Chapitre 9

Approche bayésienne

L'approche bayésienne est une approche cohérente pour la résolution des problèmes inverses car elle permet de prendre en compte et de traiter de la même manière l'information *a priori* sur la solution et celle sur les données.

En effet, cette approche permet non seulement de prendre en compte le caractère incertain des erreurs au travers de l'attribution d'une loi de probabilité aux mesures, mais aussi l'information *a priori* au travers d'une loi de probabilité *a priori* que l'on attribue aux inconnues du problème.

La fusion de ces deux sources d'information s'effectue par la règle de Bayes. On peut alors dire que cette approche généralise la régularisation déterministe.

9.1 Introduction

L'approche bayésienne correspond formellement au cheminement suivant :

1. On explicite un ensemble d'hypothèses $\mathcal{H}$ sur le problème concernant le modèle d'observation, les connaissances *a priori* sur les inconnues $\boldsymbol{f}$ et sur le bruit $\boldsymbol{b}$.

2. On attribue une loi de probabilité *a priori* $p(\boldsymbol{f}|\boldsymbol{\theta}_1, \mathcal{H})$ aux inconnues du problème pour traduire la connaissance initiale sur ces inconnues. Cette loi peut dépendre d'un certain nombre de paramètres $\boldsymbol{\theta}_1$.

3. On attribue une loi de probabilité $p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{\theta}_2; \mathcal{H})$ aux grandeurs mesurées pour traduire l'incertitude (due au bruit, aux erreurs de discrétisation et de quantification, au manque de précision de l'appareil de mesure, etc.) sur ces données. Cette loi aussi peut dépendre d'un certain nombre de paramètres $\boldsymbol{\theta}_2$. L'ensemble des paramètres $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ sont appelés les hyperparamètres du problème.

4. On utilise la règle de Bayes pour combiner l'information contenue dans les données et celle contenue dans la loi *a priori* et calculer ainsi la loi de probabilité *a posteriori*

$$p(\boldsymbol{f}|\boldsymbol{g}, \boldsymbol{\theta}; \mathcal{H}) = \frac{p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{\theta}_2; \mathcal{H})\, p(\boldsymbol{f}|, \boldsymbol{\theta}_1, \mathcal{H})}{p(\boldsymbol{g}|\boldsymbol{\theta}; \mathcal{H})}$$

où

$$p(\boldsymbol{g}|\boldsymbol{\theta}; \mathcal{H}) = \int p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{\theta}_2; \mathcal{H})\, p(\boldsymbol{f}|\boldsymbol{\theta}_1, \mathcal{H})\, \mathrm{d}\boldsymbol{f}.$$

Cette loi contient toute l'information disponible sur les inconnues $\boldsymbol{f}$ après cette fusion.

5. On peut alors calculer n'importe quelle quantité. Par exemple, la moyenne *a posteriori* de $\boldsymbol{f}$:

$$\mathrm{E}\left[\boldsymbol{f}\right] = \int \boldsymbol{f}\, p(\boldsymbol{f}|\boldsymbol{g}, \boldsymbol{\theta}; \mathcal{H})\, \mathrm{d}\boldsymbol{f},$$

ou la moyenne de n'importe quelle autre fonction $h(\boldsymbol{f})$:

$$\mathrm{E}\left[h(\boldsymbol{f})\right] = \int h(\boldsymbol{f})\, p(\boldsymbol{f}|\boldsymbol{g}, \boldsymbol{\theta}; \mathcal{H})\, \mathrm{d}\boldsymbol{f},$$

ou encore calculer la probabilité pour que $\underline{\boldsymbol{f}} < \boldsymbol{f} \leq \overline{\boldsymbol{f}}$:

$$P(\underline{\boldsymbol{f}} < \boldsymbol{f} \leq \overline{\boldsymbol{f}}) = \int_{\underline{\boldsymbol{f}}}^{\overline{\boldsymbol{f}}} p(\boldsymbol{f}|\boldsymbol{g}, \boldsymbol{\theta}; \mathcal{H})\, \mathrm{d}\boldsymbol{f}$$

ou seulement s'intéresser à une des composantes f_j de $\boldsymbol{f}$ et calculer

$$P(\underline{f}_j < f_j \leq \overline{f}_j) = \int_{\underline{f}_j}^{\overline{f}_j} p(f_j|\boldsymbol{g}, \boldsymbol{\theta}; \mathcal{H})\, \mathrm{d}f_j,$$

où

$$p(f_j|\boldsymbol{g}, \boldsymbol{\theta}; \mathcal{H}) = \int \cdots \int p(\boldsymbol{f}|\boldsymbol{g}, \boldsymbol{\theta}; \mathcal{H})\, \mathrm{d}f_1 \cdots \mathrm{d}f_{j-1}\, \mathrm{d}f_{j+1} \cdots \mathrm{d}f_n$$

est la loi *a posteriori marginale*.

Lorsque $\boldsymbol{f}$ est un scalaire ou un vecteur avec seulement deux composantes, on peut même tracer $p(f|\boldsymbol{g},\boldsymbol{\theta};\mathcal{H})$ ou $p(f_1,f_2|\boldsymbol{g},\boldsymbol{\theta};\mathcal{H})$, ce qui permettra d'avoir une idée plus précise sur la forme de ces distribution.

Mais en pratique, la manipulation d'une densité de probabilité n'est pas très commode dès lors qu'on a plus de deux variables. C'est pourquoi, on se contente souvent de fournir une description sommaire de la loi *a posteriori* en indiquant
– son mode :

$$\mathrm{M}[\boldsymbol{f}] = \arg\max_{\boldsymbol{f}} \left\{p(\boldsymbol{f}|\boldsymbol{g},\boldsymbol{\theta};\mathcal{H})\right\},$$

– sa moyenne :

$$\mathrm{E}\left[\boldsymbol{f}\right] = \int \boldsymbol{f}\, p(\boldsymbol{f}|\boldsymbol{g},\boldsymbol{\theta};\mathcal{H})\,\mathrm{d}\boldsymbol{f},$$

– les modes de ses marginales :

$$\mathrm{M}[f_j] = \arg\max_{f_j} \left\{p(f_j|\boldsymbol{g},\boldsymbol{\theta};\mathcal{H})\right\},$$

– les α-quantiles de ses marginales :

$$Q_\alpha(f_j) \ : P(f_j \leq Q_\alpha(f_j)) = \alpha,$$

– les médianes de ses marginales :

$$\mathrm{Méd}[f_j] = Q_{\frac{1}{2}}(f_j) \ : \ P\left(f_j \leq \mathrm{Méd}[f_j]\right) = \frac{1}{2},$$

– les régions de plus haute probabilité ou les supports α–interquantiles :

$$[a,b] = \left[Q_{(1-\alpha)/2}(f_j), Q_{(1+\alpha)/2}(f_j)\right] \ : P(a \leq f_j \leq b) = 1-\alpha,$$

– les variances :

$$\mathrm{Var}[f_j] = \int (f_j - \bar{f}_j)^2\, p(f_j|\boldsymbol{g},\boldsymbol{\theta};\mathcal{H}),$$

où

$$\bar{f}_j = \mathrm{E}\left[f_j\right] = \int f_j\, p(f_j|\boldsymbol{g},\boldsymbol{\theta};\mathcal{H}),$$

– les covariances :

$$\mathrm{Cov}[f_j,f_k] = \iint (f_j - \bar{f}_j)(f_k - \bar{f}_k)\, p(f_j,f_k|\boldsymbol{g},\boldsymbol{\theta};\mathcal{H})\,\mathrm{d}f_j\,\mathrm{d}f_k.$$

6. Une autre approche fondée sur la théorie de la décision consiste en la définition d'un estimateur $\widehat{\boldsymbol{f}}$ qui minimiserait la moyenne d'une fonction de coût $C(\widehat{\boldsymbol{f}},\boldsymbol{f})$. Nous reviendront sur cette approche et les choix de ces fonctions coût, ainsi que sur les liens qu'existe entre ces estimateurs et les caractéristiques de la loi *a posteriori*.

7. Finalement, une fois une solution choisie, il est impératif d'y attribuer un degré de confiance ainsi que de déterminer la sensibilité de cette solution vis-à-vis des erreurs de mesures et celles de la modélisation. Heureusement, en principe, cette approche bayésienne nous fourni tout ce qui est nécessaire pour pouvoir satisfaire cette exigence. Car disposant de la loi *a posteriori* on peut calculer, par exemple, la covariance de l'erreur de l'estimation, ce qui nous permet de mettre des barres d'erreurs sur les solutions estimées.

Voyons maintenant quelles sont les difficultés de l'utilisation de cette approche dans un problème réel et quelles sont les solutions existantes pour les résoudre. On peut classer ces difficultés dans les rubriques qui suivent :

1. choix de la forme de la loi *a priori* $p(\boldsymbol{f}|\boldsymbol{\theta}_1; \mathcal{H})$;
2. choix ou estimation de ses paramètres $\boldsymbol{\theta}_1$;
3. choix de la forme de la loi $p(\boldsymbol{g}|\boldsymbol{\beta}, \boldsymbol{\theta}_2; \mathcal{H})$;
4. choix ou estimation de ses paramètres $\boldsymbol{\theta}_2$;
5. calcul effectif de la loi *a posteriori* $p(\boldsymbol{f}|\boldsymbol{g}, \boldsymbol{\theta}; \mathcal{H})$;
6. choix d'une fonction de coût $C(\widehat{\boldsymbol{f}}, \boldsymbol{f})$ et calcul effectif de l'estimateur qu'en découle ;
7. détermination ou estimation *a posteriori* des hyperparamètres du problème, *i.e.*les paramètres $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ des lois de probabilités directes $p(\boldsymbol{f}|\boldsymbol{\theta}_1; \mathcal{H})$ et $p(\boldsymbol{g}|\boldsymbol{f}, \boldsymbol{\theta}_2; \mathcal{H})$, à partir des données, c'est-à-dire en même temps que la solution ;

L'objectif principal de ce chapitre est de faire état de l'art concernant les solutions proposées pour résoudre ces problèmes et de fournir quelques exemples concrets (étude de cas) pour illustrer les difficultés et les performances des solutions proposées.

