Approche bayésienne pour la sélection de modèles : application à la restauration d'image

Benjamin Harroue

université de BORDEAUX

THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEUR DE L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE SCIENCES PHYSIQUE ET DE L'INGÉNIEUR

SPÉCIALITÉ : AUTOMATIQUE, PRODUCTIQUE, SIGNAL ET IMAGE, INGÉNIERIE COGNITIQUE

Par **Benjamin HARROUÉ**

Approche bayésienne pour la sélection de modèles Application à la restauration d'image

Sous la direction de : **Jean-François GIOVANNELLI**
Marcelo PEREYRA

Soutenue le 30 septembre 2020

Membres du jury :

M.	**BORDÉ, Pascal**	Professeur, Univ. Bordeaux	Président du Jury
Mme	**GIREMUS, Audrey**	Maître de conférences, Univ. Bordeaux	Jury
M.	**CARFANTAN, Hervé**	Maître de conférences, Univ. Toulouse	Rapporteur
M.	**DOBIGEON, Nicolas**	Professeur, INP - ENSEEIHT Toulouse	Rapporteur
M.	**GIOVANNELLI, Jean-François**	Professeur, Univ. Bordeaux	Directeur
M.	**PEREYRA, Marcelo**	Associate Professor, Heriot-Watt Univ.	Directeur

Titre - Approche bayésienne de la sélection de modèles : application à la restauration d'image.

Résumé - L'inversion consiste à reconstruire des objets d'intérêt à partir de données acquises au travers d'un système d'observation. Dans ces travaux, nous nous penchons sur la déconvolution d'image. Les données observées constituent une version dégradée de l'objet, altéré par le système (flou et bruit). À cause de la perte d'informations engendrée, le problème devient alors mal conditionné. Une solution est de régulariser dans un cadre bayésien : en se basant sur des modèles, on introduit de l'information *a priori* sur les inconnues. Se posent alors les questions suivantes : comment comparer les modèles candidats et choisir le meilleur ? Sur quel critère faut-il s'appuyer ? À quelles caractéristiques ou quantités doit-on se fier ? Ces travaux présentent une méthode de comparaison et de sélection automatique de modèles, fondée sur la théorie de la décision bayésienne. La démarche consiste à sélectionner le modèle qui maximise la probabilité *a posteriori*. Pour calculer cette dernière, on a besoin de connaître une quantité primordiale : l'évidence. Elle s'obtient en marginalisant la loi jointe par rapport aux inconnus : l'image et les hyperparamètres. Les dépendances complexes entre les variables et la grande dimension de l'image rendent le calcul analytique de l'intégrale impossible. On a donc recours à des méthodes numériques. Dans cette première étude, on s'intéresse au cas gaussien circulant. Cela permet, d'une part, d'avoir une expression analytique de l'intégrale sur l'image, et d'autre part, de faciliter la manipulation des matrices de covariances. Plusieurs méthodes sont mises en œuvre comme l'algorithme du Chib couplé à une chaîne de Gibbs, les power posteriors, ou encore la moyenne harmonique. Les méthodes sont ensuite comparées pour déterminer lesquelles sont les plus adéquates au problème de la restauration d'image.

Mots clés - Sélection de modèles, stratégie bayésienne, évidence, problème inverse, cas gaussien, matrices circulantes, modèles de covariance, algorithme de Chib, échantillonneur de Gibbs, moyenne harmonique, power posteriors.

Title - Bayesian approach of models selection : application in image restoration.

Abstract - Inversing main goal is about reconstructing objects from data. Here, we focus on the special case of image restauration in convolution problems. The data are acquired through a altering observation system and additionnaly distorted by errors. The problem becomes ill-posed due to the loss of information. One way to tackle it is to exploit Bayesian approach in order to regularize the problem. Introducing prior information about the unknown quantities osset the loss, and it relies on stochastic models. We have to test all the candidate models, in order to select the best one. But some questions remain : how do you choose the best model ? Which features or quantities should we rely on ? In this work, we propose a method to automatically compare and choose the model, based on Bayesion decision theory : objectively compare the models based on their posterior probabilities. These probabilities directly depend on the marginal likelihood or "evidence" of the models. The evidence comes from the marginalization of the jointe law according to the unknow image and the unknow hyperparameters. This a difficult integral calculation because of the complex dependancies between the quantities and the high dimension of the image. That way, we have to work with computationnal methods and approximations. There are several methods on the test stand as Harmonic Mean, Laplace method, discrete integration, Chib from Gibbs approximation or the power posteriors. Comparing is those methods is significative step to determine which ones are the most competent in image restauration. As a first lead of research, we focus on the family of Gaussian models with circulant covariance matrices to lower some difficulties.

Keywords - Models selection,Bayesian approach, evidence, inverse problem, Gaussian case, circulant matrices, covariance models, Chib algorithm, Gibbs sampling, harmonic mean, power posteriors.

Table des matières

Notations

Vecteurs, matrices et opérations

$\boldsymbol{y}$	-	Données observées
$\boldsymbol{x}$	-	Objet/image à reconstruire
$\boldsymbol{e}$	-	Erreur/bruit
$\boldsymbol{H}$	-	Matrice de convolution
$\boldsymbol{\theta}$	-	Vecteur des hyperparamètres inconnus
$\mathscr{M}$	-	Variable de modèle
$\mathbb{E}[\ ]$	-	Espérance
$\boldsymbol{R}_{\square}$	-	Matrice de covariance associée à $\square$
$\overset{\circ}{\boldsymbol{S}}_{\square}$	-	Densité Spectrale de Puissance associée à $\square$
$^{\mathrm{t}}$	-	Transposée
†	-	Transposée conjuguée
$\boldsymbol{F}$	-	Matrice de Transformée de Fourier Discrète (TFD)
$\overset{\circ}{\square} = \boldsymbol{F}\square$	-	Transformée de Fourier Discrète de $\square$
$*$	-	Convolution

Distributions

$f(\boldsymbol{y}\|\boldsymbol{x},\boldsymbol{\theta},\mathscr{M}=k)$	-	Vraisemblance de l'objet attachée aux données
$f(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\theta}\|\mathscr{M}=k)$	-	Loi jointe
$p(\boldsymbol{x},\boldsymbol{\theta}\|\boldsymbol{y},\mathscr{M}=k)$	-	Loi *a posteriori*
$f(\boldsymbol{y}\|\mathscr{M}=k)$	-	Évidence
$p(\mathscr{M}=k\|\boldsymbol{y})$	-	Probabilité *a posteriori* du modèle k
$p(\mathscr{M}=k)$	-	Prior du modèle k
$\boldsymbol{\pi}_{\square}(\square)$	-	Loi *a priori* pour $\square$

CHAPITRE 1

Introduction

1.1 Contexte : déconvolution d'image

L'inversion [GI13, DIGMD01, Kai05, LB03] consiste à reconstruire des objets d'intérêt (signal, image, vidéo, volume) à partir de données acquises au travers d'un système d'observation. De manière plus générique, on cherche à déterminer les causes qui ont engendré les effets observés. Fondées sur l'utilisation d'ondes mécaniques ou électromagnétiques (EM), de nombreuses modalités d'acquisition ont été développées. Elles sont utilisées dans divers domaines tels que la médecine, l'industrie, la défense, l'astronomie, la météorologie, la chimie, *etc*. Parmi les ondes EM, on peut citer les rayons X qui sont utilisés en tomographie. Cette modalité, qui fournit des projections 2D d'un volume 3D, est très utilisée en médecine (scanner), dans l'industrie pour le contrôle non destructif ou encore dans le domaine de la sécurité pour le contrôle des bagages et des individus (dans les aéroports par exemple). Toujours dans la famille des ondes EM, on a également les ondes lumineuses. Des techniques comme l'interférométrie peuvent être employées pour en extraire des informations, tel que leur spectre. Son domaine d'application le plus connu est l'astronomie, utilisée dans les télescopes pour étudier les ondes provenant de l'espace. Dans les ondes mécaniques, les ultrasons sont couramment utilisés pour l'échographie. Elle est utilisée en médecine, pour étudier les organes internes, mais on la retrouve également dans l'industrie, adaptée pour la détection de défauts.

Les systèmes d'observation renvoient une version modifiée de l'objet originel. On modélise ce phénomène par une fonction mathématique reliant l'entrée et la sortie du système. En plus de cette altération, les observations sont souvent sujets à des incertitudes : par exemple, des erreurs de mesures, une modélisation simplifiée de la réalité terrain, *etc*. Ces erreurs sont généralement modélisées par un bruit.

Dans ces travaux, nous nous focalisons sur les problèmes de déconvolution d'image. On souhaite reconstruire une image de taille $N \times N$ pixels. Elle est désignée par le vecteur $\boldsymbol{x}$ de dimension $P = N^2$. On désigne les données perçues au travers du système d'observation par le vecteur $\boldsymbol{y}$, également de dimension P. L'opérateur de convolution est caractérisé par la matrice $\boldsymbol{H}$, de dimension $P \times P$. Les incertitudes/erreurs sont modélisées par un terme additif, désigné par le vecteur $\boldsymbol{e}$, de

même dimension que $\boldsymbol{y}$. Le système d'observation est schématisé sur la Figure 1.1.

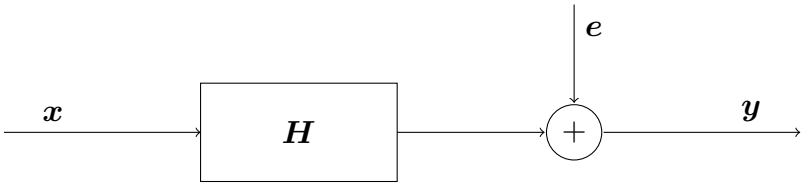
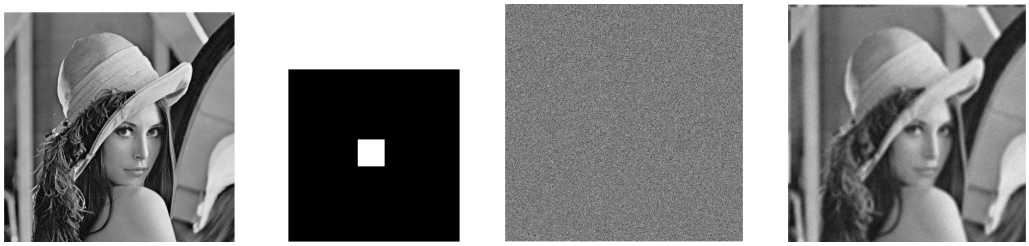


FIGURE 1.1 – Système d'observation

Mathématiquement, il est caractérisé par l'équation

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{e}. \tag{1.1}$$

Sur la Figure 1.2, on peut voir un exemple d'un tel système : l'image originale $\boldsymbol{x}$ (affichée en (a)) a subi une convolution de l'instrument $\boldsymbol{H}$ (il s'agit d'un opérateur de flou, dont la réponse impulsionelle est (b)) et l'ajout d'un bruit blanc gaussien $\boldsymbol{e}$ (image (c)). L'image résultante (d) correspond aux données observées $\boldsymbol{y}$. Le rapport signal sur bruit est d'environ 30dB.



(a) Entrée du système (b) Flou 5x5 (c) Bruit (d) Sortie du système

FIGURE 1.2 – Exemple de système d'observation : l'image (d) résulte de la convolution par le filtre (b) de l'image originale (a) et de l'ajout d'un bruit (c).

Comme nous le montre l'image (d), les dégradations subies par l'objet entraînent une perte d'informations sur l'inconnu (a) et rendent le problème **mal-conditionné**. Il devient alors délicat de retrouver l'objet (a) à partir de ces données corrompues, pour diverses raisons. Dans certains cas, il peut ne pas y avoir de solution ou, à l'inverse, y avoir plusieurs solutions. Dans le cas qui nous concerne, il y a une unique solution mais la restauration est délicate à cause de l'instabilité de cette solution face aux perturbations. Pour compenser, il est nécessaire de régulariser.

1.2 Régularisation et inférence bayésienne

Pour régulariser et reconstruire l'image originale, une possibilité est de considérer de l'information *a priori*. Cette prise en compte d'information supplémentaire se fait au travers de l'inférence bayésienne [Rob10]. Dans cette approche, les quantités inconnues sont modélisées comme des réalisations de variables aléatoires. On régularise en introduisant des distributions *a priori* sur les inconnues. On compte parmi ces inconnues l'image $\boldsymbol{x}$ et le bruit $\boldsymbol{e}$, mais pas uniquement. En effet, ces entités sont pilotées par des **hyperparamètres**, dont une partie est inaccessible. Ces derniers sont stockés dans le vecteur $\boldsymbol{\theta}$, de dimension Q. Pour ces hyperparamètres, on introduit la loi *a*

priori

$$\pi_\theta(\theta). \tag{1.2}$$

Dans nos recherches, les composantes de θ sont toujours modélisées comme indépendantes. Par conséquent, cela permet de décomposer $\pi_\theta(\theta)$ en le produit des priors de chaque hyperparamètre

$$\pi_\theta(\theta) = \prod_{q=1}^{Q} \pi_{\theta_q}(\theta_q). \tag{1.3}$$

Selon le besoin, pour chaque composante, on peut choisir entre

- un prior uniforme sur une intervalle $[\theta_q^{\,\mathrm{m}}\,,\ \theta_q^{\,\mathrm{M}}]$

$$\mathscr{U}(\theta_q; \theta_q^{\,\mathrm{m}}, \theta_q^{\,\mathrm{M}}) = \frac{1}{\theta_q^{\,\mathrm{M}} - \theta_q^{\,\mathrm{m}}}\,,$$

- un prior de type loi Gamma de paramètres $(\alpha\,,\,\beta)$

$$\Gamma(\theta_q; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\ \theta_q^{\alpha-1}\ \exp\big[-\beta\theta_q\big]\mathbb{1}_+(\theta_q)\,.$$

Comme les hyperparamètres pilotent l'image et le bruit, les distributions *a priori* pour $\boldsymbol{x}$ et $\boldsymbol{e}$ dépendent des composantes de θ. Ces lois sont choisies gaussiennes centrées, avec pour covariance respective $\Sigma_{\boldsymbol{x}}$ et $\Sigma_{\boldsymbol{e}}$, qui sont de dimension $P \times P$:

$$\pi_{\boldsymbol{x}|\theta}(\boldsymbol{x}|\theta) \;=\; \mathscr{N}(\boldsymbol{x}\,;\,\mathbf{0}, \Sigma_{\boldsymbol{x}}) \tag{1.4}$$
$$\pi_{\boldsymbol{e}|\theta}(\boldsymbol{e}|\theta) \;=\; \mathscr{N}(\boldsymbol{e}\,;\,\mathbf{0}, \Sigma_{\boldsymbol{e}}) \tag{1.5}$$

En particulier, lorsque l'on détaille la loi pour le bruit

$$\pi_{\boldsymbol{e}|\theta}(\boldsymbol{e}|\theta) = (2\pi)^{-P/2}\ \big[\det\Sigma_{\boldsymbol{e}}\big]^{-1/2}\ \exp\Big[-\frac{1}{2}\,\boldsymbol{e}^\dagger\Sigma_{\boldsymbol{e}}^{-1}\boldsymbol{e}\Big]$$

et que l'on remplace $\boldsymbol{e}$ à l'aide de l'identité (1.1), la loi pour le bruit devient alors la distribution des données $\boldsymbol{y}$, que l'on nomme **vraisemblance de l'objet attachée aux données**

$$f(\boldsymbol{y}|\boldsymbol{x}, \theta) = (2\pi)^{-P/2}\ \big[\det\Sigma_{\boldsymbol{e}}\big]^{-1/2}\ \exp\Big[-\frac{1}{2}\,(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})^\dagger\Sigma_{\boldsymbol{e}}^{-1}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})\Big]. \tag{1.6}$$

qui est également une gaussienne de moyenne $\boldsymbol{H}\boldsymbol{x}$ et de covariance $\Sigma_{\boldsymbol{e}}$.

On décompose les covariances $\Sigma_{\boldsymbol{x}}$ et $\Sigma_{\boldsymbol{e}}$ en le produit d'un paramètre γ et d'une matrice $\boldsymbol{R}$:

$$\Sigma_{\boldsymbol{x}} = \gamma_{\boldsymbol{x}}^{-1}\boldsymbol{R}_{\boldsymbol{x}} \qquad \text{et} \qquad \Sigma_{\boldsymbol{e}} = \gamma_{\boldsymbol{e}}^{-1}\boldsymbol{R}_{\boldsymbol{e}} \tag{1.7}$$

Les paramètres $\gamma_{\boldsymbol{x}}$ et $\gamma_{\boldsymbol{e}}$ quantifient les niveaux de signal et de bruit, tandis que les matrices $\boldsymbol{R}_{\boldsymbol{x}}$ et $\boldsymbol{R}_{\boldsymbol{e}}$ fixent la structure de la covariance. De plus amples détails seront donnés dans la section 1.3.

Le système d'observation est invariant par translation et on considère les modèles de l'image $\boldsymbol{x}$ et du bruit $\boldsymbol{e}$ stationnaires. De ce fait, les matrices $\Sigma_{\boldsymbol{x}}$, $\Sigma_{\boldsymbol{e}}$ et $\boldsymbol{H}$ ont une structure Tœplitz-Bloc-Tœplitz. Une matrice est qualifiée de Tœplitz lorsque toutes ses lignes sont issues d'un décalage de la première ligne. Une matrice Tœplitz-Bloc-Tœplitz est une matrice $P \times P$ où chaque bloc

$N \times N$ est une matrice de Tœplitz et la disposition de tous ces blocs suit une forme de Tœplitz (voir Annexe D).

La grande dimension des matrices de covariance et de convolution ($P^2 = N^4$ éléments) rend leur manipulation longue et difficile. Une solution pour éviter cette difficulté est le recours à une approximation Circulant-Bloc-Circulant. Une matrice circulante est un cas particulier de matrice de Tœplitz : toutes ses lignes sont issues d'un décalage, cette fois-ci circulaire, de la première ligne. Dans la même veine que les matrices Tœplitz-Bloc-Tœplitz, les matrices Circulant-Bloc-Circulant sont des matrices $P \times P$ où chaque bloc $N \times N$ est une matrice circulante et tous ces blocs sont disposés circulairement. Des descriptions plus précises de ces matrices sont également disponibles en Annexe D.

Le caractère circulant permet ainsi de diagonaliser les matrices concernées dans l'espace de Fourier. On a recours par deux fois à une transformée de Fourier discrète (TFD) :

i) une TFD pour passer dans la base de Fourier,

ii) une TFD de la première ligne de $\Sigma_{\boldsymbol{x}}$ et $\Sigma_{\boldsymbol{e}}$ pour obtenir les valeurs propres.

Calculer son inverse revient alors à inverser une matrice diagonale, tâche beaucoup plus facile.

On note $\boldsymbol{F}$ la matrice de transformée de Fourier discrète. En notant $^{\mathrm{t}}$ l'opération de transposition et † l'opération de transposition conjuguée, la matrice $\boldsymbol{F}$ vérifie $\boldsymbol{F}^{\mathrm{t}} = \boldsymbol{F}$ et $\boldsymbol{F}^{\dagger} = \boldsymbol{F}^{-1}$. En notant $p = 1, \ldots, P$ l'indice du p-ième pixel, les matrices $\Sigma_{\boldsymbol{x}}$, $\Sigma_{\boldsymbol{e}}$ et $\boldsymbol{H}$ se décomposent alors

$$
\begin{aligned}
&\Sigma_{\boldsymbol{x}} = \gamma_{\boldsymbol{x}}^{-1} \boldsymbol{F}^{\dagger} \mathring{\boldsymbol{S}}_{\boldsymbol{x}} \boldsymbol{F} && \text{où} && \mathring{\boldsymbol{S}}_{\boldsymbol{x}} = \operatorname{diag}\left\{\mathring{s}_{\boldsymbol{x}}(p)\right\} \\
&\Sigma_{\boldsymbol{e}} = \gamma_{\boldsymbol{e}}^{-1} \boldsymbol{F}^{\dagger} \mathring{\boldsymbol{S}}_{\boldsymbol{e}} \boldsymbol{F} && \text{où} && \mathring{\boldsymbol{S}}_{\boldsymbol{e}} = \operatorname{diag}\left\{\mathring{s}_{\boldsymbol{e}}(p)\right\} \\
&\boldsymbol{H} = \boldsymbol{F}^{\dagger} \mathring{\boldsymbol{H}} \boldsymbol{F} && \text{où} && \mathring{\boldsymbol{H}} = \operatorname{diag}\left\{\mathring{h}(p)\right\} .
\end{aligned}
\tag{1.8}
$$

Les matrices $\mathring{\boldsymbol{S}}_{\boldsymbol{x}}$ et $\mathring{\boldsymbol{S}}_{\boldsymbol{e}}$ sont, aux facteurs $\gamma_{\boldsymbol{x}}$ et $\gamma_{\boldsymbol{e}}$ près, les densités spectrales de puissance (DSP) de l'objet et du bruit. Ces décompositions permettent de simplifier le calcul des lois (1.4) et (1.6). Par exemple, pour (1.4), la circularité simplifie deux choses.

- Le calcul du déterminant

$$
\begin{aligned}
\det \Sigma_{\boldsymbol{x}} &= \det\left[\gamma_{\boldsymbol{x}}^{-1}\, \boldsymbol{F}^{\dagger}\, \mathring{\boldsymbol{S}}_{\boldsymbol{x}}\, \boldsymbol{F}\right] = \gamma_{\boldsymbol{x}}^{-P} \det\, \mathring{\boldsymbol{S}}_{\boldsymbol{x}} \\
&= \gamma_{\boldsymbol{x}}^{-P} \prod_{p=1}^{P} \mathring{s}_{\boldsymbol{x}}(p)\,,
\end{aligned}
\tag{1.9}
$$

- l'argument de l'exponentielle

$$
\begin{aligned}
\boldsymbol{x}^{\dagger}\, \Sigma_{\boldsymbol{x}}^{-1}\, \boldsymbol{x} = \boldsymbol{x}^{\dagger}\, (\, \gamma_{\boldsymbol{x}}^{-1}\, \boldsymbol{F}^{\dagger}\, \mathring{\boldsymbol{S}}_{\boldsymbol{x}}\, \boldsymbol{F}\,)^{-1}\, \boldsymbol{x} &= \gamma_{\boldsymbol{x}}\, (\boldsymbol{F}\boldsymbol{x})^{\dagger}\, \mathring{\boldsymbol{S}}_{\boldsymbol{x}}\, (\boldsymbol{F}\boldsymbol{x}) \\
= \gamma_{\boldsymbol{x}}\, \mathring{\boldsymbol{x}}^{\dagger}\, \mathring{\boldsymbol{S}}_{\boldsymbol{x}}^{-1}\, \mathring{\boldsymbol{x}} \;=\; \gamma_{\boldsymbol{x}} \sum_{p=1}^{P} \frac{|\mathring{x}(p)|^2}{\mathring{s}_{\boldsymbol{x}}(p)}\,. &
\end{aligned}
\tag{1.10}
$$

où $\mathring{\boldsymbol{x}} = \boldsymbol{F}\boldsymbol{x}$ est la TFD de l'image.

La loi de probabilité pour la variable spatiale $\boldsymbol{x}$ peut alors s'exprimer dans le domaine fréquentiel, dans lequel la manipulation des matrices est plus aisée :

$$\pi_{\boldsymbol{x}|\theta}(\boldsymbol{x}|\theta) = (2\pi)^{-P/2}\,\gamma_{\boldsymbol{x}}^{P/2}\,\left[\prod_{p=1}^{P}\mathring{s}_{\boldsymbol{x}}(p)^{-1/2}\right]\,\exp\left[-\frac{\gamma_{\boldsymbol{x}}}{2}\,\sum_{p=1}^{P}\frac{|\mathring{x}(p)|^2}{\mathring{s}_{\boldsymbol{x}}(p)}\right].$$

On peut réécrire la loi pour $\boldsymbol{x}$ avec une somme sur les fréquences

$$\pi_{\boldsymbol{x}|\theta}(\boldsymbol{x}|\theta) = (2\pi)^{-P/2}\gamma_{\boldsymbol{x}}^{P/2}\,\exp -\frac{1}{2}\sum_{p=1}^{P}\left[\log\mathring{s}_{\boldsymbol{x}}(p) + \gamma_{\boldsymbol{x}}\,\frac{|\mathring{x}(p)|^2}{\mathring{s}_{\boldsymbol{x}}(p)}\right]. \tag{1.11}$$

Dans cette expression apparaissent

- la DSP de l'image $\gamma_{\boldsymbol{x}}^{-1}\,\mathring{s}_{\boldsymbol{x}}(p)$,
- le terme $|\mathring{x}(p)|^2$, périodogramme de l'image $\boldsymbol{x}$ pris à la fréquence p. Ce périodogramme est un estimateur empirique de la DSP précédente

$$|\mathring{x}(p)|^2 \;\approx\; \gamma_{\boldsymbol{x}}^{-1}\,\mathring{s}_{\boldsymbol{x}}(p)\,.$$

Une étude plus approfondie leur est consacrée à la section 3.1.2 page 32.

En notant $\mathring{\boldsymbol{y}} = \boldsymbol{F}\boldsymbol{y}$ la TFD des observations, la vraisemblance de l'objet attachée aux données (1.6) se réécrit elle aussi dans le domaine fréquentiel et sous la forme une somme sur les fréquences

$$f(\boldsymbol{y}|\boldsymbol{x},\theta) = (2\pi)^{-P/2}\gamma_{\boldsymbol{e}}^{P/2}\,\exp -\frac{1}{2}\sum_{p=1}^{P}\left[\log\mathring{s}_{\boldsymbol{e}}(p) + \gamma_{\boldsymbol{e}}\,\frac{|\mathring{y}(p) - \mathring{h}(p)\mathring{x}(p)|^2}{\mathring{s}_{\boldsymbol{e}}(p)}\right]. \tag{1.12}$$

A l'instar du prior (1.11) , on retrouve, pour une fréquence donnée p,

- la DSP du bruit $\gamma_{\boldsymbol{e}}^{-1}\,\mathring{s}_{\boldsymbol{e}}(p)$,
- un périodogramme, celui du résidu, $|\mathring{y}(p) - \mathring{h}(p)\mathring{x}(p)|^2$, qui est un estimateur empirique de la DSP du bruit

$$|\mathring{y}(p) - \mathring{h}(p)\mathring{x}(p)|^2 \;\approx\; \gamma_{\boldsymbol{e}}^{-1}\,\mathring{s}_{\boldsymbol{e}}(p)\,,$$

Également, une étude de leur relation est traitée à la section 4.1 page 45.

1.3 Modèles de DSP

Le choix du type de covariance ou de DSP a un impact important sur l'information spatiale et fréquentielle prise en compte, ce qui en fait une étape difficile. Le bruit est dans de nombreuses méthodes considéré blanc, modèle qui n'est pas toujours réaliste. Pour des images naturelles, cette information fréquentielle est concentrée essentiellement autour des basses fréquences, avec une certaine décroissance vers les hautes fréquences. Nous avons donc décidé de considérer

- pour l'image, plusieurs modèles de décroissance fréquentielle ;
- pour le bruit, d'inclure des modèles autres que le bruit blanc.

Dans cette optique, on se donne alors I modèles de DSP objet et J modèles de DSP bruit. On les note respectivement par les indices $i = 1,..,I$ et $j = 1,..,J$ que l'on ajoutera par la suite aux matrices de covariances et DSP associées : $\Sigma_{\boldsymbol{x},i}$ et $\Sigma_{\boldsymbol{e},j}$, $\mathring{\boldsymbol{S}}_{\boldsymbol{x},i}$ et $\mathring{\boldsymbol{S}}_{\boldsymbol{e},j}$. Nous avons cinq premiers modèles définis directement au travers de leur DSP :

- Lorentzien séparable (LorentzS) :

$$\mathring{s}(p) = \Big[(\pi w^2)\,(1+[u_p/w]^2)\,(1+[v_p/w]^2)\Big]^{-1}$$

- Lorentzien circulaire (LorentzC) :

$$\mathring{s}(p) = \Big[\pi w^2 \log(1+1/4w^2)\Big]^{-1}\,\Big[\,1+(u_p^2+v_p^2)/w^2\,\Big]^{-1}$$

- Gaussien :

$$\mathring{s}(p) = \frac{1}{2\pi w^2}\,\exp-\Big[(u_p^2+v_p^2)/2w^2\Big]$$

- Laplacien :

$$\mathring{s}(p) = \frac{1}{4w^2}\,\exp-\Big[(|u_p|+|v_p|)/w\Big]$$

- Blanc :

$$\mathring{s}(p) = \mathbb{1}(u_p, v_p)$$

où w est la largeur de la DSP et (u_p, v_p) sont des variables parcourant l'espace des fréquences réduites : $u_p, v_p \in [-0.5\ ,\ 0.5]$.

Nous avons également deux autres modèles, qui sont définis au travers de matrices d'interaction inter-pixels

- InterPix-4 :

$$\frac{1}{\sqrt{8}}\begin{pmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{pmatrix}\ ,$$

- InterPix-8 :

$$\frac{1}{\sqrt{18}}\begin{pmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{pmatrix}\ ,$$

dont les DSP sont circulaires et ne sont pas fonctions d'une largeur w.

Il existe évidemment bien d'autres modèles de décroissance fréquentielle. Nous avons choisi ces modèles pour leurs différences de structure de dépendances, que nous avons jugées enrichissantes pour l'étude. Il s'agit d'une méthode générale de sélection de modèle et l'utilisateur est libre de constituer sa propre liste de modèles. Notre liste évoluera au cours du manuscrit et sera présentée au début des chapitres.

En Figure 1.3 sont affichées les structures énumérées ci-dessus, en échelle logarithmique, pour une largeur de $0,1$: ces images représentent les DSP aux facteurs $\gamma_{\boldsymbol{x}}$ et $\gamma_{\boldsymbol{e}}$ près, c'est-à-dire les contenus fréquentiels, dont les éléments sont stockés dans la diagonale de $\mathring{\boldsymbol{S}}_{\boldsymbol{x},i}$ et $\mathring{\boldsymbol{S}}_{\boldsymbol{e},j}$.

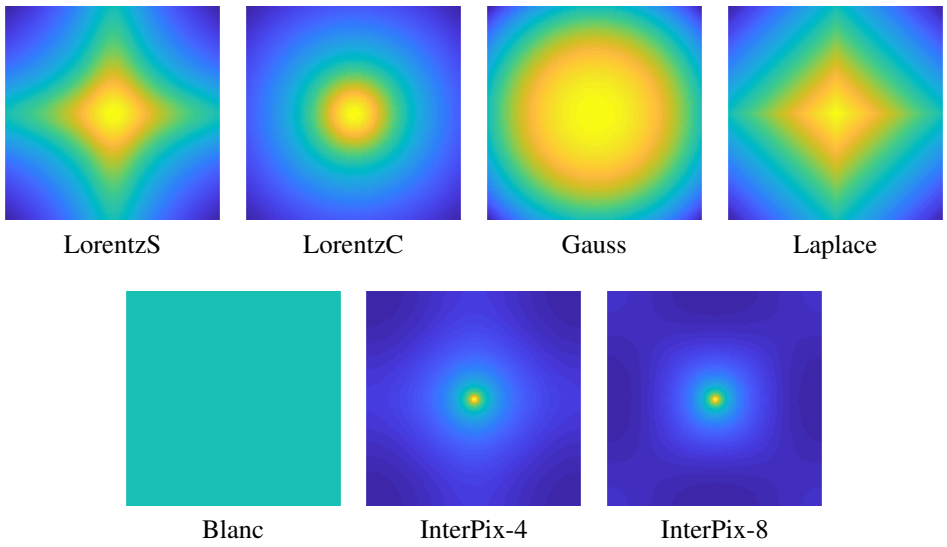


FIGURE 1.3 – Différentes formes de DSP en échelle logarithmique. La variable u_p parcourt l'axe des abscisses et v_p celui des ordonnées. La fréquence $(0,0)$ est au centre de l'image, point culminant des DSP. Les basses fréquences se trouvent près du centre et les hautes fréquences vers les bords.

Pour les quatre premiers modèles, ces matrices $\mathring{\boldsymbol{S}}_{\boldsymbol{x},i}$ et $\mathring{\boldsymbol{S}}_{\boldsymbol{e},j}$ sont, comme on a pu le voir, fonctions d'une largeur w. Cette dernière ajuste la longueur de corrélation entre les pixels : plus la largeur en fréquence est petite, plus les pixels sont corrélés. Les valeurs fluctuent alors plus lentement d'un pixel à un autre, et on peut observer des zones de pixels comme sur la Figure 1.4.

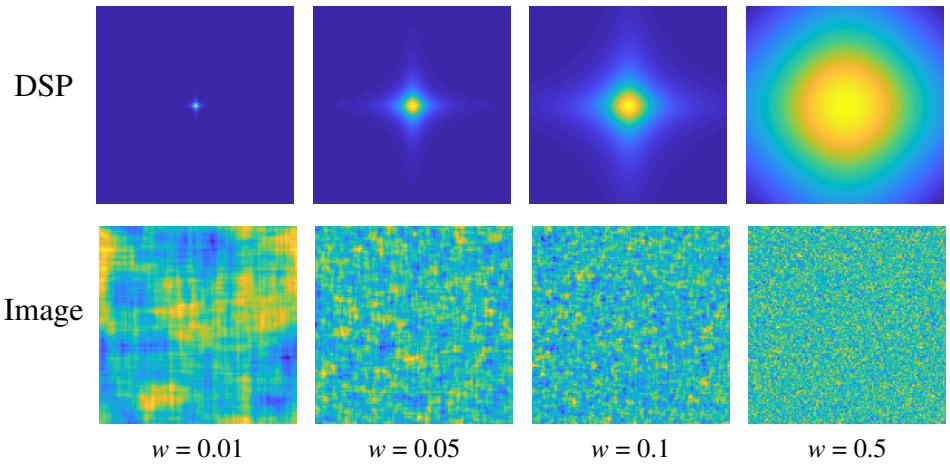


FIGURE 1.4 – Influence de la largeur spectrale sur la corrélation inter-pixels. DSP lorentzienne en échelle linéaire.

L'image et le bruit sont générés à partir de ces matrices $\mathring{\boldsymbol{S}}_{\boldsymbol{x},i}$ et $\mathring{\boldsymbol{S}}_{\boldsymbol{e},j}$ (voir la procédure en

Annexe C). Sur la Figure 1.5 sont exposées des réalisations d'image et de bruit selon les quatre premiers modèles. Elles sont ici représentées dans le domaine spatial.

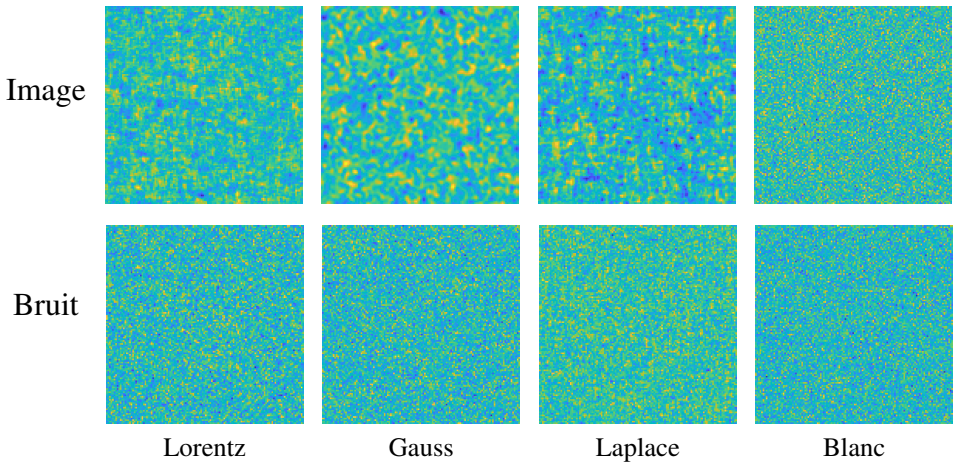


FIGURE 1.5 – Exemples de réalisation d'image et de bruit, de dimension 128×128 pixels, sous différentes DSP. Pour l'image, la largeur est fixée à $0,1$ et le vrai niveau γ_x^* vaut 6 ; pour le bruit, la largeur est à $0,5$ et le vrai niveau γ_e^* à 4.

Remarque 1. *Les réalisations des Figures 1.4 et 1.5, issues de covariances structurées, sont des images dites « texturées ». [Vac14, VGB15, VGR12, VGB14] abordent la question de la sélection de modèle pour ce type d'image. De manière générale, [Vac14] fournit une étude complète sur la déconvolution, la segmentation et l'estimation de paramètres d'images texturées, thématiques que l'on retrouve également dans les articles associés [GV17, RGGV15, VG19]. Dans ce contexte d'images texturées, nous suivons une méthodologie similaire à [OGR10a, OGR10b].*

Pour la convolution, on choisit un filtre passe-bas, dont le gain est de type sinus cardinal et de largeur 1 en fréquence réduite. Une représentation fréquentielle est donnée sur la Figure 1.6.

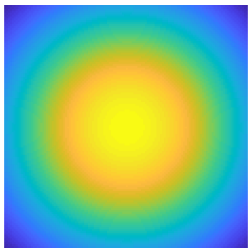


FIGURE 1.6 – Noyau de convolution : gain en sinus cardinal de largeur 1 dans l'espace des fréquences réduites. Le centre de l'image coïncide avec la fréquence (0,0).

Avec I modèles pour l'image et J pour le bruit, leur combinaison dans l'équation (1.1) offre un catalogue de $K = IJ$ modèles possibles pour les données observées $\boldsymbol{y}$. Chaque combinaison (i, j) est désignée par une valeur unique $k = 1, .., K$.

Des exemples de réalisations de $\boldsymbol{y}$ sont données sur la Figure 1.7, obtenues à partir des images vues sur la Figure 1.5. La première ligne présente les observations dans le domaine spatial et la seconde ligne les périodogrammes de ces observations.

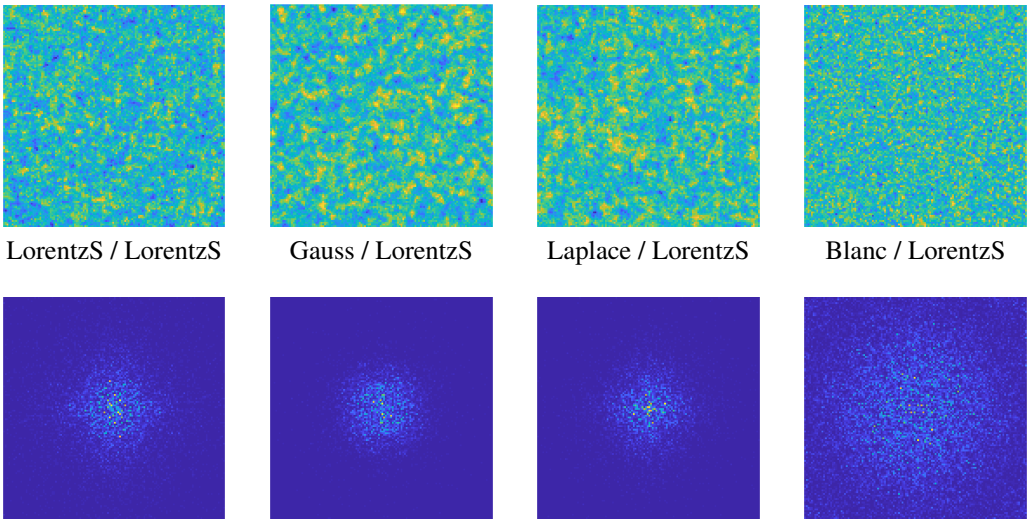


FIGURE 1.7 – Exemples de réalisation de données $\boldsymbol{y}$ (première ligne) et leurs périodogrammes associés (seconde ligne). La légende précise les modèles utilisés : modèle image / modèle bruit.

