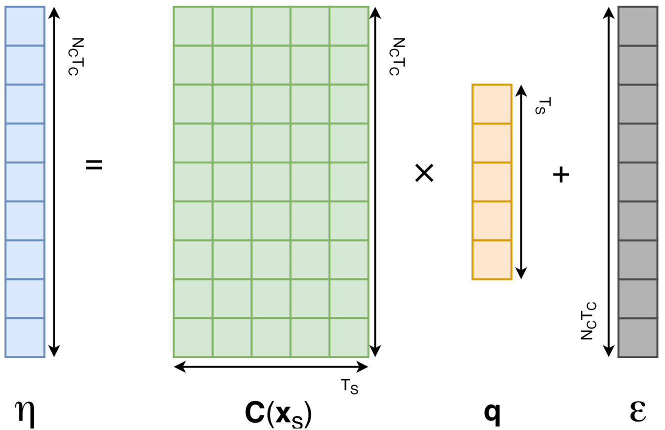
Modèle de données :

$$\boldsymbol{\eta} = \boldsymbol{C}(\boldsymbol{x_s})\boldsymbol{q} + \boldsymbol{\varepsilon}$$

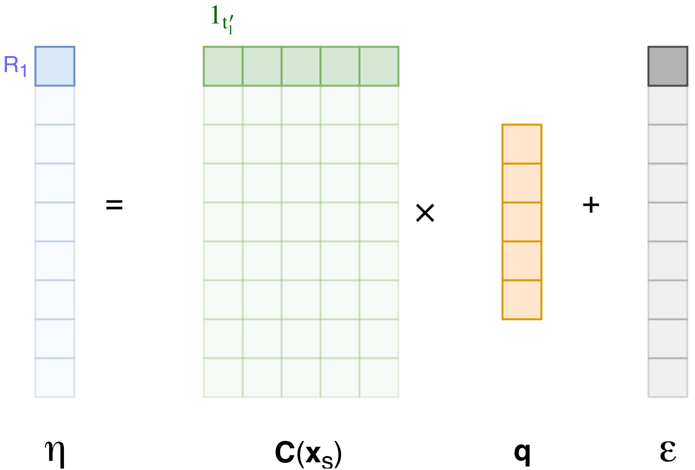
où :

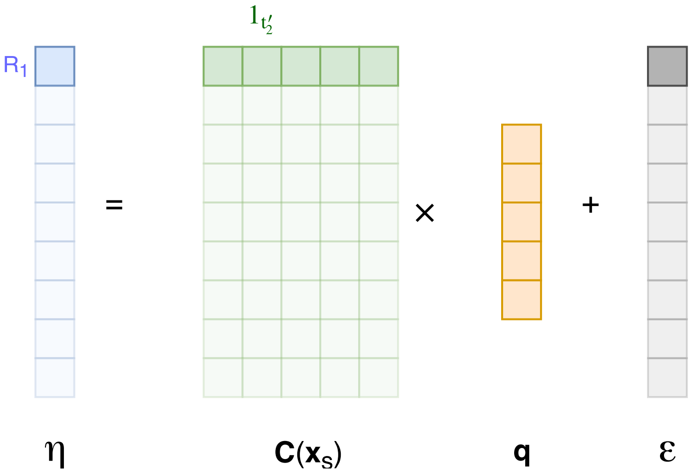
- $\boldsymbol{\eta} \in \mathbb{R}^{N_C T_C}$: observations concaténées par capteur
- $\boldsymbol{C}(\boldsymbol{x_s}) \in \mathbb{R}^{N_C T_C \times T_s}$: matrice source-récepteur construite avec un modèle de dispersion
- $\boldsymbol{q} \in \mathbb{R}^{T_s}$: profil d'émission
- $\boldsymbol{\varepsilon}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