Mais, avant d'aller plus loin prenons un exemple, désormais classique, qui est le cas linéaire gaussien qui nous permettra de comprendre l'essentiel de ces différentes étapes.

9.2 Cas linéaire gaussien

S'il y a un cas où les calculs peuvent se faire facilement jusqu'au bout et où l'on peut facilement comprendre l'approche bayésienne est le cas où le modèle reliant les inconnues $\boldsymbol{f}$ et les données $\boldsymbol{g}$ est linéaire et où on peut attribuer des lois de probabilités gaussiennes à $\boldsymbol{f}$ et aux mesures $\boldsymbol{g}$.

Nous allons suivre les étapes présentées en introduction pour la résolution du problème linéaire

$$\boldsymbol{g} = \boldsymbol{H}\boldsymbol{f} + \boldsymbol{b} \tag{9.1}$$

- Information *a priori* sur $\boldsymbol{f}$:
  Faisons l'hypothèse que nous puissions *a priori* connaître seulement la moyenne $\mathrm{E}[\boldsymbol{f}] = \boldsymbol{f}_0$ et la matrice de covariance $\mathrm{E}\left[(\boldsymbol{f}-\boldsymbol{f}_0)(\boldsymbol{f}-\boldsymbol{f}_0)^t\right] = \boldsymbol{R}_f = \sigma_f^2\boldsymbol{P}_0$ de $\boldsymbol{f}$. Nous pouvons alors faire l'hypothèse que $p(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}_0, \sigma_f^2\boldsymbol{R}_f)$:

$$p(\boldsymbol{f}) = A\exp\left[\frac{-1}{2}(\boldsymbol{f}-\boldsymbol{f}_0)^t\boldsymbol{R}_f^{-1}(\boldsymbol{f}-\boldsymbol{f}_0)\right] = A\exp\left[\frac{-1}{2\sigma_f^2}(\boldsymbol{f}-\boldsymbol{f}_0)^t\boldsymbol{P}_0^{-1}(\boldsymbol{f}-\boldsymbol{f}_0)\right] \tag{9.2}$$

  avec $A = (2\pi)^{-n/2}\,|\boldsymbol{R}_f|^{-1/2} = (2\pi\sigma_f^2)^{-n/2}\,|\boldsymbol{P}_0|^{-1/2}$.
  D'ailleurs, cette hypothèse peut être validée par le principe du maximum d'entropie.

- Information *a priori* sur $\boldsymbol{b}$:
  Faisons l'hypothèse que $\mathrm{E}[\boldsymbol{b}] = \boldsymbol{0}$ et que sa matrice de covariance $\mathrm{E}\left[\boldsymbol{b}\boldsymbol{b}^t\right] = \boldsymbol{R}_b = \sigma_b^2\boldsymbol{I}$. Cette hypothèse est bien raisonnable. En effet, supposer que $\mathrm{E}[\boldsymbol{b}] = \boldsymbol{0}$ signifie que nous faisons l'hypothèse qu'il n'y a pas d'erreur systématique et l'hypothèse $\mathrm{E}\left[\boldsymbol{b}\boldsymbol{b}^t\right] = \sigma_b^2\boldsymbol{I}$ signifie que le bruit est supposé blanc (non corrélé).

  Avec les mêmes arguments que dans le cas précédent on fait alors l'hypothèse que $p(\boldsymbol{b}) = \mathcal{N}(\boldsymbol{0}, \sigma_b^2\boldsymbol{R}_f)$. Maintenant, partant du modèle (9.1) et faisant l'hypothèse que le bruit $\boldsymbol{b}$ est aussi indépendant du $\boldsymbol{f}$ on peut facilement déduire

$$p(\boldsymbol{g}|\boldsymbol{f}) = B\exp\left[\frac{-1}{2}(\boldsymbol{g}-\boldsymbol{H}\boldsymbol{f})^t\boldsymbol{R}_b^{-1}(\boldsymbol{g}-\boldsymbol{H}\boldsymbol{f})\right] = B\exp\left[\frac{-1}{2\sigma_b^2}(\boldsymbol{g}-\boldsymbol{H}\boldsymbol{f})^t(\boldsymbol{g}-\boldsymbol{H}\boldsymbol{f})\right] \tag{9.3}$$

  avec $B = (2\pi)^{-m/2}\,|\boldsymbol{R}_b|^{-1/2} = (2\pi\sigma_b^2)^{-m/2}$.

- Règle de Bayes :
  Utilisant la règle de Bayes, il est facile de voir que

$$p(\boldsymbol{f}|\boldsymbol{g}) = C\exp\left[\frac{-1}{2}J(\boldsymbol{f})\right], \tag{9.4}$$

  avec

$$\begin{aligned} J(\boldsymbol{f}) &= (\boldsymbol{g}-\boldsymbol{H}\boldsymbol{f})^t\boldsymbol{R}_b^{-1}(\boldsymbol{g}-\boldsymbol{H}\boldsymbol{f}) + (\boldsymbol{f}-\boldsymbol{f}_0)^t\boldsymbol{R}_f^{-1}(\boldsymbol{f}-\boldsymbol{f}_0) \\ &= \frac{1}{\sigma_b^2}\left[(\boldsymbol{g}-\boldsymbol{H}\boldsymbol{f})^t(\boldsymbol{g}-\boldsymbol{H}\boldsymbol{f}) + \lambda(\boldsymbol{f}-\boldsymbol{f}_0)^t\boldsymbol{P}_0^{-1}(\boldsymbol{f}-\boldsymbol{f}_0)\right], \end{aligned}$$

  où $\lambda = \frac{\sigma_b^2}{\sigma_f^2}$.

Avec un peu de calcul il n'est pas difficile de montrer que cette loi *a posteriori* est aussi gaussienne et on a

$$p(\boldsymbol{f}|\boldsymbol{g}) = \mathcal{N}(\widehat{\boldsymbol{f}}, \widehat{\boldsymbol{P}})$$

où

$$\left\{ \begin{array}{rcl} \widehat{\boldsymbol{f}} & = & \boldsymbol{R}_f \boldsymbol{H}^t (\boldsymbol{H}\boldsymbol{R}_f\boldsymbol{H}^t + \boldsymbol{R}_b)^{-1}\boldsymbol{g} = \widehat{\boldsymbol{P}}\boldsymbol{H}^t\boldsymbol{R}_b^{-1}\boldsymbol{g}, \\ \widehat{\boldsymbol{P}} & = & \boldsymbol{R}_f - \boldsymbol{R}_f\boldsymbol{H}^t(\boldsymbol{H}\boldsymbol{R}_f\boldsymbol{H}^t + \boldsymbol{R}_b)^{-1}\boldsymbol{H}\boldsymbol{R}_f = (\boldsymbol{R}_f^{-1} + \boldsymbol{H}^t\boldsymbol{R}_b^{-1}\boldsymbol{H})^{-1}, \end{array} \right.$$

ou encore

$$\left\{ \begin{array}{rcl} \widehat{\boldsymbol{f}} & = & \left(\boldsymbol{H}^t\boldsymbol{H} + \lambda\boldsymbol{P}_0^{-1}\right)^{-1}\boldsymbol{H}^t\boldsymbol{g}`, \\ \widehat{\boldsymbol{P}} & = & \sigma_b^2\left(\boldsymbol{H}^t\boldsymbol{H} + \lambda\boldsymbol{P}_0^{-1}\right)^{-1}. \end{array} \right.$$

- Caractéristique de la solution :
  Nous avons ainsi une expression analytique pour la loi *a posteriori* et nous pouvons facilement en déduire tout ce que l'on souhaite. Par exemple nous savons que la moyenne, la médiane et la mode sont identiques et égaux à $\widehat{\boldsymbol{f}}$.

Ainsi tous ces estimateurs : le maximum *a posteriori* (MAP) ou la moyenne *a posteriori* (MP) sont identiques et égaux à $\widehat{\boldsymbol{f}}$ qui peut aussi être calculé par

$$\widehat{\boldsymbol{f}} = \arg\max_{\boldsymbol{f}} \{p(\boldsymbol{f}|\boldsymbol{g})\} = \arg\min_{\boldsymbol{f}} \{J(\boldsymbol{f}) = Q(\boldsymbol{f}) + \lambda\Omega(\boldsymbol{f})\}$$

avec

$$Q(\boldsymbol{f}) = (\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f})^t(\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f})$$

et

$$\Omega(\boldsymbol{f}) = (\boldsymbol{f} - \boldsymbol{f}_0)^t\boldsymbol{P}_0^{-1}(\boldsymbol{f} - \boldsymbol{f}_0),$$

ou encore, si on note par $\boldsymbol{P}_0^{-1} = \boldsymbol{D}^t\boldsymbol{D}$ on a

$$\begin{array}{rcl} J(\boldsymbol{f}) & = & (\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f})^t(\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}) + \lambda(\boldsymbol{f}^t - \boldsymbol{f}_0)^t\boldsymbol{D}^t\boldsymbol{D}(\boldsymbol{f} - \boldsymbol{f}_0) \\ & = & \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|^2 + \lambda\|\boldsymbol{D}(\boldsymbol{f} - \boldsymbol{f}_0)\|^2. \end{array}$$

On retrouve alors la notion de solution régularisée et la régularisation quadratique avec les avantages suivantes :

- Dans l'approche régularisation déterministe le choix des distances $\Delta_1(\boldsymbol{g}, \boldsymbol{H}\boldsymbol{f}) = \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|^2$ et de $\Delta_2(\boldsymbol{f}, \boldsymbol{f}_0) = \|\boldsymbol{D}(\boldsymbol{f} - \boldsymbol{f}_0)\|^2$ est plutôt descriptif et un peu arbitraire, alors que dans l'approche bayésienne, ces termes sont, respectivement, les conséquences des hypothèses faites sur la loi du bruit et sur la loi *a priori*.