L'information fréquentielle des images originales a clairement été altérée. D'une part, les hautes fréquences ont été filtrées par convolution et d'autre part, l'ajout du bruit a modifié l'information fréquentielle globale. Il est visuellement difficile de retrouver les modèles de DSP originels de l'image et du bruit. La section suivante présente une approche pour comparer et sélectionner automatiquement les modèles en question.

1.4 Sélection de modèles et calcul d'évidence

La plupart des travaux existants impose un modèle de DSP pour les distributions *a priori* objet et bruit et déduisent une solution. Pour enrichir l'étude, d'autres considèrent plusieurs modèles et déduisent pour chacun une solution. Tout d'abord, l'opération peut être fastidieuse. Rien que dans notre exemple, on peut avoir $K = 49$ combinaisons modèle objet/modèle bruit possibles à tester. Vient ensuite la question du critère de comparaison : à quelle quantité doit-on se fier pour sélectionner un modèle ? Pour faciliter ce travail, une solution consiste à mettre en œuvre une méthode de **comparaison** et de **sélection automatique de modèle**. L'essentiel de ces travaux est dédié à cette problématique de **sélection de modèles** (SdM) et concerne les modèles de DSP objet et bruit : il s'agira de développer une approche permettant de sélectionner le modèle objet et le modèle bruit à partir d'observations.

Le modèle est désigné par la variable discrète $\mathscr{M}$, prenant la valeur $k = 1,..,K$ vue précédemment. Dans ses travaux, on se réfère à la théorie de la décision bayésienne : on sélectionne le modèle $\widehat{\mathscr{M}}$ de plus forte probabilité *a posteriori* $p(\mathscr{M} = k|\boldsymbol{y})$

$$\widehat{\mathscr{M}} = \arg\max_{k} [p(\mathscr{M} = k|\boldsymbol{y})] \,. \tag{1.13}$$

Cette probabilité se calcule grâce à une quantité phare, appelée l'**évidence** $f(\boldsymbol{y}|\mathscr{M} = k)$ et de la

probabilité *a priori* des modèles

$$p(\mathscr{M}=k|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\mathscr{M}=k)\; p(\mathscr{M}=k)}{f(\boldsymbol{y})}$$

$$= \frac{f(\boldsymbol{y}|\mathscr{M}=k)\; p(\mathscr{M}=k)}{\sum_{k'=1}^{K} f(\boldsymbol{y}|\mathscr{M}=k')\; p(\mathscr{M}=k')}\,. \tag{1.14}$$

Cette méthodologie fonctionne avec n'importe quel prior pour le modèle (la question du choix du prior est étudiée par [LGTB13] dans un cas particulier), mais nous avons choisi un prior non-informatif uniforme, avec lequel aucun modèle n'est privilégié

$$p(\mathscr{M}=k) = \frac{1}{K}\,.$$

De ce fait, l'identité (1.14) se simplifie

$$p(\mathscr{M}=k|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\mathscr{M}=k)}{\sum_{k'=1}^{K} f(\boldsymbol{y}|\mathscr{M}=k')}, \tag{1.15}$$

et illustre clairement l'importance de connaître les évidences des modèles. L'évidence s'obtient à partir de la loi jointe $f(\boldsymbol{y},\boldsymbol{x},\theta\,|\,\mathscr{M}=k)$, conditionnée par le modèle, qu'il faut marginaliser par rapport à l'objet $\boldsymbol{x}$ et par rapport aux paramètres inconnus θ,

$$f(\boldsymbol{y}|\mathscr{M}=k) = \int_{\Theta}\int_{\mathbb{R}^P} f(\boldsymbol{y},\boldsymbol{x},\theta\,|\,\mathscr{M}=k)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\theta\,. \tag{1.16}$$

La loi jointe se décompose (d'après la loi de probabilité totale) en le produit de la vraisemblance attachée aux données $f(\boldsymbol{y}\,|\,\boldsymbol{x},\theta,\mathscr{M}=k)$ et des priors sur les inconnus $\pi_{\boldsymbol{x}|\theta}(\boldsymbol{x}|\theta,\mathscr{M}=k)$ et $\pi_{\theta}(\theta|\mathscr{M}=k)$

$$f(\boldsymbol{y}|\mathscr{M}=k) = \int_{\Theta}\int_{\mathbb{R}^P} f(\boldsymbol{y}\,|\,\boldsymbol{x},\theta,\mathscr{M}=k)\,\pi_{\boldsymbol{x}|\theta}(\boldsymbol{x}|\theta,\mathscr{M}=k)\,\pi_{\theta}(\theta|\mathscr{M}=k)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\theta\,. \tag{1.17}$$

La principale difficulté provient du calcul de ces intégrales, qui peut s'avérer compliqué aussi bien analytiquement que numériquement. Notre cadre d'étude simplifie certaines étapes de cette marginalisation laborieuse.

1. Se placer dans un modèle gaussien nous permet d'avoir une expression analytique de la marginale en $\boldsymbol{x}$. On évacue ainsi un problème d'intégration en grande dimension dû à la taille de l'image $\boldsymbol{x}$. Il n'est plus possible ensuite d'intégrer analytiquement les inconnus restantes. Des dépendances inextricables entre ces dernières (*cf.* Figure 1.8, équations (1.11) et (1.12), liste des modèles de DSP, *etc.*) apparaissent dans la loi jointe marginalisée en $\boldsymbol{x}$, qui est alors impossible à intégrer analytiquement. Il nous faut dès lors calculer l'évidence par des méthodes numériques.
2. Le caractère circulant simplifie la manipulation numérique des matrices de covariance et de convolution. De dimension importante ($P \times P$), les opérations classiques comme le calcul de

leur inverse deviennent abordables, grâce à leur diagonalisation dans le domaine de Fourier.

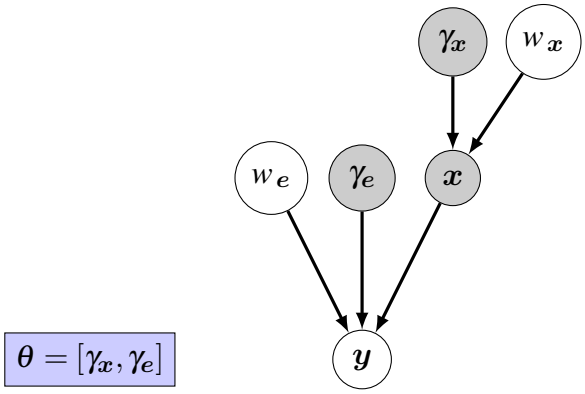


FIGURE 1.8 – Arbre des dépendances hiérarchiques de notre problème. Les cases grisées désignent les quantités inconnues.Toutes ces quantités sont conditionnées par le modèle $\mathscr{M}$, que nous n'affichons pas ici.

1.5 Présentation du document

Le **chapitre 2** est consacré à l'état de l'art des méthodes de sélection de modèles. Nous nous intéressons en particulier aux algorithmes qui visent un calcul analytique exact de l'évidence. La marginalisation de la loi jointe par rapport à l'image et aux hyperparamètres pour obtenir l'évidence, précédemment donnée en (1.16), est dans bon nombre de cas infaisable analytiquement. Grâce au caractère gaussien, on arrive à intégrer analytiquement la loi jointe par rapport à l'image

$$f(\boldsymbol{y}|\mathscr{M} = k) = \int_{\Theta} f(\boldsymbol{y}, \theta|\mathscr{M} = k)\,\mathrm{d}\theta = \int_{\Theta}\int_{\mathbb{R}^P} f(\boldsymbol{y}, \boldsymbol{x}, \theta\,|\,\mathscr{M} = k)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\theta\ .$$

Cela permet d'évacuer un certain nombre de difficultés mais, en contrepartie, la forme de la loi $f(\boldsymbol{y}, \theta\,|\,\mathscr{M} = k)$ la rend impossible à intégrer analytiquement par rapport à θ. L'évidence étant cruciale pour la sélection de modèles, il nous faut marginaliser les paramètres avec des méthodes numériques. La circularité des matrices de covariances et de convolution rendent leur manipulation numérique plus aisée, notamment pour calculer leur inverse.

Les deux premiers algorithmes abordés sont qualifiés de « naïfs ». Ils sont ainsi nommés à cause de leur faible rendement vis-à-vis des ressources utilisées. Le premier algorithme, dit de **marginalisation brute**, consiste à discrétiser l'espace des paramètres de l'intégrale ci-dessus, puis à calculer cette dernière avec une méthode basique des rectangles ou des trapèzes. Il s'agit d'un algorithme sûr et efficace, mais qui devient très rapidement gourmand en ressources lorsque l'on augmente la taille et la précision des grilles de valeurs, mais aussi le nombre d'hyperparamètres inconnus.

Le second est un algorithme d'**échantillonnage sous prior** : l'intégrale ci-dessus peut s'écrire comme une espérance sous la loi *a priori* des hyperparamètres. Par conséquent, elle peut être approchée par une moyenne empirique de la vraisemblance évaluée avec des échantillons du prior. Les lois *a priori* étant choisies peu informatives, il faut un grand nombre de tirages pour converger vers l'espérance. Tout comme le premier algorithme, le nombre d'hyperparamètres à intégrer

impacte la consommation de ressources.

L'algorithme suivant est la **moyenne harmonique**. Il est possible d'écrire l'inverse évidence comme une espérance, sous la loi *a posteriori*, de l'inverse vraisemblance. Cette espérance peut alors être approchée par une moyenne empirique de l'inverse vraisemblance, évaluée avec des tirages *a posteriori*. Bien qu'il soit facile à mettre en œuvre, il présente des problèmes de convergence : l'inverse vraisemblance peut avoir une variance infinie. C'est malheureusement le cas pour les différentes situations que nous traitons dans ce document ; la preuve analytique est développée aux sections 3.2.3 et 4.3.2.

Nous présentons ensuite l'**algorithme de Chib couplé à un échantillonneur de Gibbs**. Dans cette approche, l'idée est de calculer l'évidence avec l'identité

$$f(\boldsymbol{y}|\mathscr{M} = k) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta}, \mathscr{M} = k)\, \pi(\boldsymbol{\theta}|\mathscr{M} = k)}{p(\boldsymbol{\theta}|\boldsymbol{y}, \mathscr{M} = k)},$$

pour un $\boldsymbol{\theta}$ donné. La formule ne préconise rien sur la valeur de $\boldsymbol{\theta}$ mais dans la pratique, il est choisi à forte densité. Dans cette identité, le dénominateur n'est pas disponible analytiquement, à cause de sa constante de normalisation. Chib propose alors de calculer cette loi *a posteriori* en introduisant une variable latente, permettant d'écrire la loi sous la forme d'une espérance. On peut alors l'approcher par une moyenne empirique d'échantillons de la variable latente, que l'on obtient ici avec une chaîne de Gibbs.

Nous continuerons notre état de l'art avec l'algorithme des **power posteriors** et le **Widely Applicable BIC** (WBIC), qui est un cas particulier du premier. L'algorithme des power posteriors suit une démarche assez différente de celles qu'on a pu voir jusque là. Tout d'abord, on introduit un paramètre de température pour tempérer la vraisemblance. Les lois alors construites permettent notamment de calculer la log évidence sous forme d'une intégrale par rapport au paramètre de température. L'intégrande est quant à lui l'espérance de la log vraisemblance sous une loi *a posteriori*, elle aussi tempérée. En pratique, on a recours à un double calcul numérique. D'une part, on discrétise l'intégrale et d'autre part, on approche l'espérance par une moyenne empirique d'échantillons sous la loi *a posteriori* tempérée, pour différentes valeurs de température.

Dans le WBIC, la discrétisation de l'intégrale n'est plus nécessaire. A partir du théorème des accroissements finis, il est stipulé qu'il existe une valeur de température telle que l'intégrande évaluée en cette valeur est égale à l'intégrale. Cependant, il est difficile d'obtenir analytiquement cette valeur.

Nous présenterons pour finir une méthode classique du calcul de l'évidence : **l'approximation de Laplace**, et de son lien avec le **Bayesian Information Criterion** (BIC). L'approximation de Laplace repose sur le développement de Taylor à l'ordre 2 du logarithme de la loi jointe $\log f(\boldsymbol{y}, \boldsymbol{\theta}|\mathscr{M} = k)$. La forme quadratique qui en résulte impose une forme gaussienne à la loi jointe $\exp\left[\log f(\boldsymbol{y}, \boldsymbol{\theta}|\mathscr{M} = k)\right]$, que l'on peut facilement intégrer analytiquement. L'approximation de Laplace diffère des algorithmes présentés précédemment car elle utilise une approximation analytique pour calculer l'évidence. Quant au BIC, il se construit à partir d'une étude asymptotique de l'approximation de Laplace.

Enfin, le chapitre se conclut sur une présentation succincte d'autres méthodes et approches de la littérature que nous n'avons pas retenues pour cette étude.

****

Le **chapitre 3** présente nos travaux de sélection de modèle pour un premier cas d'étude : l'observation directe. Cette étude a pour vocation de préparer au chapitre suivant, dédié au cas en observation indirecte. Nous y présentons quelques premiers éléments de démarche, comme notamment l'augmentation progressive en complexité du problème, en considérant de plus en plus de paramètres inconnus. La cadre y est relativement simple : l'image est connue, pas de convolution ni de bruit et nous voulons comparer les modèles images i. Le niveau $\gamma_{\boldsymbol{x}}$ est considéré connu puis inconnu. Dans ces deux sous cas, l'évidence est disponible analytiquement. Nous allons profiter de cet avantage pour étudier certains des algorithmes présentés dans le cas où le paramètre de niveau est inconnu. La vraisemblance dans ce cas n'est autre que le prior $\pi(\boldsymbol{x}|\gamma_{\boldsymbol{x}}, \mathscr{M} = i)$. Nous utiliserons l'évidence analytique $\pi(\boldsymbol{x}|\mathscr{M} = i)$ comme référence pour dresser un premier bilan des performances de ces algorithmes.

Dans un premier temps, nous implémentons les deux algorithmes naïfs : la marginalisation brute et moyennage sous prior. Pour le premier, cette configuration assez commode permet d'établir une résolution et une taille de la grille « minimales » qui fournissent une bonne approximation de l'évidence. Ce cas avec seulement une variable inconnue scalaire apporte une première illustration du problème avec ces deux algorithmes : la faible quantité de ressources utiles sollicitées pour calculer l'évidence. En effet, ici, la distribution jointe $\pi(\boldsymbol{x}, \gamma_{\boldsymbol{x}}|\mathscr{M} = i)$ est très piquée, et il faut énormément de points pour quadriller / échantillonner les zones à forte probabilité.

Le dernier algorithme applicable ici est la moyenne harmonique. Cependant, comme annoncé, nous montrons que la moyenne empirique de l'inverse vraisemblance $\pi(\boldsymbol{x}|\gamma_{\boldsymbol{x}}, \mathscr{M} = i)^{-1}$ sous la loi *a posteriori* $p(\gamma_{\boldsymbol{x}}|\boldsymbol{x}, \mathscr{M} = i)$ ne converge pas vers l'inverse évidence $\pi(\boldsymbol{x}|\mathscr{M} = i)^{-1}$. Nous démontrons dans cette section que la variance de l'inverse vraisemblance est infinie dans notre cas d'étude, démonstration que nous pouvons compter parmi les contributions de ce manuscrit. La démonstration pour le cas en observation indirecte sera très similaire à celle du cas en observation directe.

La dernière section du chapitre est consacrée aux résultats de sélection de modèles obtenus

1. avec la formule de l'évidence analytique,
2. avec les algorithmes de marginalisation brute et de moyennage sous prior.

L'image étant observée directement, nous nous attendons à un taux important de bonnes sélections de modèle. Nous aurons l'occasion de faire un premier constat quant à la « naïveté » de la marginalisation brute et du moyennage sous prior. Nous constaterons également le caractère très piqué des distributions *a posteriori*, que l'on retrouve dans le cas d'étude suivant : l'observation indirecte.

****

Le **chapitre 4** est destiné au cas le plus complexe et complet de nos recherches, et nous y présentons la contribution majeure. L'image est maintenant observée indirectement, après avoir subi une convolution et l'ajout d'un bruit. Nous poursuivons cette même démarche d'augmenter progressivement la difficulté : les hyperparamètres de niveaux seront d'abord considérés connus, puis inconnus.

Grâce au caractère gaussien que nous avons choisi pour les lois *a priori* de l'image $\boldsymbol{x}$ et du bruit $\boldsymbol{e}$, il est possible de marginaliser la loi jointe par rapport à $\boldsymbol{x}$. La loi qui est en résulte est également

une gaussienne, dont on peut expliciter la moyenne et la covariance. En particulier, cette covariance sera fonction de celles de l'image $\boldsymbol{R}_{\boldsymbol{x},i}$ et du bruit $\boldsymbol{R}_{\boldsymbol{e},j}$, dont elle héritera le caractère circulant. Cette forme marginalisée de la loi jointe est utilisée par tous algorithmes que nous mettons en œuvre, d'où son importance. Nous serons même amenés pour l'algorithme de Chib à manier les deux formes de la loi jointe.

Le premier sous-cas est le dernier de notre étude où l'évidence est disponible analytiquement. À paramètres fixés, l'évidence est égale à la loi jointe marginalisée en $\boldsymbol{x}$ que nous venons d'évoquer. Nous verrons dans les résultats numériques que, tant que les paramètres de niveaux sont connus, l'altération de l'image par le système n'impacte pas la sélection de modèles, pour laquelle nous obtenons (sans surprise) de très bons résultats.

Nous traitons ensuite le second sous-cas, le plus intéressant, où les hyperparamètres de niveaux objet $\gamma_{\boldsymbol{x}}$ et bruit $\gamma_{\boldsymbol{e}}$ ne sont plus connus. Pour obtenir l'évidence, il faut désormais intégrer $\gamma_{\boldsymbol{x}}$ et $\gamma_{\boldsymbol{e}}$, ce qui n'est plus faisable analytiquement. Nous allons donc l'obtenir numériquement, avec les algorithmes que nous aurons présentés au chapitre 2. Nous commencerons par les algorithmes naïfs, la marginalisation brute et le moyennage sous prior, déjà mis en œuvre une première fois dans le chapitre dédié à l'observation directe. Nous ferons ici les mêmes constatations quant à leurs performances : ce sont des algorithmes simples et convergents, mais dont la gestion des ressources laisse à désirer. La valeur numérique de l'évidence obtenue avec ces algorithmes servira de référence pour les suivants.

Nous nous penchons ensuite sur deux algorithmes qui utilisent les ressources plus efficacement : la moyenne harmonique et l'algorithme de Chib. Ces algorithmes exploitent des échantillons *a posteriori*, qui proviennent des zones à forte densité. Ils sont en contrepartie plus compliqués à mettre en œuvre, à cause des chaînes MCMC mises en place pour obtenir ces échantillons. Comme précédemment, l'algorithme de la moyenne harmonique ne converge pas vers l'évidence, toujours à cause de la variance infinie de l'inverse vraisemblance. Nous pourrions considérer la moyenne harmonique tronquée qui élimine ce problème de variance infinie, mais nous préférons mettre en œuvre l'algorithme de Chib, qui a l'avantage ici d'être bien posé par construction. Notre hypothèse gaussienne circulante et nos choix de priors Gamma pour les hyperparamètres de niveaux permettent

1. d'expliciter la loi $p(\gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}} | \boldsymbol{y}, \boldsymbol{x}, \mathscr{M} = k)$ comme le produit de deux loi Gamma,
2. de construire des conditionnelles *a posteriori* pour $\boldsymbol{x}$, $\gamma_{\boldsymbol{x}}$ et $\gamma_{\boldsymbol{e}}$ faciles à échantillonner, utilisées dans une boucle de Gibbs pour obtenir des échantillons $\boldsymbol{x}^{(g)}$ sous la loi $p(\boldsymbol{x} | \boldsymbol{y}, \mathscr{M} = k)$.

Ce chapitre se termine sur la présentation de nos résultats numériques obtenus sur données synthétiques. L'accent est mis sur la convergence de l'algorithme de Chib vers l'évidence numérique de référence, obtenue par marginalisation brute. Nous étudions différentes situations, dans lesquelles nous mettons en regard le périodogramme des données avec la DSP du modèle sélectionné. Nous y constatons le contraste avec l'algorithme de marginalisation brute du point de vue de la complexité et de l'utilisation des ressources, ainsi que l'efficacité pour la sélection de modèle en observation indirecte. Ces différents critères appuieront notre choix d'utiliser directement l'algorithme de Chib au chapitre suivant.

****

Dans le chapitre 5, nous enrichissons notre contribution précédente en opérant la sélection de modèle sur des données réelles. Nous nous intNous considérons maintenant des images na-

turelles, observées après convolution et ajout d'un bruit blanc synthétique. La procédure reste la même qu'au chapitre 4. Nous calculons les évidences des K modèles candidats avec l'algorithme de Gibbs. Nous en déduisons ensuite les K probabilités *a posteriori* ainsi que le modèle qui les maximise.

Cependant, l'analyse est différente. En effet, l'image et les données ne sont plus issus d'un « modèle vrai » comme les données synthétiques. Nous étudions

1. quelles informations fournissent les évidences sur le degré de coïncidence des modèles candidats.
2. quelles informations délivrent le modèle sélectionné sur le contenu fréquentiel à la fois des données observées et de l'image inconnue.

Enfin, pour conclure, nous étudions l'impact de la sélection de modèle sur le processus de déconvolution. Les données sont déconvoluées avec chacun des modèles candidats, en utilisant un filtre de Wiener. Nous comparons visuellement (à l'œil nu) et quantitativement (à l'aide de distances) les images obtenues.

****

Nous terminons ce manuscrit avec le chapitre 6. Dans un premier temps, nous dressons un bilan des recherches présentées tant sur le plan de la démarche que sur le plan des résultats. Nous avons globalement obtenus des taux importants de bonnes sélection de modèles, qu'importe la configuration traitée ou l'algorithme utilisée. Les algorithmes naïfs ont démontré leur pertinence quant à leur simplicité et leur sûreté, mais ne demeurent cependant pas optimaux. L'algorithme de la moyenne harmonique dans sa forme primitive n'était pas applicable numériquement, à cause de la variance infinie de l'inverse vraisemblance. Nous avons apprécié la puissance de l'algorithme de Chib, qui cible intelligemment l'espace des paramètres et présente ainsi une meilleure gestion des ressources. Cet atout a cependant un prix : sa complexité de mis en œuvre.

Dans un second temps, nous ouvrons des perspectives sur les deux difficultés majeures que nous avons rencontrées, l'une concerne l'ajout de la largeur $w_{\boldsymbol{x}}$ aux hyperparamètres inconnus et l'autre est survenue lors de l'application de l'algorithme des power posteriors.

CHAPITRE 2

Sélection de modèles et calcul de l'évidence : état de l'art

Comme introduit précédemment, on utilise une approche bayésienne pour opérer la sélection de modèles, en se basant sur les probabilités *a posteriori* des différents modèles candidats (équations (1.14) et (1.13)). Le calcul de l'évidence $f(\boldsymbol{y}|\mathscr{M}=k)$ est l'étape primordiale de l'approche, mais aussi la plus difficile. Pour rappel, l'évidence se calcule par marginalisation de la loi jointe par rapport aux hyperparamètres $\boldsymbol{\theta}$ et à l'image $\boldsymbol{x}$:

$$\textbf{E1} \qquad f(\boldsymbol{y}|\mathscr{M}=k) = \int_{\Theta}\int_{\mathbb{R}^P} f(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\theta}\,|\,\mathscr{M}=k)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{\theta}\ . \tag{1.16}$$

Le caractère gaussien du problème permet une marginalisation analytique en $\boldsymbol{x}$ de la loi jointe (voir section 4.1). On a alors a disposition une deuxième forme de l'évidence :

$$\textbf{E2} \qquad f(\boldsymbol{y}|\mathscr{M}=k) = \int_{\Theta} f(\boldsymbol{y},\boldsymbol{\theta}\,|\,\mathscr{M}=k)\,\mathrm{d}\boldsymbol{\theta}\ . \tag{2.1}$$

La structure entre les variables rend le calcul analytique complet de l'évidence impossible et le recours à des algorithmes numériques devient nécessaire. Dans la totalité des algorithmes abordés ici, la forme **E2** sera la plus privilégiée. Cependant, on sera amené à travailler avec la loi jointe de la forme **E1**, où les variables ont des structures moins complexes.

Numériquement, on va calculer l'intégrale **E2** via deux approches. La première, que l'on peut définir comme déterministe, consiste à intégrer sur une grille de valeurs prédéfinies, comme le fait l'algorithme d'intégration brute. La seconde, stochastique, repose sur le calcul de moyennes empiriques à partir de tirages aléatoires issus ici d'une chaîne MCMC (Monte Carlo Markov Chain : [Nea93, CGM01, GS90, Per13, BDPV19, GD94]). On appliquera des algorithmes comme celui de Chib, le moyennage sous le prior, les power posteriors et la moyenne harmonique. Pour cette dernière, on démontre aux sections 3.2.3 et 4.3.2 que la variance est infinie dans notre cas d'étude et que par conséquent, elle ne converge pas vers l'évidence.

On peut également distinguer ces algorithmes sur leur manière d'explorer de l'espace des paramètres. Les algorithmes comme l'intégration brute et le moyennage sous le prior sont dites

« naïves », car l'espace est balayé « à l'aveugle » sans tenir compte des observations $\boldsymbol{y}$. A l'inverse, des algorithmes comme celui de Chib, la moyenne harmonique ou les power posteriors parcourent cet espace de manière plus « intelligente », en ciblant sur les zones de forte probabilité. On peut citer également les nombreuses méthodes de quadrature utilisant des grilles dynamiques [Dem12, MP92].

2.1 Algorithmes naïfs

2.1.1 Marginalisation brute

Cette approche consiste à calculer numériquement les intégrales par une méthode des rectangles ou des trapèzes [PvT07] sur une grille de valeurs. On favorisera la forme **E2**, car parcourir l'espace entier des images $\boldsymbol{x}$ suivant une « grille » régulière et en calculer ensuite l'intégrale serait trop coûteux et insensé. De manière plus générale, on calcule directement la version discrétisée de l'intégrale (2.1).

Considérons une grille régulière de G_q valeurs discrètes pour chaque paramètre θ_q :

$$\left(\theta_q^{\mathrm{m}} = \theta_q^{[1]}, \ldots, \theta_q^{[g]}, \ldots, \theta_q^{[G_q]} = \theta_q^{\mathrm{M}}\right).$$

Soit $\Delta_{\theta_q} = (\theta_q^{\mathrm{M}} - \theta_q^{\mathrm{m}})/G_q$ le pas d'intégration de la grille de θ_q, l'équation (2.1) se calcule numériquement ainsi :

$$f(\boldsymbol{y}|\mathscr{M} = k) = \sum_{q=1}^{Q} \left(\sum_{g=1}^{G_q} f(\boldsymbol{y}, \theta_q^{[g]} \,|\, \mathscr{M} = k) \Delta_{\theta_q} \right) \tag{2.2}$$

$$= \sum_{q=1}^{Q} \Delta_{\theta_q} \left(\sum_{g=1}^{G_q} f(\boldsymbol{y}|\theta_q^{[g]}, \mathscr{M} = k) \pi(\theta_q^{[g]}|\mathscr{M} \;=\; k) \right). \tag{2.3}$$

Il s'agit d'un algorithme simple et sûr : il permet d'obtenir une valeur précise de l'évidence pour une configuration donnée. Évidemment, selon le nombre de paramètre à estimer et la finesse des grilles à intégrer, cet algorithme est rapidement gourmand en ressources. Les résultats obtenus serviront de référence, afin de contrôler que les autres algorithmes comme celui de Chib, la moyenne harmonique ou les power posteriors fournissent la bonne valeur de l'évidence.

Il faut cependant être vigilant avec l'intervalle $[\theta_q^{\mathrm{m}}, \theta_q^{\mathrm{M}}]$ que l'on choisit, qu'importe le prior sur θ. Rigoureusement, les intégrales se calculent sur $]-\infty, +\infty[$. Il est impossible de le faire numériquement dans notre cas d'intégration simple. On est alors amené à tronquer le domaine d'intégration. Plus on choisit un intervalle étendu, plus il faudra de points pour quadriller finement le domaine à intégrer (et donc plus de temps pour calculer). A l'inverse, avec un pas d'intégration trop grand, on pourrait manquer les zones riches en information, notamment en cas de distributions très piquées. Pour récapituler, l'intégration numérique soulèvent deux points :

1. la troncature du domaine d'intégration,

2. le nombre de points pour intégrer.

Il s'agit de l'inconvénient majeur de cet algorithme, qui nécessite d'étudier au préalable d'une part l'ensemble de définition des différentes variables et d'autre part les variations de la loi jointe, avant de pouvoir construire une grille. Une amélioration possible serait d'utiliser des grilles adaptatives, où la résolution augmentent dans les zones de fortes probabilités [Dem12].

2.1.2 Moyennage sous prior

Lorsque l'on décompose la loi jointe comme le produit de la vraisemblance et du prior des hyperparamètres, l'équation (2.1) se réécrit comme une espérance sous la loi *a priori* :

$$\begin{aligned} f(\boldsymbol{y}|\mathscr{M}=k) &= \int_{\Theta} f(\boldsymbol{y},\boldsymbol{\theta}\,|\,\mathscr{M}=k)\,\mathrm{d}\boldsymbol{\theta} \\ &= \int_{\Theta} f(\boldsymbol{y}|\boldsymbol{\theta},\mathscr{M}=k)\pi(\boldsymbol{\theta}|\mathscr{M}=k)\,\mathrm{d}\boldsymbol{\theta} \\ &= \mathbb{E}_{\pi(\boldsymbol{\theta}|\mathscr{M}=k)}\Big[f(\boldsymbol{y}|\boldsymbol{\theta},\mathscr{M}=k)\Big]\,. \end{aligned} \tag{2.4}$$

Cette espérance peut être approchée par une moyenne empirique de G échantillons $\boldsymbol{\theta}^{(g)}$ tirés sous la loi *a priori* $\pi(\boldsymbol{\theta}|\mathscr{M}=k)$:

$$f(\boldsymbol{y}|\mathscr{M}=k) \approx \frac{1}{G}\sum_{g=1}^{G} f(\boldsymbol{y}|\boldsymbol{\theta}^{(g)},\mathscr{M}=k)\,. \tag{2.5}$$

Tout comme l'algorithme précédent, la forme **E1** n'est absolument pas adaptée : si l'on tire $\boldsymbol{x}$ sous son prior gaussien, il faudrait énormément d'échantillons pour parcourir suffisamment l'espace des images avant de pouvoir calculer la moyenne empirique.

Dans ce cas de figure, le prior est peu informatif (uniforme ou gamma). Il faudra donc un grand nombre de tirages pour espérer tirer suffisamment dans les zones de fortes probabilités. On est à nouveau confronté à la problématique évoquée au point 2 dans la section 2.1.1.

La question de la troncature se pose également. Pour un prior uniforme, il faut à nouveau imposer les bornes inférieure et supérieure du domaine. Ce n'est en revanche plus le cas avec un prior Gamma qui peut balayer de manière vaste l'espace des paramètres selon les valeurs $(\alpha\,,\beta)$ choisis. En contrepartie, il faut tirer davantage d'échantillons pour que la moyenne empirique se stabilise.

Les algorithmes d'intégration brute et de tirages sous prior sont dits naïfs à cause de leurs mauvais rendements vis à vis des ressources qu'ils sollicitent. Les deux nécessitent un très grand nombre de points dont seule une faible partie est « utile » pour calculer l'évidence. Une approche plus habile est de concentrer les points dans les zones de forte probabilité, ce que peuvent faire des algorithmes comme la moyenne harmonique ou encore celui de Chib et les power posteriors. En contrepartie, certaines lois seront plus complexes à échantillonner et/ou nous n'aurons pas accès analytiquement à leur constante de normalisation.

2.2 Moyenne harmonique

Le calcul par moyenne harmonique a été proposé par Newton et Raftery en 1994 [NR94, FW12, VGR12]. Elle repose sur un calcul de l'évidence par une moyenne harmonique de valeurs de la vraisemblance. En utilisant la propriété des lois de probabilités

$$\int_{\Theta} \pi(\theta|\mathscr{M} = k)\,\mathrm{d}\theta = 1,$$

on peut alors faire apparaître cette intégrale

$$\begin{aligned}
f(\boldsymbol{y}|\mathscr{M} = k)^{-1} &= f(\boldsymbol{y}|\mathscr{M} = k)^{-1} \int_{\Theta} \pi(\theta|\mathscr{M} = k)\,\mathrm{d}\theta \\
&= \int_{\Theta} f(\boldsymbol{y}|\mathscr{M} = k)^{-1}\, \pi(\theta|\mathscr{M} = k)\,\mathrm{d}\theta \quad \text{car } f(\boldsymbol{y}|\mathscr{M} = k) \text{ pas fonction de } \theta \\
&= \int_{\Theta} f(\boldsymbol{y}|\mathscr{M} = k)^{-1}\, \pi(\theta|\mathscr{M} = k)\, f(\boldsymbol{y}|\theta, \mathscr{M} = k) f(\boldsymbol{y}|\theta, \mathscr{M} = k)^{-1}\,\mathrm{d}\theta \\
&= \int_{\Theta} f(\boldsymbol{y}, \theta|\mathscr{M} = k) f(\boldsymbol{y}|\mathscr{M} = k)^{-1}\, f(\boldsymbol{y}|\theta, \mathscr{M} = k)^{-1}\,\mathrm{d}\theta \\
&= \int_{\Theta} p(\theta|\boldsymbol{y}, \mathscr{M} = k)\, f(\boldsymbol{y}|\theta, \mathscr{M} = k)^{-1}\,\mathrm{d}\theta \\
&= \mathbb{E}_{p(\theta|\boldsymbol{y}, \mathscr{M} = k)}\left[f(\boldsymbol{y}|\theta, \mathscr{M} = k)^{-1} \right]
\end{aligned}$$

On peut approcher cette inverse évidence par la moyenne empirique suivante :

$$f(\boldsymbol{y}|\mathscr{M} = k)^{-1} \approx \frac{1}{G} \sum_{g=1}^{G} f(\boldsymbol{y}|\theta^{(g)}, \mathscr{M} = k)^{-1} \tag{2.6}$$

où les $\theta^{(g)}$ sont des échantillons *a posteriori*.

Il s'agit d'une des approches les plus « simples » de la bibliographie [NRSK07]. Elle nécessite d'évaluer la vraisemblance avec des échantillons du posterior. La vraisemblance est le plus souvent connue et les échantillons peuvent s'obtenir avec une chaîne MCMC standard. Cependant, l'inverse vraisemblance peut avoir une variance (sous la loi *a posteriori*) infinie. Par conséquent, la moyenne empirique ne converge jamais vers l'espérance analytique, ce qui constitue un frein majeur pour l'application numérique.

Pour contrecarrer cet inconvénient, les auteurs de [NRSK07] proposent deux approches pour stabiliser l'algorithme. La première consiste à tronquer l'espace des hyperparamètres θ pour ne plus avoir de queues lourdes. La seconde stipule que la distribution *a posteriori* du logarithme de la vraisemblance peut être approchée par une distribution Gamma décalée.

2.3 Algorithme de Chib couplé à un échantillonneur de Gibbs

L'algorithme de Chib a été introduit en 1995 [Chi95, FW12, And10]. Sa particularité est de fournir une approximation de la loi *a posteriori* $p(\theta|\boldsymbol{y},\mathscr{M} = k)$ en se basant sur des échantillonneurs, de Gibbs ou de Metropolis-Hasting selon la version. Dans cette approche, l'évidence est extraite à partir du conditionnement

$$f(\boldsymbol{y}|\mathscr{M} = k) = \frac{f(\boldsymbol{y}|\theta,\mathscr{M} = k)\,\pi(\theta|\mathscr{M} = k)}{p(\theta|\boldsymbol{y},\mathscr{M} = k)}, \tag{2.7}$$

qui est une utilisation peu ordinaire de la règle de Bayes.

Dans cette identité, seules la vraisemblance $f(\boldsymbol{y}|\theta,\mathscr{M} = k)$ et la loi *a priori* des paramètres $\pi(\theta|\mathscr{M} = k)$ sont connus. La densité *a posteriori* des paramètres $p(\theta|\boldsymbol{y},\mathscr{M} = k)$ est connue à un coefficient près : sa constante de normalisation, qui n'est autre que ... l'évidence $f(\boldsymbol{y}|\mathscr{M} = k)$. Le cœur du travail est dédié au calcul de cette densité. On peut l'approcher par une fonction

$$\hat{p}(\theta|\boldsymbol{y},\mathscr{M} = k)$$

en utilisant des algorithmes d'échantillonnage. Nous travaillons ici avec la version qui se base sur des chaînes de Gibbs. De plus, l'équation (2.7) étant vraie pour n'importe quelle valeur de θ, on peut calculer l'évidence pour une valeur donnée $\bar{\theta}$. L'algorithme propose alors l'approximation de l'évidence suivante

$$f(\boldsymbol{y}|\mathscr{M} = k) \approx \frac{f(\boldsymbol{y}|\bar{\theta},\mathscr{M} = k)\,\pi(\bar{\theta}|\mathscr{M} = k)}{\hat{p}(\bar{\theta}|\boldsymbol{y},\mathscr{M} = k)}. \tag{2.8}$$

Nous allons voir dans la prochaine section comment approcher la distribution $\hat{p}(\bar{\theta}|\boldsymbol{y},\mathscr{M} = k)$ en utilisant un échantillonneur de Gibbs [CGM01, GG14] et comment obtenir une valeur de $\bar{\theta}$ avec les échantillons de la chaîne. Il existe également une variante de l'algorithme de Chib, plus complexe, qui utilise un échantillonneur de Metropolis-Hastings [CJ01, CG95, CGM01, And10, BGG15, VGB15].