Formalisation du problème STE

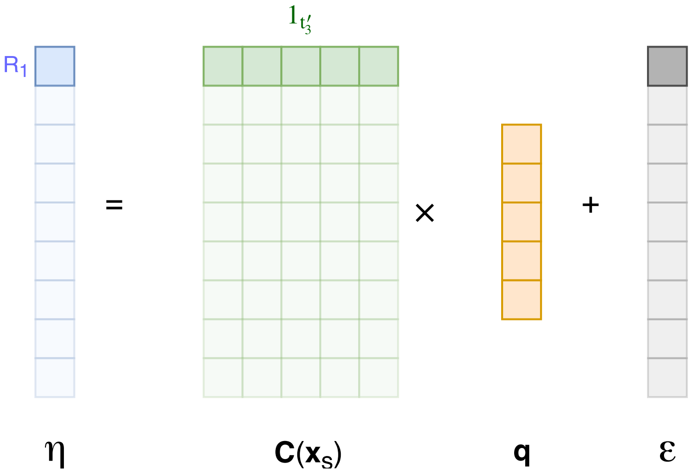


Formalisation du problème STE

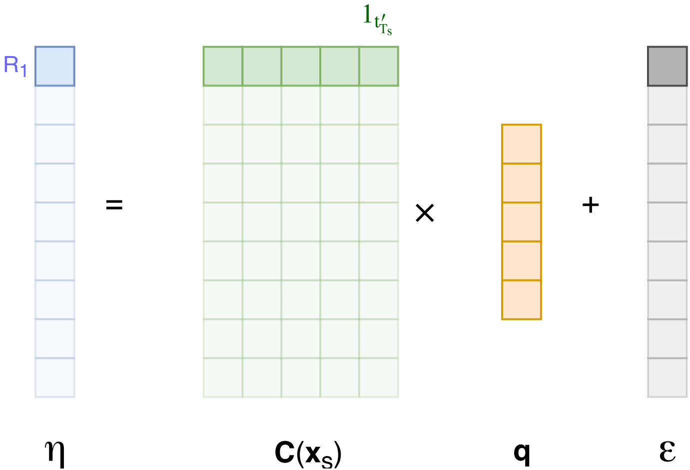



$1_{t'_2}$
R_1
$=$
$\times$
$+$
η
$\mathbf{C}(\mathbf{x}_s)$
$\mathbf{q}$
ε

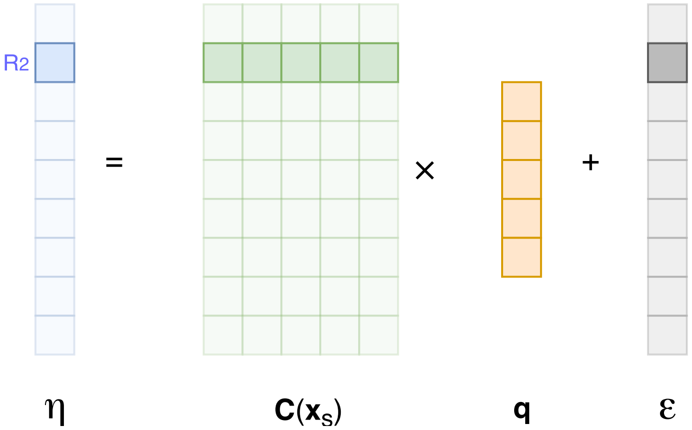
Formalisation du problème STE



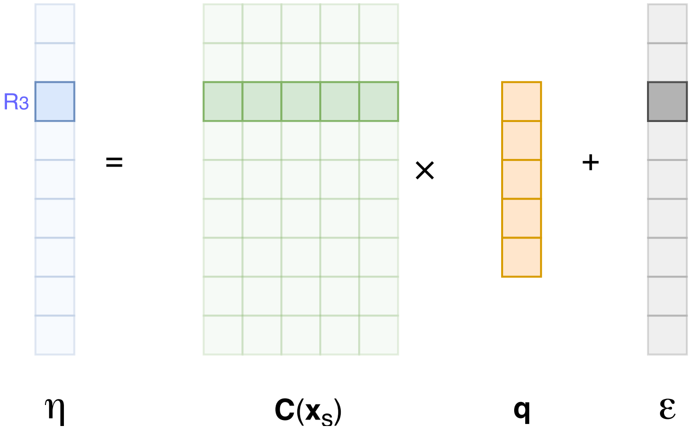
Formalisation du problème STE



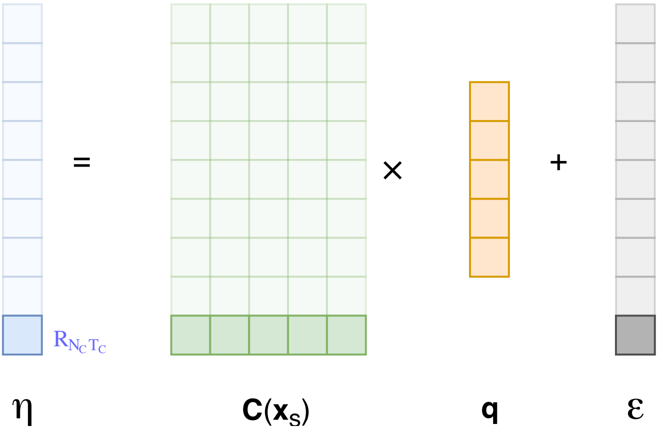
Formalisation du problème STE