- Dans régularisation déterministe, le coefficient de régularisation λ est un paramètre qui est choisi empiriquement, alors qu'ici, $\lambda = \frac{\sigma_b^2}{\sigma_f^2}$. Si l'on connait les deux variances σ_b^2 et σ_f^2, sa valeur est fixée. Mais, bien sûr, en pratique on ne connait pas ces deux valeurs et on est encore amené à les fixer par expérience. Cependant, rien que l'expession de λ nous permet de mieux comprendre son comportement : plus le niveau du bruit est important, plus il faut choisir sa valeur grande pour obtenir un résultat satisfaisant. Nous verrons que l'approche bayésienne nous donne les outils nécessaires pour la déterminer (voir le chapitre sur l'estimation des hyperparamètres).

– Dans le cadre de la régularisation déterministe, si on se demande qu'elle est le degré de confiance que l'on peut avoir dans la solution $\widehat{\boldsymbol{f}}$ proposée, nous n'avons pas d'outil convenable pour répondre. Dans l'approche bayésienne, nous pouvons répondre à cette question, par exemple en calculant la matrice de covariance *a posteriori* $\boldsymbol{P} = \mathrm{E}\left[(\boldsymbol{f} - \widehat{\boldsymbol{f}})^2\right]$. Sachant que les éléments diagonaux $P_{jj} = \mathrm{E}\left[(f_j - \widehat{f}_j)^2\right]$ sont les variances *a posteriori*, nous pouvons les utiliser pour mettre des barres d'erreur sur la solution. Les éléments non-diagonale $P_{ij} = \mathrm{E}\left[(f_i - \widehat{f}_i)(f_j - \widehat{f}_j)\right]$ peuvent aussi nous renseigner sur le lien (corrélation) qui peut exister entre l'élément f_i et l'élement f_j.

9.3 Calcul de la loi *a posteriori*

9.3.1 Cas gaussien

Nous avons vu que dans le cas d'un modèle linéaire gaussien, la loi *a posteriori* est aussi gaussienne. Sachant qu'une loi gaussienne est entièrement définie par ses deux premiers moments, le problème devient assez simple. Nous disposons des expressions analytiques pour la moyenne et la matrice de covariance *a posteriori*, expressions dont les calculs nécessitent soit l'inversion des matrices soit l'optimisation d'un critère quadratique. Il reste cependant le coût de calcul. Souvent, on peut profiter de la structure des matrices à inverser ou de la forme quadratique du critère à optimiser pour fournir des algorithmes spécifiques pour effectuer ces calculs.

9.3.2 Cas général

Dans le cas général, le calcul complet de la loi *a posteriori* peut devenir plus délicat, et souvent, on se contente, soit de l'approximer par une loi gaussienne et de calculer la moyenne et la matrice de covariance, soit de définir un estimateur ponctuel à partir de cette loi, comme par exemple le *maximum a posteriori* (MAP) :

$$\widehat{\boldsymbol{f}}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{f}} \{p(\boldsymbol{f}|\boldsymbol{g})\} = \arg\min_{\boldsymbol{f}} \{-\ln p(\boldsymbol{f}|\boldsymbol{g})\},$$

ou la *moyenne a posteriori* (en Anglais posterior mean PM) :

$$\widehat{\boldsymbol{f}}_{\mathrm{PM}} = \int \boldsymbol{f}\, p(\boldsymbol{f}|\boldsymbol{g})\, \mathrm{d}\boldsymbol{f},$$

ou encore le *maximum a posteriori marginal* (MAPmarginal) :

$$\widehat{f}_j = \arg\max_{f_j} \{p(f_j|\boldsymbol{g})\},$$

où

$$p(f_j|\boldsymbol{g},\boldsymbol{\theta};\mathcal{H}) = \int \cdots \int p(\boldsymbol{f}|\boldsymbol{g},\boldsymbol{\theta};\mathcal{H})\, \mathrm{d}f_1 \cdots \mathrm{d}f_{j-1}\, \mathrm{d}f_{j+1} \cdots \mathrm{d}f_n.$$

Notons aussi que le calcul des estimateurs PM et MAPmarginal nécessite le calcul des intégrales de dimension très élevée, alors que le calcul de l'estimateur MAP ne nécessite qu'une optimisation.

Notons aussi que dans le cas gaussien tous ces estimateurs sont identiques. En revanche, dans les autres cas, ils peuvent fournir des résultats très différents. Nous reviendrons sur les propriétés de ces estimateurs et les outils qui existent pour les calculer dans les chapitres suivants.

9.4 Choix des fonctions de coût

Dans l'approche théorie de la décision nous avons vu que, partant de la loi *a posteriori*, on peut définir une solution au problème, par l'intermédiaire d'une fonction de coût $C(\widehat{\boldsymbol{f}}, \boldsymbol{f})$. En effet, dans cette approche on définit l'estimateur optimal $\widehat{\boldsymbol{f}}$ comme l'argument qui minimise le coût moyen :

$$\widehat{\boldsymbol{f}} = \arg\min_{\boldsymbol{z}} \{\mathrm{E}[C(\boldsymbol{z}, \boldsymbol{f})]\} = \arg\min_{\boldsymbol{z}} \left\{ \int C(\boldsymbol{z}, \boldsymbol{f}) p(\boldsymbol{f}|\boldsymbol{g}) \, \mathrm{d}\boldsymbol{f} \right\}.$$

Nous allons voir que suivant le choix de la fonction de coût on obtient des estimateurs différents, et en particulier, les estimateurs MAP, PM et MAPmarginal :

- Coût quadratique et estimateur PM :

$$C(\boldsymbol{f}, \widehat{\boldsymbol{f}}) = (\boldsymbol{f} - \widehat{\boldsymbol{f}})^t \boldsymbol{Q} (\boldsymbol{f} - \widehat{\boldsymbol{f}})^t \longrightarrow \widehat{\boldsymbol{f}} = \int \boldsymbol{f} \, p(\boldsymbol{f}|\boldsymbol{g}) \, \mathrm{d}\boldsymbol{f}.$$

- Coût dirac et estimateur MAP :

$$C(\boldsymbol{f}, \widehat{\boldsymbol{f}}) = 1 - \delta(\boldsymbol{f} - \widehat{\boldsymbol{f}}) \longrightarrow \widehat{\boldsymbol{f}} = \arg\max_{\boldsymbol{f}} \{p(\boldsymbol{f}|\boldsymbol{g})\}.$$

- Coût produit de diracs et estimateur MAPmarginal :

$$C(\boldsymbol{f}, \widehat{\boldsymbol{f}}) = \prod_j 1 - \delta(x_j - \widehat{x}_j) \longrightarrow \widehat{f}_j = \arg\max_{x_j} \{p(x_j|\boldsymbol{g})\},$$

D'autres fonctions de coût peuvent aussi être utilisées, mais leur usage reste plus spécifique.

9.5 Choix de la loi *a priori*

Le problème de la conversion d'une information *a priori* en une loi de probabilité est un problème difficile et encore largement ouvert.

La principale difficulé réside dans le fait que rarement, en pratique, l'information *a priori* se présente directement sous une forme probabiliste. Par exemple, on nous dit que la grandeur recherchée est *positive ou bornée entre [0, 1]*, de variation *douce, croissante ou décroissante*, *à bande limitée ou à spectre limitée*, de variation *douce ou constante par morceaux ou par région*, de variation *impulsionnelle*, etc. La tâche, ensuite, est de construire des lois de probabilités qui puissent incorporer ces qualificatifs.

Les méthodes existantes peuvent être schématiquement regroupées en trois grandes classes.

Certaines reposent sur la théorie des *groupes de transformation* pour déterminer la mesure de référence "naturelle" pour la grandeur recherchée. On utilise ces méthodes surtout lorsqu'on sait peu de choses sur la grandeur recherchée, c'est à dire lorsque notre information *a priori* se résume à une connaissance qualitative sur la nature de la grandeur recherchée. Par exemple, nous savons que la grandeur recherchée représente un paramètre de *localisation* pouvant prendre une valeur quelconque sur l'axe des réels $\mathbb{R}$ ou un paramètre *d'échelle* qui est par nature positif et pouvant prendre une valeur quelconque sur la demi-droite des réels positifs $\mathbb{R}_+$.

Mais en pratique, cette approche a permis de justifier après coup l'emploi de la mesure de Lebesgue pour les paramètres de localisation (fournissant ainsi une extension au cas continu de la distribution uniforme résultant de l'application du "Principe d'indifférence" de Bernoulli dans le cas discret) et de la mesure de Jeffreys dans le cas des paramètres d'échelle.