Approximation de $\hat{p}(\bar{\theta}|\boldsymbol{y},\mathscr{M} = k)$ et calcul de $\bar{\theta}$

Il est préconisé dans l'approche de Chib de choisir un point $\bar{\theta}$ à forte probabilité, tel que la moyenne *a posteriori*, le maximum *a posteriori*, *etc*, bien que le choix de $\bar{\theta}$ ne soit pas crucial [Chi95]. L'approche repose également sur l'introduction d'une variable latente $\boldsymbol{a}$. Cette variable est choisie telle que $p(\bar{\theta}|\boldsymbol{y},\mathscr{M} = k)$ soit issue d'une marginalisation par rapport à $\boldsymbol{a}$:

$$\begin{aligned} p(\bar{\theta}|\boldsymbol{y},\mathscr{M} = k) &= \int_{\boldsymbol{A}} f(\bar{\theta},\boldsymbol{a}|\boldsymbol{y},\mathscr{M} = k)\,\mathrm{d}\boldsymbol{a} \\ &= \int_{\boldsymbol{A}} p(\bar{\theta}|\boldsymbol{y},\boldsymbol{a},\mathscr{M} = k)\,p(\boldsymbol{a}|\boldsymbol{y},\mathscr{M} = k)\,\mathrm{d}\boldsymbol{a} \qquad (2.9) \\ &= \mathbb{E}_{p(\boldsymbol{a}|\boldsymbol{y},\mathscr{M} = k)}\Big[p(\bar{\theta}|\boldsymbol{y},\boldsymbol{a},\mathscr{M} = k)\Big]. \qquad (2.10) \end{aligned}$$

La densité *a posteriori* jointe $p(\bar{\theta}|\boldsymbol{y},\mathscr{M}=k)$ étant une espérance, on peut l'approcher comme une moyenne empirique de G tirages :

$$\hat{p}(\bar{\theta}|\boldsymbol{y},\mathscr{M}=k)=\frac{1}{G}\sum_{g=1}^{G}p(\bar{\theta}|\boldsymbol{y},\boldsymbol{a}^{(g)},\mathscr{M}=k) \tag{2.11}$$
$$\text{où } \boldsymbol{a}^{(g)}\sim p(\boldsymbol{a}|\boldsymbol{y},\mathscr{M}=k)\,.$$

L'équation (2.11) nécessite d'une part de connaître la loi $f(\theta|\boldsymbol{y},\boldsymbol{a}^{(g)},\mathscr{M}=k)$ dans sa totalité (*i.e.* connaître la constante de normalisation) et d'autre part savoir échantillonner sous la loi $p(\boldsymbol{a}|\boldsymbol{y},\mathscr{M}=k)$. Une façon d'obtenir des échantillons $\boldsymbol{a}^{(g)}$ est de produire des échantillons $(\boldsymbol{a}^{(g)},\theta^{(g)})$ sous $f(\boldsymbol{a},\theta|\boldsymbol{y},\mathscr{M}=k)$ et de garder uniquement les tirages $\boldsymbol{a}^{(g)}$. Pour obtenir des réalisations sous la loi jointe *a posteriori*, on construit l'échantillonneur de Gibbs avec $Q+1$ densités conditionnelles *a posteriori* :

$$\left|\begin{array}{l} p(\boldsymbol{a}|\boldsymbol{y},\theta_1,\ldots,\theta_Q,\mathscr{M}=k)\\ p(\theta_1|\boldsymbol{y},\boldsymbol{a},\theta_2,\ldots,\theta_Q,\mathscr{M}=k)\\ \vdots\\ p(\theta_q|\boldsymbol{y},\boldsymbol{a},\theta_1,\ldots,\theta_{q-1},\theta_{q+1},\ldots\theta_Q,\mathscr{M}=k)\\ \vdots\\ p(\theta_Q|\boldsymbol{y},\boldsymbol{a},\theta_1,\ldots,\theta_{Q-1},\mathscr{M}=k)\end{array}\right.$$

Remarque 2. *Dans notre cas, l'image inconnue $\boldsymbol{x}$ peut jouer le rôle de la variable auxiliaire $\boldsymbol{a}$, ce que nous ferons dans la suite du manuscrit.*

Cette chaîne fournit G échantillons $(\theta^{(g)},\boldsymbol{a}^{(g)})$, à partir desquels on peut définir les valeurs

$$\bar{\theta}=\frac{1}{G}\sum_{g=1}^{G}\theta^{(g)}$$

avec lesquelles on évalue les distributions $f(\boldsymbol{y}|\bar{\theta},\mathscr{M}=k)$, $\pi(\bar{\theta}|\mathscr{M}=k)$ et $\hat{p}(\bar{\theta}|\boldsymbol{y},\mathscr{M}=k)$ pour calculer (2.8).

2.4 Power posteriors et Widely Applicable BIC (WBIC)

2.4.1 Power posteriors

Inspirés par la méthode du *thermodynamic integration scheme* [PvT07, Nea93, GAH17], Friel et Pettitt ont présenté l'algorithme des power posteriors en 2008 [FP08, FW12, FHW12]. Il s'agit d'un algorithme de calcul d'évidence qui exploite la possibilité d'écrire la log évidence sous forme d'espérance. Le principe des power posteriors est de tempérer la vraisemblance $f(\boldsymbol{y}|\theta,\mathscr{M}=k)$ par un paramètre de température λ appartenant à $[0,1]$:

$$p_\lambda=p_\lambda(\theta|\boldsymbol{y},\mathscr{M}=k)\;=\;\frac{f(\boldsymbol{y}|\theta,\mathscr{M}=k)^\lambda\,\pi(\theta|\mathscr{M}=k)}{\varepsilon_\lambda(\boldsymbol{y})} \tag{2.12}$$

$$\text{où} \quad \varepsilon_\lambda(\boldsymbol{y}) = \int_\Theta f(\boldsymbol{y}|\theta, \mathscr{M} = k)^\lambda \, \pi(\theta|\mathscr{M} = k) \, \mathrm{d}\theta \,. \tag{2.13}$$

Dans l'identité (2.13), lorsque $\lambda = 1$, on retrouve exactement l'évidence :

$$\varepsilon_1(\boldsymbol{y}) = \int_\Theta f(\boldsymbol{y}|\theta, \mathscr{M} = k) \, \pi(\theta|\mathscr{M} = k) \mathrm{d}\lambda = f(\boldsymbol{y}|\mathscr{M} = k) \,.$$

Lorsque $\lambda = 0$, il ne reste plus que le prior dans l'intégrale et on a alors $\varepsilon_0(\boldsymbol{y}) = 1$.

Pour calculer la log évidence, on utilise l'identité des power posteriors [FP08] :

$$\log f(\boldsymbol{y}|\mathscr{M} = k) = \int_0^1 \mathbb{E}_{p_\lambda}\Big[\log f(\boldsymbol{y}|\theta, \mathscr{M} = k)\Big] \mathrm{d}\lambda = \log\left[\frac{\varepsilon_1(\boldsymbol{y})}{\varepsilon_0(\boldsymbol{y})}\right] \tag{2.14}$$

dont la démonstration se trouve en annexe E.1.

En pratique, on calcule l'espérance pour différentes valeurs de $\lambda \in [0,1]$: $(\lambda_m)_{m=1..M}$. On calcule ensuite l'intégrale sur λ par une somme de ces espérances (méthodes des trapèzes) :

$$\begin{aligned} &\log f(\boldsymbol{y}|\mathscr{M} = k) \\ &= \sum_{m=2}^{M} \frac{\lambda_m - \lambda_{m-1}}{2} \Big(\mathbb{E}_{p_{\lambda_m}}\Big[\log f(\boldsymbol{y}|\theta, \mathscr{M} = k)\Big] + \mathbb{E}_{p_{\lambda_{m-1}}}\Big[\log f(\boldsymbol{y}|\theta, \mathscr{M} = k)\Big]\Big). \end{aligned} \tag{2.15}$$

Ces espérances, lorsqu'elles ne sont pas disponibles analytiquement, peuvent être approchées par une moyenne empirique de G échantillons tirés sous leur loi $p_{\lambda_m}(\theta|\boldsymbol{y}, \mathscr{M} = k)$, fournis par une chaîne MCMC :

$$\mathbb{E}_{p_{\lambda_m}}\Big[\log f(\boldsymbol{y}|\theta, \mathscr{M} = k)\Big] \approx \frac{1}{G} \sum_{g=1}^{G} \log f(\boldsymbol{y}|\theta_m^{(g)}, \mathscr{M} = k)$$

où $\theta_m^{(g)} \sim p_{\lambda_m}(\theta|\boldsymbol{y}, \mathscr{M} = k)$.

En prenant une grille de λ fine et un nombre de tirages G très grand, le calcul devient vite très gourmand en ressources. Il est justement intéressant d'étudier les différentes configurations. Par exemple, on peut construire une grille très fine pour λ mais un nombre de tirages G très faible. À l'inverse, on discrétise le domaine de λ avec très peu de points mais on tire un grand nombre de échantillons. Il faut trouver un compromis entre la résolution de la grille de λ et le nombre de tirages à effectuer.

Pour éviter ce compromis, Watanabe a fait évolué l'approche des power posteriors. Il s'est débarrassé de la question de la grille sur λ en prouvant qu'il existait une valeur unique $\lambda^\star$ pour laquelle on obtient la log évidence. On se retrouve cependant très vite confronté aux limitations d'une telle simplification.

2.4.2 WBIC

Le *WBIC* est un algorithme récent introduit par Watanabe en 2013 [Wat13, FMCP17]. Dans l'optique d'alléger le coût calculatoire, l'élément clé du WBIC, qui découle du théorème des accroissements finis, est qu'il existe une unique valeur $\lambda^\star$ telle que :

$$\log f(\boldsymbol{y}|\mathscr{M} = k) = \mathbb{E}_{p_{\lambda^\star}}\Big[\log f(\boldsymbol{y}|\theta,\mathscr{M} = k)\Big] \tag{2.16}$$

Ainsi, seulement une chaîne MCMC est nécessaire au lieu de M chaînes.

Malgré cette économie prometteuse, trouver cette valeur $\lambda^\star$ est une étape très compliquée. Même si on sait la caractériser théoriquement, il est très difficile d'en avoir une expression explicite. La valeur $\lambda^\star$ est définie telle que $p_{\lambda^\star}(\boldsymbol{y}|\theta,\mathscr{M} = k)$ soit équidistant du prior $\pi(\theta|\mathscr{M} = k)$ et du posterior $p(\theta|\boldsymbol{y},\mathscr{M} = k)$ au sens de Kullback-Leibler (voir démonstration en annexe E.2)

$$d_{KL}\big[\,p_{\lambda^\star}(\boldsymbol{y}|\theta,\mathscr{M} = k)\,,\,p(\theta|\boldsymbol{y},\mathscr{M} = k)\,\big] = d_{KL}\big[\,p_{\lambda^\star}(\boldsymbol{y}|\theta,\mathscr{M} = k)\,,\,\pi(\theta|\mathscr{M} = k)\,\big].$$

Watanabe a cependant fourni une valeur asymptotique de $\lambda^\star$. Pour un nombre n d'échantillons indépendants et identiquement distribuées (i.i.d.), lorsque n augmente, la valeur optimale de λ serait

$$\lambda^\star \sim \frac{1}{\log(n)}\,. \tag{2.17}$$

L'analyse du WBIC [FMCP17] pointe du doigt les raisons qui en font un algorithme difficile à mettre en pratique. Notamment, on retiendra que

1. la distribution $p_{\lambda^\star}(\theta|\boldsymbol{y},\mathscr{M} = k)$ n'est que rarement accessible analytiquement,

2. la valeur analytique de $\lambda^\star$ ne peut être connue et utiliser la valeur $\lambda^\star = 1/\log(n)$ introduira une erreur conséquente.

De plus, dans notre cas, on ne peut utiliser la valeur $1/\log(n)$ car les échantillons $\mathring{x}(p)$ ne sont pas identiquement distribués (voir Annexe C).

2.5 Approximation de Laplace

En 1986, Tierney et Kadane ont proposé une démarche permettant une approximation de l'évidence en utilisant la méthode de Laplace [TK86, FW12, And10, GD94]. Il s'agit d'appliquer le développement de Taylor à l'ordre 2 au logarithme de la loi jointe, ce qui permet ensuite de se ramener à une intégrale de gaussienne. Contrairement aux algorithmes précédents, qui visent exactement l'évidence analytique, cette approche fournit une approximation analytique de l'évidence. Supposons que la densité $p(\theta|\boldsymbol{y},\mathscr{M} = k)$ soit unimodale de mode *a posteriori* $\tilde{\theta}$. Certains auteurs [And10, FW12] ajoutent l'hypothèse d'une distribution piquée autour du mode, mais il n'en est rien dans l'article original [TK86]. Cette hypothèse n'est pas obligatoire pour appliquer Laplace.

On définit la fonction

$$F_k(\theta) = \log\left[f(\boldsymbol{y}|\theta, \mathscr{M} = k)\, \pi(\theta|\mathscr{M} = k)\right] \tag{2.18}$$

qui intervient dans le calcul de l'évidence de la manière suivante :

$$\begin{aligned} f(\boldsymbol{y}|\mathscr{M} = k) &= \int_\Theta f(\boldsymbol{y}|\theta, \mathscr{M} = k)\, \pi(\theta|\mathscr{M} = k)\, \mathrm{d}\theta \\ &= \int_\Theta \exp\left[F_k(\theta)\right] \mathrm{d}\theta\,. \end{aligned} \tag{2.19}$$

La valeur $\tilde{\theta}$ est également le mode de $f(\boldsymbol{y}|\theta, \mathscr{M} = k)\, \pi(\theta|\mathscr{M} = k)$ et *a fortiori*, de $F_k(\theta)$. Le gradient de $F_k(\theta)$ par rapport à θ s'annule donc en $\tilde{\theta}$. Le développement de Taylor à l'ordre 2 de (2.18) donne :

$$F_k(\theta) \approx F_k(\tilde{\theta}) - \frac{1}{2}(\theta - \tilde{\theta})^{\mathrm{t}}\, F_k''(\tilde{\theta})\, (\theta - \tilde{\theta}) \tag{2.20}$$

où $F_k''(\theta)$ est la Hessienne de $F_k(\theta)$. En appliquant cette approximation à (2.19) , on obtient :

$$\begin{aligned} f(\boldsymbol{y}|\mathscr{M} = k) &\approx \left(\int_\Theta \exp\left[-\frac{1}{2}(\theta - \tilde{\theta})^{\mathrm{t}}\, F_k''(\tilde{\theta})\, (\theta - \tilde{\theta})\right] \mathrm{d}\theta\right) \exp\left[F_k(\tilde{\theta})\right] \\ &\approx (2\pi)^{Q/2}\, \det[F_k''(\tilde{\theta})]^{-1/2}\, \exp\left[F_k(\tilde{\theta})\right] \\ &\approx (2\pi)^{Q/2}\, \det[F_k''(\tilde{\theta})]^{-1/2} f(\boldsymbol{y}|\tilde{\theta}, \mathscr{M} = k)\, \pi(\tilde{\theta}|\mathscr{M} = k) \end{aligned} \tag{2.21}$$

Il s'agit d'une méthode assez simple à exécuter car il faut seulement calculer la matrice hessienne de $F_k(\theta)$ et d'extraire le mode *a posteriori* $\tilde{\theta}$ pour approcher l'évidence. On peut également considérer une variante, où $\tilde{\theta}$ n'est pas le mode *a posteriori*, mais une valeur quelconque. Il faudra alors considérer le terme de gradient, qui ne s'annule alors plus en $\tilde{\theta}$, dans le calcul de l'approximation. Cela compliquerait significativement la démarche.

2.5.1 Bayesian Information Criterion (BIC)

Présenté par Schwarz en 1978, il se construit à partir d'une étude asymptotique de l'approximation de Laplace vue plus tôt [And10, LMH04, DTY18, KK96]. Le BIC repose sur un nombre n de réalisations pour pouvoir fournir son résultat asymptotique. On ne considérera cette approche que très peu dans nos travaux car nous travaillons sur une seule réalisation. Une piste pour pouvoir utiliser le BIC pourrait être de considérer une image dont la taille tend vers l'infini.

Toujours sous l'hypothèse que $f(\boldsymbol{y}|\theta, \mathscr{M} = k)$ est unimodale en $\tilde{\theta}$, on a d'après (2.21) et en multipliant $F_k(\theta)$ par n^{-1} :

$$f(\boldsymbol{y}|\mathscr{M} = k) \approx (2\pi)^{Q/2}\, n^{-Q/2}\, \det[F_k''(\tilde{\theta})]^{-1/2} f(\boldsymbol{y}|\tilde{\theta}, \mathscr{M} = k)\, \pi(\tilde{\theta}|\mathscr{M} = k)$$

et on en déduit

$$\log f(\boldsymbol{y}|\mathscr{M} = k) \approx \frac{Q}{2}\log(2\pi) - \frac{Q}{2}\log(n) - \frac{1}{2}\log\det[F_k''(\tilde{\theta})] + \log f(\boldsymbol{y}|\tilde{\theta}, \mathscr{M} = k) + \log\pi(\tilde{\theta}|\mathscr{M} = k)\,.$$

Lorsque n tend vers l'infini, les termes constants (c'est-à-dire les termes non-fonctions de n et de $\boldsymbol{y}$) sont alors ignorés, ce qui donne :

$$-2\log[f(\boldsymbol{y}|\mathscr{M} = k)] \approx -2\log[f(\boldsymbol{y}|\tilde{\theta}, \mathscr{M} = k)] + Q\log(n).$$

Et on pose :

$$BIC(k) = -2\log[f(\boldsymbol{y}|\tilde{\theta}, \mathscr{M} = k)] + Q\log(n)\,. \tag{2.22}$$

On retrouve une structure similaire au *AIC*, où le terme de log vraisemblance évalue la qualité d'ajustement aux données du modèle, tandis que le terme $Q\log(n)$ en pénalise la complexité. Ce terme de pénalisation est plus grand que celui du *AIC*, qui vaut $2Q$. Le modèle sélectionné est celui qui minimise le critère

$$\widehat{\mathscr{M}} = \arg\min_k [\,BIC(k)\,].$$

2.6 Autres méthodes et approches

Les algorithmes présentés plus tôt visent le calcul analytique exacte de l'évidence. Nous avons inclus également la méthode de Laplace, qui propose une approximation analytique de l'évidence, et le critère asymptotique BIC qui en découle.

Il existe évidemment d'autres méthodes qui ne reposent pas sur le calcul de l'évidence. On peut pour commencer citer les célèbres critères d'informations :

- An Information Criterion (AIC) [Aka73, DTY18, KK96, And10],
- Bayesian Information Criterion (BIC) (voir section 2.5.1),
- Generalized Information Criterion (GIC) [KK96, And10],
- Deviance Information Criterion (DIC) [SBBC02, And10],
- Bayesian Predictive Information Criterion (BPIC) [And10],
- *etc.*

Certaines de ces méthodes cherchent à pénaliser la complexité du modèle, c'est-à-dire le nombre de paramètres du modèle, dans le processus de sélection. Les travaux [PKM03, MP04, Myu00, PMZ02, RP20] traitent cette question de pénalisation, et abordent les problématiques de sur-ajustement aux données observées (*overfitting*), d'ajustement aux données observées (*goodness of fit*), de capacité d'adéquation aux données futures (*generalizability*) d'un modèle, et en particulier le compromis recherché entre ces deux dernières notions. Dans notre approche, tous les hyperparamètres sont marginalisés, ce qui évacue cette question de la complexité de nos modèles.

Il faut également mentionner la méthode MCMC à sauts réversibles (*RJMCMC*), qui permet de « sauter » d'un modèle à un autre, et ce même s'ils sont de dimension différentes[Gre95, RBF10, AD99]. Dans le cas où l'on souhaite procéder à une exploration « naïve », le nombre de rejet est important, ce qui entraine une convergence très lente.

D'autres méthodes telles que l'échantillonnage préférentiel (*Importance sampling*) et le *Nested sampling* sont testées dans [RW09, PvT07, FW12].

2.7 Bilan du chapitre

Pour opérer la sélection de modèles, nous suivons une approche bayésienne et nous nous reposons sur les probabilités *a posteriori* des différents modèles candidats. L'obtention de ces probabilités repose sur le calcul d'une quantité phare : l'évidence $f(\boldsymbol{y}|\mathscr{M} = k)$. Le nombre de paramètres inconnus du problème et leurs relations particulières rendent ce calcul très difficile. Ce chapitre a permis de balayer les différents algorithmes qui fournissent une approximation numérique de l'évidence. Nous avons vu dans un premier temps la marginalisation brute et le moyennage sous prior, que l'on qualifie de « naïfs ». Ce sont des algorithmes simples et sûrs, mais qui nécessitent énormément de ressources pour la plupart « inutilisées » pour calculer l'évidence. A l'inverse, nous avons ensuite présenté des algorithmes sollicitant les ressources de manière plus intelligentes. Les algorithmes de la moyenne harmonique, des power posteriors et de Chib ciblent les zones riches en informations, notamment avec des échantillons *a posteriori*, pour calculer numériquement l'évidence. L'espace des paramètres n'est alors plus balayé « à l'aveugle » comme les méthodes naïves. Ces algorithmes sont cependant plus complexes, autant dans l'appropriation que dans la mise en œuvre. Les chapitres suivants sont dédiés à la mise en œuvre d'une partie de ces algorithmes.

CHAPITRE 3

Sélection de modèles sur observation directe

La sélection de modèle appliquée à la déconvolution d'image est un problème encore peu exploré. Il nous est apparu judicieux d'étudier dans cette thèse un cas spécifique : le cas gaussien circulant. Il s'agit d'un cas « docile » sur les aspects calculatoires, ce qui n'empêchera pas rencontrer d'importantes difficultés de nature différente : celles propres à la sélection de modèles. Notre étude est menée en deux étapes : l'observation directe puis indirecte. La première est traitée dans ce chapitre, qui prépare le chapitre suivant, plus long, dédié à l'étape plus complexe de l'observation indirecte.

Nous considérons quatre des sept modèles présentés au chapitre 1, que nous indexons de la manière suivante :

Modèle	LorentzS	Gauss	Laplace	Blanc
Indice i	1	2	3	4

TABLE 3.1 – Tableau de correspondance entre les modèles image et leurs valeurs i.

L'objectif ici est comparer ces différents modèles image i (parmi lesquels se trouve le vrai modèle i^*) à partir de l'image $\boldsymbol{x}$, observée directement. Cette dernière dépend du paramètre de niveau $\gamma_{\boldsymbol{x}}$, pour lequel on considère alors deux sous-cas.

1. $\gamma_{\boldsymbol{x}}$ est connu : l'évidence est disponible analytiquement et est égale à la densité $\pi(\boldsymbol{x}|\mathscr{M} = i)$. Sa forme permet de fournir une interprétation qualitative de la sélection de modèle, que nous exposons à la section 3.1,
2. $\gamma_{\boldsymbol{x}}$ est inconnu : l'évidence est également disponible analytiquement , ce qui va nous permettre de tester quelques méthodes de la bibliographie comme les algorithmes naïfs et la moyenne harmonique (section 3.2). Dans ce cas, la vraisemblance de l'objet attachée aux données n'est autre que le prior pour $\boldsymbol{x}$

$$f(\boldsymbol{x}|\boldsymbol{\theta}, \mathscr{M} = i) \;\rightarrow\; \pi_{\boldsymbol{x}|\gamma_{\boldsymbol{x}}, \mathscr{M} = i}(\boldsymbol{x}|\gamma_{\boldsymbol{x}}, \mathscr{M} = i)\,.$$

3.1 Niveau γ_x connu

3.1.1 Évidence analytique

Sans paramètres inconnus, l'évidence n'est autre que la densité *a priori* $\pi(\boldsymbol{x}|\mathscr{M} = i)$. La probabilité du modèle i donnée par (1.14) page 12 devient dans ce cas

$$p(\mathscr{M} = i|\boldsymbol{x}) = \frac{\pi(\boldsymbol{x}|\mathscr{M} = i)\, p(\mathscr{M} = i)}{\pi(\boldsymbol{x})} \tag{3.1}$$

où le dénominateur est la loi jointe $f(\boldsymbol{x}, \mathscr{M} = i)$ marginalisée par rapport à la variable de modèle

$$\pi(\boldsymbol{x}) = \sum_{i'=1}^{I} \pi(\boldsymbol{x}|\mathscr{M} = i')\, p(\mathscr{M} = i')\,.$$

Par conséquent, seul le numérateur est fonction du modèle testé $\mathscr{M} = i$, au travers de la variable de contenu fréquentiel $\mathring{s}_{\boldsymbol{x},i}(p)$. Nous avons vu dans le chapitre 1 que son écriture se simplifie dans le domaine fréquentiel (équation (1.11)). Par simplicité, nous travaillons avec les logarithmes :

$$\log \pi(\boldsymbol{x}|\mathscr{M} = i) = -\frac{P}{2}\log(2\pi) + \frac{P}{2}\log(\gamma_{\boldsymbol{x}}^*) - \frac{1}{2}\sum_{p=1}^{P}\left[\log\left[\mathring{s}_{\boldsymbol{x},i}(p)\right] + \gamma_{\boldsymbol{x}}^* \frac{|\mathring{x}(p)|^2}{\mathring{s}_{\boldsymbol{x},i}(p)}\right], \tag{3.2}$$

où $\gamma_{\boldsymbol{x}}^*$ est le vrai niveau objet.

Les deux premiers termes étant constants qu'importe le modèle testé, la sélection de modèle se joue au niveau de la somme sur les fréquences p. Nous détaillons ce point dans la section suivante.

3.1.2 Une interprétation de la sélection de modèle

Pour mieux comprendre comment la sélection de modèle opère ici, nous avons isolé la variable de DSP ${\gamma_{\boldsymbol{x}}^*}^{-1}\mathring{s}_{\boldsymbol{x},i}(p)$ dans l'identité (3.2)

$$\log \pi(\boldsymbol{x}|\mathscr{M} = i) \# -\frac{1}{2}\sum_{p=1}^{P}\left[\log\left[\,{\gamma_{\boldsymbol{x}}^*}^{-1}\mathring{s}_{\boldsymbol{x},i}(p)\,\right] + |\mathring{x}(p)|^2\left[\,{\gamma_{\boldsymbol{x}}^*}^{-1}\mathring{s}_{\boldsymbol{x},i}(p)\,\right]^{-1}\right]. \tag{3.3}$$

Parmi les modèles candidats, celui qui maximise cette équation maximise la loi $\pi(\boldsymbol{x}, \mathscr{M} = i)$ et donc, *a fortiori*, la probabilité *a posteriori* (3.1).

Dans cette équation apparaît une fonction ϕ_p

$$\phi_p(u) = \log(u) + \eta_p\, u^{-1}\,. \tag{3.4}$$

où $\eta_p = |\mathring{x}(p)|^2$. Cette fonction évalue une pseudo-distance entre la DSP testée et le périodogramme de l'image, aux différentes fréquences p. Pour rappel, ce périodogramme est un estimateur empirique de la DSP du modèle vrai i^* : ${\gamma_{\boldsymbol{x}}^*}^{-1}\mathring{s}_{\boldsymbol{x},i^*}(p)$.

Par conséquent, le modèle dont les valeurs de DSP ${\gamma_{\boldsymbol{x}}^*}^{-1}\,\mathring{s}_{\boldsymbol{x},i}(p)$ sont les plus proches de celles du périodogramme $|\mathring{x}(p)|^2$ minimise la fonction ϕ_p. Autrement dit, le modèle de DSP candidat i le plus proche du modèle vrai i^* aux différentes fréquences p minimise ϕ_p.

3.2 Niveau γ_x inconnu

Pour le sous-cas où le niveau objet n'est pas connu, les notations introduites à la section 1.2 s'écrivent $\theta = \gamma_{\boldsymbol{x}}$ et $Q = 1$. On attribue à $\gamma_{\boldsymbol{x}}$, que l'on choisit indépendant de la variable de modèle $\mathscr{M}$, une loi *a priori*. On la choisit également peu informative : une loi Gamma, de paramètres $(\alpha_{\boldsymbol{x}}, \beta_{\boldsymbol{x}})$ très petits

$$\pi_{\gamma_{\boldsymbol{x}}}(\gamma_{\boldsymbol{x}}) = \Gamma(\gamma_{\boldsymbol{x}}\,;\,\alpha_{\boldsymbol{x}}, \beta_{\boldsymbol{x}}), \tag{3.5}$$

et qui permet de manipuler des formes conjuguées.

Pour alléger les écritures, nous travaillerons avec les notations suivantes

$$\pi(\boldsymbol{x}|\gamma_{\boldsymbol{x}}, \mathscr{M} = i) \quad \text{et} \quad \pi(\gamma_{\boldsymbol{x}}),$$

tout en gardant à l'esprit que ces lois π ne sont pas les mêmes.

3.2.1 Évidence analytique

Dans le cas présent, l'évidence se calcule en marginalisant la loi jointe de $\boldsymbol{x}$ et $\gamma_{\boldsymbol{x}}$ par rapport à $\gamma_{\boldsymbol{x}}$

$$\begin{aligned} f(\boldsymbol{x}|\mathscr{M} = i) &= \int_{\gamma_{\boldsymbol{x}}} f(\boldsymbol{x}, \gamma_{\boldsymbol{x}}|\mathscr{M} = i)\ \mathrm{d}\gamma_{\boldsymbol{x}} \\ &= \int_{\gamma_{\boldsymbol{x}}} \pi(\boldsymbol{x}|\gamma_{\boldsymbol{x}}, \mathscr{M} = i)\,\pi(\gamma_{\boldsymbol{x}})\ \mathrm{d}\gamma_{\boldsymbol{x}} \end{aligned} \tag{3.6}$$

L'intégrande ci-dessus s'explicite

$$(2\pi)^{-P/2}\,\left[\det \mathring{\boldsymbol{S}}_{\boldsymbol{x}}\right]^{-1/2}\,\frac{\beta_{\boldsymbol{x}}^{\alpha_{\boldsymbol{x}}}}{\Gamma(\alpha_{\boldsymbol{x}})}\,\gamma_{\boldsymbol{x}}^{P/2+\alpha_{\boldsymbol{x}}-1}\,\exp\left[-\gamma_{\boldsymbol{x}}\left(\beta_{\boldsymbol{x}} + \frac{1}{2}\,\mathring{\boldsymbol{x}}^{\dagger}\,{\mathring{\boldsymbol{S}}_{x,i}}^{-1}\,\mathring{\boldsymbol{x}}\right)\right]. \tag{3.7}$$

et en posant

$$A_{\boldsymbol{x}} = \frac{P}{2} + \alpha_{\boldsymbol{x}} \quad \text{et} \quad B_{\boldsymbol{x},i} = \beta_{\boldsymbol{x}} + \frac{1}{2}\,\mathring{\boldsymbol{x}}^{\dagger}\,{\mathring{\boldsymbol{S}}_{x,i}}^{-1}\,\mathring{\boldsymbol{x}}\,, \tag{3.8}$$

on reconnaît dans l'intégrale une loi gamma non normalisée de paramètre $(A_{\boldsymbol{x}}, B_{\boldsymbol{x},i})$. L'évidence vaut donc

$$f(\boldsymbol{x}|\mathscr{M} = i) \;=\; (2\pi)^{-P/2}\,\beta_{\boldsymbol{x}}^{\alpha_{\boldsymbol{x}}}\,\Gamma(\alpha_{\boldsymbol{x}})^{-1}\,\Gamma(A_{\boldsymbol{x}})\,\left[\det \mathring{\boldsymbol{S}}_{\boldsymbol{x},i}\right]^{-1/2} {B_{\boldsymbol{x},i}}^{-A_{\boldsymbol{x}}}\,. \tag{3.9}$$

Nous avons donc une expression analytique de l'évidence. Nous profitons de cette occasion pour tester quelques premières méthodes de calcul d'évidence et comparer avec cette évidence de référence. Les prochaines sections sont consacrées aux deux algorithmes naïfs (la marginalisation brute et le tirage sous prior) et à la moyenne harmonique. Les deux premiers algorithmes fourniront des résultats convaincants, contrairement au dernier, dont on prouvera analytiquement la divergence dans notre cas.

3.2.2 Algorithmes naïfs

3.2.2.1 Marginalisation brute

On a choisi comme premier algorithme à tester celui décrit en section 2.1.1, où l'on marginalise numériquement à l'aide de la méthode des rectangles. On se donne une grille de G points pour $\gamma_{\boldsymbol{x}}$ sur un intervalle $[\gamma_{\boldsymbol{x}}^{\mathrm{m}}, \gamma_{\boldsymbol{x}}^{\mathrm{M}}]$

$$\left(\gamma_{\boldsymbol{x}}^{\mathrm{m}} = \gamma_{\boldsymbol{x}}^{[1]}, \ldots, \gamma_{\boldsymbol{x}}^{[g]}, \ldots, \gamma_{\boldsymbol{x}}^{[G]} = \gamma_{\boldsymbol{x}}^{\mathrm{M}}\right) \tag{3.10}$$

avec le pas d'intégration associé

$$\Delta_{\gamma_{\boldsymbol{x}}} = \frac{\gamma_{\boldsymbol{x}}^{\mathrm{M}} - \gamma_{\boldsymbol{x}}^{\mathrm{m}}}{G}\,. \tag{3.11}$$

On applique ensuite la formule (2.3) à l'intégrale (3.6)

$$f(\boldsymbol{x}|\mathscr{M} = i)$$

$$= \Delta_{\gamma_{\boldsymbol{x}}} \sum_{g=1}^{G} \pi(\boldsymbol{x}|\gamma_{\boldsymbol{x}}^{[g]}, \mathscr{M} = i)\, \pi(\gamma_{\boldsymbol{x}}^{[g]}) \tag{3.12}$$

$$= (2\pi)^{-P/2} \left[\det \mathring{\boldsymbol{S}}_{\boldsymbol{x}}\right]^{-1/2} \beta_{\boldsymbol{x}}^{\alpha_{\boldsymbol{x}}}\, \Gamma(\alpha_{\boldsymbol{x}})^{-1}\, \Delta_{\gamma_{\boldsymbol{x}}} \sum_{g=1}^{G} \left(\gamma_{\boldsymbol{x}}^{[g]}\right)^{A_{\boldsymbol{x}}-1} \exp\left[-B_{\boldsymbol{x},i}\, \gamma_{\boldsymbol{x}}^{[g]}\right]. \tag{3.13}$$

et on obtient ainsi la valeur de l'évidence. Connaître l'évidence analytique (3.9) nous permettra de fixer une résolution et une taille de domaine au delà desquelles on peut considérer que la somme à converger et que la valeur numérique obtenue de l'évidence est fiable.

3.2.2.2 Moyennage sous prior

Tout comme en (2.4), l'intégrale (3.6) permet d'écrire l'évidence sous forme d'espérance. Cette dernière peut être approchée par une moyenne empirique de la vraisemblance $\pi(\boldsymbol{x}|\gamma_{\boldsymbol{x}}, \mathscr{M} = i)$ évaluée pour des tirages de $\gamma_{\boldsymbol{x}}$ sous son prior $\pi(\gamma_{\boldsymbol{x}})$. En notant G le nombre de tirages :

$$f(\boldsymbol{x}|\mathscr{M} = i) = \frac{1}{G} \sum_{g=1}^{G} \pi(\boldsymbol{x}|\gamma_{\boldsymbol{x}}^{(g)}, \mathscr{M} = i)$$

$$= (2\pi)^{-P/2} \left[\det \overset{\circ}{\boldsymbol{S}}_{\boldsymbol{x},i}\right]^{-1/2} \frac{1}{G} \sum_{g=1}^{G} \left(\gamma_{\boldsymbol{x}}^{(g)}\right)^{P/2} \exp\left[-\frac{\gamma_{\boldsymbol{x}}^{(g)}}{2} ||\overset{\circ}{\boldsymbol{x}}||^2_{\overset{\circ}{\boldsymbol{S}}_{\boldsymbol{x},i}}\right].$$

Pour un nombre G suffisamment grand, la moyenne empirique ci-dessus tend vers l'espérance sous prior de $\pi(\boldsymbol{x}|\gamma_{\boldsymbol{x}}, \mathscr{M} = i)$, *i.e.* l'évidence, et donc vers la valeur exacte (3.9).

3.2.3 Moyenne harmonique

Dans le cas présent, nous pouvons étudier analytiquement la convergence de la moyenne harmonique. Dans cette section, les espérances sont sous la loi $p(\gamma_{\boldsymbol{x}}|\boldsymbol{x}, \mathscr{M} = i)$. Cette loi, proportionnelle à la loi jointe $f(\boldsymbol{x}, \gamma_{\boldsymbol{x}}|\mathscr{M} = i)$, est la loi gamma de paramètres données par (3.8) :

$$p(\gamma_{\boldsymbol{x}}|\boldsymbol{x}, \mathscr{M} = i) \;=\; \Gamma\big(\gamma_{\boldsymbol{x}}\,;\, A_{\boldsymbol{x}}\,,\, B_{\boldsymbol{x},i}\big)\,. \tag{3.14}$$

La moyenne harmonique s'écrit alors :

$$f(\boldsymbol{x}|\mathscr{M} = i)^{-1} = \mathbb{E}\Big[\, f(\boldsymbol{x}|\gamma_{\boldsymbol{x}}, \mathscr{M} = i)^{-1}\,\Big] \tag{3.15}$$

$$\approx \frac{1}{G} \sum_{g=1}^{G} f(\boldsymbol{x}|\gamma_{\boldsymbol{x}}, \mathscr{M} = i)^{-1} \qquad \text{où} \qquad \gamma_{\boldsymbol{x}}^{(g)} \sim p(\gamma_{\boldsymbol{x}}|\boldsymbol{x}, \mathscr{M} = i)\,. \tag{3.16}$$

Nous allons démontrer que, dans notre cas, cette moyenne empirique ne converge pas en moyenne quadratique vers l'espérance (3.15). Cette divergence est causée par la variance infinie de l'inverse vraisemblance. La variance de cette dernière se calcule

$$\mathbb{V}\Big[f(\boldsymbol{x}|\gamma_{\boldsymbol{x}}, \mathscr{M} = i)^{-1}\Big] = \mathbb{E}\Big[\Big(f(\boldsymbol{x}|\gamma_{\boldsymbol{x}}, \mathscr{M} = i)^{-1}\Big)^2\Big] - \mathbb{E}\Big[f(\boldsymbol{x}|\gamma_{\boldsymbol{x}}, \mathscr{M} = i)^{-1}\Big]^2 \tag{3.17}$$

Son deuxième terme est l'évidence élevée au carré, qui est fini. Analysons le premier terme :

$$\begin{aligned}
&\mathbb{E}\Big[f(\boldsymbol{x}|\gamma_{\boldsymbol{x}}, \mathscr{M} = i)^{-2}\Big] \\
&\quad = \int_{\mathbb{R}^+} p(\gamma_{\boldsymbol{x}}|\boldsymbol{x}, \mathscr{M} = i)\, f(\boldsymbol{x}|\gamma_{\boldsymbol{x}}, \mathscr{M} = i)^{-2}\, \mathrm{d}\gamma_{\boldsymbol{x}} \\
&\quad = (2\pi)^P \left[\det \overset{\circ}{\boldsymbol{S}}_{\boldsymbol{x},i}\right] \beta_{\boldsymbol{x}}^{\alpha_{\boldsymbol{x}}}\, \Gamma(\alpha_{\boldsymbol{x}})^{-1} \int_0^{+\infty} \gamma_{\boldsymbol{x}}^{-P/2+\alpha_{\boldsymbol{x}}-1} \exp\left[\gamma_{\boldsymbol{x}} \left(\frac{1}{2}\, \overset{\circ}{\boldsymbol{x}}^{\dagger}\, (\overset{\circ}{\boldsymbol{S}}_{x,i})^{-1}\, \overset{\circ}{\boldsymbol{x}} - \beta_{\boldsymbol{x}}\right)\right] \mathrm{d}\gamma_{\boldsymbol{x}} \\
&\quad \propto \int_0^{+\infty} \gamma_{\boldsymbol{x}}^{-U_{\boldsymbol{x}}} \exp\Big[\gamma_{\boldsymbol{x}}\, V_{\boldsymbol{x}}\Big]\, \mathrm{d}\gamma_{\boldsymbol{x}}
\end{aligned} \tag{3.18}$$

$$\text{avec} \quad \begin{aligned} U_{\boldsymbol{x}} &= P/2 - \alpha_{\boldsymbol{x}} + 1\,, \\ V_{\boldsymbol{x}} &= \frac{1}{2}\, \overset{\circ}{\boldsymbol{x}}^{\dagger}\, (\overset{\circ}{\boldsymbol{S}}_{x,i})^{-1}\, \overset{\circ}{\boldsymbol{x}} - \beta_{\boldsymbol{x}}\,. \end{aligned}$$

Avec $U_{\boldsymbol{x}} \gg 1$ et $V_{\boldsymbol{x}} \gg 1$, une étude asymptotique révèle que l'intégrande de (3.18) tend vers $+\infty$ lorsque $\gamma_{\boldsymbol{x}} \to +\infty$: ${\gamma_{\boldsymbol{x}}}^{-U_{\boldsymbol{x}}} \to 0$ et $\exp[\gamma_{\boldsymbol{x}} V_{\boldsymbol{x}}] \to +\infty$.

