Segmentation par champ de Markov (1/2)

Retour sur la notion de variable aléatoire

Les images numériques constituent une bonne illustration de la notion de *variable aléatoire*. En chaque « site » s d'une image en niveaux de gris (pour le moment, un site désigne un pixel), on définit une variable aléatoire (VA), notée X_s, qui prend ses valeurs dans l'ensemble d'entiers $\{0, 1, 2, \ldots, 255\}$. Le niveau de gris x_s de s est une *réalisation* de X_s.

Une VA X est dite « normale » si elle suit une *loi normale*, ce qui s'écrit de la manière suivante en dimension 1 (nous noterons parfois $p(x)$ au lieu de $p(X = x)$, afin de simplifier les notations) :

$$p(X = x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\,\sigma^2}\right\} \tag{1}$$

où μ désigne la moyenne et σ l'écart-type (donc σ^2 la variance). La loi (1) est sommable à 1, ce qui sous-entend que $x \in \mathbb{R}$. On remarque donc d'emblée que la VA X_s ne peut pas réellement constituer une VA normale, puisque le niveau de gris x_s est entier et borné. Néanmoins, la loi normale modélise bien de nombreux phénomènes naturels. On peut donc supposer que X_s constitue une VA normale discrète, tronquée.

En chaque site s d'une image, on ne dispose que d'une seule réalisation de la VA X_s. On ne peut donc pas estimer les paramètres (μ, σ) de la loi. Mais si n pixels d'une image sont i.i.d., c'est-à-dire *indépendants et identiquement distribués*, ce qui implique qu'ils suivent la même loi normale, alors il devient possible d'estimer les paramètres de cette loi. Or, la méthode la plus générale d'estimation est celle du *maximum de vraisemblance*.

La vraisemblance des réalisations x_1, x_2, …, x_n de n VA normales i.i.d. s'écrit :

$$L(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x_i-\mu)^2}{2\,\sigma^2}\right\} \tag{2}$$

Cette expression prend la forme d'un produit car les n VA sont i.i.d. Plutôt que la vraisemblance elle-même, on préfère souvent maximiser la log-vraisemblance, car la fonction logarithme transforme le produit en somme, et qu'une somme est plus facile à maximiser qu'un produit. L'estimation des paramètres (μ, σ) s'écrit donc :

$$(\widehat{\mu}, \widehat{\sigma}) = \underset{(\mu,\sigma)\in\mathbb{R}\times\mathbb{R}^+}{\arg\max} \left\{ \sum_{i=1}^{n} \left[-\ln\left(\sigma\sqrt{2\pi}\right) - \frac{(x_i-\mu)^2}{2\,\sigma^2} \right] \right\} \tag{3}$$

soit encore :

$$(\widehat{\mu}, \widehat{\sigma}) = \underset{(\mu,\sigma)\in\mathbb{R}\times\mathbb{R}^+}{\arg\min} \left\{ n\ln\sqrt{2\pi} + n\ln\sigma + \frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \right\} \tag{4}$$

Les conditions nécessaires d'optimalité du premier ordre (dérivées partielles en μ et en σ égales à 0) s'écrivent :

$$\begin{cases} \displaystyle\sum_{i=1}^{n}(x_i-\mu) = 0 \\ \displaystyle\frac{n}{\sigma} - \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 = 0 \end{cases} \iff \begin{cases} \displaystyle\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \\ \displaystyle\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 \end{cases} \tag{5}$$

On retrouve, sans surprise, les définitions de la moyenne et de la variance.

Segmentation d'une image par classification

Les pixels de l'image de la figure 1 sont indépendants, mais pas identiquement distribués. Ils sont répartis en $N = 4$ zones homogènes, ou « segments ». La *segmentation*, qui consiste à effectuer une partition des pixels permettant de séparer les segments, est une des tâches les plus courantes de la vision par ordinateur.

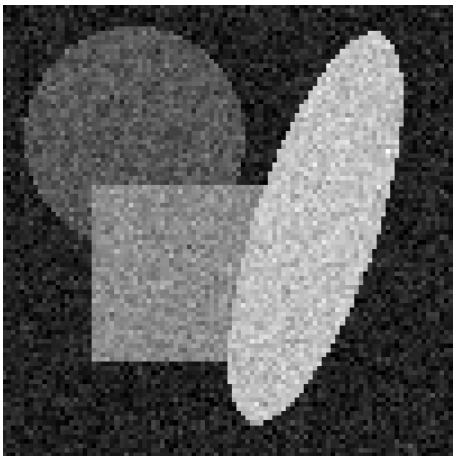

FIGURE 1 – Image de synthèse comportant $N = 4$ zones homogènes, ou « segments ».