D'autres méthodes reposent sur des principes informationnel. Il s'agit principalement des méthodes dites "à maximum d'entropie" dans lesquelles on recherche une distribution qui soit la plus proche (au sens d'une distance de Kullback) d'une distribution de référence (souvent choisie par l'approche précédente) tout en vérifiant une information incomplète connue *a priori* sous la forme des moments (des contraintes linéaires) sur la loi recherchée. Mais là encore, cette approche a surtout permis de justifier après coup certains choix. De plus, elle n'est véritablement praticable que lorsque cette information *a priori* est faite de contraintes linéaires sur la distribution recherchée. On trouve alors la famille des *distributions exponentielles* comme nous l'avons vu dans l'étude du principe du maximum d'entropie.

Il existe enfin une dernière classe très importante, celle des construction faites "à la main". C'est dans cette catégorie qu'entrent par exemple les modèles markoviens qui ont connu un développement spectaculaire depuis 1984 en traitement d'image, et qui permettent d'incorporer dans une distribution *a priori* des propriétés locales essentielles que doit posséder l'objet. La construction de ces modèles demande beaucoup de savoir-faire.

Bien entendu, il est hors de question de vouloir détailler toutes ces méthodes dans le cadre de ce document. Nous nous limiterons, dans la section suivante, à une description succinte des deux premières approches, plutôt à travers de quelques exemples que d'une manière formelle. Mais, en ce qui concerne la modélisation markovienne, nous lui réservons un chapitre tout entier.

9.5.1 Maximum d'entropie

Nous avons vu dans la section (8.3) que le principe du maximum d'entropie peut être utilisé pour attribuer une loi de probabilité à une grandeur traduisant une information *a priori* sous forme de moments. Nous allons illustrer ceci à travers quelques exemples.

Supposons que nous ayons un voltmètre et que nous cherchions à mesurer la tension de la ville $X(t)$ à différents instants. Supposons que notre instrument soit un voltmètre idéal et ce que nous mesurons $Y(t)$ peut être modélisé par

$$Y(t) = X(t) + B(t)$$

où $B(t)$ représent à la fois les erreurs de lecture et le bruit externe. Nous voudrions appliquer l'approche bayésienne pour fournir une estimation $\widehat{X}(t)$ de $X(t)$. Considérons tout d'abord que nous avons fait juste une mesure y_1 à partir de laquelle que nous voudrions estimer X. Nous avons vu que la première étape avant appliquer la régle de Bayes est d'attribuer les lois de probabilité $p(x)$ et $p(y|x)$.

Concernant X, nous savons que sa moyenne est de 220 volts et que sa variation autour de cette moyenne est ± 5 volts (ceci dépend sûrement de pays et de la qualité des installations, mais supposons que nous soyons en France!). Comment traduire alors cette information en une loi de probabilité? C'est exactement dans ce cadre que le principe du ME peut fournir une réponse satisfaisante:

$$\mathrm{E}[X] = x_0 = 220 \text{ volts}, \quad \mathrm{Var}[X] = \sigma_x^2 = 25 \text{ volts}^2 \quad \overset{\text{ME}}{\longrightarrow} \quad X \simeq \mathcal{N}(x_0, \sigma_x^2)$$

Concernant B, il est naturel de faire l'hypothèse que sa moyenne est nulle (il n'y a pas d'erreur systématique) et supposons que l'on puisse avoir une idée de sa puissance. En langage d'ingénieur : "le bruit est d'environ 10 dB". Là aussi, le principe du ME peut fournir une réponse satisfaisante:

$$\mathrm{E}[B] = 0, \quad \mathrm{Var}[B] = \sigma_b^2 = 1 \text{ volts}^2 \quad \overset{\text{ME}}{\longrightarrow} \quad B \simeq \mathcal{N}(0, \sigma_b^2).$$

Nous avons maintenant tous les ingrédients pour calculer la loi *a posteriori*:

$$X|Y \simeq \mathcal{N}(\widehat{x}, \widehat{\sigma}^2),$$

et il n'est pas difficile de calculer $\widehat{x}$ et $\widehat{\sigma}^2$:

$$\widehat{x} = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_b^2} y + \frac{\sigma_b^2}{\sigma_x^2 + \sigma_b^2} x_0$$

$$\widehat{\sigma}^2 = \frac{\sigma_x^2 \sigma_b^2}{\sigma_x^2 + \sigma_b^2}.$$

Notons que

si $\sigma_b^2 = 0 \longrightarrow \widehat{x} = y, \quad \widehat{\sigma}^2 = 0$ et

si $\sigma_x^2 = 0 \longrightarrow \widehat{x} = x_0, \quad \widehat{\sigma}^2 = 0.$

Considérons maintenant que la grandeur X que nous voulons mesurer est l'énergie ou la puissance consommée dans un circuit et l'appareil de mesure est un wattmètre. Mais, avant même de choisir notre instrument nous savons que la grandeur à mesuer est positive et que son ordre de grandeur est d'environ 200 Watts:

$$X \geq 0, \quad \mathrm{E}[X] = x_0 = 200 \text{ Watts} \quad \overset{\text{ME}}{\longrightarrow} \quad X \simeq \mathcal{E}(x_0)$$

ce qui signifie :

$$p(x) = x_0 \exp\left[-x/x_0\right], \quad x > 0.$$

Si nous gardons les mêmes hypothèses sur B alors nous avons

$$p(y|x) \propto \exp\left[-\frac{1}{2\sigma_b^2}(y-x)^2\right]$$

ce qui nous donne pour la loi *a posteriori* :

$$p(x|y) \propto \exp\left[-\frac{1}{2\sigma_b^2}\left((y-x)^2 + \frac{2\sigma_b^2}{x_0}x\right)\right], \quad x > 0.$$

En réarrangeant le terme en exponentiel, on peut facilement voir que cette loi est une loi gaussienne tronquée :

$$p(x|y) \propto \exp\left[-\frac{1}{2\sigma_b^2}(x-\widehat{x})^2\right], \quad x > 0, \quad \text{avec} \quad \widehat{x} = y - \frac{\sigma_b^2}{x_0}$$

Revenons maintenant sur le premier exemple et considérons que nous avons fait M mesures aux intervalles réguliers $\boldsymbol{Y} = \{y_1, \cdots, y_M\}$ et que nous voulons estimer $\boldsymbol{X} = \{X_1, \cdots, X_M\}$.

Une première hypothèse consiste à dire qu'*a priori*, il n'y a pas de raison de faire l'hypothèse que la loi de X_i soit différente de la loi de X_j et que X_i soit dépendante de X_j (hypothèse dite i.i.d.) et d'attribuer alors une loi de probabilité jointe :

$$p(\boldsymbol{x}) = \prod_{j=1}^{M} p(x_j) \longrightarrow \boldsymbol{X} \simeq \mathcal{N}(\boldsymbol{x}_0, \sigma_x^2 \boldsymbol{I}).$$

La même hypothèse concernant les éléments de $\boldsymbol{B} = \{B_1, \cdots, B_M\}$ nous permet d'écrire :

$$p(\boldsymbol{b}) = \prod_{j=1}^{M} p(b_j) \longrightarrow \boldsymbol{B} \simeq \mathcal{N}(\boldsymbol{0}, \sigma_b^2 \boldsymbol{I}).$$

Il n'est pas difficile de montrer que la loi *a posteriori* est :

$$p(\boldsymbol{x}|\boldsymbol{y}) = \prod_{j=1}^{M} p(x_j|y_j) \longrightarrow \boldsymbol{X}|\boldsymbol{Y} \simeq \mathcal{N}(\boldsymbol{Y}, \sigma^2 \boldsymbol{I})$$

Une deuxième hypothèse plus réaliste est de dire qu'*a priori*, X_j dépend de X_{j-1} (la valeur de tension à un instant ne peut pas être très différente de sa valeur à un pas d'échantillonnage plus loin). Comment peut-on alors traduire cette information? Supposons que nous puissions avoir une idée *a priori* sur le coéfficient de corrélation entre X_j et X_{j-1}. Ainsi, si on résume tout ce que nous connaissons sur $\boldsymbol{X}$:

$$\mathrm{E}\left[X_j\right] = x_0, \quad \mathrm{Var}\left[X_j\right] = \sigma_x^2, \quad \mathrm{E}\left[(X_j - x_0)(X_{j-1} - x_0)\right] = \beta\sigma_x^2.$$

Utilisant de nouveau le principe du ME, nous obtenons :

$$p(\boldsymbol{x}) \propto \exp\left[\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)^t \boldsymbol{P}^{-1}(\boldsymbol{x}-\boldsymbol{x}_0)\right], \quad \text{avec} \quad \boldsymbol{P} = \sigma_x^2 \begin{pmatrix} 1 & \beta & & & & \\ \beta & 1 & \beta & & & \\ & \ddots & \ddots & \ddots & & \\ & & & \beta & 1 & \beta \\ & & & & \beta & 1 \end{pmatrix}$$

Ainsi, d'une manière générale, l'utilisation du principe du maximum d'entropie nous conduit vers les lois exponentielles généralisées. En effet, si on suppose que notre information *a priori* est de la forme :

$$\mathrm{E}\left[\phi_k(\boldsymbol{X})\right] = d_k, \quad k = 1, \cdots, K,$$

nous avons vu que la loi à entropie maximale qui satisfait ces contraintes est de la forme :

$$p(\boldsymbol{x}) = \frac{1}{Z} \exp\left[-\sum_{k=1}^{K} \lambda_k \phi_k(\boldsymbol{x})\right],$$

où

$$Z(\lambda_1, \cdots, \lambda_m) = \int p(\boldsymbol{x}) \exp\left[-\sum_{k=1}^{K} \lambda_k \phi_k(\boldsymbol{x})\right] \mathrm{d}\boldsymbol{x}$$

et où les $\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$ sont déterminés par :

$$-\frac{\partial \ln Z(\lambda_1, \cdots, \lambda_m)}{\partial \lambda_k} = d_k, \quad k = 1, \cdots, K.$$

Un cas particulier intéressant est lorsque les fonctions $\phi_k(x)$ sont séparables :

$$\phi_k(\boldsymbol{x}) = \sum_{j=1}^{N} \phi_{k_j}(x_j)$$

on a alors

$$p(\boldsymbol{x}) = \prod_{j=1}^{N} p(x_j), \quad \text{avec} \quad p(x_j) = \frac{1}{Z_j} \exp\left[-\sum_{k=1}^{K} \lambda_k \phi_{k_j}(x_j)\right],$$

ce qui signifie que les x_j sont indépendants, et si de plus $\phi_{k_j} = \phi_k$ on a

$$p(\boldsymbol{x}) = \prod_{j=1}^{N} p(x_j), \quad \text{avec} \quad p(x_j) = \frac{1}{z} \exp\left[-\sum_{k=1}^{K} \lambda_k \phi_k(x_j)\right],$$

ce qui signifie que les x_j sont i.i.d (indépendants et de même loi).