En conséquence, l'intégrale (3.18) diverge et la variance $\mathbb{V}\left[f(\boldsymbol{x}|\gamma_{\boldsymbol{x}}, \mathscr{M} = i)^{-1}\right]$ est infinie : la moyenne empirique (3.16) ne converge donc pas en moyenne quadratique vers l'espérance (3.15). On se retrouve ici dans une configuration qui illustre la critique autour de la moyenne harmonique.

3.3 Résultats numériques

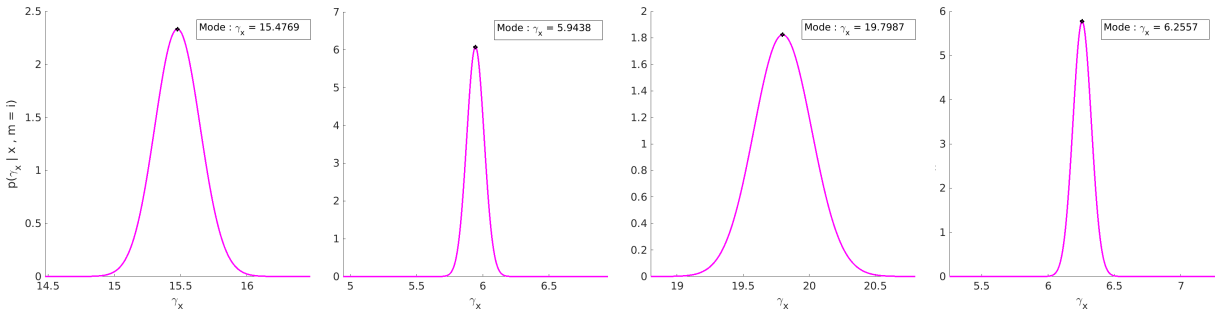
Nous allons présenter ici les différents résultats numériques pour le cas en observation directe. Pour chaque modèle i, nous avons généré un jeu de 50 images $\boldsymbol{x}$, de dimension $P = 128^2$ pixels, avec les valeurs de paramètres suivantes :

- Vrai niveau : $\gamma_{\boldsymbol{x}}^* = 6$,
- Vraie largeur : $w_{\boldsymbol{x}} = 0,1$.

Ces images ont été obtenues en suivant la procédure décrite en Annexe C.

Pour le sous-cas où le niveau $\gamma_{\boldsymbol{x}}$ est connu, nous avons calculé la log-évidence analytique (3.2) page 32 pour chaque modèle i sur toutes les images. Sans surprise, nous avons obtenus 100% de bonnes sélections.

Dans le sous-cas où $\gamma_{\boldsymbol{x}}$ est inconnu, nous avons la possibilité de calculer l'évidence analytiquement et numériquement. L'évidence s'obtient désormais en marginalisant la loi $f(\boldsymbol{x}, \gamma_{\boldsymbol{x}}|\mathscr{M} = i)$ par rapport à $\gamma_{\boldsymbol{x}}$. Nous avons tracé sur la Figure 3.1 cette loi $f(\boldsymbol{x}, \gamma_{\boldsymbol{x}}|\mathscr{M} = i)$ pour les différents modèles i et nous y avons aussi indiqué la valeur du mode *a posteriori*. L'image traitée a pour vrai modèle la DSP gaussienne $i^* = 2$.



$i = 1$ $i = 2$ $i = 3$ $i = 4$

FIGURE 3.1 – Loi jointe $f(\boldsymbol{x}, \gamma_{\boldsymbol{x}}|\mathscr{M} = i)$ en fonction du modèle objet i testé. Le vrai modèle ici est la DSP gaussienne ($i^* = 2$).

On y reconnaît les lois Gamma que nous avions explicité page 33. On constate que ces quatre distributions sont très piquées, ce qui va compliquer la tâche des algorithmes naïfs pour cibler ces zones à fortes densités.

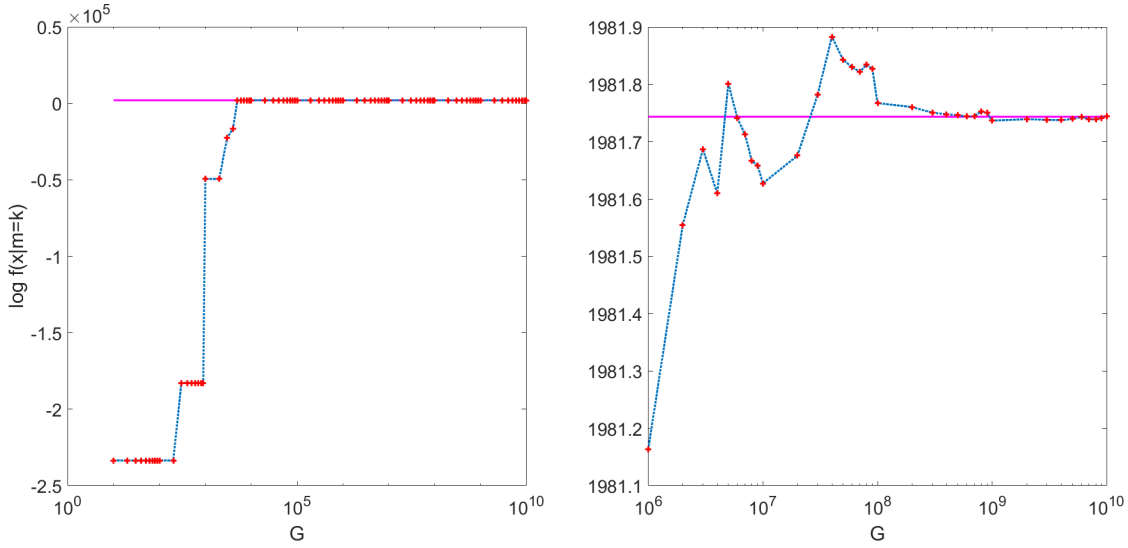
Remarque 3. *Nous invitons le lecteur à prendre connaissance de l'annexe B qui explique comment sont calculées en pratique les évidences et les probabilités sans subir d'explosions numériques.*

3.3.1 Algorithmes naïfs

Pour que la loi *a priori* de $\gamma_{\boldsymbol{x}}$ soit peu informative, les paramètres de la loi Gamma sont fixés à

- $\alpha_{\boldsymbol{x}} = 10^{-5}$,
- $\beta_{\boldsymbol{x}} = 10^{-5}$.

Pour le moyennage sous le prior, nous avons étudié la convergence de la log-évidence numérique vers la valeur exacte, présentée sur la Figure 3.2. La simplicité calculatoire de ce sous-cas a permis de traiter $G = 10^{10}$ échantillons *a priori*. La Figure (b) est un agrandissement de la Figure (a) sur l'intervalle $[10^6, 10^{10}]$ pour mieux observer le comportement de la moyenne empirique.



(a) Convergence de $\log f(\boldsymbol{x}|\mathscr{M} = k)$ en fonction de G.

(b) Agrandissement de (a) de 10^6 à 10^{10} tirages.

FIGURE 3.2 – Courbe de convergence de l'algorithme de moyennage sous prior en fonction du nombre G de tirages, représentée par la courbe bleue à croix rouges. La courbe magenta situe la valeur analytique de la log-évidence.

Dans un premier temps, on observe des paliers pour des valeurs de G inférieures à 10^6 tirages. On les doit au fait que les échantillons tirés se trouvent dans les zones à densité faible, voire nulle, de la loi jointe, et qu'aucun ne soit « tombé » suffisamment proche du pic de densité.

Ensuite, on constate que la log-évidence se stabilise à partir de 10^8 échantillons. Pour mieux quantifier la convergence, nous avons calculé l'erreur relative entre les valeurs numériques de la log-évidence et la valeur analytique en fonction du nombre de tirage G, tracée en vert sur la Figure 3.3, ainsi que l'erreur relative entre les évidences associées, tracée en pointillés bleus sur cette même Figure.

Nous y constatons que l'erreur relative d'évidence devient inférieure à 1% pour un nombre de tirages supérieur à 3.10^8. Nous souhaitons attirer l'attention sur la question de la précision de l'approximation numérique. Il faut en effet tenir compte du passage à l'exponentielle entre la log-évidence et l'évidence. On constate en effet sur la Figure 3.3 qu'une erreur de 5.10^{-2}%, que l'on pourrait qualifier de raisonnable, pour les log-évidences engendre une erreur relative supérieure à 10% pour les évidences. Pour espérer une erreur inférieure à 1%, l'erreur entre les log-évidences doit être inférieure à 5.10^{-3}%, précision que l'obtient pour sûr avec $G > 3.10^8$.

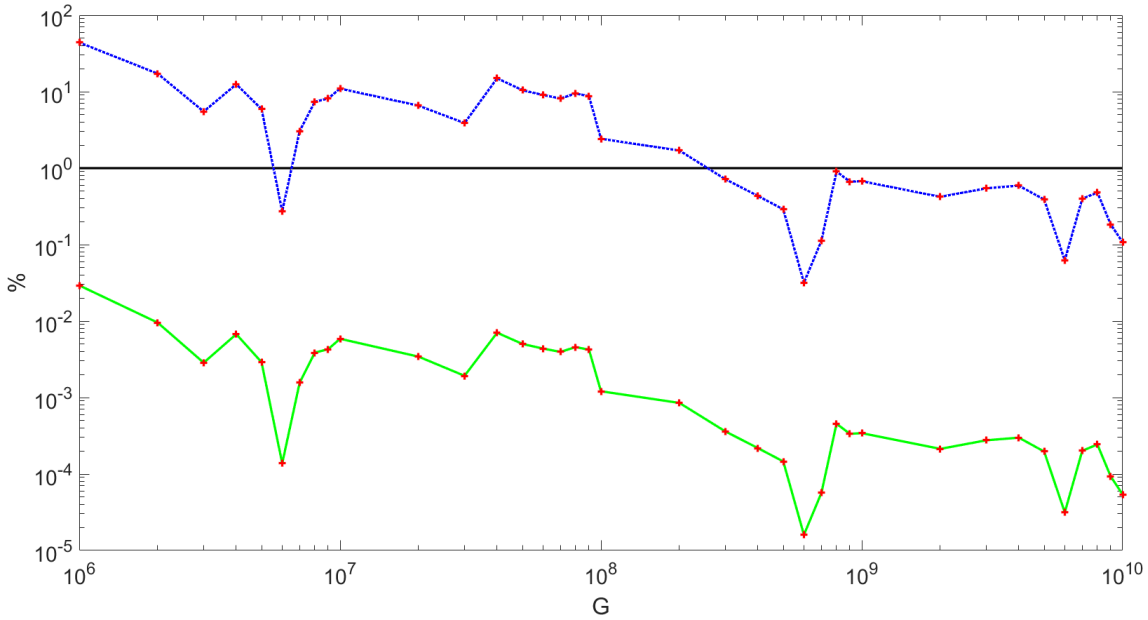


FIGURE 3.3 – Évolution de l'erreur relative pour un nombre de tirages G supérieur à 10^6. La courbe verte représente l'erreur relative entre la log-évidence analytique et les valeurs obtenues par moyennage sous prior. La courbe en pointillés bleus représente quant à elle l'erreur relative entre l'évidence analytique et les évidences obtenues avec l'algorithme.

Dans la suite de la section, nous fixerons le nombre de tirages G à 10^{10} échantillons.

Pour la marginalisation brute, nous avons quadrillé l'espace de γ_x sur l'intervalle $[10^{-6}, 30]$ avec $G = 10^6$ points.

Nous avons affichées dans la Table 3.2 les valeurs des log-évidences obtenues avec la marginalisation brute (MargiBrute) et le moyennage sous prior (MoyenPrior) pour les quatre modèles objets candidats, en regard de la valeur analytique $\log f(\boldsymbol{y}|m=k)$. Nous avons également calculé les probabilités *a posteriori* $p(\mathscr{M} = i|\boldsymbol{y})$ associées, affichées dans la dernière ligne.

Modèle i	$i=1$	$i=2$	$i=3$	$i=4$
$\log f(\boldsymbol{y}\|\mathscr{M} = i)$	9 582,0262	36 955,8816	15 565,0694	-8 710,4663
MargiBrute	9 582,02623	36 955,8816	15 565,0694	-8 710,46627
MoyenPrior	9 582,02284	36 955,8781	15 565,0661	-8 710,46974
$p(\mathscr{M} = i\|\boldsymbol{y})$	0	1	0	0

TABLE 3.2 – Valeurs de la log-évidence calculées pour les différents modèles candidats. L'image traitée a pour vrai modèle $i^* = 2$, repéré par la couleur magenta. La valeur maximale de $p(\mathscr{M} = i|\boldsymbol{y})$ est repérée par la couleur bleue.

On obtient des résultats typiques : le modèle vrai a la plus grande valeur de l'évidence, et par conséquent la plus forte probabilité *a posteriori*. Les valeurs obtenues avec les algorithmes naïfs présentent moins de $0,5\%$ d'erreur et les taux de bonnes sélections atteignent 100%.

Lors du calcul des évidences avec le moyennage sous le prior, $99,9991\%$ des échantillons sont tombés dans les zones à densité nulle. Avec la marginalisation brute, ce nombre diminue à $93,3333\%$, mais uniquement parce nous avons restreint le domaine d'intégration. Dans la pra-

tique, la marginalisation brute est plus économique car l'espace à parcourir est tronquée, grâce à une étude préalable de la distribution. Nous privilégierons de ce fait cet algorithme pour le cas en observation indirecte avec niveaux inconnus, où le coût calculatoire est plus conséquent. Il faut cependant garder à l'esprit qu'en théorie, ces deux algorithmes sollicitent la même quantité effarante de ressources, sans se démarquer quant à leur efficacité.

3.3.2 Moyenne harmonique

Nous savons d'ores et déjà que l'algorithme de la moyenne harmonique ne fonctionne pas, à cause de la variance infinie de l'inverse vraisemblance. Nous profitons cependant de la simplicité calculatoire de ce sous-cas pour illustrer la divergence de la moyenne empirique

$$-\log \frac{1}{G}\sum_{g=1}^{G} f(\boldsymbol{x}|\boldsymbol{\gamma}_{\boldsymbol{x}}^{(g)}, \mathscr{M} = i)^{-1}.$$

Nous avons réussi à traiter jusqu'à $G = 10^{11}$ échantillons *a posteriori*. La Figure 3.4 présente l'évolution de la moyenne harmonique en fonction du nombre d'échantillons G.

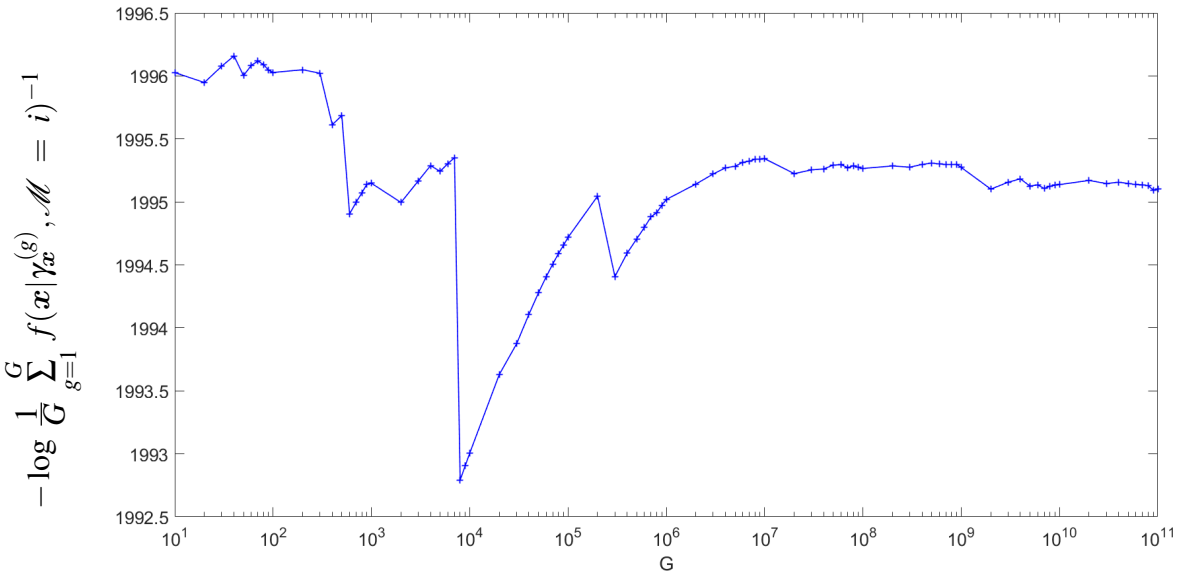


FIGURE 3.4 – Divergence de la moyenne harmonique. Même après un nombre important de tirages, la moyenne empirique ne converge pas. La valeur exacte de la log évidence est 1 970,266.

On constate bien que même avec un très grand nombre de tirages, la moyenne empirique ne converge pas vers la valeur analytique de l'évidence, qui vaut ici 1 970,266.

3.4 Bilan du chapitre

Que le niveau soit connu ou non, il est possible d'obtenir une expression analytique de l'évidence. Dans le second cas, elle a pu servir de référence pour tester quelques premiers algorithmes. Pour les deux algorithmes naïfs, nous avons pu illustrer les assertions énoncés dans le chapitre 2.

- Ce sont des algorithmes simples et sûrs : nous avons obtenu des valeurs numériques très proches de l'évidence exacte (3.9).
- Ils consomment énormément de ressources. Pour la marginalisation brute, il faut définir pour γ_x une grille suffisamment grande et fine, pour approcher l'évidence. Pour le moyennage sous prior, il faut un nombre de tirages suffisamment grand pour que la moyenne empirique converge vers l'espérance.
- Ces ressources sont mal utilisées. En effet, ces deux algorithmes parcourent l'espace du paramètre γ_x à l'aveugle, et une grande partie des points utilisés sont à densité nulle, à cause de la forme très piquée des lois à marginaliser.

Quant au troisième algorithme, nous avons démontré la moyenne harmonique ne converge pas vers l'évidence, à cause de la variance infinie de l'inverse vraisemblance. La simplicité (relative) de ce cas nous a permis de calculer avec un grand nombre d'échantillons et constater la divergence. La moyenne harmonique fait partie des algorithmes qui parcourt l'espace des paramètres plus intelligemment, mais à cause de la divergence de la moyenne empirique, on ne peut pas calculer numériquement l'évidence.

De manière générale, cette consommation de ressources va devenir d'autant plus importante en observation indirecte, à cause de la convolution et du nombre plus important de paramètres. Dans le chapitre suivant, nous re-testerons seulement la marginalisation brute. Nous démontrerons aussi à nouveau que la moyenne harmonique ne converge pas vers l'évidence. En revanche, nous allons mettre en œuvre un nouvel algorithme, celui de Chib couplé à un échantillonneur de Gibbs, et constater ses performances.

CHAPITRE 4

Sélection de modèles sur observation indirecte

Ce chapitre concerne le cas de l'observation indirecte. L' observation $\boldsymbol{y}$ sont une version altérée de l'image, à partir desquelles nous voulons sélectionner et le modèle de l'image et le modèle du bruit. On se donne ici $I = 4$ modèles de DSP objet et $J = 4$ modèles de DSP bruit, offrant ainsi un catalogue de $K = IJ = 16$ modèles pour les données observées $\boldsymbol{y}$. Chaque combinaison (i, j) est désignée par une valeur unique $k = 1, .., K$ répertoriée dans la Table 4.1.

Objet	**Bruit**	**Indice** k	**Objet**	**Bruit**	**Indice** k
LorentzS	LorentzS	1	Laplace	LorentzS	9
LorentzS	Gauss	2	Laplace	Gauss	10
LorentzS	Laplace	3	Laplace	Laplace	11
LorentzS	Blanc	4	Laplace	Blanc	12
Gauss	LorentzS	5	Blanc	LorentzS	13
Gauss	Gauss	6	Blanc	Gauss	14
Gauss	Laplace	7	Blanc	Laplace	15
Gauss	Blanc	8	Blanc	Blanc	16

TABLE 4.1 – Tableau de correspondance entre les différentes combinaisons de modèles et leurs valeurs k.

Par soucis de clarté, nous travaillons ici avec un sous-ensemble de quatre modèles. Le but premier est d'illustrer les performances de la méthode, qui ne changent pas selon la liste considérée. Libre à l'utilisateur de construire sa propre liste.

Dans cette partie, nous nous concentrons sur les paramètres de niveaux $\gamma_{\boldsymbol{x}}$ et $\gamma_{\boldsymbol{e}}$, que nous traiterons dans les configurations suivantes

1. $\gamma_{\boldsymbol{x}}$ et $\gamma_{\boldsymbol{e}}$ sont connus,
2. $\gamma_{\boldsymbol{x}}$ et $\gamma_{\boldsymbol{e}}$ sont inconnus et on aura alors $\theta = [\gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}}]$ et $Q = 2$,

pour mettre en avant l'augmentation de complexité de la première à la deuxième configuration.

Dans le premier cas, l'évidence analytique est disponible en marginalisant la loi jointe par

rapport à $\boldsymbol{x}$

$$f(\boldsymbol{y}|\mathscr{M}=k) = \int_{\mathbb{R}^P} f(\boldsymbol{y},\boldsymbol{x}\,|\,\mathscr{M}=k)\,\mathrm{d}\boldsymbol{x}\;. \tag{4.1}$$

Plus précisément, pour marginaliser en $\boldsymbol{x}$, il faut intégrer par rapport à tous les pixels de l'image, soit $P = N^2$ paramètres. Le caractère gaussien permet une marginalisation analytique, allégeant ainsi du calcul numérique d'une intégrale en très grande dimension. La circularité facilite la manipulation des matrices de covariances, notamment le calcul des inverses et des déterminants.

Dans le second cas, il faut en plus marginaliser par rapport à $\boldsymbol{\theta} = [\gamma_{\boldsymbol{x}}, \gamma_e]$

$$f(\boldsymbol{y}|\mathscr{M}=k) = \int_{\Theta}\int_{\mathbb{R}^P} f(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\theta}\,|\,\mathscr{M}=k)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{\theta}\;. \tag{1.16}$$

Or une fois l'image intégrée, la marginalisation analytique en $\gamma_{\boldsymbol{x}}$ et / ou en γ_e n'est plus possible, à cause de la forme non-usuelle de l'intégrande. On doit cette difficulté aux dépendances compliquées entre variables qui rendent le problème plus délicat à gérer. On utilise alors des algorithmes de calcul numérique, dont on évaluera la force et la pertinence dans cette étude.

Ces configurations vont permettre

- d'une part de se concentrer sur la sélection de modèle,
- d'autre part de mettre en évidence les difficultés propres au problème et de les analyser.

On précise maintenant les notations des différentes lois introduites au chapitre 1, en commençant par la vraisemblance et le prior sur l'image. On réécrit les lois (1.4) et (1.6) ainsi

$$\begin{aligned} \pi_{\boldsymbol{x}|\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{\theta}) &\;\rightarrow\; \pi_{\boldsymbol{x}|\gamma_{\boldsymbol{x}},\gamma_e,\mathscr{M}=k}(\boldsymbol{x}|\gamma_{\boldsymbol{x}},\gamma_e,\mathscr{M}=k) \\ f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) &\;\rightarrow\; f(\boldsymbol{y}|\boldsymbol{x},\gamma_{\boldsymbol{x}},\gamma_e,\mathscr{M}=k)\,. \end{aligned}$$

Il est à noter, au regard de la hiérarchie des variables Figure 1.8 page 13, que les données $\boldsymbol{y}$ et le niveaux $\gamma_{\boldsymbol{x}}$ sont indépendants conditionnellement à $\boldsymbol{x}$ et que l'image $\boldsymbol{x}$ et le niveau γ_e sont indépendants entre eux conditionnellement. De ce fait, la vraisemblance n'est pas fonction de $\gamma_{\boldsymbol{x}}$ et le prior sur $\boldsymbol{x}$ n'est pas fonction de γ_e :

$$\begin{aligned} \pi_{\boldsymbol{x}|\gamma_{\boldsymbol{x}},\gamma_e,\mathscr{M}=k}(\boldsymbol{x}|\gamma_{\boldsymbol{x}},\gamma_e,\mathscr{M}=k) &\;\rightarrow\; \pi_{\boldsymbol{x}|\gamma_{\boldsymbol{x}},\mathscr{M}=k}(\boldsymbol{x}|\gamma_{\boldsymbol{x}},\mathscr{M}=k) \\ f(\boldsymbol{y}|\boldsymbol{x},\gamma_{\boldsymbol{x}},\gamma_e,\mathscr{M}=k) &\;\rightarrow\; f(\boldsymbol{y}|\boldsymbol{x},\gamma_e,\mathscr{M}=k)\,. \end{aligned}$$

Pour la deuxième configuration, nous avons besoin de lois *a priori* pour $\gamma_{\boldsymbol{x}}$ et γ_e. On construit leurs priors avec les mêmes propriétés que le prior (3.5) page 33

1. de type Gamma, pour manipuler de formes conjuguées,
2. peu informatifs, *i.e.* de paramètres respectifs $(\alpha_{\boldsymbol{x}},\beta_{\boldsymbol{x}})$ et (α_e,β_e) très petits,
3. $\gamma_{\boldsymbol{x}}$ et γ_e sont choisis *a priori* indépendants du modèle $\mathscr{M}$.

Ces deux loi se notent

$$\pi_{\gamma_{\boldsymbol{x}}}(\gamma_{\boldsymbol{x}}) = \Gamma(\gamma_{\boldsymbol{x}}\,;\,\alpha_{\boldsymbol{x}},\beta_{\boldsymbol{x}})\;, \tag{4.2}$$

$$\pi_{\gamma_e}(\gamma_e) = \Gamma(\gamma_e \,;\, \alpha_e, \beta_e)\,. \tag{4.3}$$

Pour alléger les formules, nous adopterons les écritures suivantes

$$\begin{aligned} \pi_{\boldsymbol{x}|\gamma_{\boldsymbol{x}},\mathscr{M}\,=\,k}(\boldsymbol{x}|\gamma_{\boldsymbol{x}},\mathscr{M} \;=\; k) \;&\rightarrow\; \pi(\boldsymbol{x}|\gamma_{\boldsymbol{x}},\mathscr{M} \;=\; k),\\ \pi_{\gamma_{\boldsymbol{x}}}(\gamma_{\boldsymbol{x}}) \;&\rightarrow\; \pi(\gamma_{\boldsymbol{x}}),\\ \pi_{\gamma_e}(\gamma_e) \;&\rightarrow\; \pi(\gamma_e)\,. \end{aligned}$$

La prochaine section est consacrée à la marginalisation analytique de l'image, déjà évoquée plus tôt, qui est une étape primordiale dans notre démarche de sélection de modèle.

4.1 Marginalisation de l'image $\boldsymbol{x}$

La marginalisation de l'image $\boldsymbol{x}$ est une étape nécessaire à tous les algorithmes présentés. D'après la hiérarchie des variables de la Figure 1.8, la loi jointe $f(\boldsymbol{y},\boldsymbol{x},\gamma_{\boldsymbol{x}},\gamma_e|\mathscr{M} \;=\; k)$ se décompose de la manière suivante

$$f(\boldsymbol{y},\boldsymbol{x},\gamma_{\boldsymbol{x}},\gamma_e|\mathscr{M} \;=\; k) = f(\boldsymbol{y}|\boldsymbol{x},\gamma_e,\mathscr{M} \;=\; k)\;\pi(\boldsymbol{x}|\gamma_{\boldsymbol{x}},\mathscr{M} \;=\; k)\;\pi(\gamma_{\boldsymbol{x}})\;\pi(\gamma_e)\,, \tag{4.4}$$

où les priors $\pi(\gamma_{\boldsymbol{x}})$ et $\pi(\gamma_e)$ ne sont pas fonctions du modèle $\mathscr{M}$. Lorsque l'on marginalise la loi jointe en $\boldsymbol{x}$, on obtient

$$\begin{aligned} \int_{\mathbb{R}^P} f(\boldsymbol{y},\boldsymbol{x},\gamma_{\boldsymbol{x}},\gamma_e|\mathscr{M} \;=\; k)\,\mathrm{d}\boldsymbol{x} &= f(\boldsymbol{y},\gamma_{\boldsymbol{x}},\gamma_e|\mathscr{M} \;=\; k)\\ &= f(\boldsymbol{y}|\gamma_{\boldsymbol{x}},\gamma_e\mathscr{M} \;=\; k)\;\pi(\gamma_{\boldsymbol{x}})\;\pi(\gamma_e)\,. \end{aligned}$$

Comme on a pu le voir au chapitre 1, les variables $\boldsymbol{x}$ et $\boldsymbol{y}$ suivent toutes deux des lois normales, dont les covariances sont fonctions de $\gamma_{\boldsymbol{x}}$ et γ_e

$$\pi(\boldsymbol{x}|\gamma_{\boldsymbol{x}},\mathscr{M} \;=\; k) = \mathscr{N}(\boldsymbol{x}\,;\,\boldsymbol{0},\Sigma_{\boldsymbol{x},i}) \quad \text{où} \quad \Sigma_{\boldsymbol{x},i} = \gamma_{\boldsymbol{x}}^{-1}\boldsymbol{R}_{\boldsymbol{x},i}$$

$$f(\boldsymbol{y}|\boldsymbol{x},\gamma_e,\mathscr{M} \;=\; k) = \mathscr{N}(\boldsymbol{y}\,;\,\boldsymbol{H}\boldsymbol{x},\Sigma_{e,j}) \quad \text{où} \quad \Sigma_{e,j} = \gamma_e^{-1}\boldsymbol{R}_{e,j}\,.$$

Les données $\boldsymbol{y}$ sont liées à $\boldsymbol{x}$ et $\boldsymbol{e}$ par l'équation linéaire (1.1). Par conséquent, la distribution $f(\boldsymbol{y}|\gamma_{\boldsymbol{x}},\gamma_e,\mathscr{M} \;=\; k)$ est également une loi gaussienne dont on peut déterminer la moyenne et la covariance :

$$\begin{aligned} \bullet\; \mathbb{E}[\boldsymbol{y}] \;&=\; \mathbb{E}[\boldsymbol{H}\boldsymbol{x}+\boldsymbol{e}] = \boldsymbol{H}\mathbb{E}[\boldsymbol{x}]+\mathbb{E}[\boldsymbol{e}] = 0.\\ \bullet\; \boldsymbol{R}_{\boldsymbol{y},k} \;&=\; \mathbb{E}\left[\boldsymbol{y}\boldsymbol{y}^\dagger\right] = \mathbb{E}\left[(\boldsymbol{H}\boldsymbol{x}+\boldsymbol{e})(\boldsymbol{H}\boldsymbol{x}+\boldsymbol{e})^\dagger\right]\\ &=\; \mathbb{E}\left[\boldsymbol{H}\boldsymbol{x}\boldsymbol{x}^\dagger\boldsymbol{H}^\dagger\right]+\mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}^\dagger\right]+\mathbb{E}\left[\boldsymbol{H}\boldsymbol{x}\boldsymbol{e}^\dagger\right]+\mathbb{E}\left[\boldsymbol{e}\boldsymbol{x}^\dagger\boldsymbol{H}^\dagger\right] \end{aligned} \tag{4.5}$$

Les deux derniers termes sont les inter-covariances des variables $\boldsymbol{x}$ et $\boldsymbol{e}$. Étant indépendantes, leur inter-covariance vaut le produit de leurs espérances. Ces dernières étant toutes deux nulles, ces deux termes sont donc nuls. On en déduit son expression finale

$$\boldsymbol{R}_{\boldsymbol{y},k} \;=\; \boldsymbol{H}\boldsymbol{R}_{\boldsymbol{x},i}\boldsymbol{H}^{\dagger} + \boldsymbol{R}_{\boldsymbol{e},j}\,.$$

Pour rappel, chaque valeur k correspond à une combinaison unique (i,j), i désignant le modèle objet et j le modèle bruit comme indiqué Table 4.1.

La covariance $\boldsymbol{R}_{\boldsymbol{y},k}$ hérite également du caractère circulant des matrices $\boldsymbol{R}_{\boldsymbol{x},i}$, $\boldsymbol{R}_{\boldsymbol{e},j}$ et $\boldsymbol{H}$. Elle se décompose de la même manière dans l'espace de Fourier et on peut expliciter la matrice des valeurs propres associée

$$\begin{aligned}\boldsymbol{R}_{\boldsymbol{y},k} &= \boldsymbol{H}\boldsymbol{R}_{\boldsymbol{x},i}\boldsymbol{H}^{\dagger} + \boldsymbol{R}_{\boldsymbol{e},j} = \gamma_{\boldsymbol{x}}^{-1}(\mathbf{F}^{\dagger}\mathring{\boldsymbol{H}}\mathbf{F})(\mathbf{F}^{\dagger}\mathring{\boldsymbol{S}}_{x,i}\mathbf{F})(\mathbf{F}^{\dagger}\mathring{\boldsymbol{H}}\mathbf{F})^{\dagger} + \gamma_{\boldsymbol{e}}^{-1}\mathbf{F}^{\dagger}\mathring{\boldsymbol{S}}_{e,j}\mathbf{F}\\ &= \boldsymbol{F}^{\dagger}(\gamma_{\boldsymbol{x}}^{-1}\mathring{\boldsymbol{H}}\mathring{\boldsymbol{S}}_{\boldsymbol{x},i}\mathring{\boldsymbol{H}}^{\dagger} + \gamma_{\boldsymbol{e}}^{-1}\mathring{\boldsymbol{S}}_{\boldsymbol{e},j})\boldsymbol{F}\end{aligned}$$

et on pose

$$\mathring{\boldsymbol{S}}_{\boldsymbol{y},k} = \gamma_{\boldsymbol{x}}^{-1}\mathring{\boldsymbol{H}}\mathring{\boldsymbol{S}}_{\boldsymbol{x},i}\mathring{\boldsymbol{H}}^{\dagger} + \gamma_{\boldsymbol{e}}^{-1}\mathring{\boldsymbol{S}}_{\boldsymbol{e},j}\,, \tag{4.6}$$

qui représente ainsi la DSP des données. On peut exprimer chacun des coefficients diagonaux $\mathring{s}_{\boldsymbol{y},k}(p)$ en fonction des cœfficients des matrices $\mathring{\boldsymbol{S}}_{\boldsymbol{x},i}$, $\mathring{\boldsymbol{S}}_{\boldsymbol{e},j}$ et $\mathring{\boldsymbol{H}}$

$$\mathring{s}_{\boldsymbol{y},k}(p) = \gamma_{\boldsymbol{x}}^{-1}\,|\mathring{h}(p)|^2\,\mathring{s}_{\boldsymbol{x},i}(p) + \gamma_{\boldsymbol{e}}^{-1}\,\mathring{s}_{\boldsymbol{e},j}(p)\,. \tag{4.7}$$

La vraisemblance marginale $f(\boldsymbol{y}|\gamma_{\boldsymbol{x}},\gamma_{\boldsymbol{e}},\mathscr{M} \;=\; k)$ s'écrit donc dans le domaine fréquentiel. Cela permet de simplifier le calcul du déterminant

$$\det \boldsymbol{R}_{y,k} = \det(\mathring{\boldsymbol{S}}_{y,k}) = \prod_{p=1}^{P}\mathring{s}_{y,k}(p) \tag{4.8}$$

et de la norme de $\boldsymbol{y}$

$$\begin{aligned}||\boldsymbol{y}||^2_{\boldsymbol{R}_{\boldsymbol{y},k}} = \boldsymbol{y}^{\dagger}{\boldsymbol{R}_{\boldsymbol{y},k}}^{-1}\boldsymbol{y} \;&=\; (\boldsymbol{F}\boldsymbol{y})^{\dagger}{\mathring{\boldsymbol{S}}_{\boldsymbol{y},k}}^{-1}\boldsymbol{F}\boldsymbol{y}\\ = \mathring{\boldsymbol{y}}^{\dagger}{\mathring{\boldsymbol{S}}_{\boldsymbol{y},k}}^{-1}\mathring{\boldsymbol{y}} \;&=\; \sum_{p=1}^{P}\frac{|\mathring{y}(p)|^2}{\mathring{s}_{\boldsymbol{y},k}(p)}\,.\end{aligned} \tag{4.9}$$

La vraisemblance s'écrit alors, en considérant (4.8) et (4.9)

$$\begin{aligned}f(\boldsymbol{y}|\gamma_{\boldsymbol{x}},\gamma_{\boldsymbol{e}},\mathscr{M} \;=\; k) &= (2\pi)^{-P/2}\,\det(\boldsymbol{R}_{\boldsymbol{y},k})^{-1/2}\,\exp -\frac{1}{2}||\boldsymbol{y}||^2_{\boldsymbol{R}_{\boldsymbol{y},k}}\\ &= (2\pi)^{-P/2}\left[\prod_{p=1}^{P}\mathring{s}_{\boldsymbol{y},k}(p)\right]^{-1/2}\exp\left[-\frac{1}{2}\sum_{p=1}^{P}\frac{|\mathring{y}(p)|^2}{\mathring{s}_{\boldsymbol{y},k}(p)}\right].\end{aligned}$$

A l'image de la vraisemblance de l'objet attachée aux données (1.12) et du prior associé à l'image

(1.11), la loi $f(\boldsymbol{y}|\gamma_{\boldsymbol{x}},\gamma_{\boldsymbol{e}},\mathscr{M} = k)$ s'écrit sous forme d'une somme sur les fréquences

$$f(\boldsymbol{y}|\gamma_{\boldsymbol{x}},\gamma_{\boldsymbol{e}},\mathscr{M} = k) = (2\pi)^{-P/2} \exp -\frac{1}{2}\sum_{p=1}^{P}\left[\log \mathring{s}_{\boldsymbol{y},k}(p) + \frac{|\mathring{y}(p)|^2}{\mathring{s}_{\boldsymbol{y},k}(p)}\right]. \tag{4.10}$$

En pratique, nous travaillons avec le logarithme de cette vraisemblance :

$$\log f(\boldsymbol{y}|\gamma_{\boldsymbol{x}},\gamma_{\boldsymbol{e}},\mathscr{M} = k) = -\frac{P}{2}\log(2\pi) - \frac{1}{2}\sum_{p=1}^{P}\left[\log \mathring{s}_{\boldsymbol{y},k}(p) + \frac{|\mathring{y}(p)|^2}{\mathring{s}_{\boldsymbol{y},k}(p)}\right], \tag{4.11}$$

afin de faciliter les écritures.

Si on remplace la variable $\mathring{s}_{\boldsymbol{y},k}(p)$ par sa valeur (4.7) , le terme entre crochet devient

$$\log\left[\,\gamma_{\boldsymbol{x}}^{-1}|\mathring{h}(p)|^2\mathring{s}_{\boldsymbol{x},i}(p) + \gamma_{\boldsymbol{e}}^{-1}\mathring{s}_{\boldsymbol{e},j}(p)\,\right] + \frac{|\mathring{y}(p)|^2}{\gamma_{\boldsymbol{x}}^{-1}|\mathring{h}(p)|^2\mathring{s}_{\boldsymbol{x},i}(p) + \gamma_{\boldsymbol{e}}^{-1}\mathring{s}_{\boldsymbol{e},j}(p)},$$

on constate très vite qu'une marginalisation analytique des hyperparamètres est impossible.