Formalisation du problème STE



Formalisation du problème STE



Démarche de résolution

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

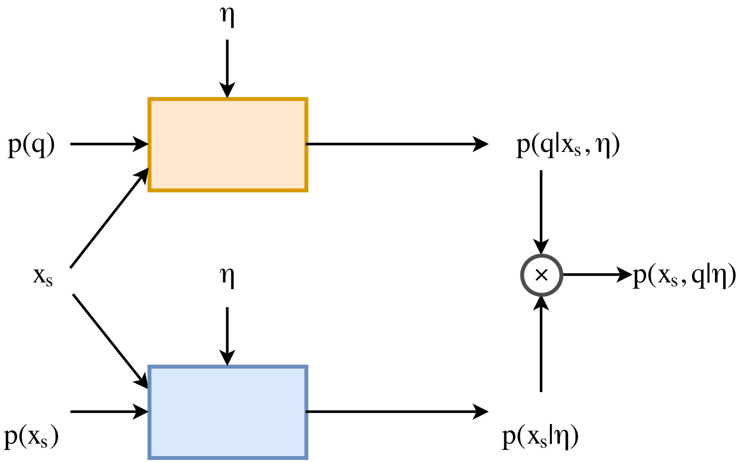
Marginalisation du profil d'émission

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

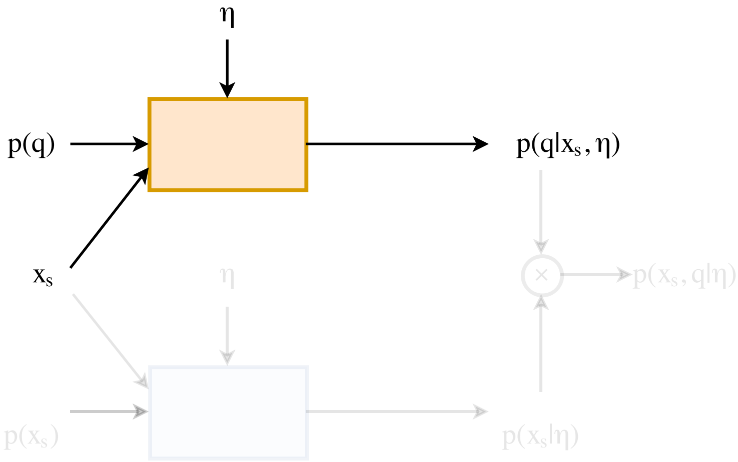
$$p(\boldsymbol{x_s}, \boldsymbol{q}|\boldsymbol{\eta}) = p(\boldsymbol{q}|\boldsymbol{x_s}, \boldsymbol{\eta})p(\boldsymbol{x_s}|\boldsymbol{\eta})$$