On trouve parfois l'appellation *lois entropiques* pour les lois appartenant à ces deux derniers cas. Quelques cas particuliers de ces lois utilisées fréquemment sont :

– Gaussienne généralisée :

$$\phi(x_j) = |x_j|^r, 1 < r < 2, \quad \longrightarrow \quad p(x_j) \propto \exp\left[-\alpha |x_j|^r\right].$$

La loi conjointe est alors de la forme :

$$p(\boldsymbol{x}) \propto \exp\left[-\alpha \sum_{j=1}^{N} |x_j|^r\right] \quad \longrightarrow \quad -\ln p(\boldsymbol{x}) = cte + \alpha \sum_{j=1}^{N} |x_j|^r.$$

– Loi Gamma:

$$\phi(x_j) = -\alpha \ln(\frac{x_j}{m_j}) + (\frac{x_j}{m_j}), \quad \longrightarrow \quad p(x_j) \propto (\frac{x_j}{m_j})^{-\alpha} \exp\left[\frac{x_j}{m_j}\right].$$

et

$$p(\boldsymbol{x}) = \prod_{j=1}^{N} p(x_j) \propto \left(\prod_{j=1}^{N} (\frac{x_j}{m_j})^{-\alpha}\right) \exp\left[\sum_{j=1}^{N} \frac{x_j}{m_j}\right]$$

et par conséquent

$$-\ln p(\boldsymbol{x}) = cte - \alpha \sum_{j=1}^{N} \ln(\frac{x_j}{m_j}) + \sum_{j=1}^{N} \frac{x_j}{m_j}.$$

– Loi Béta:

$$\phi(x_j) = \alpha \ln(x_j) + \beta \ln(1 - x_j), \quad \longrightarrow \quad p(x_j) \propto x_j^{\alpha}(1 - x_j)^{\beta}.$$

et la loi conjointe est:

$$p(\boldsymbol{x}) \propto \prod_{j=1}^{N} x_j^{\alpha}(1 - x_j)^{\beta} \quad \longrightarrow \quad -\ln p(\boldsymbol{x}) = cte + \alpha \sum_{j=1}^{N} \ln(x_j) + \beta \sum_{j=1}^{N} \ln(1 - x_j).$$

Nous avons vu que dans le cas i.i.d., la forme générale de ces lois est:

$$p(\boldsymbol{x}) \propto \exp\left[-\sum_{k=1}^{K} \lambda_k \phi_k(\boldsymbol{x})\right], \quad \text{avec} \quad \phi_k(\boldsymbol{x}) = \sum_{j=1}^{N} \phi_k(x_j).$$

On peut se poser alors une question sur le choix des fonctions $\phi_k(x_j)$. Par exemple, on peut se demander s'il existe des familles de fonctions ϕ_k de telle sorte que la lois $p(x_j)$ et par conséquent la loi $p(\boldsymbol{x})$ aient une propriété, par exemple d'invariance par changement d'échelle. Dans une étude récente, certains auteurs se sont posés de telles questions et ont essayé d'y répondre. Ici, nous résumons quelques uns de ces résultats.

L'idée de base se trouve dans la recherche des fonctions ϕ_k de telle sorte que la famille de lois $p(x_j)$ correspondantes soit fermée pour un groupe de transformations comme le changement d'échelle par exemple.

Prenons le cas du changement d'échelle pour illustrer cette idée. Si x_j représente une quantité et $p(x_j)$ sa loi de probabilité, on voudrait que lors d'un changement d'échelle αx_j avec $\alpha > 0$, la loi $p(\alpha x_j)$ reste dans la même famille que $p(x_j)$. Autrement dit:

$$p(x_j) \propto \exp\left[-\sum_{k=1}^{K} \lambda_k \phi_k(x_j)\right] \longrightarrow p(\alpha x_j) \propto \exp\left[-\sum_{k=1}^{K} \lambda'_k \phi_k(x_j)\right]$$

où λ'_k ne doivent dépendre que des λ_k et de facteur d'échelle α.

Il est alors facile de montrer que pour satisfaire à cette propriété il suffit d'avoir:

$$\forall x_j \ \& \ \forall \alpha > 0, \quad \sum_{k=1}^{K} \lambda_k \phi_k(x_j) = cte + \sum_{k=1}^{K} \lambda'_k \phi_k(x_j)$$

L'étude de cette propriété pour $K = 1$ et $K = 2$ donne les familles suivantes:

Lois à un paramètre $K = 1$:

$$\left\{\phi(x)\right\} = \left\{x^r, \ln x\right\}, \quad r > 0.$$

Lois à deux paramètres $K = 2$:

$$\left\{(\phi_1(x), \phi_2(x))\right\} = \left\{(x^{r_1}, x^{r_2}), (x^{r_1}, \ln x), (x^{r_1}, x^{r_1} \ln x), (\ln x, \ln^2 x)\right\},$$

où r, r_1 et r_2 sont des réels positifs.

On retrouve quelques cas particuliers intéressants en choisissant $\phi_2(x) = x$ et pour différents choix de $\phi_1(x)$:

- $\phi_1(x) = x^2$; on retrouve la loi gaussienne :

$$p(x) \propto \exp\left[-\lambda x^2 - \mu x\right] \propto \exp\left[-\lambda\left(x + \frac{\mu}{2\lambda}\right)^2\right] = \mathcal{N}\left(m = \frac{-\mu}{\lambda}, \sigma^2 = \frac{1}{2\lambda}\right).$$

- $\phi_1(x) = \ln x$; on retrouve la loi gamma :

$$p(x) \propto \exp\left[-\lambda \ln x - \mu x\right] = x^{-\lambda} \exp\left[-\mu x\right].$$

- $\phi_1(x) = x \ln x$; on retrouve une loi dite de la forme entropie de Shannon :

$$p(x) \propto \exp\left[-\lambda x \ln x - \mu x\right].$$

9.5.2 Modèles markoviens

La modélisation markovienne a pris une importance sans égale en traitement d'image depuis les travaux de Geman et Geman. L'objet de ce document n'est pas de donner une présentation complète de la modélisation markovienne, mais plutôt une présentation différente.

Nous avons vu dans la section précédente que, d'une manière générale, l'utilisation du principe du maximum d'entropie avec une information *a priori* de la forme :

$$\mathrm{E}\left[\phi_k(\boldsymbol{x})\right] = d_k, \quad k = 1, \cdots, K,$$

nous conduit aux lois de la forme :

$$p(\boldsymbol{x}) = \frac{1}{Z} \exp\left[-\sum_{k=1}^{K} \lambda_k \phi_k(\boldsymbol{x})\right].$$

Considérons maintenant le cas où les fonctions $\phi_k(\boldsymbol{x})$ sont de la forme

$$\phi_k(\boldsymbol{x}) = \sum_j \sum_{i \in V_j} \phi_k(x_j, x_i)$$

où V_j signifie les (échantillons ou les pixels) voisins de j. Les fonctions ϕ_k sont alors appelées *potentiels* et

$$U(\boldsymbol{x}) = \sum_j \sum_{i \in V_j} \phi_k(x_j, x_i)$$

l'énergie.

Un tel choix pour ces fonctions signifie que nous faisons l'hypothèse qu'il y a une corrélation locale entre les valeurs du signal aux instants successifs ou de l'image sur des pixels voisins.

Un cas particulier très courant est lorsque $k = 1$ et $\phi_k(x_j, x_i) = \phi_k(x_j - x_i)$. On a alors

$$p(\boldsymbol{x}) \propto \exp\left[-\lambda \sum_j \sum_{i \in V_j} \phi(x_j - x_i)\right]$$

Le choix de la fonction potentiel ϕ est crucial : les fonctions convexes permettront de modéliser des signaux et des champs continus et les fonctions non convexes permettrons de modéliser des signaux et des champs qui contiennent des discontinuitées. Une liste de ces fonctions classiquement utilisées en traitement du signal et des images est fournie en annexe D.

9.6 Quelques exemples de construction à la main

Dans cette section, nous allons montrer comment on peut choisir une loi de probabilité *a priori* à la main et avec du bon sens. La figure qui suit montre un ensemble de signaux très différents.
Nous allons, pour chacune de ces formes de signaux proposer une famille de lois convenable.