Il existe trois familles de méthodes de segmentation, qui reposent sur trois principes très différents : recherche de contours, recherche de régions, ou classification. C'est cette dernière approche qui nous intéresse, car la *segmentation par classification* (cf. TP3) illustre bien l'intérêt des champs de Markov, qui ont néanmoins de nombreuses autres applications (cf. TP4).

Classification par le maximum de vraisemblance

En chaque site s, nous définissons une deuxième VA, notée K_s, qui prend ses valeurs dans l'ensemble d'entiers $E = \{1, 2, \ldots, N\}$, où N désigne le nombre de « classes », chaque classe étant censée coïncider avec un segment. L'indice $k_s \in E$ de la classe à laquelle est affecté le pixel s est une réalisation de K_s.

En supposant que chacune des N classes suit une loi normale, nous pouvons écrire la vraisemblance de x_s, relativement à la classe k_s, sous la forme d'une probabilité conditionnelle (là encore, on note parfois $p(x_s|k_s)$ au lieu de $p(X_s = x_s|K_s = k_s)$, afin de simplifier les notations) :

$$p(X_s = x_s|K_s = k_s) = \frac{1}{\sigma_{k_s}\sqrt{2\pi}} \exp\left\{-\frac{(x_s - \mu_{k_s})^2}{2\,\sigma_{k_s}^2}\right\} \tag{6}$$

La classification par le *maximum de vraisemblance* revient à affecter s à la classe qui maximise sa vraisemblance :

$$\begin{aligned} \widehat{k_s} &= \arg\max_{k_s \in E} \{p(X_s = x_s|K_s = k_s)\} \\ &= \arg\min_{k_s \in E} \left\{\ln \sigma_{k_s} + \frac{(x_s - \mu_{k_s})^2}{2\,\sigma_{k_s}^2}\right\} \end{aligned} \tag{7}$$

Pour pouvoir utiliser cette méthode de classification, il est nécessaire de connaître non seulement le nombre N de classes, mais également les paramètres $(\mu_{k_s}, \sigma_{k_s})_{k_s \in E}$. Cela peut être fait en sélectionnant un ensemble de pixels appartenant à la même classe, et en utilisant les estimations (5), ce qui nécessite l'intervention d'un utilisateur. En ce sens, on dit qu'il s'agit de *classification supervisée*, par opposition à d'autres méthodes de classification, généralement moins précises, mais qui sont entièrement automatiques.

Sur l'image de la figure 1, pour $N = 4$, cette méthode permet d'obtenir environ 94% de bonnes classifications. Mais comme il est élémentaire de segmenter une telle image « à main levée », on subodore qu'il sera possible d'atteindre des scores bien plus élevés (cf. TP3).

Classification par le maximum a posteriori

Une variante légèrement plus sophistiquée de la méthode précédente est la classification par le *maximum a posteriori* (MAP). Le théorème de Bayes donne l'expression de la *probabilité a posteriori* de x_s, c'est-à-dire de la probabilité que le pixel s de niveau de gris x_s appartienne à la classe k_s :

$$p(K_s = k_s | X_s = x_s) = \frac{p(X_s = x_s | K_s = k_s)\, p(K_s = k_s)}{p(X_s = x_s)} \tag{8}$$

Dans cette expression, $p(X_s = x_s | K_s = k_s)$ désigne la *vraisemblance*, c'est-à-dire la probabilité que le pixel s appartenant à la classe k_s ait un niveau de gris égal à x_s, $p(K_s = k_s)$ désigne la *probabilité a priori* de la classe k_s, et $p(X_s = x_s)$ la *probabilité a priori* du niveau de gris x_s. La classification par le MAP consiste à affecter le pixel s à la classe qui maximise sa probabilité a posterirori :

$$\begin{aligned} \widehat{k_s} &= \arg\max_{k_s \in E} \{p(K_s = k_s | X_s = x_s)\} \\ &= \arg\max_{k_s \in E} \{p(X_s = x_s | K_s = k_s)\, p(K_s = k_s)\} \\ &= \arg\min_{k_s \in E} \{\ln \sigma_{k_s} + \frac{(x_s - \mu_{k_s})^2}{2\,\sigma_{k_s}^2} - \ln p(K_s = k_s)\} \end{aligned} \tag{9}$$