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

La marginalisation permet de n'échantillonner que les $\boldsymbol{x_s}$:



Loi conditionnelle de q



Loi conditionnelle de $\boldsymbol{q}$

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

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

A priori gaussien sur q

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

$$p(\boldsymbol{q}|\boldsymbol{x_s}, \boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{q}|\widetilde{\boldsymbol{\mu}}_q, \widetilde{\boldsymbol{\Sigma}}_q)$$

avec $\widetilde{\boldsymbol{\mu}}_q$ et $\widetilde{\boldsymbol{\Sigma}}_q$ 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 $\boldsymbol{q}$!

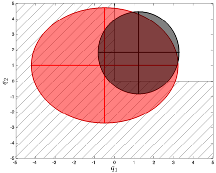
Loi conditionnelle de $\boldsymbol{q}$

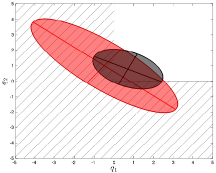
Contrainte de positivité par troncature de la densité

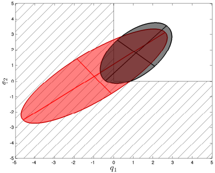
Objectif : restreindre $p(\boldsymbol{q}|\boldsymbol{x_s}, \boldsymbol{\eta})$ à des valeurs positives en conservant la nature gaussienne de la densité d'origine.

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

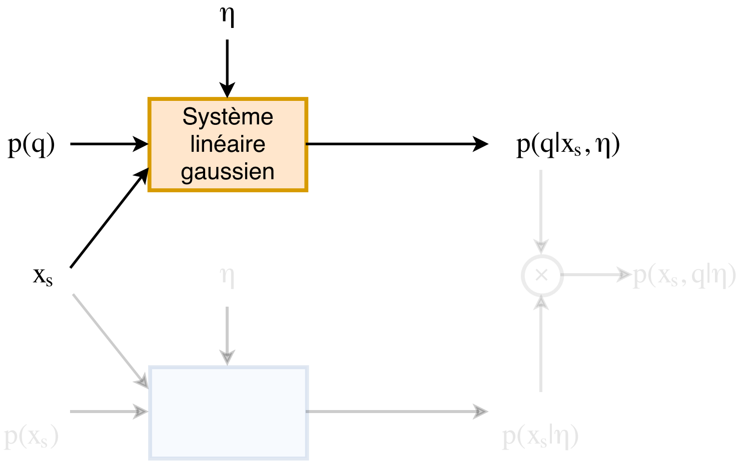
Exemples sur lois gaussiennes bivariées :



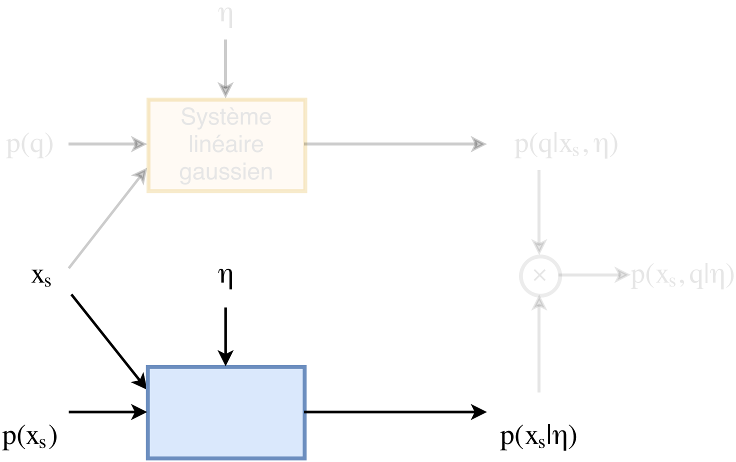




Loi conditionnelle de q



Loi conditionnelle de q



Loi a posteriori marginale de $\boldsymbol{x}_s$

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

- 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