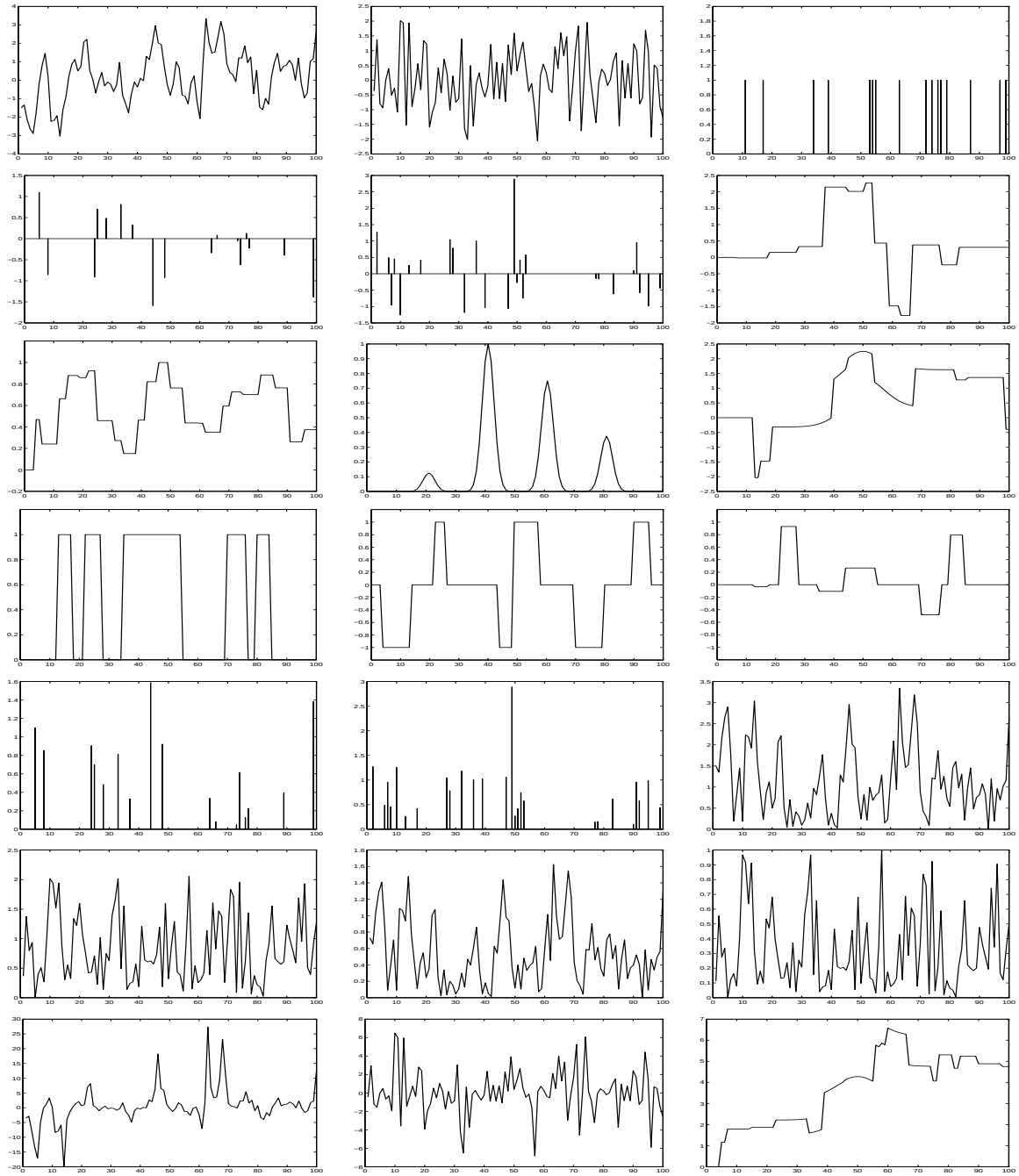
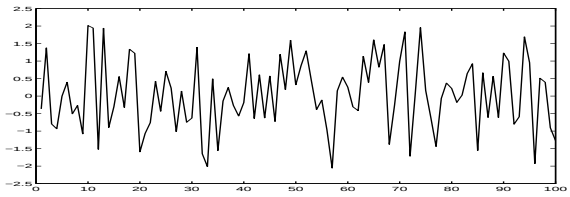


FIG. 9.1 - *Différents types de signaux.*

1. **Modèle gaussien blanc pour des signaux continus à variation rapide**



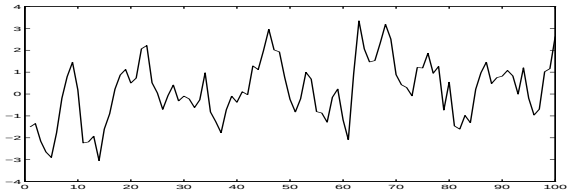
En regardant la forme de ce signal et en constatant que les échantillons du signal sont réparties d'une manière symétrique autours de zéro, on peut proposer une loi gaussienne centrée :

$$\boldsymbol{x} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{R}_x = \sigma_x^2 \boldsymbol{I}\right) \longrightarrow p(\boldsymbol{x}) \propto \exp\left[-\frac{1}{2\sigma_x^2}\|\boldsymbol{x}\|^2\right]$$

ou encore

$$p(\boldsymbol{x}) \propto \exp\left[-\frac{1}{2\sigma_x^2}\sum_j x_j^2\right].$$

2. **Modèle gaussien coloré pour des signaux continus à variation lente :**



En regardant la forme de ce signal, et en la comparant au cas précédent, on peut proposer une loi centré, gaussienne mais colorée :

$$\boldsymbol{x} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{R}_x = \sigma_x^2 \boldsymbol{P}_0\right) \longrightarrow p(\boldsymbol{x}) \propto \exp\left[-\frac{1}{2\sigma_x^2}\boldsymbol{x}^t \boldsymbol{P}_0^{-1}\boldsymbol{x}\right],$$

avec $\boldsymbol{P}_0$ une matrice symétrique et définie positive, ce qui permet d'écrire $\boldsymbol{P}_0^{-1} = \boldsymbol{D}^t\boldsymbol{D}$ et on a alors

$$p(\boldsymbol{x}) \propto \exp\left[-\frac{1}{2\sigma_x^2}\|\boldsymbol{D}\boldsymbol{x}\|^2\right].$$

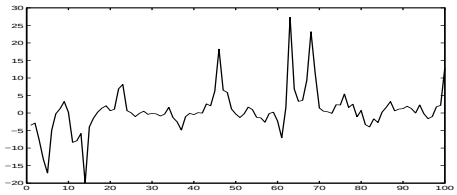
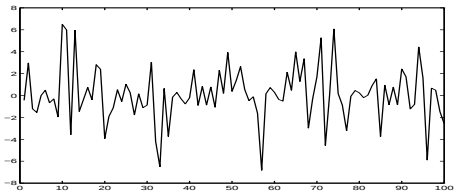
Un choix possible pour $\boldsymbol{D}$ est la matrice de différences finies d'ordre un :

$$\boldsymbol{D}_1 = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -1 & 1 & \ddots & & \vdots \\ 0 & -1 & 1 & \ddots & \vdots \\ \vdots & \ddots & & & 0 \\ 0 & \cdots & 0 & -1 & 1 \end{pmatrix}$$

ce qui permet d'écrire :

$$p(\boldsymbol{x}) \propto \exp\left[-\frac{1}{2\sigma_x^2}\sum_j (x_j - x_{j-1})^2\right]$$

3. **Modèle gaussien généralisé pour des signaux continus plutôt impulsionnels**

En regardant la forme de ce signal, et en la comparant au cas précédent, on peut constater que la proportion des échantillons ayant des valeurs proches de zéro est plus importante que celle des échantillons ayant des valeurs plus grandes et que cette proportion semble décroître très rapidement. On peut alors proposer une loi gaussienne généralisée:

$$p(\boldsymbol{x}) \propto \exp\left[-\beta \|\boldsymbol{D}_k \boldsymbol{x}\|^p\right], \quad 1 < p \leq 2,$$

avec $\boldsymbol{D}_0 = \boldsymbol{I}$ et $\boldsymbol{D}_k$ la matrice des différences finies d'ordre k.

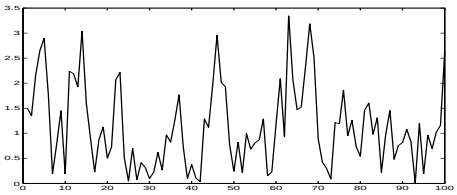
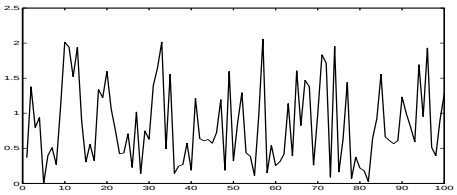
Dans le cas $\boldsymbol{D}_0 = \boldsymbol{I}$ on a

$$p(\boldsymbol{x}) \propto \exp\left[-\beta \sum_j |x_j|^p\right]$$

et dans le cas $\boldsymbol{D}_1$ on a

$$p(\boldsymbol{x}) \propto \exp\left[-\beta \sum_j |x_j - x_{j-1}|^p\right].$$

4. **Modèle des signaux positifs**

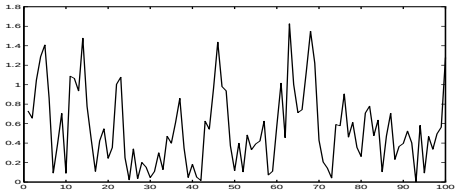
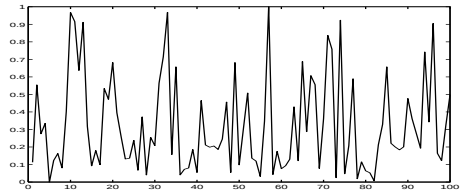
Loi gaussienne généralisée tronquée:

$$p(\boldsymbol{x}) \propto \exp\left[-\beta \sum_j x_j^p\right], \quad x_j > 0.$$