Une interprétation de la sélection de modèle : suite

On retrouve dans cette équation la fonction ϕ_p vue page 32, évaluée ici pour la DSP des données $\mathring{s}_{\boldsymbol{y},k}(p)$. Le cœfficient η_p est ici la valeur du périodogramme des données $|\mathring{y}(p)|^2$ à la fréquence p. A l'instar de $|\mathring{x}(p)|^2$, $|\mathring{y}(p)|^2$ renvoie une estimation empirique de la DSP vraie $\mathring{s}_{\boldsymbol{y},k^*}(p)$ des données observées $\boldsymbol{y}$

$$|\mathring{y}(p)|^2 \approx \mathring{s}_{\boldsymbol{y},k^*}(p) = \gamma_{\boldsymbol{x}}^{*\,-1}|\mathring{h}(p)|^2\mathring{s}_{\boldsymbol{x},i^*}(p) + \gamma_{\boldsymbol{e}}^{*\,-1}\mathring{s}_{\boldsymbol{e},j^*}(p)$$

qui dépend des vrais modèles objet i^* et bruit j^* ainsi que des vrais niveaux $\gamma_{\boldsymbol{x}}^*$ et $\gamma_{\boldsymbol{e}}^*$.

Comme dans le cas en observation directe, la discrimination entre les modèles s'effectue dans l'argument de l'exponentielle : le modèle k dont les valeurs de DSP minimisent les fonctions ϕ_p maximise la probabilité *a posteriori* $p(\mathscr{M} = k|\boldsymbol{y})$.

4.2 Niveaux γ_x et γ_e connus

Dans cette section nous travaillons dans le cas où les paramètres de niveaux sont connus. L'évidence prend alors la forme de la vraisemblance marginale calculée à la section 4.1, où les hyperparamètres sont fixés à leurs vraies valeurs $\gamma_{\boldsymbol{x}}^*$ et $\gamma_{\boldsymbol{e}}^*$. Les cœfficients diagonaux de la matrice $\mathring{\boldsymbol{S}}_{\boldsymbol{y},k}$ valent alors

$$\mathring{s}_{\boldsymbol{y},k}(p) = \gamma_{\boldsymbol{x}}^{*\,-1}|\mathring{h}(p)|^2\mathring{s}_{\boldsymbol{x},i}(p) + \gamma_{\boldsymbol{e}}^{*\,-1}\mathring{s}_{\boldsymbol{e},j}(p)\,.$$

On choisit d'écrire l'évidence sous la forme (4.10) :

$$f(\boldsymbol{y}|\mathscr{M} = k) = (2\pi)^{-P/2} \exp -\frac{1}{2} \sum_{p=1}^{P} \left[\log \mathring{s}_{\boldsymbol{y},k}(p) + \frac{|\mathring{y}(p)|^2}{\mathring{s}_{\boldsymbol{y},k}(p)} \right]. \tag{4.12}$$

Il s'agit du dernier cas d'étude du manuscrit dans lequel l'évidence est disponible analytiquement. En effet, comme vu fin de la section 4.1, la forme de la loi $f(\boldsymbol{y}|\gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}}, \mathscr{M} = k)$ rend la marginalisation de la loi jointe impossible lorsque $\gamma_{\boldsymbol{x}}$ et $\gamma_{\boldsymbol{e}}$ sont considérés inconnus.

4.3 Niveaux $\gamma_{\boldsymbol{x}}$ et $\gamma_{\boldsymbol{e}}$ inconnus

4.3.1 Algorithmes naïfs

Nous allons mettre en œuvre les algorithmes dit « naïfs » présentés aux sections 2.1.1 et 2.1.2 page 20, dont nous précisons au préalable les notations.

4.3.1.1 Marginalisation brute

Comme pour l'observation directe, on commence l'étude avec l'algorithme de marginalisation brute. La démarche est identique à celle de la section 3.2.2.1 : on définit une grille de points pour chacun des hyperparamètres $\gamma_{\boldsymbol{x}}$ et $\gamma_{\boldsymbol{e}}$. On considère respectivement $G_{\boldsymbol{x}}$ points sur un intervalle $[\gamma_{\boldsymbol{x}}^{\mathrm{m}}, \gamma_{\boldsymbol{x}}^{\mathrm{M}}]$ et $G_{\boldsymbol{e}}$ points sur un intervalle $[\gamma_{\boldsymbol{e}}^{\mathrm{m}}, \gamma_{\boldsymbol{e}}^{\mathrm{M}}]$

$$\gamma_{\boldsymbol{x}}^{\mathrm{m}} = \gamma_{\boldsymbol{x}}^{[1]}, \ldots, \gamma_{\boldsymbol{x}}^{[g_{\boldsymbol{x}}]}, \ldots, \gamma_{\boldsymbol{x}}^{[G_{\boldsymbol{x}}]} = \gamma_{\boldsymbol{x}}^{\mathrm{M}} \qquad \text{et} \qquad \gamma_{\boldsymbol{e}}^{\mathrm{m}} = \gamma_{\boldsymbol{e}}^{[1]}, \ldots, \gamma_{\boldsymbol{e}}^{[g_{\boldsymbol{e}}]}, \ldots, \gamma_{\boldsymbol{e}}^{[G_{\boldsymbol{e}}]} = \gamma_{\boldsymbol{e}}^{\mathrm{M}}$$

avec les pas d'intégration associés :

$$\Delta_{\gamma_{\boldsymbol{x}}} = \frac{\gamma_{\boldsymbol{x}}^{\mathrm{M}} - \gamma_{\boldsymbol{x}}^{\mathrm{m}}}{G_{\boldsymbol{x}}} \qquad \text{et} \qquad \Delta_{\gamma_{\boldsymbol{e}}} = \frac{\gamma_{\boldsymbol{e}}^{\mathrm{M}} - \gamma_{\boldsymbol{e}}^{\mathrm{m}}}{G_{\boldsymbol{e}}}.$$

La double intégrale qui permet d'obtenir l'évidence

$$f(\boldsymbol{y}|\mathscr{M} = k) = \int_{\mathbb{R}^+} \int_{\mathbb{R}^+} f(\boldsymbol{y}|\gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}}, \mathscr{M} = k)\, \pi(\gamma_{\boldsymbol{x}})\, \pi(\gamma_{\boldsymbol{e}})\, \mathrm{d}\gamma_{\boldsymbol{x}}\, \mathrm{d}\gamma_{\boldsymbol{e}} \tag{4.13}$$

s'approche alors

$$f(\boldsymbol{y}|\mathscr{M} = k) \approx \sum_{g_{\boldsymbol{x}}=1}^{G_{\boldsymbol{x}}} \sum_{g_{\boldsymbol{e}}=1}^{G_{\boldsymbol{e}}} f(\boldsymbol{y}|\gamma_{\boldsymbol{x}}^{[g_{\boldsymbol{x}}]}, \gamma_{\boldsymbol{e}}^{[g_{\boldsymbol{e}}]}, \mathscr{M} = k)\, \pi(\gamma_{\boldsymbol{x}}^{[g_{\boldsymbol{x}}]})\, \pi(\gamma_{\boldsymbol{e}}^{[g_{\boldsymbol{e}}]})\, \Delta_{\gamma_{\boldsymbol{x}}}\, \Delta_{\gamma_{\boldsymbol{e}}}\,. \tag{4.14}$$

Pour chaque valeur de $\gamma_{\boldsymbol{x}}$ et $\gamma_{\boldsymbol{e}}$, il faut évaluer les DSP (4.7) ainsi que les priors $\pi(\gamma_{\boldsymbol{x}})$ et $\pi(\gamma_{\boldsymbol{e}})$, en déduire la log vraisemblance (4.11) et en prendre l'exponentielle. La partie suivante est dédiée à la mise en œuvre du second algorithme naïf : le moyennage sous le prior.

4.3.1.2 Moyennage sous priors

Une autre manière de calculer l'évidence est de tout simplement moyenner la vraisemblance sous le prior. Avec les nouvelles notations, les équations (2.4) et (2.5) deviennent

$$f(\boldsymbol{y}|\mathscr{M}=k) = \int_{\mathbb{R}^+}\int_{\mathbb{R}^+} f(\boldsymbol{y}|\gamma_{\boldsymbol{x}},\gamma_e,\mathscr{M}=k)\pi(\gamma_e)\,\pi(\gamma_{\boldsymbol{x}})\,\mathrm{d}\gamma_e\,\mathrm{d}\gamma_{\boldsymbol{x}} \tag{4.15}$$

On considère un nombre G de tirages, suffisamment élevé pour que les zones à forte densité soient densément explorées. L'évidence ci-dessus s'approche alors

$$f(\boldsymbol{y}|\mathscr{M}=k) \approx \frac{1}{G}\sum_{g=1}^{G} f(\boldsymbol{y}|\gamma_{\boldsymbol{x}}^{(g)},\gamma_e^{(g)},\mathscr{M}=k) \tag{4.16}$$
$$\text{avec}\quad \gamma_{\boldsymbol{x}}^{(g)}\sim\pi(\gamma_{\boldsymbol{x}})\quad\text{et}\quad \gamma_e^{(g)}\sim\pi(\gamma_e)\,.$$

4.3.2 Moyenne harmonique

Dans cette section, nous montrons à nouveau que la moyenne harmonique ne peut être appliquée, cette fois-ci en observation indirecte. Dans le même esprit que la section 3.2.3, nous démontrons que la variance *a posteriori* de l'inverse vraisemblance est infinie. Ici, les espérances sont sous la loi $p(\boldsymbol{x},\gamma_{\boldsymbol{x}},\gamma_e|\mathscr{M}=k)$. La moyenne harmonique

$$f(\boldsymbol{y}|\mathscr{M}=k)^{-1} = \mathbb{E}\Big[f(\boldsymbol{y}|\boldsymbol{x},\gamma_e,\mathscr{M}=k)^{-1}\Big] \tag{4.17}$$

et cette espérance s'approche

$$f(\boldsymbol{y}|\mathscr{M}=k)^{-1} \approx \frac{1}{G}\sum_{g=1}^{G} f(\boldsymbol{y}|\boldsymbol{x}^{(g)},\gamma_e^{(g)},\mathscr{M}=k)^{-1} \tag{4.18}$$
$$\text{où}\quad (\boldsymbol{x}^{(g)},\gamma_e^{(g)})\sim p(\boldsymbol{x},\gamma_{\boldsymbol{x}},\gamma_e|\mathscr{M}=k)\,.$$

Pour étudier la variance de l'inverse vraisemblance

$$\mathbb{V}\Big[f(\boldsymbol{y}|\boldsymbol{x},\gamma_e,\mathscr{M}=k)^{-1}\Big] = \mathbb{E}\Big[f(\boldsymbol{y}|\boldsymbol{x},\gamma_e,\mathscr{M}=k)^{-2}\Big] - \mathbb{E}\Big[f(\boldsymbol{y}|\boldsymbol{x},\gamma_e,\mathscr{M}=k)^{-1}\Big]^2, \tag{4.19}$$

on se concentre à nouveau sur le premier terme, le second étant toujours l'évidence au carré (et donc fini !) :

$$\mathbb{E}\Big[f(\boldsymbol{y}|\boldsymbol{x},\gamma_e,\mathscr{M}=k)^{-2}\Big]$$
$$= \int_{\mathbb{R}^P}\int_{\mathbb{R}^+}\int_{\mathbb{R}^+} p(\boldsymbol{x},\gamma_{\boldsymbol{x}},\gamma_e|\mathscr{M}=k)\,f(\boldsymbol{y}|\boldsymbol{x},\gamma_e,\mathscr{M}=k)^{-2}\,\mathrm{d}\gamma_e\,\mathrm{d}\gamma_{\boldsymbol{x}}\,\mathrm{d}\boldsymbol{x}\,.$$

On s'intéresse plus particulièrement à l'intégrale de γ_e :

$$\begin{aligned}&\int_{\mathbb{R}^+} f(\boldsymbol{y}|\boldsymbol{x},\gamma_e,\mathscr{M}\ =\ k)\,\pi(\gamma_e)\,f(\boldsymbol{y}|\boldsymbol{x},\gamma_e,\mathscr{M}\ =\ k)^{-2}\,\mathrm{d}\gamma_e\\ &=\int_{\mathbb{R}^+} f(\boldsymbol{y}|\boldsymbol{x},\gamma_e,\mathscr{M}\ =\ k)^{-1}\,\pi(\gamma_e)\,\mathrm{d}\gamma_e\end{aligned}$$

On remplace $f(\boldsymbol{y}|\boldsymbol{x},\gamma_e,\mathscr{M}\ =\ k)^{-1}$ et $\pi(\gamma_e)$ par leur expression respective, puis on isole les termes fonctions de γ_e :

$$\begin{aligned}&=\int_0^{+\infty}(2\pi)^{P/2}\det\mathring{\boldsymbol{S}}_{e,j}{}^{1/2}\gamma_e^{-P/2}\exp\left[\frac{\gamma_e}{2}||\mathring{\boldsymbol{y}}-\mathring{\boldsymbol{H}}\mathring{\boldsymbol{x}}||^2_{\mathring{\boldsymbol{S}}_{e,j}}\right]\beta_e^{\alpha_e}\Gamma(\alpha_e)^{-1}\gamma_e^{\alpha_e-1}\exp\left[-\beta_e\gamma_e\right]\mathrm{d}\gamma_e\\ &\propto\int_0^{+\infty}\gamma_e^{-P/2+\alpha_e-1}\exp\left[\gamma_e\left(\frac{1}{2}||\mathring{\boldsymbol{y}}-\mathring{\boldsymbol{H}}\mathring{\boldsymbol{x}}||^2_{\mathring{\boldsymbol{S}}_{e,j}}-\beta_e\right)\right]\mathrm{d}\gamma_e\\ &\propto\int_0^{+\infty}\gamma_e^{-U_e}\,\exp\left[\,\gamma_e\,V_e\right]\mathrm{d}\gamma_e \qquad (4.20)\end{aligned}$$

$$\begin{aligned}\text{avec}\quad U_e\ &=P/2-\alpha_e+1 &&\gg 1\ ,\\ V_e\ &=\tfrac{1}{2}\,(\mathring{\boldsymbol{y}}-\mathring{\boldsymbol{H}}\mathring{\boldsymbol{x}})^{\dagger}\,(\mathring{\boldsymbol{S}}_{e,j})^{-1}\,(\mathring{\boldsymbol{y}}-\mathring{\boldsymbol{H}}\mathring{\boldsymbol{x}})-\beta_e &&\gg 1\ .\end{aligned}$$

On retrouve la même forme d'intégrale qu'à l'équation (3.18) avec les mêmes conditions sur U_e et V_e. On peut donc en conclure que l'intégrale (4.20) diverge car son intégrande tend vers $+\infty$ lorsque $\gamma_e \to +\infty$, et donc que la variance (4.19) est infinie. Même conclusion qu'à la section 3.2.3 : la moyenne empirique (4.18) ne peut donc pas converger vers l'espérance analytique (4.17).

4.3.3 Algorithme de Chib couplé à un échantillonneur de Gibbs

Nous allons maintenant nous pencher sur l'approche introduite à la section 2.3, l'algorithme de Chib qui repose sur une chaîne de Gibbs. Dans le cas présent, le vecteur des hyperparamètres est toujours $\boldsymbol{\theta} = [\gamma_x, \gamma_e]$ et on manipule l'image $\boldsymbol{x}$ comme variable latente $\boldsymbol{a}$ de l'algorithme.

Remarque 4. *On pourrait choisir un des deux niveaux γ comme variable latente plutôt que $\boldsymbol{x}$, mais les calculs analytiques et numériques deviennent plus difficiles, voire impossibles (voir la discussion Annexe A.2).*

L'équation (2.8) page 23 donnant l'évidence devient

$$\begin{aligned}\log f(\boldsymbol{y}|\mathscr{M}\ =\ k) \approx \log f(\boldsymbol{y}|\bar{\gamma}_x,\bar{\gamma}_e,\mathscr{M}\ =\ k)+\log\pi(\bar{\gamma}_x)+\log\pi(\bar{\gamma}_e)\\ -\log\hat{p}(\bar{\gamma}_x,\bar{\gamma}_e|\boldsymbol{y},\mathscr{M}\ =\ k)\,, \qquad (4.21)\end{aligned}$$

où $\hat{p}(\bar{\gamma}_x,\bar{\gamma}_e|\boldsymbol{y},\mathscr{M}\ =\ k)$ est l'approximation de la loi *a posteriori* $p(\bar{\gamma}_x,\bar{\gamma}_e|\boldsymbol{y},\mathscr{M}\ =\ k)$. Elle se

calcule, d'après (2.11),

$$\hat{p}(\bar{\gamma}_{\boldsymbol{x}}, \bar{\gamma}_{\boldsymbol{e}} | \boldsymbol{y}, \mathscr{M} = k) = \frac{1}{G} \sum_{g=1}^{G} f(\bar{\gamma}_{\boldsymbol{x}}, \bar{\gamma}_{\boldsymbol{e}} | \boldsymbol{y}, \boldsymbol{x}^{(g)}, \mathscr{M} = k) \tag{4.22}$$
$$\text{où} \quad \boldsymbol{x}^{(g)} \sim p(\boldsymbol{x} | \boldsymbol{y}, \mathscr{M} = k).$$

Pour pouvoir calculer cette somme empirique, nous avons besoin de deux choses.

1. Connaître exactement la loi $f(\gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}} | \boldsymbol{y}, \boldsymbol{x}, \mathscr{M} = k)$,
2. Savoir échantillonner la loi $p(\boldsymbol{x} | \boldsymbol{y}, \mathscr{M} = k)$.

Ces deux points sont décrits respectivement aux sections 4.3.3.1 et 4.3.3.2.

4.3.3.1 La distribution $f(\gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}} | \boldsymbol{y}, \boldsymbol{x}, \mathscr{M} = k)$

D'après la règles de Bayes, cette distribution s'écrit

$$f(\gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}} | \boldsymbol{y}, \boldsymbol{x}, \mathscr{M} = k) = \frac{f(\gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}}, \boldsymbol{y}, \boldsymbol{x} | \mathscr{M} = k)}{f(\boldsymbol{y}, \boldsymbol{x} | \mathscr{M} = k)}. \tag{4.23}$$

D'après l'identité (4.4) page 43, la loi jointe au numérateur se décompose de la manière suivante

$$f(\gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}}, \boldsymbol{y}, \boldsymbol{x} | \mathscr{M} = k) = f(\boldsymbol{y} | \boldsymbol{x}, \gamma_{\boldsymbol{e}}, \mathscr{M} = k)\, \pi(\boldsymbol{x} | \gamma_{\boldsymbol{x}}, \mathscr{M} = k)\, \pi(\gamma_{\boldsymbol{x}})\, \pi(\gamma_{\boldsymbol{e}}).$$

Le dénominateur lui s'obtient en marginalisant le numérateur par rapport à $\gamma_{\boldsymbol{x}}$ et $\gamma_{\boldsymbol{e}}$. On factorise ce qui dépend de $\gamma_{\boldsymbol{x}}$ et ce qui dépend de $\gamma_{\boldsymbol{e}}$ en deux facteurs distincts, pour pouvoir ensuite séparer les intégrales :

$$\begin{aligned}
f(\boldsymbol{y}, \boldsymbol{x} | \mathscr{M} = k) &= \int_{\mathbb{R}^+} \int_{\mathbb{R}^+} f(\boldsymbol{y} | \boldsymbol{x}, \gamma_{\boldsymbol{e}}, \mathscr{M} = k)\, \pi(\boldsymbol{x} | \gamma_{\boldsymbol{x}}, \mathscr{M} = k)\, \pi(\gamma_{\boldsymbol{x}})\, \pi(\gamma_{\boldsymbol{e}})\, \mathrm{d}\gamma_{\boldsymbol{e}}\, \mathrm{d}\gamma_{\boldsymbol{x}} \\
&= \int_{\mathbb{R}^+} \int_{\mathbb{R}^+} f(\boldsymbol{y}, \gamma_{\boldsymbol{e}} | \boldsymbol{x}, \mathscr{M} = k)\ \pi(\boldsymbol{x}, \gamma_{\boldsymbol{x}} | \mathscr{M} = k)\, \mathrm{d}\gamma_{\boldsymbol{e}}\, \mathrm{d}\gamma_{\boldsymbol{x}} \\
&= \int_{\mathbb{R}^+} f(\boldsymbol{y}, \gamma_{\boldsymbol{e}} | \boldsymbol{x}, \mathscr{M} = k)\, \mathrm{d}\gamma_{\boldsymbol{e}} \int_{\mathbb{R}^+} \pi(\boldsymbol{x}, \gamma_{\boldsymbol{x}} | \mathscr{M} = k)\, \mathrm{d}\gamma_{\boldsymbol{x}} \\
&= f(\boldsymbol{y} | \boldsymbol{x}, \mathscr{M} = k)\, \pi(\boldsymbol{x} | \mathscr{M} = k)
\end{aligned}$$

La distribution (4.23) se réécrit

$$f(\gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}} | \boldsymbol{y}, \boldsymbol{x}, \mathscr{M} = k) = \frac{f(\boldsymbol{y} | \boldsymbol{x}, \gamma_{\boldsymbol{e}}, \mathscr{M} = k)\, \pi(\gamma_{\boldsymbol{e}})}{f(\boldsymbol{y} | \boldsymbol{x}, \mathscr{M} = k)} \times \frac{\pi(\boldsymbol{x} | \gamma_{\boldsymbol{x}}, \mathscr{M} = k)\, \pi(\gamma_{\boldsymbol{x}})}{\pi(\boldsymbol{x} | \mathscr{M} = k)}$$

et on distingue alors deux facteurs.

1. Le premier facteur

$$\frac{f(\boldsymbol{y} | \boldsymbol{x}, \gamma_{\boldsymbol{e}}, \mathscr{M} = k)\, \pi(\gamma_{\boldsymbol{e}})}{f(\boldsymbol{y} | \boldsymbol{x}, \mathscr{M} = k)} = p(\gamma_{\boldsymbol{e}} | \boldsymbol{y}, \boldsymbol{x}, \mathscr{M} = k) \tag{4.24}$$

est la loi conditionnelle *a posteriori* de γ_e.

2. Le second

$$\frac{\pi(\boldsymbol{x}|\gamma_{\boldsymbol{x}},\mathscr{M}=k)\,\pi(\gamma_{\boldsymbol{x}})}{\pi(\boldsymbol{x}|\mathscr{M}=k)}=p(\gamma_{\boldsymbol{x}}|\boldsymbol{x},\mathscr{M}=k)$$

correspond à la conditionnelle *a priori* de $\gamma_{\boldsymbol{x}}$. Nous démontrons en Annexe A.1 qu'elle est égale à la loi conditionnelle *a posteriori* de $\gamma_{\boldsymbol{x}}$, malgré le fait qu'elle ne soit pas fonction des données $\boldsymbol{y}$. Nous garderons par conséquent la terminologie « conditionnelle *a posteriori* » pour cette distribution dans la suite du manuscrit.

La loi $f(\gamma_{\boldsymbol{x}},\gamma_e|\boldsymbol{y},\boldsymbol{x},\mathscr{M}=k)$ est donc séparable en le produit des lois conditionnelles *a posteriori* de $\gamma_{\boldsymbol{x}}$ et γ_e, preuve de leur indépendance *a posteriori*.

La configuration que nous avons choisie $\big(\theta=[\gamma_{\boldsymbol{x}},\gamma_e]$ et $\boldsymbol{a}=\boldsymbol{x}\big)$ permet de calculer analytiquement les constantes de normalisation $K_{\gamma_{\boldsymbol{x}}|*}$ de la loi $p(\gamma_{\boldsymbol{x}}|\boldsymbol{x},\mathscr{M}=k)$ et $K_{\gamma_e|*}$ de la loi $p(\gamma_e|\boldsymbol{y},\boldsymbol{x},\mathscr{M}=k)$. On va ici expliciter le calcul pour la constante $K_{\gamma_e|*}$, qui sera le même pour la constante $K_{\gamma_{\boldsymbol{x}}|*}$.

$$\begin{aligned}
p(\gamma_e|\boldsymbol{y},\boldsymbol{x},\mathscr{M}=k) &\propto f(\boldsymbol{y}|\boldsymbol{x},\gamma_e,\mathscr{M}=k)\,\pi(\gamma_e) \quad \text{d'après (4.24)}\\
&\propto (2\pi)^{-P/2}\,\gamma_e^{P/2}\,\det\big[\mathring{\boldsymbol{S}}_{e,j}\big]^{-1/2}\,\exp\Big[-\frac{\gamma_e}{2}(\mathring{\boldsymbol{y}}-\mathring{\boldsymbol{H}}\mathring{\boldsymbol{x}})^\dagger\mathring{\boldsymbol{S}}_{e,j}^{-1}(\mathring{\boldsymbol{y}}-\mathring{\boldsymbol{H}}\mathring{\boldsymbol{x}})\Big]\\
&\quad \beta_e^{\alpha_e}\,\Gamma(\alpha_e)^{-1}\,\gamma_e^{\alpha_e-1}\,\exp\big[-\beta_e\gamma_e\big]\,\mathbb{1}_+(\gamma_e)\\
&\propto \gamma_e^{P/2+\alpha_e-1}\,\exp\left[-\gamma_e\Big(\beta_e+\frac{1}{2}(\mathring{\boldsymbol{y}}-\mathring{\boldsymbol{H}}\mathring{\boldsymbol{x}})^\dagger\mathring{\boldsymbol{S}}_{e,j}^{-1}(\mathring{\boldsymbol{y}}-\mathring{\boldsymbol{H}}\mathring{\boldsymbol{x}})\Big)\right]. \qquad (4.25)
\end{aligned}$$

On reconnaît ci-dessus la forme d'une distribution Gamma à un facteur près, de paramètres

$$A_e=P/2+\alpha_e \quad \text{et} \quad B_e=\beta_e+\frac{1}{2}(\mathring{\boldsymbol{y}}-\mathring{\boldsymbol{H}}\mathring{\boldsymbol{x}})^\dagger\mathring{\boldsymbol{S}}_{e,j}{}^{-1}(\mathring{\boldsymbol{y}}-\mathring{\boldsymbol{H}}\mathring{\boldsymbol{x}}) \qquad (4.26)$$

On a donc

$$\begin{aligned}
p(\gamma_e|\boldsymbol{y},\boldsymbol{x},\mathscr{M}=k) &= K_{\gamma_e|*}\,\gamma_e^{A_e-1}\,\exp\big[-\gamma_e\,B_e\big]\\
&= \Gamma(\gamma_e\,;A_e,B_e), \qquad (4.27)
\end{aligned}$$

où la constante de normalisation vaut $K_{\gamma_e|*}=B_e^{A_e}\,\Gamma(A_e)^{-1}$.

La densité $p(\gamma_{\boldsymbol{x}}|\boldsymbol{x},\mathscr{M}=k)$ est elle aussi une loi Gamma de paramètres

$$A_{\boldsymbol{x}}=P/2+\alpha_{\boldsymbol{x}} \quad \text{et} \quad B_{\boldsymbol{x}}=\beta_{\boldsymbol{x}}+\frac{1}{2}\mathring{\boldsymbol{x}}^\dagger\mathring{\boldsymbol{S}}_{\boldsymbol{x},i}{}^{-1}\mathring{\boldsymbol{x}} \qquad (4.28)$$

et de constante de normalisation $K_{\gamma_{\boldsymbol{x}}|*}=B_{\boldsymbol{x}}^{A_{\boldsymbol{x}}}\,\Gamma(A_{\boldsymbol{x}})^{-1}$.

En conclusion, la distribution $f(\gamma_{\boldsymbol{x}},\gamma_e|\boldsymbol{y},\boldsymbol{x},\mathscr{M}=k)$ est connue exactement et se calcule comme le produit de deux lois Gamma :

$$f(\gamma_{\boldsymbol{x}},\gamma_e|\boldsymbol{y},\boldsymbol{x},\mathscr{M}=k)=\Gamma(\gamma_e\,;A_e,B_e)\,\Gamma(\gamma_{\boldsymbol{x}}\,;A_{\boldsymbol{x}},B_{\boldsymbol{x}}). \qquad (4.29)$$

La possibilité de calculer exactement la distribution $f(\gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}}|\boldsymbol{y}, \boldsymbol{x}, \mathscr{M} = k)$ est une conséquence avantageuse de nos choix de lois *a priori*. Considérer des priors gaussiens pour $\boldsymbol{x}$ et $\boldsymbol{e}$ d'une part et des priors Gamma pour $\gamma_{\boldsymbol{x}}$ et $\gamma_{\boldsymbol{e}}$ d'autre part nous a permis de manipuler des formes conjuguées. Le choix de ces lois *a priori* a également un impact dans la section suivante, où nous pouvons extraire des lois faciles à échantillonner.

4.3.3.2 Tirages sous $p(\boldsymbol{x}|\boldsymbol{y}, \mathscr{M} = k)$

Dans cette sous-partie, on traite l'échantillonnage de la loi $p(\boldsymbol{x}|\boldsymbol{y}, \mathscr{M} = k)$, à l'aide d'une chaîne de Gibbs. La loi *a posteriori* $p(\boldsymbol{x}|\boldsymbol{y}, \mathscr{M} = k)$ est difficile à échantillonner à cause de sa forme non-usuelle. On la doit à la marginalisation de la loi *a posteriori* par rapport à $\gamma_{\boldsymbol{x}}$ et $\gamma_{\boldsymbol{e}}$

$$p(\boldsymbol{x}|\boldsymbol{y}, \mathscr{M} = k) = \int_{\mathbb{R}^+} \int_{\mathbb{R}^+} p(\boldsymbol{x}, \gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}}|\boldsymbol{y}, \mathscr{M} = k) \, \mathrm{d}\gamma_{\boldsymbol{e}} \, \mathrm{d}\gamma_{\boldsymbol{x}} \, .$$

Une solution est de simuler des échantillons de la loi $p(\boldsymbol{x}, \gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}}|\boldsymbol{y}, \mathscr{M} = k)$, et les échantillons $\boldsymbol{x}^{(g)}$ obtenus seront marginalement sous la densité $p(\boldsymbol{x}|\boldsymbol{y}, \mathscr{M} = k)$. On parle de *demarginalization* ou *completion construction* en anglais [RC05].

De plus, la loi $p(\boldsymbol{x}, \gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}}|\boldsymbol{y}, \mathscr{M} = k)$ est telle que ses conditionnelles *a posteriori* de $\boldsymbol{x}$, $\gamma_{\boldsymbol{x}}$ et $\gamma_{\boldsymbol{e}}$ sont faciles à échantillonner. On peut simuler le trio $(\gamma_{\boldsymbol{x}}^{(g)}, \gamma_{\boldsymbol{e}}^{(g)}, \boldsymbol{x}^{(g)})$ à l'aide d'une chaîne de Gibbs, en échantillonnant itérativement sous chacune des conditionnelles *a posteriori* :

$$\left| \begin{array}{lcl} \gamma_{\boldsymbol{x}}^{(g)} & \sim & p(\gamma_{\boldsymbol{x}}|\boldsymbol{x}^{(g-1)}, \mathscr{M} = k) \\ \gamma_{\boldsymbol{e}}^{(g)} & \sim & p(\gamma_{\boldsymbol{e}}|\boldsymbol{y}, \boldsymbol{x}^{(g-1)}, \mathscr{M} = k) \\ \boldsymbol{x}^{(g)} & \sim & p(\boldsymbol{x}|\boldsymbol{y}, \gamma_{\boldsymbol{x}}^{(g)}, \gamma_{\boldsymbol{e}}^{(g)}, \mathscr{M} = k) \end{array} \right.$$

Comme nous avons pu le voir à la section 4.3.3.1, nous connaissons les conditionnelles *a posteriori* de $\gamma_{\boldsymbol{x}}$ et $\gamma_{\boldsymbol{e}}$, qui sont des lois Gamma. Pour rappel, la loi de $\gamma_{\boldsymbol{x}}$ n'est fonction ni de $\boldsymbol{y}$ ni de $\gamma_{\boldsymbol{e}}$, et la loi de $\gamma_{\boldsymbol{e}}$ n'est pas fonction de $\gamma_{\boldsymbol{x}}$. Il ne reste plus qu'à définir la loi conditionnelle *a posteriori* pour la variable $\boldsymbol{x}$

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{y}, \gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}}, \mathscr{M} = k) &\propto f(\boldsymbol{y}|\boldsymbol{x}, \gamma_{\boldsymbol{e}}, \mathscr{M} = k) \, \pi(\boldsymbol{x}|\gamma_{\boldsymbol{x}}, \mathscr{M} = k) \\ &\propto \exp\left[-\frac{\gamma_{\boldsymbol{e}}}{2} (\mathring{\boldsymbol{y}} - \mathring{\boldsymbol{H}}\mathring{\boldsymbol{x}})^{\dagger} \mathring{\boldsymbol{S}}_{\boldsymbol{e}}^{-1} (\mathring{\boldsymbol{y}} - \mathring{\boldsymbol{H}}\mathring{\boldsymbol{x}}) \right] \exp\left[-\frac{\gamma_{\boldsymbol{x}}}{2} \mathring{\boldsymbol{x}}^{\dagger} \mathring{\boldsymbol{S}}_{\boldsymbol{x}}^{-1} \mathring{\boldsymbol{x}} \right] \\ &\propto \exp -\frac{1}{2} \mathring{\boldsymbol{x}}^{\dagger} \left[\gamma_{\boldsymbol{e}} \mathring{\boldsymbol{H}}^{\dagger} \mathring{\boldsymbol{S}}_{\boldsymbol{e}}^{-1} \mathring{\boldsymbol{H}} + \gamma_{\boldsymbol{x}} \mathring{\boldsymbol{S}}_{\boldsymbol{x}}^{-1} \right] \mathring{\boldsymbol{x}} + \mathrm{Re}\left[\left(\gamma_{\boldsymbol{e}} \mathring{\boldsymbol{H}}^{\dagger} \mathring{\boldsymbol{S}}_{\boldsymbol{e}}^{-1} \mathring{\boldsymbol{y}} \right)^{\dagger} \mathring{\boldsymbol{x}} \right] . \end{aligned} \tag{4.30}$$

On reconnaît ici la forme d'une loi normale à un facteur près. Exprimées dans le domaine de Fourier, on peut identifier la covariance

$$\mathring{\Sigma}_{\boldsymbol{x}|*} = \left[\gamma_{\boldsymbol{e}} \mathring{\boldsymbol{H}}^{\dagger} \mathring{\boldsymbol{S}}_{\boldsymbol{e}}^{-1} \mathring{\boldsymbol{H}} + \gamma_{\boldsymbol{x}} \mathring{\boldsymbol{S}}_{\boldsymbol{x}}^{-1} \right]^{-1}$$

et la moyenne

$$\mathring{\mu}_{\boldsymbol{x}|*} = \gamma_{\boldsymbol{e}} \mathring{\Sigma}_{\boldsymbol{x}|*} \mathring{\boldsymbol{H}}^{\dagger} \mathring{\boldsymbol{S}}_{\boldsymbol{e}}^{-1} \mathring{\boldsymbol{y}}$$

dans laquelle on y reconnaît le filtre de Wiener-Hunt [Wie49, OGR10a].

La loi conditionnelle *a posteriori* pour $\boldsymbol{x}$ est donc une gaussienne

$$p(\boldsymbol{x}|\boldsymbol{y},\gamma_{\boldsymbol{x}},\gamma_{\boldsymbol{e}},\mathscr{M}=k) = \mathcal{N}\left(\mathring{\boldsymbol{x}}\,;\,\mathring{\mu}_{\boldsymbol{x}|*},\mathring{\Sigma}_{\boldsymbol{x}|*}\right). \tag{4.31}$$

Comme les gaussiennes précédentes, elle s'exprime et se sépare dans le domaine fréquentiel

$$\prod_{p=1}^{P}\mathcal{N}\left(\mathring{x}(p)\,;\,\mathring{\mu}_{\boldsymbol{x}|*}(p),\mathring{\Sigma}_{\boldsymbol{x}|*}(p)\right),$$

et peut donc être échantillonnée suivant la démarche de l'Annexe C.

Une fois la chaîne de Gibbs terminée, on peut choisir de définir $\bar{\gamma}_{\boldsymbol{x}}$ et $\bar{\gamma}_{\boldsymbol{e}}$ comme les moyennes empiriques des échantillons

$$\bar{\gamma}_{\boldsymbol{x}}=\frac{1}{G}\sum_{g=1}^{G}\gamma_{\boldsymbol{x}}^{(g)} \qquad \text{et} \qquad \bar{\gamma}_{\boldsymbol{e}}=\frac{1}{G}\sum_{g=1}^{G}\gamma_{\boldsymbol{e}}^{(g)}\,. \tag{4.32}$$

On évalue la distribution (4.29) avec $\bar{\gamma}_{\boldsymbol{x}}$ et $\bar{\gamma}_{\boldsymbol{e}}$, pour tous les échantillons $\boldsymbol{x}^{(g)}$. On notera les paramètres d'intensité de (4.26) et (4.28)

$$B_{\boldsymbol{e},j}^{(g)}=\beta_{\boldsymbol{e}}+\frac{1}{2}\,||\mathring{\boldsymbol{y}}-\mathring{\boldsymbol{H}}\,\mathring{\boldsymbol{x}}^{(g)}||^2_{\mathring{\boldsymbol{S}}_{\boldsymbol{e},j}} \quad \text{et} \quad B_{\boldsymbol{x},i}^{(g)}=\beta_{\boldsymbol{x}}+\frac{1}{2}\,||\mathring{\boldsymbol{x}}^{(g)}||^2_{\mathring{\boldsymbol{S}}_{\boldsymbol{x},i}} \tag{4.33}$$

pour signaler que ces termes dépendent des tirages $\boldsymbol{x}^{(g)}$ et alléger les écritures. La moyenne empirique (4.22) se réécrit

$$\begin{aligned}
&\hat{p}(\bar{\gamma}_{\boldsymbol{x}},\bar{\gamma}_{\boldsymbol{e}}|\boldsymbol{y},\mathscr{M}=k)\\
&\quad=\frac{1}{G}\sum_{g=1}^{G}\Gamma(\bar{\gamma}_{\boldsymbol{x}}\,;\,A_{\boldsymbol{x}},B_{\boldsymbol{x},i}^{(g)})\,\Gamma(\bar{\gamma}_{\boldsymbol{e}}\,;\,A_{\boldsymbol{e}},B_{\boldsymbol{e},j}^{(g)})\\
&\quad=\frac{{\bar{\gamma}_{\boldsymbol{x}}}^{A_{\boldsymbol{x}}-1}}{\Gamma(A_{\boldsymbol{x}})}\,\frac{{\bar{\gamma}_{\boldsymbol{e}}}^{A_{\boldsymbol{e}}-1}}{\Gamma(A_{\boldsymbol{e}})}\,\frac{1}{G}\sum_{g=1}^{G}\exp\left[A_{\boldsymbol{x}}\log B_{\boldsymbol{x},i}^{(g)}-\bar{\gamma}_{\boldsymbol{x}}\,B_{\boldsymbol{x},i}^{(g)}+A_{\boldsymbol{e}}\log B_{\boldsymbol{e},j}^{(g)}-\bar{\gamma}_{\boldsymbol{e}}\,B_{\boldsymbol{e},j}^{(g)}\right].
\end{aligned}$$

Pour terminer, il faut évaluer les autres termes de (4.21) avec $\bar{\gamma}_{\boldsymbol{x}}$ et $\bar{\gamma}_{\boldsymbol{e}}$. Le premier terme de la somme est la log vraisemblance (4.11) page 45, dont il faut au préalable évaluer le cœfficient

$$\overline{\mathring{s}_{\boldsymbol{y},k}}(p)={\bar{\gamma}_{\boldsymbol{x}}}^{-1}\,|\mathring{h}(p)|^2\,\mathring{s}_{\boldsymbol{x},i}(p)+{\bar{\gamma}_{\boldsymbol{e}}}^{-1}\,\mathring{s}_{\boldsymbol{e},j}(p)\,. \tag{4.34}$$

Enfin, il ne reste plus que les deuxième et troisième termes à calculer, qui sont les log priors des hyperparamètres

$$\begin{aligned}
\log\pi(\bar{\gamma}_{\boldsymbol{x}})&=\alpha_{\boldsymbol{x}}\log\beta_{\boldsymbol{x}}-\log\Gamma(\alpha_{\boldsymbol{x}})+(\alpha_{\boldsymbol{x}}-1)\log(\bar{\gamma}_{\boldsymbol{x}})-\beta_{\boldsymbol{x}}\bar{\gamma}_{\boldsymbol{x}}\,,\\
\log\pi(\bar{\gamma}_{\boldsymbol{e}})&=\alpha_{\boldsymbol{e}}\log\beta_{\boldsymbol{e}}-\log\Gamma(\alpha_{\boldsymbol{e}})+(\alpha_{\boldsymbol{e}}-1)\log(\bar{\gamma}_{\boldsymbol{e}})-\beta_{\boldsymbol{e}}\bar{\gamma}_{\boldsymbol{e}}\,.
\end{aligned}$$

La section suivante est dédiée aux résultats numériques obtenus avec les différents algorithmes

en observation indirecte, avec paramètres connus puis inconnus.