En guise d'a priori $p(K_s = k_s)$ sur la classe k_s, nous n'avons pas tellement d'autre choix que d'utiliser la proportion de pixels appartenant à cette classe. Le problème est qu'il est nécessaire de connaître le résultat de la segmentation pour pouvoir comptabiliser les pixels appartenant à chacune des classes ! Mais cette idée n'est pas aussi absurde qu'il y paraît. Commençons par calculer l'histogramme de l'image de la figure 1 :

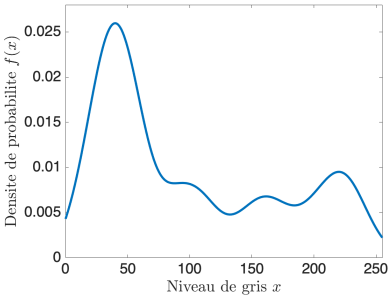

FIGURE 2 – Histogramme de l'image de la figure 1.

Comme cette densité de probabilité $f(x)$ présente $N = 4$ maxima locaux, il semble légitime de la modéliser par un « mélange » de $N = 4$ gaussiennes :

$$f(x) = \sum_{i=1}^{N} \frac{p_i}{\sigma_i \sqrt{2\pi}} \exp\left\{-\frac{(x - \mu_i)^2}{2\,\sigma_i^2}\right\} \tag{10}$$

où μ_i, σ_i et p_i désignent, respectivement, la moyenne, l'écart-type et le poids de la $i^{\text{ème}}$ gaussienne. Si nous trouvons les valeurs de ces paramètres pour lesquelles le mélange de gaussiennes (10) est égal à l'histogramme, nous pourrons appliquer la méthode de classification (9), en identifiant μ_{k_s} à μ_i, σ_{k_s} à σ_i, et $p(K_s = k_s)$ à p_i.

Bien qu'elle soit prometteuse, cette piste est difficile à mettre en œuvre. En effet, il est difficile d'estimer les paramètres de la loi de mélange (10), estimation qui prend la forme suivante en moindres carrés :

$$(\widehat{\mu_i}, \widehat{\sigma_i}, \widehat{p_i})_{i \in E} = \arg\min_{(\mu_i, \sigma_i, p_i)_{i \in E}} \sum_{x=0}^{255} \left[f(x) - \sum_{i=1}^{N} \frac{p_i}{\sigma_i \sqrt{2\pi}} \exp\left\{-\frac{(x - \mu_i)^2}{2\,\sigma_i^2}\right\} \right]^2 \tag{11}$$

Un autre problème est que, même en utilisant les valeurs exactes de ces paramètres, ce qui est possible puisqu'il s'agit d'une image de synthèse, il reste encore environ 5% de pixels mal classés (cf. TP3).

Plutôt que rejeter la classification par MAP, nous allons nous en inspirer, mais au lieu de classer les pixels séparément, nous allons les classer simultanément, ce qui nécessite d'introduire la notion de *champs de Markov*.

Champs de Markov

En chaque pixel s, nous avons introduit deux variables aléatoires X_s et K_s. Comment traiter simultanément l'ensemble de $2n$ VA correspondant aux n pixels $s \in \mathcal{S}$, où $\mathcal{S}$ désigne l'ensemble des sites ? Il existe deux extensions de la notion de VA à un ensemble de VA :

- Un *processus aléatoire* (PA) est un ensemble de VA définies à différents instants.
- Un *champ aléatoire* (CA) est un ensemble de VA définies en différentes positions.

Comme pour une VA, un PA ou un CA peuvent être continus ou discrets. Nous nous intéressons aux CA discrets. En effet, une image $\mathbf{x} = (x_s)_{s\in\mathcal{S}}$ peut être considérée comme la réalisation d'un CA $\mathbf{X} = (X_s)_{s\in\mathcal{S}}$. De même, vis-à-vis de notre problème de segmentation par classification, une *configuration* $\mathbf{k} = (k_s)_{s\in\mathcal{S}}$ peut être considérée comme la réalisation d'un CA $\mathbf{K} = (K_s)_{s\in\mathcal{S}}$. Sachant que les VA K_s prennent leurs valeurs dans l'*ensemble d'état* E (dans le cas présent : $E = \{1, \ldots, N\}$), l'*espace des configurations* est donc $\Omega = E \times \cdots \times E = E^n$.