Loi Gamma:

$$p(\boldsymbol{x}) \propto \exp\left[-\alpha \sum_j \ln x_j - \beta \sum_j x_j\right], \quad x_j > 0.$$

5. **Modèle des signaux continus mais bornés entre 0 et 1 : Loi Béta**

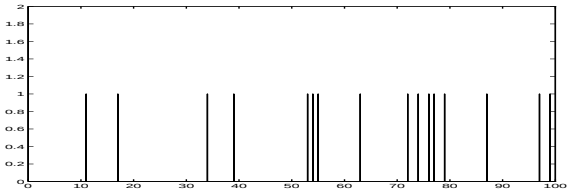
$$p(\boldsymbol{x}) = \prod_{j=1}^{N} p(x_j),$$

avec

$$\begin{aligned} p(x_j) &\propto x_j^{-\alpha} - (1-x_j)^{-\beta} \\ &\propto \exp\left[-\alpha \ln x_j - \beta \ln(1-x_j)\right] \end{aligned}$$

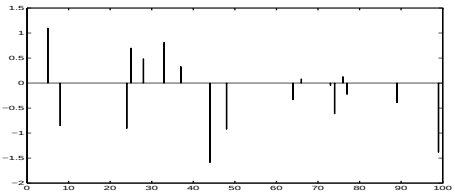
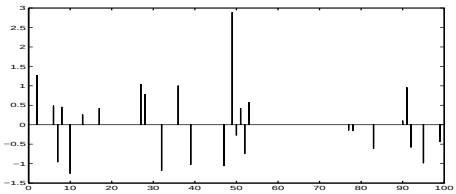
$$-\ln p(\boldsymbol{x}) = cte + \alpha \ln x_j + \beta \ln(1-x_j)$$

6. **Modèle de Bernoulli :**



$$\begin{cases} P(X_j = 1) = \alpha, \\ P(X_j = 0) = 1 - \alpha \end{cases} \longrightarrow P(\boldsymbol{X} = \boldsymbol{x}) = \alpha^s (1-\alpha)^{1-s} \text{ avec } s = \sum(x_j)$$

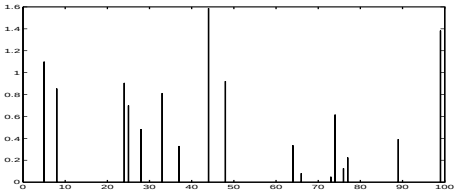
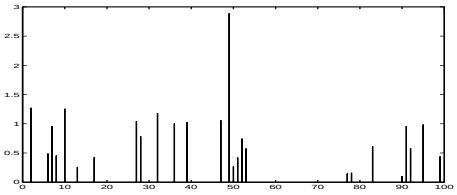
7. **Modèle Bernoulli-Gaussien :**

$$p(x_j|q_j) = \begin{cases} 0 & \text{si } \; q_j = 0, \\ \mathcal{N}(0, \sigma_x^2) & \text{si } \; q_j = 1 \end{cases} \qquad \text{et} \qquad \begin{cases} P(q_j = 1) = \lambda, \\ P(q_j = 0) = 1 - \lambda \end{cases}$$

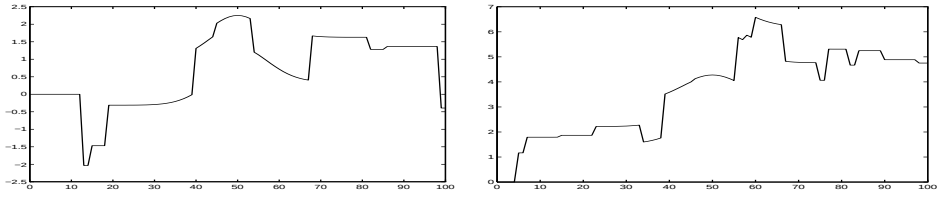
$$p(\boldsymbol{x}|\boldsymbol{q}) \propto \exp\left[\frac{-1}{2q_j\sigma_x^2}\sum_j q_j x_j^2\right]$$

8. **Modèle Bernoulli-Gamma :**

$$p(x_j|q_j) = \begin{cases} 0 & \text{si } \; q_j = 0, \\ \text{Gamma}(\alpha, \beta) & \text{si } \; q_j = 1 \end{cases} \qquad \text{et} \qquad \begin{cases} P(q_j = 1) = \lambda, \\ P(q_j = 0) = 1 - \lambda \end{cases}$$

9. **Modèles pour les signaux continus par morceaux**

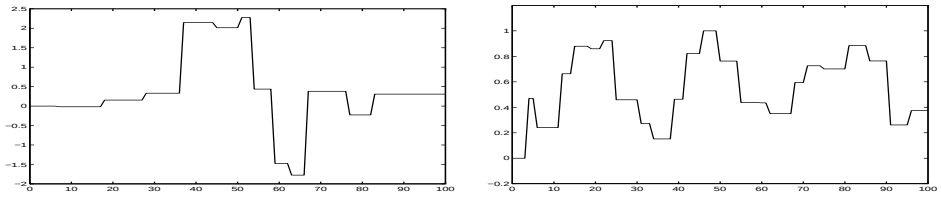


$$p(x_j|x_{j-1},q_j) = \begin{cases} \mathcal{N}(x_{j-1},\sigma_1^2) & \text{si } q_j = 0 \\ \mathcal{N}(0,\sigma_2^2) & \text{si } q_j = 1 \end{cases}$$

ou

$$p(\boldsymbol{x}|\boldsymbol{q}) \propto \exp\left[-\frac{1}{\sigma_1^2}\sum_j q_j(x_j - x_{j-1})\right]$$

10. **Modèles pour les signaux constants par morceaux**



$$p(x_j|x_{j-1},q_j) = \begin{cases} \delta(x_j - x_{j-1}) & \text{si } q_j = 0 \\ \mathcal{N}(0,\sigma_2^2) & \text{si } q_j = 1 \end{cases}$$

9.7 Estimation des paramètres d'une loi à partir des observations directes

Nous avons vu que le PME peut nous permettre d'attribuer une loi de probabilité $p(x)$ à une quantité X pour traduire une information incomplète sur X si cette information porte sur des moments $\mathrm{E}[\phi_k(X)] = d_k$. Malheureusement, en pratique, rarement nous connaissons les valeurs numériques de d_k. C'est pourquoi, en général, l'attribution d'une loi de probabilité se fait en deux étapes :

1. Choix d'une famille de loi qui est équivalent au choix des fonctions ϕ_k ;

2. Détermination des valeurs des paramètres à partir d'un certain nombres d'observations directes ou indirectes du X.

Dans cette section, nous allons nous limiter au cas où nous avons la possibilité d'observer directement X. Supposons maintenant que l'étape 1 est faite et que nous avons choisi une loi $p(x;\boldsymbol{\theta})$ dépendant des paramètres $\boldsymbol{\theta} = [\theta_1, \cdots, \theta_K]$. Considérons alors le cas où nous avons N échantillons $\{x_1, \cdots, x_N\}$. Nous cherchons à déterminer $\boldsymbol{\theta}$ à partir de ces échantillons. Deux méthodes classiquement utilisées sont : la méthode des moments (MM) et la méthode du maximum de vraisemblance (MV).

9.7.1 Méthode des moments

L'idée de base dans cette méthode est de déterminer $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K)$ en utilisant un système d'équations reliant les moments théoriques $\mathrm{E}\left[X^k\right]$ et les moments empiriques $\overline{X^k} = \frac{1}{N}\sum_{j=1}^{N} x_j^k$:

$$\mathrm{E}\left[X^k\right] = \int x^k \, p(x;\boldsymbol{\theta}) \, \mathrm{d}x = \frac{1}{N}\sum_{j=1}^{N} x_j^k, \quad k = 1, \cdots, M$$

avec $M \geq K$ et en espérant qu'il existe une solution unique à ce système d'équations (non linéaires en $\boldsymbol{\theta}$).

Pour illustrer cette méthode considérons les deux exemples suivants :

Exemple 1 :
Cas d'une loi gaussienne à deux paramètres $\boldsymbol{\theta} = [m, \sigma^2]$

$$p(x; m, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-m)^2\right].$$

Si on forme le système d'équations :

$$\begin{aligned} \mathrm{E}[X] &= m = \frac{1}{N}\sum_{j=1}^{N} x_j \\ \mathrm{E}\left[X^2\right] &= \sigma^2 + m^2 = \frac{1}{N}\sum_{j=1}^{N} x_j^2 \end{aligned}$$

il est alors facile de voir qu'il y a une solution qui est :

$$\begin{aligned} m &= \overline{X} = \frac{1}{N}\sum_{j=1}^{N} x_j \\ \sigma^2 &= \overline{(X-m)^2} = \frac{1}{N}\sum_{j=1}^{N}(x_j - m)^2. \end{aligned}$$

Exemple 2 :
Cas d'une loi gamma à deux paramètres $\boldsymbol{\theta} = [\alpha, \beta]$

$$p(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma\alpha} x^{\alpha-1}\exp\left[-\beta x\right].$$

Si on forme le système d'équations :

$$\begin{aligned} \mathrm{E}\left[X\right] &= \frac{\alpha}{\beta} = \frac{1}{N}\sum_{j=1}^{N} x_j, \\ \mathrm{E}\left[X^2\right] &= \frac{\alpha(1+\alpha)}{\beta^2} = \frac{1}{N}\sum_{j=1}^{N} x_j^2, \end{aligned}$$

et si on note par $m = \overline{X} = \frac{1}{N}\sum_{j=1}^{N} x_j$ et par $v = \overline{(X-m)^2} = \frac{1}{N}\sum_{j=1}^{N}(x_j - m)^2$, il est alors facile de voir qu'il y a une solution qui est :

$$\begin{aligned} \alpha &= \frac{2v - m^2}{v}, \\ \beta &= \frac{m}{v}. \end{aligned}$$

Les inconvenients majeurs de cette méthode sont :

- l'absence de base théorique pour cette méthode ;
- la sensibilité au nombre de données N de ces estimateurs.