4.4 Résultats numériques

Nous présentons dans cette section les aboutissements de nos travaux concernant la sélection de modèles sur observation indirecte. Dans un premier temps, nous travaillerons à modèle fixé : pour tous les algorithmes, nous utiliserons sur la même observation $\boldsymbol{y}$, résultant du modèle vrai $k^* = 4$. Nous présenterons alors quelques graphes et résultats pour dépeindre le comportement et les performances de l'algorithme étudié. Nous nous concentrerons ensuite sur un algorithme en particulier, celui de Chib. Le modèle et les données $\boldsymbol{y}$ ne seront alors plus fixés et nous opérerons la sélection de modèles dans toutes les configurations offertes. Les statistiques de sélection de modèles obtenues seront présentées sous formes de matrices de confusion.

Pour chaque modèle vrai k^*, nous avons construit des jeux de 50 observations $\boldsymbol{y}$, selon la procédure décrite en Annexe C. Les images $\boldsymbol{x}$ générées au chapitre 3 ont été convoluées puis bruitées. Les paramètres pour le bruit $\boldsymbol{e}$ sont

- Vrai niveau : $\gamma_e^* = 4$,
- Vraie largeur : $w_e = 0,5$ (sauf pour le modèle Blanc).

La matrice de convolution $\boldsymbol{H}$ est celle présentée en Figure 1.6 page 10 :

- Type de gain fréquentiel : sinus cardinal,
- Largeur de gain : $L = 1$.

Les deux prochaines sections exposent les résultats numériques obtenus lorsque les niveaux sont respectivement connus puis inconnus.

4.4.1 Niveaux γ_x et γ_e connus

Ici, l'évidence est connue analytiquement. Les données $\boldsymbol{y}$ traitées dans cet exemple résultent du vrai modèle $k^* = 4$, ce qui correspond à une DSP Lorentzienne séparable pour l'image ($i^* = 1$) et blanche pour le bruit ($j^* = 4$). Nous avons rapporté dans la Table 4.2 page 54 les valeurs de log-évidence données par l'équation (4.12) page 46 et les probabilités *a posteriori* associées pour les seize modèles candidats.

On constate comme à la section 3.3 que le bon modèle est sélectionné, avec une probabilité *a posteriori* extrêmement proche de 1. Les taux de bonnes sélections atteignent également 100% pour toutes les configurations de modèles.

Modèle k	$\log f(\boldsymbol{y}\|\mathscr{M}=k)$	$p(\mathscr{M}=k\|\boldsymbol{y})$	k	$\log f(\boldsymbol{y}\|\mathscr{M}=k)$	$p(\mathscr{M}=k\|\boldsymbol{y})$
1	-27 510,3592	0	9	-27 642,4809	0
2	-16 894,2148	0	10	-16 922,0129	0
3	-20 929,4374	0	11	-21 013,8969	0
4	**-13 434,0876**	**1**	12	-13 474,9742	10^{-18}
5	-28 671,0386	0	13	-25 433,2611	0
6	-17 216,5601	0	14	-16 951,4817	0
7	-21 407,6841	0	15	-19 810,1321	0
8	-13 537,7587	10^{-46}	16	-13 846,389	10^{-180}

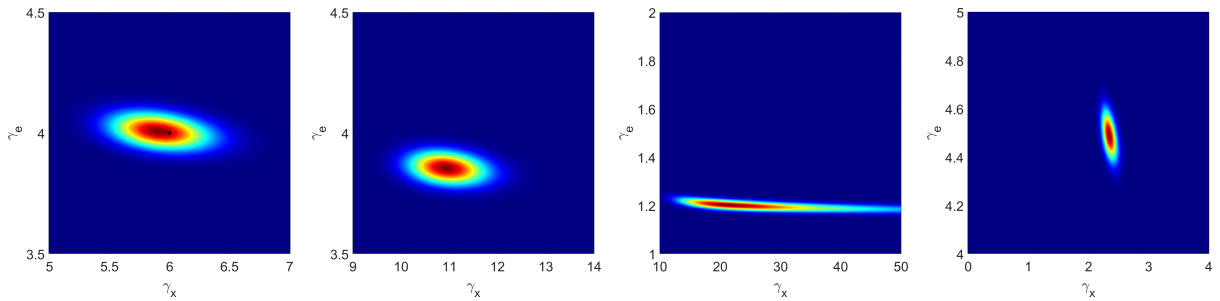
TABLE 4.2 – Valeurs de la log-évidence calculées pour les différents modèles candidats, parmi lesquels le vrai modèle est $k^* = 4$. Il est repéré par la couleur magenta et la valeur maximale de $p(\mathscr{M}=k|\boldsymbol{y})$ par la couleur bleue.

4.4.2 Niveaux γ_x et γ_e inconnus

L'évidence s'obtient ici en marginalisant par rapport à γ_x et γ_e la loi jointe $f(\boldsymbol{y},\gamma_x,\gamma_e|\mathscr{M}=k)$, que l'on construit à partir des densités $f(\boldsymbol{y}|\gamma_x,\gamma_e,\mathscr{M}=k)$, $\pi(\gamma_x)$ et $\pi(\gamma_e)$. Les paramètres des loi Gamma *a priori* $\pi(\gamma_x)$ et $\pi(\gamma_e)$ valent

- $\alpha_x = \beta_x = 10^{-10}$,
- $\alpha_e = \beta_e = 10^{-10}$.

La Figure 4.1 montre cette loi jointe au facteur $f(\boldsymbol{y}|\mathscr{M}=k)$ près pour quelques modèles k. Sur l'axe vertical, $f(\gamma_x,\gamma_e|\boldsymbol{y},\mathscr{M}=k)$ demeure relativement piquée : la largeur de la zone à forte densité ne dépasse pas 1, ce qui est relativement petit lorsque l'on doit parcourir $\mathbb{R}$ entier pour trouver cette zone. En revanche, sur l'axe horizontal, la largeur de la loi $f(\gamma_x,\gamma_e|\boldsymbol{y},\mathscr{M}=k)$ varie plus fortement d'un modèle à l'autre. Lorsqu'il s'agit du vrai modèle $k=k^*=4$, le mode de la loi $f(\boldsymbol{y},\gamma_x,\gamma_e|\mathscr{M}=k)$ se trouve aux coordonnées $(6,4)$, repérées sur le graphe (a) par une étoile noire.



(a) $k=4$ (b) $k=8$ (c) $k=15$ (d) $k=16$

FIGURE 4.1 – Loi jointe $p(\boldsymbol{y},\gamma_x,\gamma_e|\mathscr{M}=k)$, au facteur $f(\boldsymbol{y}|\mathscr{M}=k)$ près, pour quelques modèles k. Le vrai modèle ici est $k^*=4$ (DSP lorenztienne pour l'objet ($i=2$) et DSP blanche pour le bruit ($j=4$)). La couleur bleue nuit correspond à des valeurs inférieures à 10^{-50} et la couleur rouge foncée au maximum de la distribution.

Chronologiquement, nous allons exposer les valeurs de log-évidences obtenues par marginalisation brute, puis les résultats obtenus avec l'algorithme de Chib et ses performances.

4.4.2.1 Algorithmes naïfs

Les deux algorithmes naïfs donnant les mêmes valeurs, nous choisissons de montrer uniquement la marginalisation brute. On construit une grille

- de $G_{\boldsymbol{x}} = 2.10^3$ points sur l'intervalle $[\gamma_{\boldsymbol{x}}^{\mathrm{m}}, \gamma_{\boldsymbol{x}}^{\mathrm{M}}] = [10^{-1}, 10^2]$ pour $\gamma_{\boldsymbol{x}}$,
- de $G_{\boldsymbol{e}} = 2.10^2$ points sur l'intervalle $[\gamma_{\boldsymbol{e}}^{\mathrm{m}}, \gamma_{\boldsymbol{e}}^{\mathrm{M}}] = [10^{-1}, 7]$ pour $\gamma_{\boldsymbol{e}}$.

Pour chacun des seize modèles, nous avons calculé les log-évidences données par l'équation (4.14) page 46, exhibées dans la Table 4.3. On constate que le modèle sélectionné est le modèle vrai $k^* = 4$, avec une probabilité extrêmement proche de 1.

Modèle k	$\log f(\boldsymbol{y}\|\mathscr{M} = k)$	$p(\mathscr{M} = k\|\boldsymbol{y})$	k	$\log f(\boldsymbol{y}\|\mathscr{M} = k)$	$p(\mathscr{M} = k\|\boldsymbol{y})$
1	-13 667,8154	10^{-81}	9	-13 657,2062	10^{-77}
2	-13 569,5701	10^{-39}	10	-13 561,7282	10^{-35}
3	-13 781,5808	10^{-131}	11	-13 773,9223	10^{-127}
4	**-13 481,822**	**0,99896**	12	-13 488,6908	0,0010386
5	-13 648,5389	10^{-73}	13	-13 797,8499	10^{-138}
6	-13 553,9952	10^{-32}	14	-13 755,5958	10^{-119}
7	-13 767,361	10^{-125}	15	-13 818,6692	10^{-147}
8	-13 499,9165	10^{-8}	16	-13 760,3541	10^{-95}

TABLE 4.3 – Valeurs de la log-évidence calculées par (4.14) pour les différents modèles candidats. Les données $\boldsymbol{y}$ traitées ont pour vrai modèle $k^* = 4$. Le vrai modèle est repéré par la couleur magenta et la valeur maximale de $p(\mathscr{M} = k|\boldsymbol{y})$ par la couleur bleue.

Comme attendu, cet algorithme naïf s'est montré sûr et simple à implémenter. Malgré notre troncature de l'espace des paramètres, il nous a fallu 4.10^5 points pour quadriller l'espace, et seule une faible partie des points se situaient dans les zones à fortes densités. L'algorithme de Chib lui utilise essentiellement des échantillons *a posteriori*, qui tombent dans les zones à forte densité. Les ressources sont clairement mieux utilisées, mais il est plus compliqué à mettre en œuvre, à cause des processus MCMC.

4.4.2.2 Algorithme de Chib

Nous fixons le nombre de tirages à $G = 5.10^5$ et nous initialisons la chaîne de Gibbs construite page 51 en posant $\boldsymbol{x}^{(0)} = \boldsymbol{y}$. Des échantillons de $\gamma_{\boldsymbol{x}}$ et $\gamma_{\boldsymbol{e}}$ issus de la chaîne sont montrés en Figures 4.2, ainsi que leurs histogrammes. Le modèle testé à ce moment est le vrai modèle $k = k^* = 4$. On constate que les échantillons fluctuent autour des vraies valeurs $\gamma_{\boldsymbol{x}}^* = 6$ et $\gamma_{\boldsymbol{e}}^* = 4$.

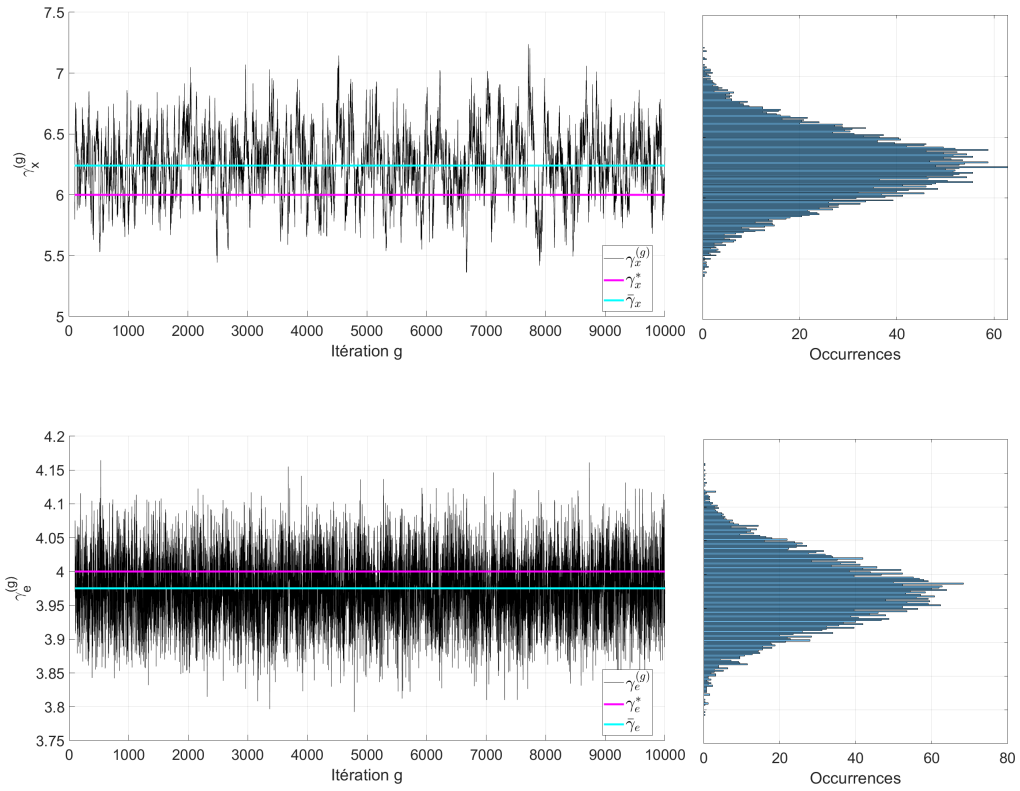


FIGURE 4.2 – Évolution (à gauche) et histogramme (à droite) des échantillons issus de la chaîne de Gibbs. La première ligne montre les tirages $\gamma_{\boldsymbol{x}}^{(g)}$ et la seconde les tirages $\gamma_{\boldsymbol{e}}^{(g)}$. Les échantillons sont tracés en noir, leur moyenne en cyan et les vrais niveaux en magenta. Les 100 premiers échantillons de burn-in ont été retirés.

Ces échantillons sont présentés en regard de la loi *a posteriori* dont ils sont issus en Figure 4.3. Ils sont concentrés autour de la zone à forte densité de la loi $p(\gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}} | \boldsymbol{y}, \mathscr{M} = k)$, contrairement à des échantillons du prior, dispersés dans tout l'espace. Pour une même quantité de points sollicités, l'algorithme de Chib compte très peu d'échantillons à densité nulle. L'ensemble des échantillons est alors « utile » au calcul de la log-évidence.

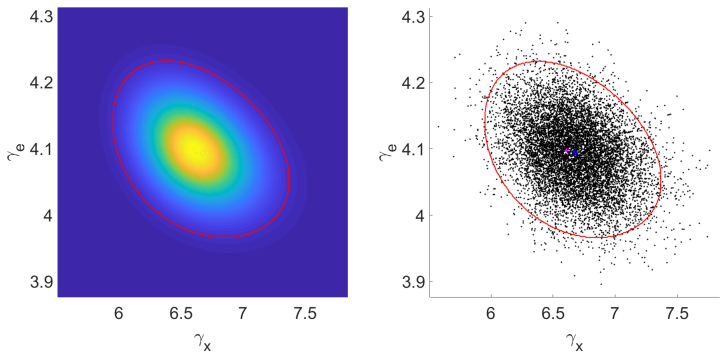


FIGURE 4.3 – A gauche : loi jointe *a posteriori* $p(\gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}} | y, \mathscr{M} = k)$ pour le modèle $k = k^* = 4$ à un facteur près. A droite : répartition des échantillons de $(\gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}})$. La courbe rouge sur les deux figures délimite la région où se concentrent 95% du volume de $p(\gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}} | y, \mathscr{M} = k)$.

Nous avons ensuite contrôlé l'algorithme de Chib sur plusieurs aspects. D'abord, nous nous

sommes assurés que l'algorithme converge vers la valeur numérique de référence, obtenue par marginalisation brute dans la section précédente. On peut voir sur la Figure 4.4 que les valeurs (4.21) se rapproche de la référence à partir de 10^4 tirages et se stabilise.

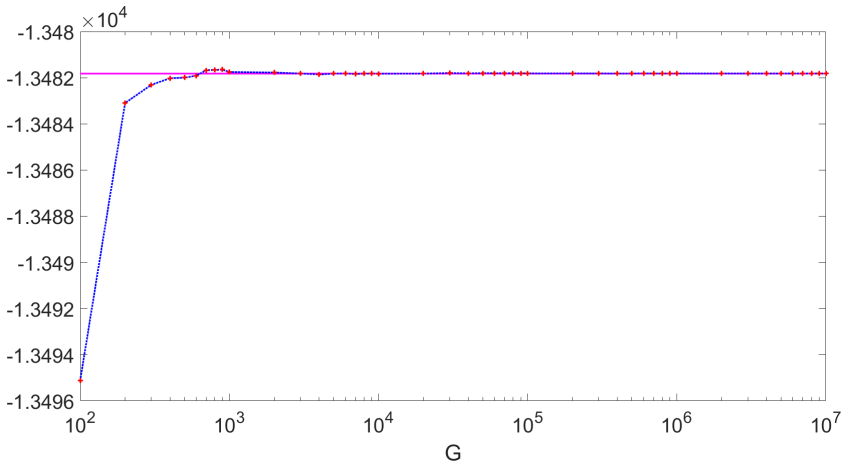


FIGURE 4.4 – Convergence de l'approximation (4.21) selon le nombre de tirages de la chaîne de Gibbs. La droite magenta correspond au logarithme de (4.14) obtenue par marginalisation brute.

Une version agrandie de cette courbe est affichée à gauche sur la Figure 4.5a, pour un nombre de tirages G supérieur à 10^3. La courbe (b) lui faisant face quantifie la différence, sur le même intervalle, entre la courbe bleue à croix rouge et la droite magenta.

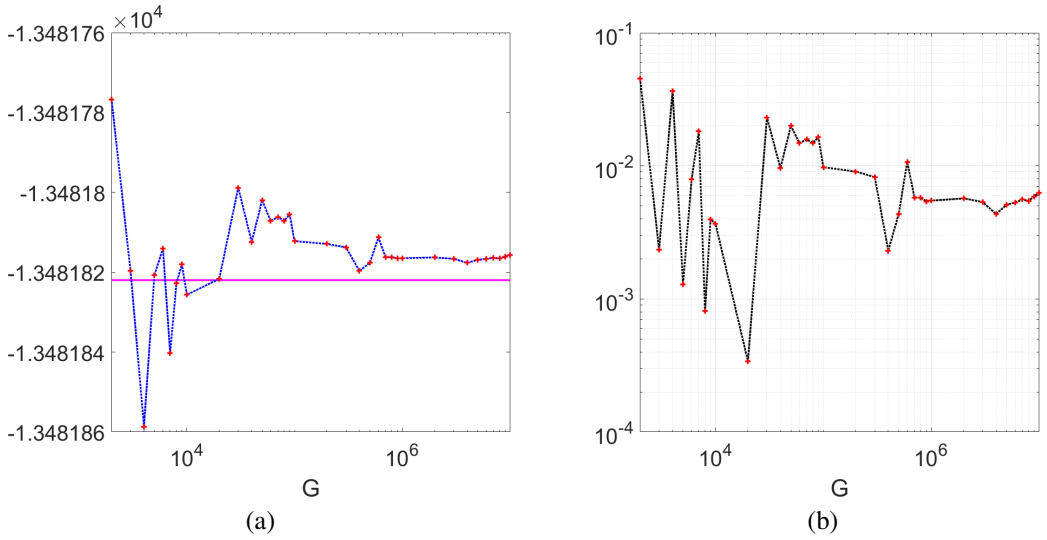


FIGURE 4.5 – L'image (a) est un agrandissement de la Figure 4.4 pour un nombre G supérieur à 10^3. La droite magenta correspond à la valeur numérique (4.14) obtenue par marginalisation brute. L'image (b) représente la différence entre les valeurs (4.21) et le logarithme de (4.14) sur le même intervalle.

On constate que l'algorithme de Chib a convergé mais aussi qu'il subsiste un écart entre les deux courbes (la courbe en 4.5b stagne autour de 10^{-2}). Les approches avec chaînes MCMC étant garanties convergentes, cette différence proviendrait de l'algorithme de marginalisation brute. Malgré notre grille de 4.10^5 points, nous avons manqué de l'information, en partie dû à la troncature des domaines d'intégration. Nous avions exposé cet inconvénient à la section 2.1.1, nous ne

sommes donc pas surpris de le constater en application numérique. Cela constitue un bon exemple et nous encourage à privilégier l'algorithme de Chib pour calculer les évidences.

Pour consolider nos observations, nous avons regardé le rapport des valeurs consécutives de $f(\boldsymbol{y}|\mathscr{M} = k)$, qui doit tendre vers 1 lorsque l'algorithme a convergé. La Figure 4.6 confirme que l'algorithme a convergé au delà de 10^5 tirages *a posteriori*.

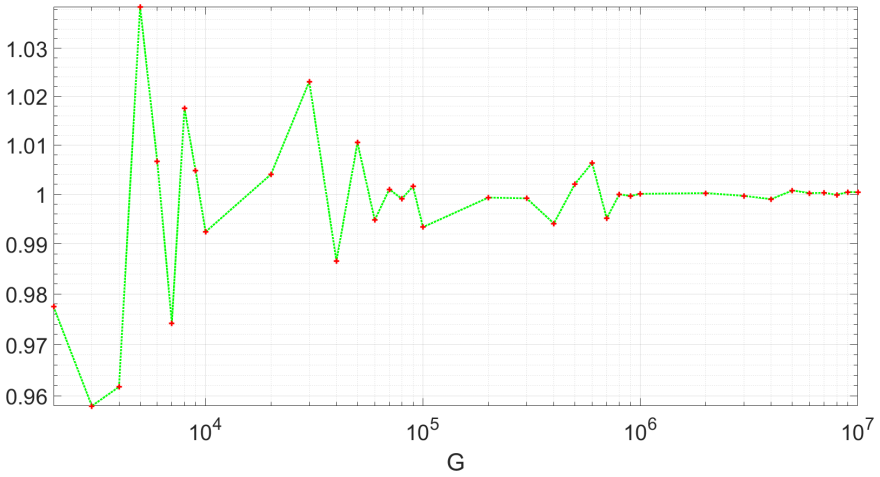


FIGURE 4.6 – Rapports consécutifs des valeurs (4.21) obtenues avec l'algorithme de Chib.

Nous allons maintenant mener nos expériences de sélections de modèle avec cet algorithme. Le graphique sur la Figure 4.7 présente les résultats de sélection de modèles pour les $K = 16$ configurations. Pour chaque modèle vrai k^*, nous avons traité les 50 données en comparant tous les modèles k et compté chaque fois que l'algorithme sélectionnait le vrai modèle k^*. Ces nombres de bonnes sélections sont présentés dans une matrice sous forme de pourcentage.

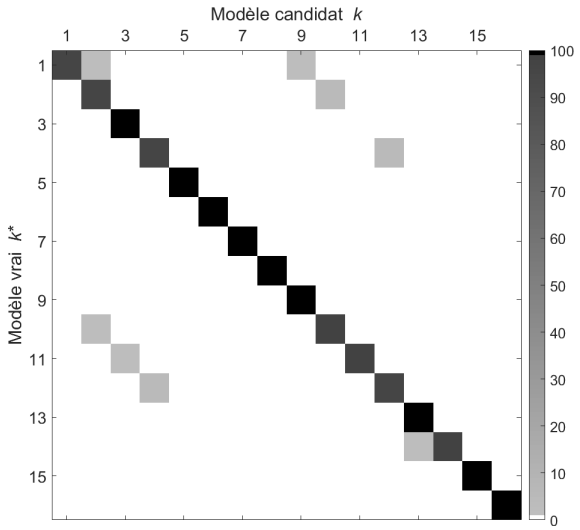


FIGURE 4.7 – Matrice de confusion : résultats de la sélection de modèles pour l'algorithme de Chib. L'axe des ordonnées donne l'indice du modèle vrai et l'axe des abscisses l'indice du modèle testé. Les valeurs sont exprimées en pourcentage de bonnes sélections.

On peut voir que pour les seize configurations, on obtient de très bons résultats. Certaines configurations atteignent 100% de bonnes sélections, lorsque les modèles objets sont gaussiens et presque pour tous les modèles objets blancs. Pour le reste, les proportions de bonnes sélections

s'élèvent à plus de 90%. Il est à noter qu'il y a aucun cas où les modèles objet et bruit sont tous deux mal sélectionnés.

4.4.2.3 Exemple de mauvaise sélection de modèles

On constate quelques cas isolés où la probabilité *a posteriori* est plus forte pour un modèle autre que le vrai modèle, c'est-à-dire lorsqu'il y a une erreur de sélection. Nous allons comparer les contenus fréquentiels issues d'une sélection erronée à ceux d'une sélection correcte. La Table 4.4 présente les évidences calculées pour ce cas de mauvaise sélection, où le modèle $k = 12$ a été sélectionné plutôt que le vrai modèle $k^* = 4$.

Modèle k	$\log f(\boldsymbol{y}\|\mathscr{M} = k)$	$p(\mathscr{M} = k\|\boldsymbol{y})$	k	$\log f(\boldsymbol{y}\|\mathscr{M} = k)$	$p(\mathscr{M} = k\|\boldsymbol{y})$
1	-13 716,3941	10^{-101}	9	-13 702,331	10^{-95}
2	-13 606,5623	10^{-54}	10	-13 593,8135	10^{-48}
3	-13 838,9836	10^{-155}	11	-13 829,7562	10^{-150}
4	**-13 485,3651**	**0,3745**	**12**	**-13 484,8522**	**0,6255**
5	-13 698,1182	10^{-93}	13	-13 836,3589	10^{-153}
6	-13 590,4788	10^{-47}	14	-13 786,1317	10^{-132}
7	-13 826,4286	10^{-149}	15	-13 868,1098	10^{-167}
8	-13 498,9043	10^{-7}	16	-13 710,0887	10^{-99}

TABLE 4.4 – Valeurs de la log-évidence calculées pour les différents modèles candidats. Les données $\boldsymbol{y}$ traitées ont pour vrai modèle $k^* = 4$. Le vrai modèle est repéré par la couleur magenta et la valeur maximale de $p(\mathscr{M} = k|\boldsymbol{y})$ par la couleur bleue.

Dans cette situation, la probabilité du modèle $k = 12$ est deux fois plus grande que celle du modèle vrai $k^* = 4$, respectivement égales à $0,63$ et $0,37$. A l'issu de la chaîne de Gibbs, on obtient des valeurs différentes de $\bar{\gamma}_{\boldsymbol{x}}$ et $\bar{\gamma}_{\boldsymbol{e}}$.

- Pour le modèle $k = 4$, on a $\bar{\gamma}_{\boldsymbol{x}} = 6,26$ et $\bar{\gamma}_{\boldsymbol{e}} = 9,97$.
- Pour le modèle $k = 12$, on a $\bar{\gamma}_{\boldsymbol{x}} = 9,24$ et $\bar{\gamma}_{\boldsymbol{e}} = 9,92$.

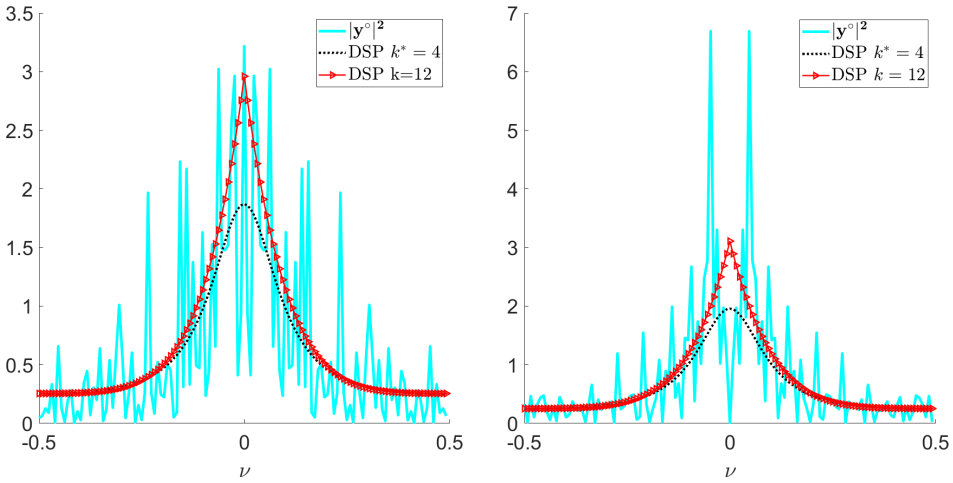
Les deux valeurs de $\bar{\gamma}_{\boldsymbol{e}}$ sont très proches, car dans les deux configurations, il s'agit d'une DSP blanche pour le bruit ($j = 4$). Les modèles de DSP objet étant quant à eux différents (Lorentz pour $k = 4$ et Laplace pour $k = 12$), les valeurs de $\bar{\gamma}_{\boldsymbol{x}}$ le sont également.

Comparons le périodogramme des données $|\mathring{\boldsymbol{y}}|^2$ et la DSP des données des deux modèles concernés, évaluée avec leurs $\bar{\gamma}_{\boldsymbol{x}}$ et $\bar{\gamma}_{\boldsymbol{e}}$ respectifs. Pour rappel, les cœfficients de ces DSP se calculent

$$\overline{\mathring{s}_{\boldsymbol{y},k}}(p) = \bar{\gamma}_{\boldsymbol{x}}{}^{-1}\,|\mathring{h}(p)|^2\,\mathring{s}_{\boldsymbol{x},i}(p) + \bar{\gamma}_{\boldsymbol{e}}{}^{-1}\,\mathring{s}_{\boldsymbol{e},j}(p)\,.$$

Sur la Figure 4.8, la courbe cyan représente le périodogramme des données, la courbe noire la DSP des données pour $k = 4$ et la courbe rouge à triangles la DSP des données pour $k = 12$. Le périodogramme en (a) provient de l'observation pour laquelle on obtient les évidences de la Table 4.4, ayant mené à une mauvaise sélection. A l'opposé, le périodogramme en (b) provient de l'observation pour laquelle on obtient les évidences de la Table 4.3, ayant conduit à une bonne

sélection.



(a) Cas de mauvaise sélection.

(b) Cas de bonne sélection.

FIGURE 4.8 – Comparaison des coupes à la fréquence nulle du périodogramme des données $|\mathring{\boldsymbol{y}}|^2$ (ici en cyan) avec les DSP de cœfficients $\overline{\mathring{s}_{\boldsymbol{y},k}(p)}$ pour $k = k^* = 4$ (courbe noire) et $k = 12$ (courbe rouge à triangles).

On constate que sur l'image (a), la DSP de $k = 12$ semble visuellement plus proche du périodogramme que la DSP de $k = 4$. Autrement dit, la distance entre le périodogramme et la DSP des données serait minime pour le modèle $k = 12$ dans le cas présent. Ce constat fait écho à l'interprétation de la sélection de modèle énoncée à la section 4.1 : les valeurs pour $k = 12$ minimisent les fonctions ϕ_p, et par conséquent maximise les probabilités *a posteriori*.

Sur l'image (b), dans le cas d'une bonne sélection, les deux DSP ne concordent pas vraiment le périodogramme. On observe cependant que la distance est plus faible entre la courbe cyan et la courbe noire qu'avec la courbe rouge.

4.5 Bilan du chapitre

Nous avons vu à l'œuvre dans ce chapitre les différents algorithmes d'approximation d'évidence dans le cas d'observation indirecte. Il s'agit d'un cas bien plus complexe (et donc plus riche) que l'observation directe.

- L'image n'est plus connue et il en résulte l'ajout de $P = N^2$ paramètres inconnus au problème.
- On peut marginaliser analytiquement l'image, mais la forme obtenue ne permet pas de marginaliser analytiquement les paramètres de niveaux.
- On considère un second paramètre de niveau, celui de bruit γ_e, qui peut être connu ou non.
- On veut également sélectionner le modèle de bruit.

Pour faciliter quelques opérations calculatoires, nous nous sommes placés dans le cas gaussien circulant. Le caractère gaussien des loi *a priori* nous a permis de marginaliser analytiquement l'image $\boldsymbol{x}$, *i.e.* les P nouveaux paramètres inconnus. Quant à la circularité, elle nous a permis de faciliter la manipulation numérique des matrices de covariances et de convolution. Nous avons ensuite étudié deux cas

- les niveaux $\gamma_{\boldsymbol{x}}$ et $\gamma_{\boldsymbol{e}}$ sont connus,
- les niveaux $\gamma_{\boldsymbol{x}}$ et $\gamma_{\boldsymbol{e}}$ sont inconnus.

Dans le premier cas, nous étions en mesure de calculer exactement l'évidence, donnée par (4.12) page 46. Dans le second cas, il n'est plus possible d'obtenir analytiquement l'évidence, à cause de la forme de la loi $f(\boldsymbol{y}, \gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}} | \mathscr{M} = k)$. Nous avons donc eu recours aux algorithmes présentés plus tôt :

- les algorithmes naïfs,
- la moyenne harmonique,
- l'algorithme de Chib.

Les algorithmes naïfs ont fourni d'excellents résultats. Cependant, leur utilisation importante de ressources vient nuancer leur simplicité d'application et leur fiabilité. Ces deux algorithmes se retrouvent vite limités par le nombre de paramètres inconnus en jeu.

Tout comme en observation directe, nous avons pu démontrer analytiquement que la moyenne harmonique ne converge pas en moyenne quadratique vers l'évidence. On retrouve la même forme d'intégrale divergente, responsable de la variance infinie de l'inverse vraisemblance.

L'algorithme de Chib couplé à une chaîne de Gibbs a également donné d'excellentes proportions de sélection de modèles. Cet algorithme plus complexe est cependant bien plus adapté pour traiter les problèmes avec plus d'inconnus. Nos choix de lois *a priori* ont facilité l'étape d'approximation de la loi $p(\gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}} | \boldsymbol{y}, \mathscr{M} = k)$. D'une part, nous avons pu expliciter analytiquement la loi $f(\gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}} | \boldsymbol{y}, \boldsymbol{x}, \mathscr{M} = k)$. D'autre part, les lois *a posteriori* conditionnelles étaient faciles / classiques à échantillonner, grâce à la conjugaison. Une des puissances de cet algorithme est sa capacité à cibler les zones à forte densité. Il est également robuste, nous verrons cependant dans le chapitre de conclusion que certaines configurations sont problématiques.

Nous avons présenté dans ce chapitre le cœur de notre contribution dans la sélection de modèle. Nous avons traité un cas très complet, l'observation indirecte avec deux paramètres inconnus. On

doit la richesse de ce cas à sa complexité, que nous avons graduellement augmenté au cours de ce manuscrit. Nous avons appréhendé et testé des algorithmes d'approximations d'évidence, rarement utilisés dans un tel cadre. Nous avons pu les approuver ou les réfuter aux vues de nos analyses. Leurs applications étaient encore « théoriques », car nous avons travaillé jusque-là sur données synthétiques. Le prochain chapitre est dédié à leur déploiement sur données réelles, que nous mettrons en regard de nos résultats numériques théoriques.

CHAPITRE 5

Sélection de modèles sur données réelles

Dans le chapitre précédent, nous avons présenté notre contribution majeure : la sélection de modèles à partir d'observation indirecte. Nous avons travaillé sur données synthétiques et notre approche a fourni d'excellents résultats. Pour enrichir l'étude, nous testons maintenant l'approche sur des données réelles.

Dans ce chapitre, les données $\boldsymbol{y}$ sont une observation indirecte de l'image « bateau ». Nous l'avons artificiellement dégradée avec le noyau de convolution vu en Figure 1.6 page 10 et un bruit blanc gaussien de paramètres $\gamma_e^* = 72$, similaire au chapitre 4. L'image vraie du bateau $\boldsymbol{x}^*$ et les données observées $\boldsymbol{y}$ résultantes sont affichées sur la Figure 5.1.

(a) Image vraie $\boldsymbol{x}^*$ (b) Données observées $\boldsymbol{y}$

FIGURE 5.1 – L'image (b) résulte d'une convolution de l'image (a) par le noyau page 10 et de l'ajout d'un bruit blanc gaussien (de paramètre $\gamma_e^* = 72$).

Jusqu'à présent, l'image $\boldsymbol{x}^*$ était générée synthétiquement à partir d'un modèle $\mathscr{M}^*$, que l'on qualifiait de « vrai modèle ». De plus, ce dernier faisait partie de la liste des modèles à comparer. Ici, l'image, et donc les données, ne proviennent pas d'un des modèles de notre liste. Il n'existe donc pas à proprement parlé de « vrai modèle » : les modèles candidats sont par conséquent mal spécifiés.

De ce fait, la philosophie est un peu différente dans ce chapitre : il ne s'agit pas de voir si l'approche sélectionne le « vrai modèle », mais plutôt de comparer les modèles et voir quelles informations nous délivrent les évidences sur leur degré d'adéquation avec les données [PM16]. En d'autres termes, tous les modèles sont faux et plus l'évidence est grande, plus le modèle est proche des données. Cela permet d'identifier les modèles dits forts, c'est-à-dire ceux qui concordent le mieux avec les données, et à l'opposé, les modèles faibles.

Nous obtenons l'évidence de chaque modèle avec l'algorithme de Chib, dont les équations demeurent celles de la section 4.3.3. Nous calculons ensuite les $K = IJ$ probabilités *a posteriori* de la page 12, et nous sélectionnons le modèle qui maximise ces dernières. Les résultats laissent ensuite place à une analyse fréquentielle des quantités en jeux. En particulier, nous comparons

1. périodogramme des données et DSP des données du modèle testé, paramétrée par les valeurs $(\bar{\gamma}_x, \bar{\gamma}_e)$ estimées,
2. périodogramme de l'image et DSP du modèle image testé, paramétrée par la valeur $\bar{\gamma}_x$ estimée.

Pour conclure ce chapitre, nous analysons l'impact de la sélection de modèle sur la déconvolution. L'image déconvoluée est obtenue par filtrage de Wiener, filtre paramétré avec les modèles image et bruit sélectionnés précédemment et les hyperparamètres estimés. Nous évaluons par la suite la qualité de la déconvolution, à la fois visuellement et quantitativement.

5.1 Comparaison de modèles sur des images réelles

Dans cette section, nous considérons uniquement $J = 1$ modèle de DSP bruit : Blanc. Nous avons vu que l'algorithme parvenait à sélectionner aussi bien le modèle de DSP bruit que le modèle de DSP image. Cela permet de focaliser l'analyse sur la sélection d'un modèle d'image pour une image qui n'est pas issue d'une réalisation stochastique.