Une nouvelle formulation de la classification par MAP traitant tous les pixels simultanément est donc :

$$\begin{aligned} \widehat{\mathbf{k}} &= \arg\max_{\mathbf{k}\in\Omega} \{p(\mathbf{K} = \mathbf{k}|\mathbf{X} = \mathbf{x})\} \\ &= \arg\max_{\mathbf{k}\in\Omega} \{p(\mathbf{X} = \mathbf{x}|\mathbf{K} = \mathbf{k})\, p(\mathbf{K} = \mathbf{k})\} \end{aligned} \tag{12}$$

L'hypothèse d'indépendance des VA X_s permet d'écrire la vraisemblance de $\mathbf{x}$ sous la forme d'un produit :

$$p(\mathbf{X} = \mathbf{x}|\mathbf{K} = \mathbf{k}) = \prod_{s\in\mathcal{S}} p(X_s = x_s|K_s = k_s) \tag{13}$$

Faire la même hypothèse d'indépendance des VA K_s reviendrait à écrire $p(\mathbf{K} = \mathbf{k}) = \prod_{s\in\mathcal{S}} p(K_s = k_s)$. Sous cette hypothèse, le problème (12) se réécrirait alors :

$$(\widehat{k_s})_{s\in\mathcal{S}} = \arg\max_{(k_s)_{s\in\mathcal{S}}} \prod_{s\in\mathcal{S}} p(X_s = x_s|K_s = k_s)\, p(K_s = k_s) \tag{14}$$

problème dont les inconnues $(k_s)_{s\in\mathcal{S}}$ peuvent être découplées. Résoudre le problème (14) revient donc à résoudre la version (9) de la classification par MAP, en chaque pixel s de l'image. Or, le découplage des pixels est justement ce que nous voulons éviter. Dès lors, nous devons reformuler la probabilité a priori $p(\mathbf{K} = \mathbf{k})$ de manière à imposer des couplages entre pixels voisins. Cette propriété de couplage entre voisins caractérise les *champs de Markov*, qui sont des CA particuliers.

Avant de décrire plus en détail les champs de Markov, il est bon de faire un point sur la terminologie. De manière générale, on appelle *modèle graphique probabiliste* un graphe décrivant les relations de dépendance entre VA. Chaque nœud du graphe est une VA, tandis que chaque arête représente une dépendance entre deux VA. Il existe deux types de modèles graphiques probabilistes : si le graphe est orienté, on parle de *réseau bayésien* ; au contraire, dans le cas d'un graphe non orienté, on parle de *champ de Markov*, que l'on devrait d'ailleurs appeler champ *aléatoire* de Markov (en anglais : *Markov Random Fields*, ou MRF).

Terminons enfin cette introduction par quelques éléments d'histoire des sciences liés aux champs de Markov :

- Le Russe Andreï Markov (1856-1922) a étudié la probabilité d'apparition d'une lettre en fonction de la lettre précédente, dans le roman *Eugène Onéguine* d'Alexandre Pouchkine. Cette étude est l'ancêtre des *chaînes de Markov*, qui sont des réseaux bayésiens. En effet, une chaîne de Markov est un processus aléatoire modélisé par un graphe orienté (à cause de la disymétrie entre passé et futur).
- L'Américain William Gibbs (1839-1903), l'Écossais James Maxwell (1831-1879) et l'Autrichien Ludwig Boltzmann (1844-1906) sont les trois fondateurs de la physique statistique.
- Le Russe Roland Dobrushin (1929-1995) a été le premier à faire le lien entre champ de Markov et champ de Gibbs.
- L'Allemand Ernst Ising (1900-1998) a été le premier à utiliser un champ de Markov pour modéliser le ferromagnétisme, appelé *modèle d'Ising* (1924).
- Enfin, l'Australien Renfrey Potts (1925-2005) a proposé un champ de Markov plus général que le modèle d'Ising, appelé *modèle de Potts* (1952).