9.7.2 Méthode du maximum de vraisemblance

Estimation au sens du MV consiste à définir la fonction de vraisemblance $V(\boldsymbol{\theta}) = p(x_1, \cdots, x_n; \boldsymbol{\theta})$ ou plutôt la fonction log-vraisemblance $L(\boldsymbol{\theta}) = -\ln V(\boldsymbol{\theta})$ et à calculer l'argument $\widehat{\boldsymbol{\theta}}$ qui la maximise :

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \{V(\boldsymbol{\theta})\} = \arg\min_{\boldsymbol{\theta}} \{L(\boldsymbol{\theta})\}$$

Le principal avantage de cet estimateur est qu'il est efficace.

Pour illustrer cette méthode et la comparer à la méthode des moments nous allons considérer les deux exemples de la section précédente.

Exemple 1 :
Cas d'une loi gaussienne à deux paramètres $\boldsymbol{\theta} = [m, \sigma^2]$

$$p(x; m, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-m)^2\right]$$

La fonction du log-vraisemblance s'écrit :

$$L(m, \sigma^2) = \frac{N}{2}\ln(2\pi) + \frac{N}{2}\ln\sigma^2 + \frac{1}{2\sigma^2}\sum_{j=1}^{N}(x_j - m)^2.$$

Pour calculer l'argument qui la minimise, il faut annuler simultanément la dérivée par rapport à m et la dérivée par rapport à σ^2, ce qui permet d'obtenir un résulte bien connu :

$$\begin{aligned} m &= \frac{1}{N}\sum_{j=1}^{N} x_j, \\ \sigma^2 &= \frac{1}{N-1}\sum_{j=1}^{N}(x_j - m)^2. \end{aligned}$$

Notons la seule différence par rapport aux résultats de la méthode des moments apparaît dans l'estimation du σ^2.

Exemple 2 :
Cas d'une loi gamma à deux paramètres $\boldsymbol{\theta} = [\alpha, \beta]$

$$p(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp\left[-\beta x\right].$$

La fonction du log-vraisemblance s'écrit :

$$L(\alpha, \beta) = N[\alpha \ln\beta - \ln\Gamma(\alpha)] - (\alpha - 1)\sum_{j=1}^{N}\ln x_j - \beta\sum_{j=1}^{N} x_j.$$

Pour calculer l'argument qui la minimise, il faut annuler simultanément la dérivée par rapport à α et la dérivée par rapport à β :

$$\begin{aligned} \frac{\partial L}{\partial \alpha} &= N\left[\ln\beta - \frac{\partial \ln\Gamma(\alpha)}{\partial\alpha}\right] - \sum_{j=1}^{N}\ln x_j = 0 \\ \frac{\partial L}{\partial \beta} &= \frac{\alpha}{\beta} - \sum_{j=1}^{N} x_j = 0 \end{aligned}$$

Comme on peut le constater, il n'y a pas de solution analytique à ces deux équations, mais on peut la calculer numériquement.

Pour montrer cependant qu'il peut y avoir un lien entre ces deux méthodes, on peut reécrire les relations différemment.

Supposons que la loi $p(x; \boldsymbol{\theta})$ soit de la forme générale :

$$p(x; \boldsymbol{\theta}) \propto \exp\left[-\sum_{k=1}^{K}\theta_k \phi_k(x)\right].$$

On a alors

$$L(\boldsymbol{\theta}) = -\ln p(x_1, \cdots, x_N; \boldsymbol{\theta}) = cte + \sum_{j=1}^{N} \sum_{k=1}^{K} \theta_k \phi_k(x_j).$$

On peut alors montrer facilement que, dans la méthode des moments nous avons à résoudre

$$\int x^k \, p(x; \boldsymbol{\theta}) \, \mathrm{d}x = \frac{1}{N} \sum_{j=1}^{N} x_j^k, \quad k = 0, \cdots, K,$$

et dans la méthode du MV nous avons à résoudre

$$\int \phi_k(x) \, p(x; \boldsymbol{\theta}) \, \mathrm{d}x = \frac{1}{N} \sum_{j=1}^{N} \phi_k(x_j), \quad k = 0, \cdots, K,$$

où nous avons noté $\phi_0(x) = 1$.
On peut alors remarquer que les deux méthodes sont équivalentess pour la famille des lois exponentielles où $\phi_k(x) = x^k$.

9.8 Estimation des hyperparamètres

Il existe assez peu de méthodes de détermination des hyperparamètres. Dans le cas où ceux-ci se limitent au seul coefficient de régularisation, et où le régulariseur est quadratique, les méthodes de *validation croisée* fournissent des solutions acceptables. Mais il s'agit de méthodes déterministes par essence, qui minimisent un critère de risque dépendant de l'objet inconnu $\boldsymbol{f}$, ce qui ne peut être effectué qu'asymptotiquement puisque cet objet est évidemment inconnu.

Les méthodes bayésienne, qui attribuent une distribution de probabilité *a priori* à l'objet, ne présentent pas cette limitation. Les hyperparamètres θ constituent un second niveau de description du problème, indispensable pour "rigidifier" le premier niveau constitué par les paramètres eux-mêmes— c'est-à-dire l'objet $\boldsymbol{f}$. Dans un problème mal-posé, la valeur des hyperparamètres est importante pour obtenir une solution acceptable, mais ne présente pas d'intérêt en soi. Dans une approche bayésienne, on peut donc distinguer deux niveaux d'inférence. Le premier infère sur $\boldsymbol{f}$, pour une valeur donnée de θ, au travers de la distribution *a posteriori*. Le second infère sur θ grâce à une relation analogue :

$$p(\theta \mid \boldsymbol{g}, \boldsymbol{H}) = \frac{p(\theta \mid \boldsymbol{H})\, p(\boldsymbol{g} \mid \theta, \boldsymbol{H})}{p(\boldsymbol{g} \mid \boldsymbol{H})} \,. \tag{9.5}$$

On retrouve là une caractéristique de l'utilisation de la règle de Bayes : la vraisemblance $p(\boldsymbol{g} \mid \theta, \boldsymbol{H})$ attachée aux données dans le second niveau est le coefficient de normalisation dans le premier.

Si, comme cela est souvent le cas, ce terme est suffisamment "piqué", l'influence de la distribution *a priori* $p(\theta \mid \boldsymbol{H})$ est négligeable, et le second niveau d'inférence peut être résolu par maximisation de cette vraisemblance. Mais il faut pour cela résoudre un problème de marginalisation :

$$p(\boldsymbol{g} \mid \theta, \boldsymbol{H}) = \int p(\boldsymbol{f}, \boldsymbol{g} \mid \theta, \boldsymbol{H})\, \mathrm{d}\boldsymbol{f} = \int p(\boldsymbol{g} \mid \boldsymbol{f}, \theta, \boldsymbol{H})\, p(\boldsymbol{f} \mid \theta)\, \mathrm{d}\boldsymbol{f} \,. \tag{9.6}$$

Une telle intégrale conduit très rarement à une forme explicite.

Pour contourner cette difficulté, on peut introduire des "variables cachées" $\boldsymbol{z}$ qui viennent compléter les observations $\boldsymbol{g}$ de manière à ce que la nouvelle vraisemblance $p(\boldsymbol{g}, \boldsymbol{z} \mid \theta, \boldsymbol{H})$ soit plus simple à calculer. On est alors conduit à maximiser des espérances conditionnelles par des techniques itératives, déterministes ou stochastiques (algorithmes EM et SEM).

On peut aussi remarquer que la *vraisemblance généralisée*

$$p(\boldsymbol{g}, \boldsymbol{f} \mid \theta, \boldsymbol{H}) = p(\boldsymbol{f} \mid \boldsymbol{g}, \theta, \boldsymbol{H})\, p(\boldsymbol{g} \mid \theta, \boldsymbol{H}) = p(\boldsymbol{g} \mid \boldsymbol{f}, \theta, \boldsymbol{H})\, p(\boldsymbol{f} \mid \theta) \tag{9.7}$$

résume toute l'information propre au premier niveau d'inférence, et vouloir en faire la maximisation conjointe par rapport à $\boldsymbol{f}$ et θ. Le problème d'intégration soulevé par (9.6) est évidemment évacué. A θ fixé, le maximum de vraisemblance généralisée (MVG) coïncide avec le MAP. Par contre, à $\boldsymbol{f}$ fixé, la situation est beaucoup moins favorable : la vraisemblance généralisée n'est pas, en général, majorée sur son domaine de définition, et il n'existe même pas toujours un maximum local.

A COMPLETER

Chapitre 10

Modélisation markovienne

Dans ce chapitre nous donnons un apérçu général de la modélisation markovienne des signaux et des images. Les modèles markoviens et particulèrement les champs de Markov et les champs de Gibbs sont devenus des outils indispensables pour la traduction d'une information *a priori* plus élaborée que la douceuer ou la positivité en une loi de probabilité. En effet, la modélisation markovienne permet de construire des modèles qui peuvent prendre en compte les discontinuités dans un signal ou une image. Cette outil prend particulièrement son importance en traitement d'image où l'on peut construire les modèles composites (bords, contours, intensité) pour les images et les uitliser lors de la segmentation, restauration ou reconstruction d'image.