Une analyse empirique préalable nous a fourni une connaissance sur la largeur fréquentielle de l'image (a), qui vaut $w_x = 10^{-3}$. Cependant, deux des modèles présentés au chapitre 1 ne sont pas adaptés pour cette largeur : avec cette valeur, les modèles Laplacien et Gaussien s'annulent sur le domaine, ce qui engendre des problèmes numériques. Nous considérons par conséquent $I = 5$ modèles de la liste du chapitre 1. Leur indexation est donnée en Table 5.1.

Modèle	LorentzS	LorentzC	InterPix-4	InterPix-8	Blanc
Indice	1	2	3	4	5

TABLE 5.1 – Tableau de correspondance entre les modèles image et leurs indices.

Ne considérant qu'un seul modèle de DSP bruit, la variable de modèle k suit cette même indexation.

Les paramètres conservent les mêmes valeurs numériques qu'à la section 4.4. Avant de calculer les évidences, nous avons étudié le comportement des échantillons issues des boucles de Gibbs, afin de les comparer à celles du chapitre 4. Les échantillons de γ_x et γ_e, ainsi que leurs histogrammes, sont affichés sur la Figure 5.2 pour le modèle $k = 2$.

Les échantillons fluctuent de manière relativement similaire aux chaînes de la Figure 4.2 page

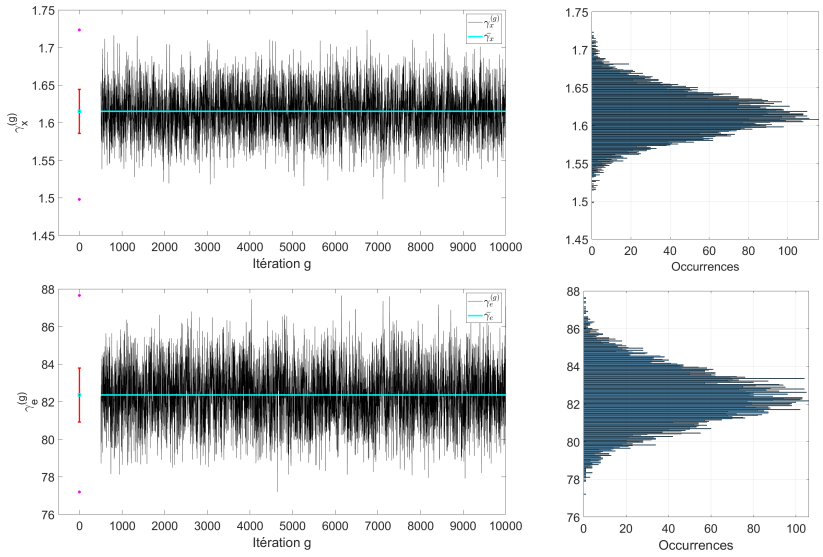


FIGURE 5.2 – Évolution (à gauche) et histogramme (à droite) des échantillons issus de la chaîne de Gibbs pour le modèle $k = 2$. La première ligne montre les tirages de γ_x et la seconde les tirages de γ_e. Les échantillons sont tracés en noir et leur moyenne en cyan. Les 500 premiers échantillons de burn-in ont été retirés.

56. Sans surprise, les couples d'échantillons de (γ_x, γ_e) se concentrent autour du point $(\bar{\gamma}_x, \bar{\gamma}_e)$, zone de forte densité de $p(\gamma_x, \gamma_e | \boldsymbol{y}, \mathscr{M} = k)$, comme en témoigne les graphes de la Figure 5.3.

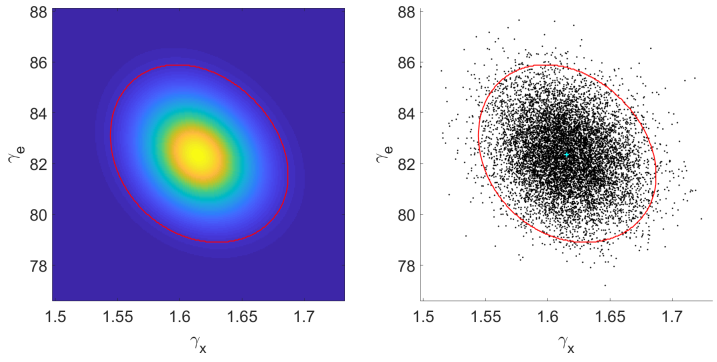


FIGURE 5.3 – A gauche : loi jointe *a posteriori* $p(\gamma_x, \gamma_e | y, \mathscr{M} = k)$ pour le modèle $k = 2$ à un facteur près. A droite : répartition des échantillons de (γ_x, γ_e). La courbe rouge sur les deux figures délimite la région où se concentrent 95% du volume de $p(\gamma_x, \gamma_e | y, \mathscr{M} = k)$.

Les valeurs $(\bar{\gamma}_x, \bar{\gamma}_e)$ obtenues pour chaque modèle sont affichées dans la Table 5.2. On remarque que les ordres de grandeurs des valeurs de $\bar{\gamma}_x$ sont disparates d'un modèle à l'autre : de 10^{-3} à 10^{0}. Pour $\bar{\gamma}_e$, les ordres de grandeurs sont plus homogènes. On constate ensuite que les modèles 2, 3 et 4 sont très proches de la valeur vraie $\gamma_e^* = 72$, en particulier les modèles 3 et 4. Les modèles 1 et 5 en sont quant à eux très éloignés. Ici, connaissant la vraie valeur γ_e^*, les valeurs de $\bar{\gamma}_e$ fournissent un premier indice sur les modèles enclin à manifester les plus fortes évidences.

Ces intuitions sont en effet vérifiées. La Table 5.3 fournit les valeurs de log-évidences obtenues avec l'algorithme de Chib (équation (4.21)) page 48 et les probabilités *a posteriori* correspondantes

k	1	2	3	4	5
$\bar{\gamma}_{\boldsymbol{x}}$	$3,04.10^{-3}$	1,615	$1,57.10^{-2}$	$1,91.10^{-2}$	0,335
$\bar{\gamma}_{\boldsymbol{e}}$	58,64	82,36	70,03	73,85	123,20

TABLE 5.2 – Valeurs $(\bar{\gamma}_{\boldsymbol{x}}, \bar{\gamma}_{\boldsymbol{e}})$ estimées avec l'algorithme de Chib sur données réelles. La vraie valeur $\gamma_{\boldsymbol{e}}^*$ vaut 72.

pour les cinq modèles candidats.

Nous retrouvons les mêmes valeurs de log-évidences en utilisant la marginalisation brute. Grâce aux échantillons de l'algorithme de Chib, nous avons pu définir une grille adaptée, centrée autour de $(\bar{\gamma}_{\boldsymbol{x}}, \bar{\gamma}_{\boldsymbol{e}})$, pour chaque modèle. Sinon, il aurait fallu créer une grille suffisamment grande et précise pour balayer toutes les zones de forte densités des lois $p(\gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}}|y, |\mathscr{M} = k)$. La consommation de ressources pour intégrer sur une telle grille aurait été gargantuesque. Il est donc préférable d'utiliser l'algorithme de Chib.

Modèle k	$\log f(\boldsymbol{y}\|\mathscr{M} = k)$	$p(\mathscr{M} = k\|\boldsymbol{y})$
1	1 286.49338	0
2	**2 544.82857**	**1**
3	2 524,53629	10^{-9}
4	2 341,18512	10^{-89}
5	-12 769,583	0

TABLE 5.3 – Valeurs des log-évidence. Les valeurs maximales sont repérées par la couleur bleue.

Le modèle 2 maximise la log-évidence. Il en découle que ce dernier coïncide le plus avec les données $\boldsymbol{y}$, suivi de près par les modèles 3 et 4 et, loin derrière, les modèles 1 et 5. Nous reviendrons sur ce constat dans la section suivante, lorsque nous étudierons la déconvolution.

On constate que les modèles 2, 3 et 4 ont les plus fortes log-évidences, dont les valeurs sont relativement proches. Avec un score plus faible, on trouve le modèle 1. Bien qu'il soit de même nature que le modèle 2, on remarque une très nette différence de valeur entre les deux modèles. Enfin, loin derrière, le modèle 5 donne la plus faible valeur. Ce résultat n'est pas surprenant, car on se doutait qu'une image réelle telle que (a) ne pouvait avoir une DSP blanche.

En ce qui concerne les probabilités $p(\mathscr{M} = k|\boldsymbol{y})$, le résultat est sans appel : le modèle sélectionné ici est le modèle $k = 2$, *i.e.* le modèle de DSP Lorentzienne circulaire pour l'image. Nous allons maintenant comparer les périodogrammes avec les DSP des différents modèles.

5.1.1 DSP et périodogramme des données $\boldsymbol{y}$

Tout d'abord, nous nous concentrons sur les données $\boldsymbol{y}$. Ici, la DSP des données s'écrit

$$\mathring{\boldsymbol{S}}_{\boldsymbol{y},k} = \bar{\gamma}_{\boldsymbol{x}}^{-1} \mathring{\boldsymbol{H}} \mathring{\boldsymbol{S}}_{\boldsymbol{x},i} \mathring{\boldsymbol{H}}^{\dagger} + \bar{\gamma}_{\boldsymbol{e}}^{-1} \mathring{\boldsymbol{S}}_{e,j} \ . \tag{5.1}$$

et elle est tracée en rouge sur la Figure 5.4 pour chaque modèle k avec le couple $(\bar{\gamma}_{\boldsymbol{x}}, \bar{\gamma}_{\boldsymbol{e}})$) associés. Nous y avons également tracé en bleu le périodogramme des données. À chaque colonne correspond un modèle : de $k = 1$ à gauche à $k = 5$ à droite.

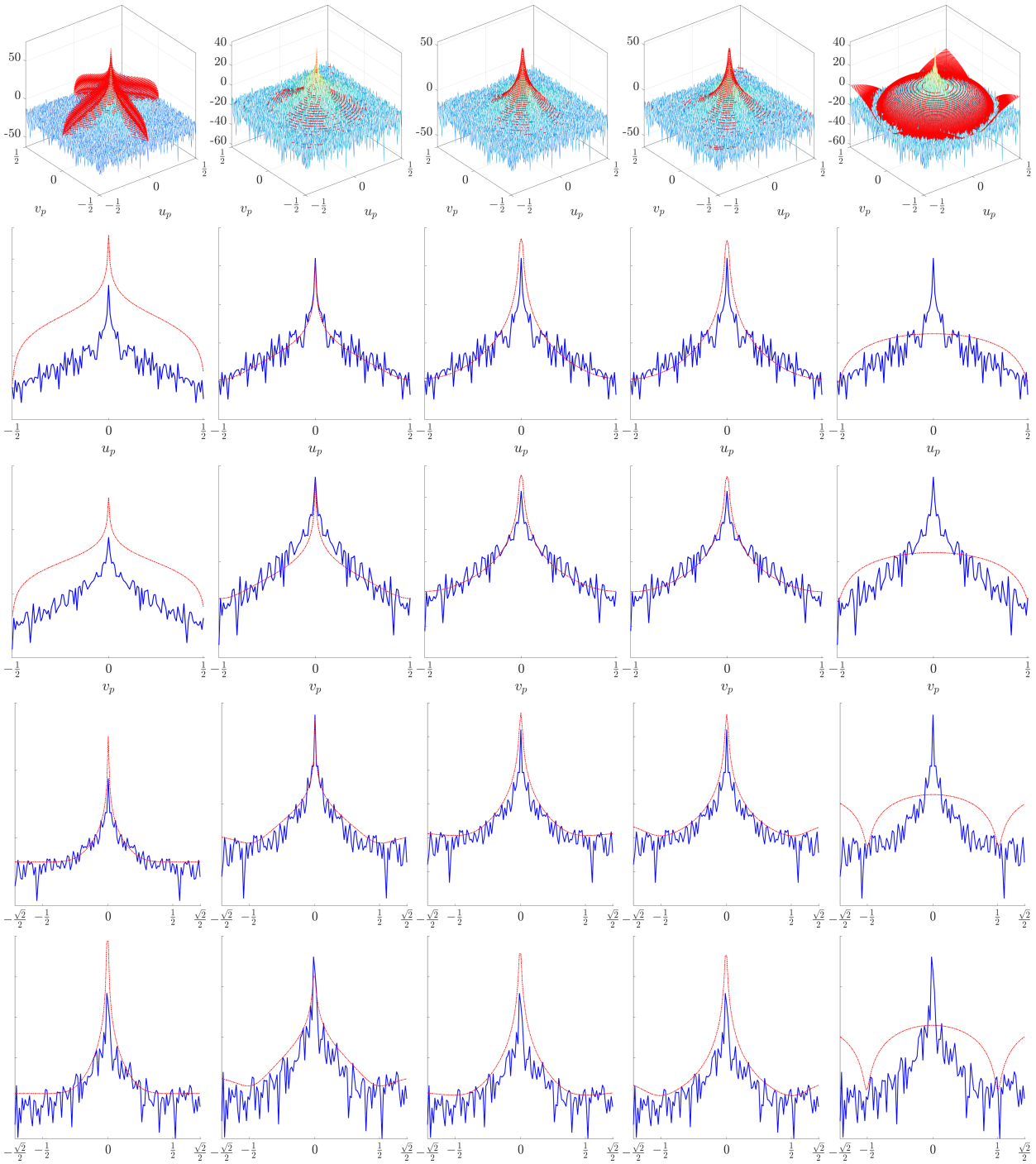


FIGURE 5.4 – Comparaison de la DSP de chaque modèle k (en rouge) et du périodogramme des données (en bleu). Le modèle ayant la plus grande probabilité est le LorentzC, en deuxième colonne. De haut en bas : superposition du périodogramme et de la DSP d'un modèle, coupe à la fréquence $v_p = 0$, coupe à la fréquence $u_p = 0$, coupe $u_p = -v_p$ (diagonale), coupe $u_p = v_p$ (anti-diagonale).

Ces graphes sont à étudier conjointement aux analyses faites à propos de la Table 5.3. On constate en effet les comportements prédits par les valeurs des log-évidences : les modèles forts (LorentzC, InterPix-4 et InterPix-8) ont une DSP qui s'accorde mieux avec le périodogramme des données que les modèles faibles (LorentzS et Blanc).

1. Le modèle $k = 2$ qui maximise la log-évidence est affiché dans la deuxième colonne. La DSP des données est très proche du périodogramme, dans les hautes et les basses fréquences. En regardant en détails, on voit que c'est celui qui coïncide le plus, comparé aux autres modèles.

2. Les deux colonnes suivantes sont dédiées aux modèles $k = 3$ et $k = 4$. On observe la bonne adéquation de leurs DSP au périodogramme, mais qu'elles sont cependant légèrement décalées, contrairement au modèle $k = 2$. Bien que la différence avec le modèle $k = 2$ soit visible sur les graphes, il est cependant impossible d'identifier à l'œil nu lequel coïncide le plus avec les données entre le modèle 3 et le modèle 4. Les log-évidences nous certifient quantitativement qu'il s'agit du modèle $k = 3$.

3. Affiché dans la première colonne, on constate pour le modèle $k = 1$ des écarts flagrants entre la DSP et le périodogramme, en particulier sur les lignes 2 et 3. Malgré cet éloignement, il est à noter que la DSP suit globalement la forme du périodogramme. Ces distances plus importantes par région expliquent une plus faible valeur de log-évidence pour ce modèle.

4. A l'opposé de ces trois modèles se dresse le modèle $k = 5$. En effet, la DSP de ce modèle ne suit absolument pas la forme du périodogramme des données. De cette aberration résulte un écart important en toutes les fréquences p. En conséquences, la log-évidence de ce modèle est très faible et également très éloignée des autres valeurs.

Naturellement, le modèle $\mathring{\boldsymbol{S}}_{\boldsymbol{y},k}$ sélectionné est celui que concorde au mieux avec le périodogramme des données. Qu'en est-il cependant de l'image inconnue $\boldsymbol{x}^*$ et du modèle $\mathring{\boldsymbol{S}}_{\boldsymbol{x},i}$ sélectionné ? Quel est le degré d'adéquation entre ces deux quantités ? Les réponses sont apportées dans la section suivante.

5.1.2 DSP et périodogramme de l'image $\boldsymbol{x}$

Nous avons affiché en Figure 5.5 le périodogramme de l'image $\boldsymbol{x}^*$, comparé aux DSP $\mathring{\boldsymbol{S}}_{\boldsymbol{x},i}$ des quatre premiers modèles objet i (nous avons jugé peu pertinent d'inclure le modèle de DSP blanc).

1. Analysons en premier le modèle $k = 2$. Sur les lignes 2 et 3, la DSP coïncide avec le périodogramme. Sur les lignes 4 et 5, la DSP suit la forme du périodogramme dans les basses fréquences, et un écart se creuse lorsqu'on se déplace vers les hautes fréquences. Malgré ces écarts, il demeure le modèle qui épouse le mieux le périodogramme.

2. Pour les modèles $k = 3$ et $k = 4$, on constate à nouveau que la DSP est légèrement au dessus du périodogramme dans les basses fréquences. Dans les hautes fréquences, comme pour le modèle $k = 2$, les deux courbes sont en adéquation sur les lignes 2 et 3, alors que l'on y observe un écart sur 4 et 5. On note cependant que cet écart est moins prononcé pour le modèle $k = 3$ que pour les modèles $k = 2$ et $k = 4$.

3. Sur les courbes du modèle $k = 1$, on retrouve, sans surprise, l'importante différence entre la DSP et le périodogramme sur les lignes 2 et 3. Sur les lignes 4 et 5, la DSP épouse cependant mieux le périodogramme que les autres modèles dans les hautes fréquences, mais pas dans les basses fréquences.

On a pu constater en passant des Figures 5.4 à 5.5 que les DSP objet suivaient moins bien le périodogramme de $\boldsymbol{x}^*$ que la DSP des modèles k avec le périodogramme de $\boldsymbol{y}$. Ce phénomène s'explique ainsi : le modèle objet est sélectionné à partir d'une version altérée de l'image $\boldsymbol{x}$, et non à partir de l'image elle-même. L'approche de sélection de modèle tend à faire coïncider le modèle et le périodogramme des données avant tout.

Dans la prochaine section, nous allons reconstruire l'image à partir des données observées en faisant intervenir un filtre de Wiener et analyser l'impact de la sélection de modèle sur le rendu de la déconvolution.

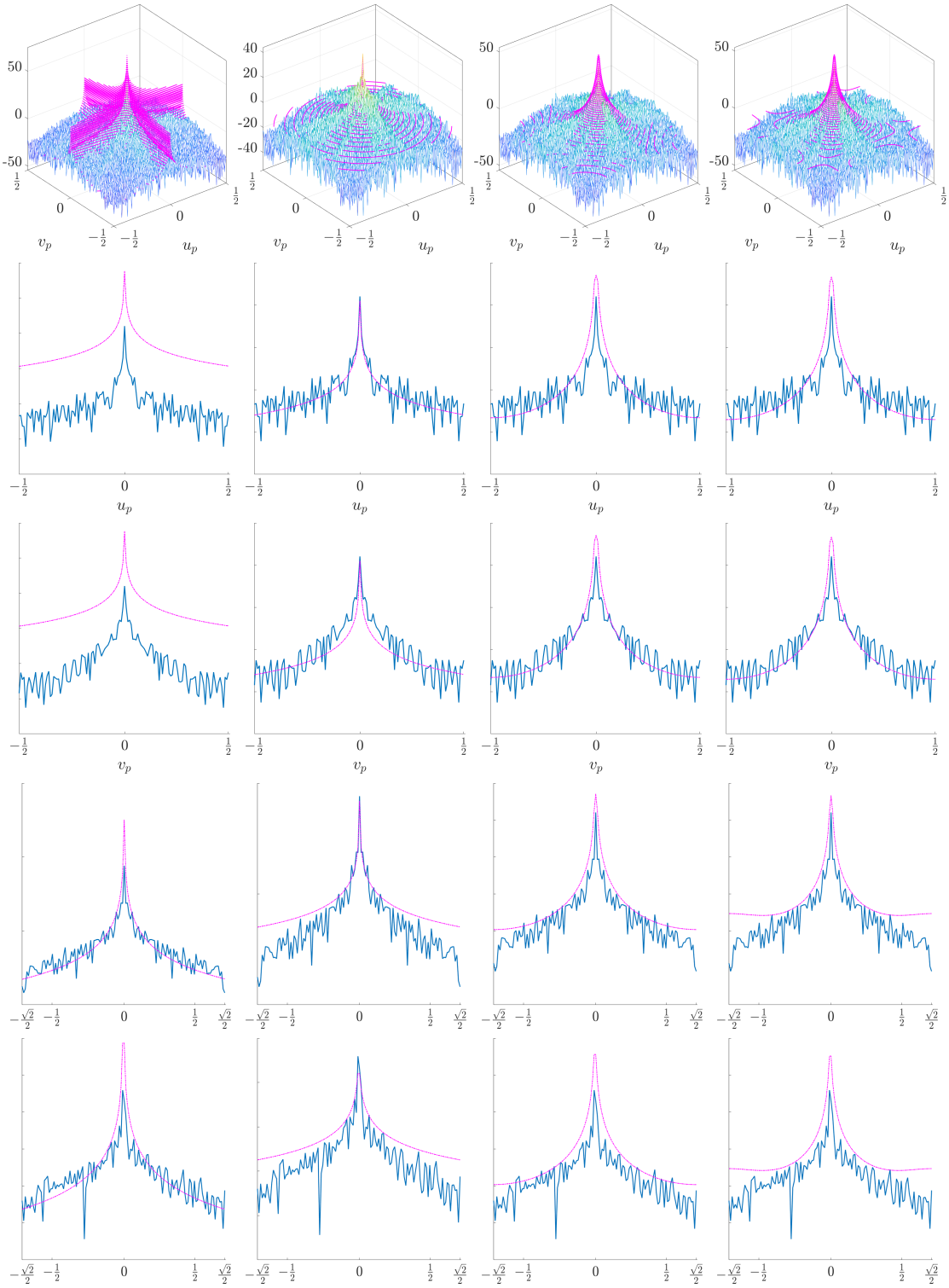


FIGURE 5.5 – Comparaison de la DSP objet i de chaque modèle k (en magenta) et du périodogramme de l'image vraie (en cyan). Le modèle ayant la plus forte probabilité est le LorentzC, en deuxième colonne.De haut en bas : superposition d'un périodogramme et de la DSP d'un modèle, coupe à la fréquence $v_p = 0$, coupe à la fréquence $u_p = 0$, coupe $u_p = -v_p$ (diagonale), coupe $u_p = v_p$ (anti-diagonale).

5.2 Déconvolution par filtrage de Wiener

Nous étudions dans cette section la manière dont la sélection de modèle affecte la déconvolution des données. Il n'existe pas de garantie que les modèles avec les plus fortes évidences donnent les meilleures images restaurées en terme de qualité. On s'attend cependant à ce qu'un modèle fort fournisse une déconvolution de meilleure qualité qu'un modèle faible. Dans la suite de cette section, nous exposons des preuves empiriques d'un telle relation sur deux plans distincts.

Visuellement : nous comparons à l'œil nu les images déconvoluées de chaque modèle à l'image vraie $\boldsymbol{x}^*$.

Quantitativement : nous quantifions l'erreur d'estimation de ces mêmes images déconvoluées à l'aide de trois mesures d'erreurs : les distances ℓ_1, ℓ_2 et l'indice de dissemblance Structural Dissimilarity (DSSIM).

Ici, la moyenne de $p(\boldsymbol{x}|\boldsymbol{y}, \bar{\gamma}_{\boldsymbol{x}}, \bar{\gamma}_{\boldsymbol{e}}, \mathscr{M} = k)$ correspond à un filtre de Wiener. Pour rappel, nous l'avons déjà défini page 52, lors du calcul des paramètres de la loi normale $p(\boldsymbol{x}|\boldsymbol{y}, \gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}}, \mathscr{M} = k)$:

$$\left| \begin{array}{rcl} \mathring{\Sigma}_{\boldsymbol{x}|*} & = & \left[\gamma_{\boldsymbol{e}} \mathring{\boldsymbol{H}}^{\dagger} \mathring{\boldsymbol{S}}_{e,j}{}^{-1} \mathring{\boldsymbol{H}} + \gamma_{\boldsymbol{x}} \mathring{\boldsymbol{S}}_{\boldsymbol{x},i}{}^{-1} \right]^{-1} \\ \mathring{\mu}_{\boldsymbol{x}|*} & = & \gamma_{\boldsymbol{e}} \mathring{\Sigma}_{\boldsymbol{x}|*} \mathring{\boldsymbol{H}}^{\dagger} \mathring{\boldsymbol{S}}_{e,j}{}^{-1} \mathring{\boldsymbol{y}} , \end{array} \right.$$

où $\mathring{\mu}_{\boldsymbol{x}|*}$ correspond à l'image déconvoluée. Il existe plusieurs stratégies de déconvolution.

1. Une déconvolution conditionnellement à $\mathscr{M} = k$, $\bar{\gamma}_{\boldsymbol{x}}$ et $\bar{\gamma}_{\boldsymbol{e}}$: une fois la chaîne de Gibbs terminée, on évalue le couple d'identités ci-dessus avec les valeurs $\bar{\gamma}_{\boldsymbol{x}}$ et $\bar{\gamma}_{\boldsymbol{e}}$ (ce que nous faisons par la suite).
2. Une déconvolution conditionnellement à $\mathscr{M} = k$ et marginalement à $\gamma_{\boldsymbol{x}}$ et $\gamma_{\boldsymbol{e}}$: les images déconvoluées tirées au cours de la chaîne de Gibbs sont moyennées

$$\bar{\boldsymbol{x}} = \frac{1}{G} \sum_{g=1}^{G} \boldsymbol{x}^{(g)} \qquad \text{où} \qquad \boldsymbol{x}^{(g)} \sim p(\boldsymbol{x}|\boldsymbol{y}, \mathscr{M} = k) .$$

Dans notre situation, ces deux stratégies donnent des images restaurées similaires. On le doit à la très faible variance des échantillons de $(\gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}})$ autour de leurs valeurs moyenne $(\bar{\gamma}_{\boldsymbol{x}}, \bar{\gamma}_{\boldsymbol{e}})$. Pour les deux stratégies, nous avons constaté empiriquement que

- d'une part, les mesures d'erreur entre leur image déconvoluée respective et l'image vraie sont identiques (à 10^{-5} près),
- d'autre part, les mesures d'erreur entre leurs deux images déconvoluées sont très proches de 0 (de 10^{-4} à 10^{-9}).

Remarque 5. *On aurait pu également déconvoluer marginalement à la variable de modèle $\mathscr{M}$, avec un algorithme de type RJMCMC [Gre95] par exemple. Dans notre cas, tout le poids des probabilités est souvent accaparé par un seul des modèles. Une déconvolution, marginalement ou non par rapport à ces derniers, donnerait approximativement le même résultat.*

La Table 5.4 exhibe les mesures d'erreur obtenues entre l'image vraie et l'image déconvoluée

provenant des différents modèles candidats. Nous y rappelons les valeurs de log-évidences et de probabilités *a posteriori* obtenues pour chaque modèle.

Modèle k	ℓ_1	ℓ_2	DSSIM	$\log f(\boldsymbol{y}\|\mathscr{M} = k)$	$p(\mathscr{M} = k\|\boldsymbol{y})$
1	0,09961	0,12181	0,13199	1 286,49338	0
2	0,092644	0,10899	0,12146	**2 544,82857**	**1**
3	**0,085501**	**0,10521**	**0,1038**	2 524,53629	10^{-9}
4	0,090721	0,10945	0,11552	2 341,18512	10^{-89}
5	0,29945	0,34425	0,30199	-12 769,583	0

TABLE 5.4 – Mesures d'erreur entre les images déconvoluées à partir des différents modèles et la vraie. La valeur minimale de chaque distance est repérée par la couleur cyan. Les log-évidences et les probabilités *a posteriori* des modèles sont rappelées dans les deux dernières colonnes. La valeur maximale de ces deux colonnes est repérée par la couleur bleue.

On constate les mêmes tendances que les valeurs de la log-évidence : les modèles 2, 3 et 4 ont des valeurs très proches et présentent les meilleurs résultats, le modèle 1 est un peu plus éloigné, mais pas autant que le modèle 5, qui présente les pires résultats..

Constat remarquable, on observe d'une part que le modèle $k = 3$ (InterPix-4) minimise les trois mesures et d'autre part qu'il ne s'agit pas du modèle sélectionné plus tôt $k = 2$. Cependant, il convient de remarquer que ce modèle $k = 2$, ainsi que le modèle $k = 4$, ont des valeurs très proches des valeurs minimales. Ce sont ces mêmes trois modèles dont les log-évidences sont avoisinantes.

Les images déconvoluées des trois modèles avec les plus faibles mesures d'erreur sont affichées sur la Figure 5.6, en regard de l'image vraie et des données. Le flou a été annihilé et on perçoit un résidu de bruit. Il demeure très difficile de distinguer les trois modèles à l'œil nu, et encore plus difficile de déterminer lequel est le plus proche de l'image vraie.

À l'inverse, comme en témoigne la Figure 5.7, les images déconvoluées des modèles LorentzS et Blanc présentent de gros défauts visuels, comme notamment l'apparition de motifs. Le plus marqué est le modèle Blanc, dont on comprend les mesures d'erreur plus importantes.

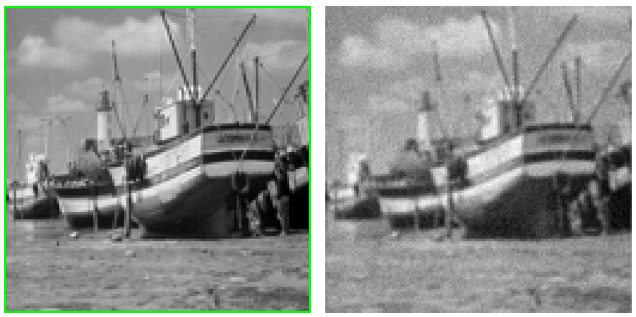
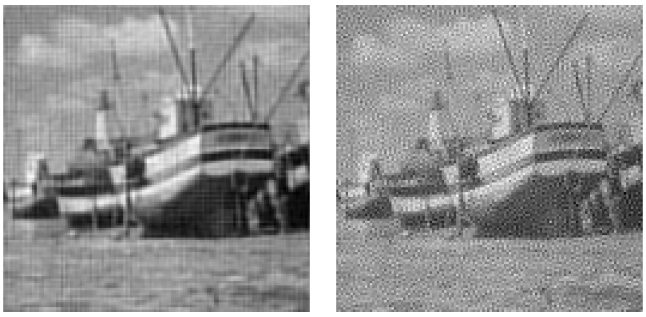


Image vraie $\boldsymbol{x}^*$ Données observées $\boldsymbol{y}$

LorentzC InterPix-4 InterPix-8

FIGURE 5.6 – Images déconvoluées provenant des modèles avec les plus faibles mesures d'erreur. L'image cadrée de vert correspond à l'image vraie.



LorentzS Blanc

FIGURE 5.7 – Images déconvoluées provenant des modèles avec les plus fortes mesures d'erreur.

5.3 Bilan du chapitre

Ce chapitre enrichit l'étude du chapitre 4, en appliquant l'approche sur des données réelles. L'image $\boldsymbol{x}$ et les données $\boldsymbol{y}$ ne sont plus des réalisations stochastiques d'un modèle spécifique de DSP. De ce fait, il n'existe pas de modèle vrai et les modèles que nous considérons sont par conséquent mal spécifiés. L'état d'esprit de ce chapitre est différent des chapitres précédents : il s'agissait de voir ici quelles informations délivraient les évidences sur le degré d'adéquation des différents modèles testés avec les données observées d'une part, puis avec l'image originale d'autre part.

Bien que le contexte soit différent, l'approche ne change pas. Nous avons appliqué à l'identique la méthode et l'algorithme de Chib tel que présenté au chapitre 4 pour calculer les évidences. Nous

avons constaté que trois modèles se démarquaient par leurs fortes évidences : LorentzC, InterPix-4 et InterPix-8. Puis nous avons calculé les probabilités *a posteriori* des modèles et sélectionné celui qui les maximisaient.

L'étude s'est ensuite portée sur une comparaison des contenus fréquentiels.

1. Dans un premier temps, nous avons comparé le périodogramme des données aux différentes DSP des modèles testés. Les courbes observées étaient en adéquation avec les valeurs de log-évidences : les modèles ayant les plus fortes évidences sont très proches du périodogramme des données, et inversement.
2. Nous avons ensuite comparé le périodogramme de l'image originale avec la DSP objet des différents modèles testés. Sans surprise, on y fit les mêmes constats quand au degré d'adéquation. On y observe une adéquation moins forte, à cause du fait que l'approche de sélection de modèle vise à sélectionner une modèle à partir des données, et non de l'image.

Pour conclure, nous nous sommes intéressé au lien qu'il pouvait y avoir entre la sélection de modèle et la déconvolution des données. Pour ce faire, nous avons utilisé le filtre de Wiener qui apparait dans l'algorithme de Chib. Pour évaluer la qualité de la déconvolution, nous avons analysé visuellement les images déconvoluées et quantifié l'erreur à l'aide de mesures de distances. Nous avons fait d'intéressants constats.

- Le modèle qui minimise les erreurs (InterPix-4) n'est pas celui qui a la plus forte évidence (LorentzC).
- Cependant, les trois modèles avec les plus fortes évidences (LorentzC, InterPix-4 et InterPix-8) sont également ceux qui conduisent aux plus faibles erreurs. Que ce soit les évidences ou les erreurs, les valeurs pour ces trois modèles sont relativement proches. De plus, les images déconvoluées avec ces modèles sont très similaires. Il est impossible de définir à l'œil nu laquelle présente le moins d'erreurs.
- A l'inverse, on a observé que les deux modèles dont les log-évidences sont faibles (LorentzS et Blanc) donnaient de plus grandes erreurs. Visuellement, leurs images déconvoluées sont de mauvaises qualités, comparé aux modèles précédents.

Ce chapitre conclut l'avancée des travaux de ce manuscrit. Le chapitre suivant présente un bilan global et évoque les principales difficultés rencontrées. Nous discutons de quelques perspectives pour continuer à enrichir ces travaux.

CHAPITRE 6

Conclusion : bilan et perspectives

6.1 Bilan

Ce manuscrit fait état de l'avancée de nos travaux sur la sélection de modèle appliquée à la restauration d'image. Pour aborder ce problème encore ouvert, nous nous sommes placés dans le cas gaussien circulant. Les simplifications calculatoires qu'il amène ont permis d'apprécier au mieux les difficultés liées au calcul de l'évidence : les dépendances complexes entre les variables, leurs marginalisations, la gestion des ressources, *etc*. Nous avons travaillé sur observation directe, puis indirecte, avec un certain nombre de paramètres inconnus. Dans ces cas d'étude, nous avons apportés quelques contributions.

- La contribution majeure est l'application de l'algorithme de Chib pour calculer des évidences et sélectionner des modèles dans le cas en observation indirecte avec paramètres de niveaux inconnus. Il s'agit du cas le plus complexe que nous ayons traitée et nous y avons obtenus d'excellents taux de bonnes sélections. Nous avons réussi à appliquer un algorithme stochastique complexe et plus efficace que les algorithmes dit naïfs. Sa fiabilité en fait l'algorithme a privilégier pour opérer la sélection de modèles.

- Nous avons fourni une interprétation aussi bien qualitative que quantitative de la sélection de modèles de DSP d'image et de bruit. Des pseudos-distances entre le périodogramme des données observées et sa DSP sont calculées et les modèles qui ne minimisent pas ces distances sont discriminés.

- Nous avons prouvé analytiquement que la moyenne harmonique, dans sa forme la plus naïve, ne converge pas vers l'évidence, que ce soit en observation directe ou indirecte.

La section suivante discute de quelques perspectives et pistes de recherche pour enrichir ces travaux.

6.2 Perspectives

Le problème étant encore ouvert, les perspectives de recherche sont nombreuses sur différents aspects du problème.

À propos de la sélection de modèle : sans trop s'éloigner de la configuration actuelle, on pourrait également inclure le modèle de gain aux modèles inconnus, il n'y aurait en soit aucune difficulté supplémentaire. En revanche, il serait intéressant mais plus ardu de considérer des modèles plus complexes, comme par exemples des modèles emboîtés. Le problème récurrent de ces types de modèle est la tendance des algorithmes à sélectionner le modèle le plus riche. Le point fort de l'approche bayésienne suivie ici est qu'elle pénalise la complexité du modèle lors du calcul de l'évidence $f(\boldsymbol{y}|\mathscr{M} = k)$. De manière plus générale, on pourrait se concentrer sur des configurations où le nombre d'hyperparamètres inconnus varie d'un modèle à l'autre.

À propos des hyperparamètres : considérer d'autres hyperparamètres inconnus de nature différentes, comme la largeur de DSP image $w_{\boldsymbol{x}}$, la largeur de DSP bruit $w_{\boldsymbol{e}}$, ou encore la largeur du gain utilisé, qui donnent des informations essentielles sur la décroissance fréquentielle.

À propos des algorithmes : d'autres algorithmes de calcul d'évidence peuvent se révéler prometteurs. La moyenne harmonique tronquée, connue pour évacuer le problème de convergence rencontrée avec la MH naïve, est à considérer. Également, il existe une autre version de l'algorithme de Chib, couplé cette fois-ci à un échantillonneur de Metropolis-Hastings. Enfin, l'étude des power posteriors n'a pas été terminée, entravée par des difficultés encore obscures.

Nous allons maintenant aborder plus en détails les problèmes rencontrés avec les deux contributions inachevées.

6.2.1 Discussion sur la largeur $w_{\boldsymbol{x}}$

Une fois l'algorithme de Chib validé pour le cas observation indirecte avec $\gamma_{\boldsymbol{x}}$ et $\gamma_{\boldsymbol{e}}$ inconnus, nous voulions progresser en complexité en ajoutant un troisième paramètre : la largeur $w_{\boldsymbol{x}}$. L'inclusion de ce paramètre dans la boucle d'échantillonnage de Gibbs ne pose aucun souci. Le coût calculatoire augmente à cause de la conditionnelle *a posteriori* de $w_{\boldsymbol{x}}$, qui n'a pas la forme d'une loi usuelle. Nous avons choisi d'obtenir des échantillons de $w_{\boldsymbol{x}}$ par inversion de la fonction de répartition. Les échantillons obtenus fluctuent « normalement » autour de la valeur vraie $w_{\boldsymbol{x}}^*$ (lorsque le modèle vrai est testé).

Cependant, on observe une certaine instabilité au moment du calcul de la quantité

$$p(\bar{\gamma}_{\boldsymbol{x}}, \bar{\gamma}_{\boldsymbol{e}}, \bar{w}_{\boldsymbol{x}}|\boldsymbol{y}, \mathscr{M} = k) = \frac{1}{G}\sum_{g=1}^{G} f(\bar{\gamma}_{\boldsymbol{x}}, \bar{\gamma}_{\boldsymbol{e}}, \bar{w}_{\boldsymbol{x}}|\boldsymbol{y}, \boldsymbol{x}^{(g)}, \mathscr{M} = k).$$

À cause de la forte variance des échantillons $\log f(\bar{\gamma}_{\boldsymbol{x}}, \bar{\gamma}_{\boldsymbol{e}}, \bar{w}_{\boldsymbol{x}}|\boldsymbol{y}, \boldsymbol{x}^{(g)}, \mathscr{M} = k)$, la moyenne ci-dessus explose numériquement. Des pistes ont été explorées pour tenter de diminuer cette variance,

comme notamment choisir une valeur de $\bar{w}_{\boldsymbol{x}}$ autre que la moyenne empirique des tirages du Gibbs ou encore changer de variable latente.

Face à cette difficulté, nous avons pris du recul et tenté de l'appliquer en observation directe. Dans cette configuration, nous avons considéré $w_{\boldsymbol{x}}$ comme variable latente. Cela permettait d'avoir une expression analytique de $f(\gamma_{\boldsymbol{x}}|\boldsymbol{x}, w_{\boldsymbol{x}}, \mathscr{M} = i)$, qui s'avère être une loi Gamma. Malgré ceci, les valeurs obtenues de $\log f(\boldsymbol{x}|\mathscr{M} = k)$ n'étaient pas consistantes. A l'inverse, en considérant $\gamma_{\boldsymbol{x}}$ comme variable auxiliaire, on obtenait des valeurs plus stables, mais cela sollicite des intégrations numériques supplémentaires.

6.2.2 Discussion sur les power posteriors

Cet algorithme a présenté quelques problèmes de convergence. Dans un premier temps, nous avons traité le cas observation indirecte avec paramètres de niveaux connus. Cette configuration permet d'avoir une expression analytique de l'espérance. Seule l'intégrale sur λ est à calculer numériquement et nous obtenons des valeurs stables. Avec cette valeur de référence, on peut alors étudier les propriétés de convergence de l'algorithme, en fonction des différents facteurs tels que la finesse de la grille et le nombre de tirages.

Les difficultés se sont manifestées dès la considération des paramètres de niveaux dans les inconnus, que ce soit en observation directe ou indirecte. Lorsque la variable de température λ tend vers 0, la distribution $p_\lambda(\boldsymbol{\theta}|\boldsymbol{y}, \mathscr{M} = k)$ tend à se comporter comme un prior, prior que nous avons choisi non-informatif. Pour ces valeurs de λ, il faudrait donc un nombre de tirages plus conséquent, afin de parcourir suffisamment l'espace des hyperparamètres et que la moyenne empirique converge. À l'opposé, lorsque λ se rapproche de 1, la loi p_λ tend vers la loi *a posteriori*. Les échantillons *a posteriori* ciblent alors les zones a forte densité, et nécessite moins de tirages que pour un moyennage sous prior. Une solution pour assurer la convergence serait

- une grille adaptative de λ, avec un nombre de points plus importants lorsque λ tend vers 1,
- un nombre de tirage dynamique, qui serait plus important pour les valeurs de λ proches de 0 que celles proches de 1.

ANNEXE A

Algorithme de Chib en observation indirecte (section 4.3.3)

A.1 Loi conditionnelle *a posteriori* de $\gamma_{\boldsymbol{x}}$

Nous allons démontrer ici que la conditionnelle *a posteriori* de $\gamma_{\boldsymbol{x}}$ n'est pas fonction des données $\boldsymbol{y}$ (section 4.3.3.1) :

$$p(\gamma_{\boldsymbol{x}}|\boldsymbol{y},\boldsymbol{x},\gamma_{\boldsymbol{e}},\mathscr{M} = k) = \frac{f(\boldsymbol{y},\boldsymbol{x},\gamma_{\boldsymbol{x}},\gamma_{\boldsymbol{e}}|\mathscr{M} = k)}{f(\boldsymbol{y},\boldsymbol{x},\gamma_{\boldsymbol{e}}|\mathscr{M} = k)}.$$

Nous avons déjà explicité la décomposition de la loi jointe au numérateur à l'équation (4.4) page 43 suivant la hiérarchie des dépendances Figure 1.8 page 13. Toujours suivant cette hiérarchie, le dénominateur se décompose

$$f(\boldsymbol{y},\boldsymbol{x},\gamma_{\boldsymbol{e}}|\mathscr{M} = k) = f(\boldsymbol{y}\,|\boldsymbol{x},\gamma_{\boldsymbol{e}},\mathscr{M} = k)\,\pi(\boldsymbol{x}|\mathscr{M} = k)\pi(\gamma_{\boldsymbol{e}})\,.$$

Des termes communs apparaissent au numérateur et au dénominateur et se simplifient :

$$\begin{aligned} p(\gamma_{\boldsymbol{x}}|\boldsymbol{y},\boldsymbol{x},\gamma_{\boldsymbol{e}},\mathscr{M} = k) &= \frac{\pi(\boldsymbol{x}|\gamma_{\boldsymbol{x}},\mathscr{M} = k)\,\pi(\gamma_{\boldsymbol{x}})}{\pi(\boldsymbol{x}|\mathscr{M} = k)}\,\frac{f(\boldsymbol{y}|\boldsymbol{x},\gamma_{\boldsymbol{e}},\mathscr{M} = k)}{f(\boldsymbol{y}|\boldsymbol{x},\gamma_{\boldsymbol{e}},\mathscr{M} = k)}\,\frac{\pi(\gamma_{\boldsymbol{e}})}{\pi(\gamma_{\boldsymbol{e}})} \\ &= \frac{\pi(\boldsymbol{x},\gamma_{\boldsymbol{x}}|\mathscr{M} = k)}{\pi(\boldsymbol{x}|\mathscr{M} = k)} \\ &= p(\gamma_{\boldsymbol{x}}|\boldsymbol{x},\mathscr{M} = k)\,. \end{aligned}$$

Comme annoncée, la conditionnelle *a posteriori* $p(\gamma_{\boldsymbol{x}}|\boldsymbol{y},\boldsymbol{x},\gamma_{\boldsymbol{e}},\mathscr{M} = k)$, qui est aussi égale à la conditionnelle *a priori* de $\gamma_{\boldsymbol{x}}$, n'est pas fonction des données $\boldsymbol{y}$.

A.2 Choix de la variable latente

Dans cette annexe, nous allons aborder la question du choix de la variable latente pour l'algorithme de Chib et étayer notre choix. En observation indirecte (chapitre 4) avec les paramètres de niveaux inconnus (section 4.3), nous avons en théorie trois possibilités

1. $\theta = [\boldsymbol{x}, \gamma_{\boldsymbol{e}}]$ et $a = \gamma_{\boldsymbol{x}}$,
2. $\theta = [\boldsymbol{x}, \gamma_{\boldsymbol{x}}]$ et $a = \gamma_{\boldsymbol{e}}$,
3. $\theta = [\gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}}]$ et $a = \boldsymbol{x}$ (configuration finalement choisie).

Nous avons vu à la section 4.3.3.1 que la configuration 3 permettait d'expliciter exactement la distribution d'intérêt $p(\theta|\boldsymbol{y}, a, \mathscr{M} = k)$, grâce notamment à l'indépendance *a posteriori* des variables $\gamma_{\boldsymbol{x}}$ et $\gamma_{\boldsymbol{e}}$. Nous allons maintenant étudier la distribution $p(\theta|\boldsymbol{y}, a, \mathscr{M} = k)$ pour les configurations 1 et 2.

1. Avec $\theta = [\boldsymbol{x}, \gamma_{\boldsymbol{e}}]$ et $a = \gamma_{\boldsymbol{x}}$, la loi $p(\theta|\boldsymbol{y}, a, \mathscr{M} = k)$ devient

$$\begin{aligned} p(\boldsymbol{x}, \gamma_{\boldsymbol{e}}|\boldsymbol{y}, \gamma_{\boldsymbol{x}}, \mathscr{M} = k) &= \frac{f(\boldsymbol{y}, \boldsymbol{x}, \gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}}|\mathscr{M} = k)}{f(\boldsymbol{y}, \gamma_{\boldsymbol{x}}|\mathscr{M} = k)} \\ &= \frac{f(\boldsymbol{y}|\boldsymbol{x}, \gamma_{\boldsymbol{e}}, \mathscr{M} = k)\, \pi(\boldsymbol{x}|\gamma_{\boldsymbol{x}}, \mathscr{M} = k)\, \pi(\gamma_{\boldsymbol{x}})\, \pi(\gamma_{\boldsymbol{e}})}{f(\boldsymbol{y}|\gamma_{\boldsymbol{x}}, \mathscr{M} = k)\, \pi(\gamma_{\boldsymbol{x}})} \\ &= \frac{f(\boldsymbol{y}|\boldsymbol{x}, \gamma_{\boldsymbol{e}}, \mathscr{M} = k)\, \pi(\boldsymbol{x}|\gamma_{\boldsymbol{x}}, \mathscr{M} = k)\, \pi(\gamma_{\boldsymbol{e}})}{f(\boldsymbol{y}|\gamma_{\boldsymbol{x}}, \mathscr{M} = k)} \end{aligned}$$

   où le dénominateur est la loi $f(\boldsymbol{y}, \gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}}|\mathscr{M} = k)$ marginalisé par rapport à $\gamma_{\boldsymbol{e}}$.

2. Avec $\theta = [\boldsymbol{x}, \gamma_{\boldsymbol{x}}]$ et $a = \gamma_{\boldsymbol{e}}$, elle s'explicite

$$\begin{aligned} p(\boldsymbol{x}, \gamma_{\boldsymbol{x}}|\boldsymbol{y}, \gamma_{\boldsymbol{e}}, \mathscr{M} = k) &= \frac{f(\boldsymbol{y}, \boldsymbol{x}, \gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}}|\mathscr{M} = k)}{f(\boldsymbol{y}, \gamma_{\boldsymbol{e}}|\mathscr{M} = k)} \\ &= \frac{f(\boldsymbol{y}|\boldsymbol{x}, \gamma_{\boldsymbol{e}}, \mathscr{M} = k)\, \pi(\boldsymbol{x}|\gamma_{\boldsymbol{x}}, \mathscr{M} = k)\, \pi(\gamma_{\boldsymbol{x}})\, \pi(\gamma_{\boldsymbol{e}})}{f(\boldsymbol{y}|\gamma_{\boldsymbol{e}}, \mathscr{M} = k)\, \pi(\gamma_{\boldsymbol{e}})} \\ &= \frac{f(\boldsymbol{y}|\boldsymbol{x}, \gamma_{\boldsymbol{e}}, \mathscr{M} = k)\, \pi(\boldsymbol{x}|\gamma_{\boldsymbol{x}}, \mathscr{M} = k)\, \pi(\gamma_{\boldsymbol{e}})}{f(\boldsymbol{y}|\gamma_{\boldsymbol{e}}, \mathscr{M} = k)} \end{aligned}$$

   où le dénominateur est la loi $f(\boldsymbol{y}, \gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}}|\mathscr{M} = k)$ marginalisé par rapport à $\gamma_{\boldsymbol{x}}$.

Dans les deux cas, les numérateurs sont connus analytiquement, ce qui n'est en revanche pas le cas de leurs dénominateurs. Nous avons vu page 45 en remplaçant la valeur de $\mathring{s}_{\boldsymbol{y}}(p)$ dans l'identité (4.10) qu'une marginalisation analytique des paramètres n'était plus possible. Pour obtenir les dénominateurs, il faudrait donc intégrer numériquement la loi $f(\boldsymbol{y}, \gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}}|\mathscr{M} = k)$ ce qui, dans notre cas, ferait perdre l'intérêt de l'algorithme. La configuration 3 présente cet avantage majeur d'avoir une expression analytique de la distribution $p(\gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{e}}|\boldsymbol{y}, \boldsymbol{x}, \mathscr{M} = k)$.

ANNEXE B

Approche LogSumExp

Dans notre étude, nous faisons face à des valeurs numérique importantes, dépassant les capacités de la machine. À de nombreuses reprises, le calcul de l'évidence fait intervenir une ou plusieurs sommes discrètes. Nous présentons la technique de manière générique, où l'évidence se calcule comme une somme de termes $\Upsilon_g(k)$, fonction du modèle k. Par simplicité calculatoire, on calcule dans un premier temps le logarithme, puis l'exponentielle :

$$f(\boldsymbol{y}|\mathscr{M} = k) = \sum_g \Upsilon_g(k) = \sum_g \exp\left[\log \Upsilon_g(k)\right]. \tag{B.1}$$

Pour éviter les explosions numériques lors du passage à l'exponentielle (à cause de trop grandes valeurs), on a recours à une astuce mathématique. On commence par extraire le maximum des $\log \Upsilon_g(k)$:

$$\Delta(k) = \max_g \{ \log \Upsilon_g(k) \}.$$

Il ne dépend pas de l'indice g, mais dépend en revanche de k, le modèle considéré.

Le calcul de l'évidence (B.1) peut alors se reformuler

$$\begin{aligned} f(\boldsymbol{y}|\mathscr{M} = k) &= \sum_g \exp\left[\log \Upsilon_g(k) - \Delta(k)\right] \exp\left[\Delta(k)\right] \\ &= \exp\left[\Delta(k)\right] \sum_g \exp\left[\log \Upsilon_g(k) - \Delta(k)\right] \end{aligned} \tag{B.2}$$

Pour alléger les écritures, on notera

$$\tilde{e}(k) = \sum_g \exp\left[\log \Upsilon_g(k) - \Delta(k)\right].$$

Mathématiquement, les identités (B.1) et (B.2) sont strictement égales. Numériquement, la formule (B.2) permet de ramener l'argument de l'exponentielle dans l'intégrale sur l'intervalle $]-\infty, 0]$, et donc de borner les valeurs de cette même exponentielle entre 0 et 1. Cela permet d'éviter de

sommer des valeurs à $+\infty$.

Après avoir déterminé séparément $\Delta(k)$ et $\tilde{e}(k)$ pour chaque modèle candidat, on calcule les probabilités *a posteriori* (1.15)

$$\begin{aligned} p(\mathscr{M}=k|\boldsymbol{y}) &= \frac{f(\boldsymbol{y}|\mathscr{M}=k)}{\sum_{k'=1}^{K} f(\boldsymbol{y}|\mathscr{M}=k')} = \frac{\tilde{e}(k)\exp\left[\Delta(k)\right]}{\sum_{k'=1}^{K} \tilde{e}(k')\exp\left[\Delta(k')\right]} \\ &= \frac{\tilde{e}(k)}{\sum_{k'=1}^{K} \tilde{e}(k')\exp\left[\Delta(k')-\Delta(k)\right]}, \end{aligned} \tag{B.3}$$

évitant ainsi toute explosion numérique.

Remarque 6. *Les valeurs toujours élevées de $\Delta(k)$ empêchent de calculer directement l'évidence, et il est nécessaire de pouvoir contrôler le bon déroulement des calculs. Nous travaillerons de ce fait avec le logarithme de l'évidence :*

$$\log f(\boldsymbol{y}|\mathscr{M} = k) = \Delta(k) + \log \tilde{e}(k).$$

Il s'agit de résultats cruciaux pour la sélection de modèles. Ils ont été utilisés pour obtenir tous les résultats numériques qui se trouvent aux sections 3.3 et 4.4, ainsi qu'au chapitre 5.

ANNEXE C

Simulation de l'image, du bruit et des données

Pour simuler l'image, on génère $\mathring{\boldsymbol{x}} = \boldsymbol{F}\boldsymbol{x}$ puis on obtient $\boldsymbol{x}$ par FFT inverse. Pour rappel, la variable $\boldsymbol{x}$ suit la loi

$$\mathscr{N}(\boldsymbol{x}\,;\,\boldsymbol{0},\Sigma_{\boldsymbol{x}})\,.$$

La TFD étant une opération linéaire, la loi de $\mathring{\boldsymbol{x}}$ est également une gaussienne. On peut déterminer sa moyenne

$$\mathbb{E}[\mathring{\boldsymbol{x}}] = \boldsymbol{F}\mathbb{E}[\boldsymbol{x}] = 0$$

et sa covariance (d'après (1.8))

$$\Sigma_{\mathring{\boldsymbol{x}}} = \mathbb{E}[\mathring{\boldsymbol{x}}\mathring{\boldsymbol{x}}^{\dagger}] = \boldsymbol{F}\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\dagger}]\boldsymbol{F}^{\dagger} = \boldsymbol{F}\,\Sigma_{\boldsymbol{x}}\,\boldsymbol{F}^{\dagger} = \gamma_{\boldsymbol{x}}^{-1}\,\mathring{\boldsymbol{S}}_{\boldsymbol{x}}.$$

Comme $\Sigma_{\boldsymbol{x}}$ est une matrice Circulante-Bloc-Circulante, $\mathring{\boldsymbol{S}}_{\boldsymbol{x}}$ est diagonale. Par conséquent, les covariances inter-fréquences sont nulles, ce qui implique que les variables $\mathring{x}(p)$ sont indépendantes. Il y a ici équivalence entre indépendance et décorrélation car nous sommes dans le cas gaussien. Elles ne sont en revanche pas identiquement distribuées et suivent chacune une distribution normale centrée, de variance respective $\gamma_{\boldsymbol{x}}^{-1}\,\mathring{s}_{\boldsymbol{x}}(p)$.

Pour obtenir $\mathring{\boldsymbol{x}}$, il suffit de simuler pour toutes les fréquences p

$$\mathring{x}(p) \sim \mathscr{N}(0, \gamma_{\boldsymbol{x}}^{-1}\mathring{s}_{\boldsymbol{x}}(p))\,.$$

Enfin, on déduit l'image $\boldsymbol{x} = \boldsymbol{F}^{\dagger}\mathring{\boldsymbol{x}}$ par FFT inverse.

Pour obtenir une réalisation de bruit $\boldsymbol{e}$, on procède de la même manière. Le bruit $\boldsymbol{e}$ est, tout comme $\boldsymbol{x}$, une variable gaussienne centrée de covariance (toujours d'après (1.8))

$$\boldsymbol{R}_{\mathring{\boldsymbol{e}}} = \gamma_{\boldsymbol{x}}^{-1}\,\mathring{\boldsymbol{S}}_{\boldsymbol{e},j}.$$

dont chaque fréquence p se simule

$$\mathring{e}(p) \sim \mathscr{N}(0, \gamma_{\boldsymbol{e}}^{-1}\mathring{s}_{\boldsymbol{e},j}(p))\,.$$

Pour générer les données observées, nous allons procéder de même. On construit la FFT des données à partir de l'équation

$$\mathring{\boldsymbol{y}} = \mathring{\boldsymbol{H}}\mathring{\boldsymbol{x}} + \mathring{\boldsymbol{e}}\,,$$

on multiplie l'image $\mathring{\boldsymbol{x}}$ par le gain $\mathring{\boldsymbol{H}}$ et on ajoute le bruit $\mathring{e}$. Enfin, on calcule $\boldsymbol{y} = \boldsymbol{F}^{\dagger}\mathring{\boldsymbol{y}}$ par FFT inverse.

ANNEXE D

Matrices de Tœplitz et matrices circulantes

Dans les problèmes mono-dimensionnels, une matrice est dite de Tœplitz lorsque ses diagonales sont à coefficients constants, c-à-d que chaque ligne est issue d'un décalage de la précédente. Une telle matrice est de la forme

$$\begin{bmatrix} b_{i,0} & b_{i,1} & \dots & \dots & b_{i,N-1} \\ b_{i,-1} & b_{i,0} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & b_{i,1} \\ b_{i,-(N-1)} & \dots & \dots & b_{i,-1} & b_{i,0} \end{bmatrix}.$$

Il existe un cas particulier de matrice de Tœplitz où chaque ligne est issue un décalage **circulaire** de la première : la matrice circulante. Elle est de la forme

$$\begin{bmatrix} b_{i,0} & b_{i,1} & \dots & \dots & b_{i,N-1} \\ b_{i,N-1} & b_{i,0} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & b_{i,1} \\ b_{i,1} & \dots & \dots & b_{i,N-1} & b_{i,0} \end{bmatrix}.$$

Pour les problèmes de dimension 2, il existe des matrices où les cœfficients peuvent être regroupés en sous-matrice. Parmi ces matrices, il en existe un type où

- les blocs sont agencés sous la forme d'une matrice de Tœplitz, c'est-à-dire que chaque ligne de bloc est une version décalée de la première ligne de bloc,
- chaque bloc est une matrice de Tœplitz.

On les nomme matrice Tœplitz-Bloc-Tœplitz et sont de la forme suivante :

$$\begin{pmatrix} \boldsymbol{B}_0 & \boldsymbol{B}_1 & \dots & \dots & \boldsymbol{B}_{N-1} \\ \boldsymbol{B}_{-1} & \boldsymbol{B}_0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \boldsymbol{B}_1 \\ \boldsymbol{B}_{-(N-1)} & \dots & \dots & \boldsymbol{B}_{-1} & \boldsymbol{B}_0 \end{pmatrix}$$

A l'instar des matrices de Tœplitz simples, il existe également un cas particulier de matrice Tœplitz-Bloc-Tœplitz, où

- les blocs sont agencés sous la forme d'une matrice circulante, avec un décalage circulaire de la première ligne de bloc,
- chaque bloc est une matrice circulante.

Ce sont des matrices dites Circulantes-Bloc-Circulantes et sont de la forme

$$\begin{pmatrix} \boldsymbol{B}_0 & \boldsymbol{B}_1 & \dots & \dots & \boldsymbol{B}_{N-1} \\ \boldsymbol{B}_{N-1} & \boldsymbol{B}_0 & \ddots & & \boldsymbol{B}_{N-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \boldsymbol{B}_1 \\ \boldsymbol{B}_1 & \dots & \dots & \boldsymbol{B}_{N-1} & \boldsymbol{B}_0 \end{pmatrix}$$

où les $\boldsymbol{B}_n$ sont des matrices circulantes.

ANNEXE E

Power posteriors et WBIC

E.1 Démonstration de l'identité (2.14)

Nous détaillons ici la démonstration de l'identité phare de l'algorithme des power posteriors et de WBIC. Pour obtenir l'équation

$$\log f(\boldsymbol{y}|\mathscr{M} = k) = \int_0^1 \mathbb{E}_{p_\lambda}\Big[\log f(\boldsymbol{y}|\theta,\mathscr{M} = k)\Big]\mathrm{d}\lambda = \log\frac{\varepsilon_1(\boldsymbol{y})}{\varepsilon_0(\boldsymbol{y})}, \tag{2.14}$$

il faut dériver le logarithme de la fonction

$$\varepsilon_\lambda(\boldsymbol{y}) = \int_\Theta f(\boldsymbol{y}|\theta,\mathscr{M} = k)^\lambda\, \pi(\theta|\mathscr{M} = k)\,\mathrm{d}\theta$$

par rapport au paramètre de température λ :

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\lambda}\log\varepsilon_\lambda(\boldsymbol{y}) &= \frac{1}{\varepsilon_\lambda(\boldsymbol{y})}\frac{\mathrm{d}}{\mathrm{d}\lambda}\,\varepsilon_\lambda(\boldsymbol{y}) \\ &= \frac{1}{\varepsilon_\lambda(\boldsymbol{y})}\frac{\mathrm{d}}{\mathrm{d}\lambda}\int_\theta f(\boldsymbol{y}|\theta,\mathscr{M} = k)^\lambda\, \pi(\theta|\mathscr{M} = k)\,\mathrm{d}\theta\,.\end{aligned}$$

On dérive le terme $f(\boldsymbol{y}|\theta,\mathscr{M} = k)^\lambda$ par rapport λ

$$\begin{aligned}&= \frac{1}{\varepsilon_\lambda(\boldsymbol{y})}\int_\theta \frac{\mathrm{d}}{\mathrm{d}\lambda}\Big[f(\boldsymbol{y}|\theta,\mathscr{M} = k)^\lambda\Big]\, \pi(\theta|\mathscr{M} = k)\,\mathrm{d}\theta \\ &= \frac{1}{\varepsilon_\lambda(\boldsymbol{y})}\int_\theta f(\boldsymbol{y}|\theta,\mathscr{M} = k)^\lambda\, \log\big[f(\boldsymbol{y}|\theta,\mathscr{M} = k)\big]\, \pi(\theta|\mathscr{M} = k)\,\mathrm{d}\theta\end{aligned}$$

$$= \int_{\theta} \frac{f(\boldsymbol{y}|\theta, \mathscr{M} = k)^{\lambda}\, \pi(\theta|\mathscr{M} = k)}{\varepsilon_{\lambda}(\boldsymbol{y})} \log\left[f(\boldsymbol{y}|\theta, \mathscr{M} = k)\right] \mathrm{d}\theta\,.$$

On combine ensuite $f(\boldsymbol{y}|\theta, \mathscr{M} = k)^{\lambda}$, $\pi(\theta|\mathscr{M} = k)$ et $\varepsilon_{\lambda}(\boldsymbol{y})$ pour construire la distribution $p_{\lambda}(\theta|\boldsymbol{y}, \mathscr{M} = k)$ et faire apparaître une espérance

$$\begin{aligned} &= \int_{\theta} p_{\lambda}(\theta|\boldsymbol{y}, \mathscr{M} = k) \log\left[f(\boldsymbol{y}|\theta, \mathscr{M} = k)\right] \mathrm{d}\theta \\ &= \mathbb{E}_{p_{\lambda}}\Big[\log f(\boldsymbol{y}|\theta, \mathscr{M} = k)\Big]\,. \end{aligned}$$

En intégrant ensuite l'identité précédente sur l'intervalle de définition de λ, on retrouve l'équation (2.14)

$$\log \varepsilon_1(\boldsymbol{y}) - \log \varepsilon_0(\boldsymbol{y}) = \int_0^1 \mathbb{E}_{p_{\lambda}}\Big[\log f(\boldsymbol{y}|\theta, \mathscr{M} = k)\Big] \mathrm{d}\lambda\,.$$

E.2 Démonstration de l'identité (2.16)

La valeur $\lambda^{\star}$ est définie telle que la distribution $p_{\lambda^{\star}} = p_{\lambda^{\star}}(\boldsymbol{y}|\theta, \mathscr{M} = k)$ soit équidistant du prior $\pi(\theta|\mathscr{M} = k)$ et du posterior $p(\theta|\boldsymbol{y}, \mathscr{M} = k)$ au sens de Kullback-Leibler. Les lignes suivantes sont équivalentes :

$$\begin{aligned} d_{KL}\big[\, p_{\lambda^{\star}}\,,\, p(\theta|\boldsymbol{y}, \mathscr{M} = k)\,\big] &= d_{KL}\big[\, p_{\lambda^{\star}}\,,\, \pi(\theta|\mathscr{M} = k)\,\big] \\ \int_{\Theta} p_{\lambda^{\star}} \log\left[\frac{p_{\lambda^{\star}}}{p(\theta|\boldsymbol{y}, \mathscr{M} = k)}\right] \mathrm{d}\theta &= \int_{\Theta} p_{\lambda^{\star}} \log\left[\frac{p_{\lambda^{\star}}}{\pi(\theta|\mathscr{M} = k)}\right] \mathrm{d}\theta \end{aligned}$$

En développant les logarithmes, on obtient de part et d'autre de l'identité le cœfficient

$$\int_{\Theta} p_{\lambda^{\star}} \log p_{\lambda^{\star}}\, \mathrm{d}\theta\,,$$

qui se simplifie et on a

$$\int_{\Theta} p_{\lambda^{\star}} \log\left[p(\theta|\boldsymbol{y}, \mathscr{M} = k)\right] \mathrm{d}\theta = \int_{\Theta} p_{\lambda^{\star}} \log\left[\pi(\theta|\mathscr{M} = k)\right] \mathrm{d}\theta\,.$$

Le logarithme de loi *a posteriori* de θ s'explicite $\log p(\theta|\boldsymbol{y}, \mathscr{M} = k) = \log f(\boldsymbol{y}|\theta, \mathscr{M} = k) + \log \pi(\theta|\mathscr{M} = k) - \log f(\boldsymbol{y}|\mathscr{M} = k)$ et on obtient

$$\int_{\Theta} p_{\lambda^{\star}} \log f(\boldsymbol{y}|\theta, \mathscr{M} = k)\, \mathrm{d}\theta - \int_{\Theta} p_{\lambda^{\star}} \log f(\boldsymbol{y}|\mathscr{M} = k)\, \mathrm{d}\theta = 0$$

$$\log f(\boldsymbol{y}|\mathscr{M} = k) \underbrace{\int_{\Theta} p_{\lambda^\star} \, \mathrm{d}\theta}_{=1} = \int_{\Theta} p_{\lambda^\star} \log \Big[f(\boldsymbol{y}|\theta, \mathscr{M} = k) \Big] \, \mathrm{d}\theta$$

$$\log f(\boldsymbol{y}|\mathscr{M} = k) = \mathbb{E}_{p_{\lambda^\star}} \Big[\log f(\boldsymbol{y}|\theta, \mathscr{M} = k) \Big].$$

On retrouve ainsi l'identité phare de l'algorithme WBIC.

Bibliographie

[AD99] C. Andrieu and A. Doucet. Joint bayesian model selection and estimation of noisy sinusoids via Reversible Jump MCMC. *IEEE Transactions on Signal Processing*, 47(10) :2667–2676, 1999. [Citée page 29]

[Aka73] H. Akaike. Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, pages 267–281, 1973. [Citée page 28]

[And10] T. Ando. *Bayesian Model Selection and Statistical Modeling*. Chapman and Hall/CRC, 2010. [Citée pages 23, 26, 27 et 28]

[BDPV19] V. De Bortoli, A. Durmus, M. Pereyra, and A. F. Vidal. Efficient stochastic optimisation by unadjusted Langevin Monte Carlo. Application to maximum marginal likelihood and empirical Bayesian estimation, 2019. [Citée page 19]

[BGG15] A. C. Bărbos, A. Giremus, and J.-F. Giovannelli. Bayesian noise model selection and system identification using Chib's approximation based on the Metropolis-Hastings sampler. *XXVème Colloque GRETSI*, September 2015. [Citée page 23]

[CG95] S. Chib and E. Greenberg. Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49(4) :327–335, 1995. [Citée page 23]

[CGM01] H. A. Chipman, E. I. George, and R. E. McCulloch. The practical implementation of Bayesian model selection. 2001. [Citée pages 19 et 23]

[Chi95] S. Chib. Marginal likelihood from the Gibbs output. *Journal Of the American Statistical Association*, 90(432) :1313–1321, 1995. [Citée page 23]

[CJ01] S. Chib and I. Jeliazkov. Marginal likelihood from the Metroplis-Hastings output. *Journal Of the American Statistical Association*, 96(453) :270–281, 2001. [Citée page 23]

[Dem12] J. P. Demailly. *Analyse numérique et équations différentielles*. Grenoble Sciences. EDP Sciences, 2012. [Citée pages 20 et 21]

[DIGMD01] G. Demoment, J. Idier, J.-F. Giovannelli, and A. Mohammad-Djafari. *Problèmes inverses en traitement du signal et de l'image*. Traité Télécoms, Hermès, 2001. [Citée page 3]

[DMP18] A. Durmus, É. Moulines, and M. Pereyra. Efficient Bayesian computation by proximal Markov chain Monte Carlo : when Langevin meets Moreau. *SIAM Journal on Imaging Sciences*, 11(1) :473–506, 2018.

[DTY18] J. Ding, V. Tarokh, and Y. Yang. Model selection techniques : An overview. *IEEE Signal Processing Magazine*, 35(6) :16–34, 2018. [Citée pages 27 et 28]

[FHW12] N. Friel, M. Hurn, and J. Wyse. Improving power posterior estimation of statistical evidence. *Statistics and computing*, 24(5), 2012. [Citée page 24]

[FMCP17] N. Friel, J.P. McKeone, C.J.Õates, and A.N. Pettitt. Investigation of the widely applicable Bayesian information criterion. *Statistics and Computing*, 27 :833–844, 2017. [Citée page 26]

[FP08] N. Friel and A.N. Pettitt. Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 70 :589–607, 2008. [Citée pages 24 et 25]

[FW12] N. Friel and J. Wyse. Estimation the model evidence – a review. *Statistica Neerlandica*, 66(3) :288–308, 2012. [Citée pages 22, 23, 24, 26 et 29]

[GAH17] M. Grzegorczyk, A. Aderhold, and D. Husmeier. Targeting bayes factors with direct-path non-equilibrium thermodynamic integration. *Computational Statistics*, 32(2) :717–761, June 2017. [Citée page 24]

[GB16] J.-F. Giovannelli and A. Barbos. Unsupervised segmentation of piecewise constant images from incomplete, distorted and noisy data. pages 1–5, 2016.

[GD94] A. E. Gelfandand and D. K. Dey. Bayesian Model Choice : Asymptotics and Exact Calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56, 01 1994. [Citée pages 19 et 26]

[GG14] J.-F. Giovannelli and A. Giremus. Bayesian noise model selection and system identification based on approximation of the evidence. *2014 IEEE Workshop on Statistical Signal Processing (SSP)*, pages 125–128, June 2014. [Citée page 23]

[GI13] J.-F. Giovannelli and J. Idier. *Méthodes d'inversion appliquées au traitement du signal et de l'image*. Traité IC2, Hermès, 2013. [Citée page 3]

[Gre95] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4) :711–732, 1995. [Citée pages 29 et 71]

[GS90] A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410) :398–409, 1990. [Citée page 19]

[GV17] J.-F. Giovannelli and C. Vacar. Deconvolution-segmentation for textured images. *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 191–195, 2017. [Citée page 10]

[Kai05] J. Kaipio. *Statistical and Computational Inverse Problems (Applied Mathematical Sciences)*. Springer, 2005. [Citée page 3]

[KK96] S. Konishi and G. Kitagawa. Generalised information criteria in model selection. *Biometrika*, 83(4) :875–890, 12 1996. [Citée pages 27 et 28]

[LB03] P. Lopes and J. Xavierand V. Barroso. Exploiting 2nd-order statistics in Bayesian signal reconstruction problems. *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, 4 :IV – 632, 2003. [Citée page 3]

[LGTB13] J. D. Lawrence, R. B. Gramacy., L. Thomas, and S. T. Buckland. The importance of prior choice in model selection : a density dependence example. *Methods in Ecology and Evolution*, 4(1) :25–33, 2013. [Citée page 12]

[LMH04] E. Lebarbier and T. Mary-Huard. Le critère BIC : fondements théoriques et interprétation. Technical report, INRIA, 2004. [Citée page 27]

[MP92] A. M. Mathai and S. B. Provost. *Quadratic Forms in Random Variables*. Statistics : A Series of Textbooks and Monographs. Taylor & Francis, 1992. [Citée page 20]

[MP04] J. Myung and M.A. Pitt. Model comparison methods. *Methods in Enzymology*, Numerical Computer Methods - Part D :351–366, 2004. [Citée page 28]

[Myu00] I. Myung. The importance of complexity in model selection. *Journal of Mathematical Psychology*, 44(1) :190 – 204, 2000. [Citée page 28]

[Nea93] Radford M. Neal. Probabilistic inference using markov chain monte carlo methods, 1993. [Citée pages 19 et 24]

[NR94] M. A. Newton and A. E. Raftery. Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1) :3–48, 1994. [Citée page 22]

[NRSK07] M. A. Newton, A. E. Raftery, J. M. Satagopan, and P. N. Krivitsky. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity (with discussion). *Bayesian statistics*, 8 :1–45, 2007. [Citée page 22]

[OFG12] F. Orieux, O. Féron, and J.-F. Giovannelli. Sampling high-dimensional gaussian distributions for general linear inverse problems. *IEEE Signal Processing Letters - IEEE SIGNAL PROCESS LETT*, 19 :251–254, 2012.

[OGR10a] F. Orieux, J.-F. Giovannelli, and T. Rodet. Bayesian estimation of regularization and PSF parameters for Wiener-Hunt deconvolution. 2010. [Citée pages 10 et 52]

[OGR10b] F. Orieux, J.-F. Giovannelli, and T. Rodet. Deconvolution with gaussian blur parameter and hyperparameters estimation. *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference*, pages 1350 – 1353, 2010. [Citée page 10]

[Per13] Marcelo Pereyra. Proximal markov chain monte carlo algorithms. *Statistics and Computing*, 26, 06 2013. [Citée page 19]

[PKM03] M. Pitt, W. Kim, and I. Myung. Flexibility versus generalizability in model selection. *Psychonomic bulletin & review*, 10 :29–44, 04 2003. [Citée page 28]

[PM16] M. Pereyra and S. McLaughlin. Comparing Bayesian models in the absence of ground truth. *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 528–532, Aug 2016. [Citée page 64]

[PMZ02] M. Pitt, I. Myung, and S. Zhang. Toward a method of selecting among computational models of cognition. *Psychological Review*, pages 472–491, 2002. [Citée page 28]

[PvT07] R. Preuss and U. von Toussaint. Comparison of numerical methods for evidence calculation. *Bayesian Inference and Maximum Entropy Methods in Science and Engineering : 27th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, 2007. [Citée pages 20, 24 et 29]

[RBF10] A. Roodaki, J. Bect, and G. Fleury. Comparison of fully bayesian and empirical Bayes approaches for joint bayesian model selection and estimation of sinusoids via Reversible Jump MCMC. *Proceedings of the European Signal Processing Conference (EUSIPCO'10)*, 2010. [Citée page 29]

[RC05] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag, 2005. [Citée page 51]

[RGGV15] R. Rosu, J.-F. Giovannelli, A. Giremus, and C. Vacar. Potts model parameter estimation in Bayesian segmentation of piecewise constant images. 2015. [Citée page 10]

[Rob10] C. P. Robert. *Le choix bayésien - Principes et pratique*. Springer Editions, 2010. [Citée page 4]

[RP20] J. Rougier and C. E. Priebe. The Exact Form of the "Ockham Factor" in Model Selection. *The American Statistician*, pages 1–6, 2020. [Citée page 28]

[RW09] C. P. Robert and D. Wraith. Computational methods for Bayesian model choice. *Proc. AIP*, 1193, 2009. [Citée page 29]

[SBBC02] D. Spiegelhalter, N. Best, and A. van der Linde B. Carlin. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, pages 583–616, 2002. [Citée page 28]

[TK86] L. Tierney and J.B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal Of the American Statistical Association*, 81(393) :82–86, 1986. [Citée page 26]

[Vac14] Cornelia Vacar. *Inversion for textured images : unsupervised myopic deconvolution, model selection, deconvolution-segmentation*. 2014. [Citée page 10]

[VBPD19] A. F. Vidal, V. De Bortoli, M. Pereyra, and A. Durmus. Maximum likelihood estimation of regularisation parameters in high-dimensional inverse problems : an empirical Bayesian approach, 2019.

[VG19] C. Vacar and J.-F. Giovannelli. Unsupervised joint deconvolution and segmentation method for textured images : a bayesian approach and an advanced sampling algorithm. *EURASIP Journal on Advances in Signal Processing*, 2019, 2019. [Citée page 10]

[VGB11] C. Vacar, J.-F. Giovannelli, and Y. Berthoumieu. Langevin and Hessian with Fisher approximation stochastic sampling for parameter estimation of structured covariance. *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference*, pages 3964 – 3967, 2011.

[VGB14] C. Vacar, J.-F. Giovannelli, and Y. Berthoumieu. Bayesian texture and instrument parameter estimation from blurred and noisy images using MCMC. *Signal Processing Letters, IEEE*, 21 :707–711, 2014. [Citée page 10]

[VGB15] C. Vacar, J.-F. Giovannelli, and Y. Berthoumieu. Bayesian texture classification from indirect observations using fast sampling. *IEEE Transactions on Signal Processing*, 64 :1–1, 2015. [Citée pages 10 et 23]

[VGR12] C. Vacar, J.-F. Giovannelli, and A. Roman. Bayesian texture model selection by harmonic mean. *2012 19th IEEE International Conference on Image Processing*, pages 2533–2536, Sep. 2012. [Citée pages 10 et 22]

[Wat13] S. Watanabe. A widely applicable Bayesian Information Criterion. *Journal of Machine Learning Research*, 14(1) :867–897, 2013. [Citée page 26]

[Wie49] N. Wiener. *Extrapolation, interpolation, and smoothing of stationary time series : with engineering applications*. MIT, Cambridge, MA, 1949. [Citée page 52]