Chapitre 6

Indicateurs statistiques de l'apprentissage des données multimodales

Sommaire

6.1 Introduction

Dans le chapitre 4, nous avons présenté différents tracés permettant d'évaluer comment l'architecture de cartes extrait une représentation interne du modèle d'entrée lors de l'apprentissage, sur une architecture de cartes 1D et des entrées en une dimension également. Nous nous intéressons au développement d'indicateurs permettant de quantifier l'apprentissage du modèle d'entrée. L'existence de tels indicateurs nous permettrait de comparer plusieurs expériences entre elles de façon numérique, par exemple pour effectuer l'optimisation des paramètres d'apprentissage et de remplacer les tracés lorsque U est en plus grande dimension. Nous analysons et adaptons dans ce chapitre plusieurs méthodes permettant cette évaluation.

Nous avons défini les entrées et éléments des cartes en termes de variables aléatoires. D'autre part, nous avons observé que l'étude de l'apprentissage revient à une étude des dépendance entre les éléments des cartes et la variable latente du modèle, U. Aussi, nous nous intéressons à des mesures statistiques de relation entre signaux. Plusieurs méthodes permettent d'évaluer une telle relation statistique. Parmi ceux-ci, citons le coefficient de corrélation (Pearsons' R), le ratio de corrélation, illustrés en figure 6.1, ainsi que l'information mutuelle. Le coefficient de corrélation mesure une dépendance linéaire entre des échantillons X et Y. Il est défini par le rapport de la covariance des variables et le produit de leurs écarts-type :

$$r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \tag{6.1}$$

Ce coefficient est symétrique, vaut 0 si les variables sont indépendantes et 1 s'il existe une relation linéaire entre X et Y. Il mesure spécifiquement une relation linéaire entre les variables. D'un point de vue apprentissage, r mesure la qualité de l'approximation des valeurs de Y par une fonction linéaire sur X. Sur la figure 6.1, les valeurs obtenues par une régression linéaire aux moindres carrés sont indiquées en rouge. Le coefficient r mesure comment ces valeurs approximent le couple (X,Y). Le ratio de corrélation $\eta(Y;X)$ est une autre mesure statistique de relation entre X et Y. Ce coefficient n'est pas symétrique. Il mesure à quel point les valeurs de Y sont bien approximées par une fonction de X et permet donc de mesurer une dépendance fonctionnelle non linéaire entre deux échantillons. Il est défini par :

$$\eta(Y;X) = 1 - \frac{\mathbb{E}(Var(Y|X)))}{Var(Y)} \tag{6.2}$$

Nous détaillerons le calcul de ce coefficient dans la suite de ce chapitre. Notons seulement que la fonction $\varphi(x) = \mathbb{E}(Y|X=x)$, utilisée pour le calcul de $Var(Y|X)$, est la fonction approximant

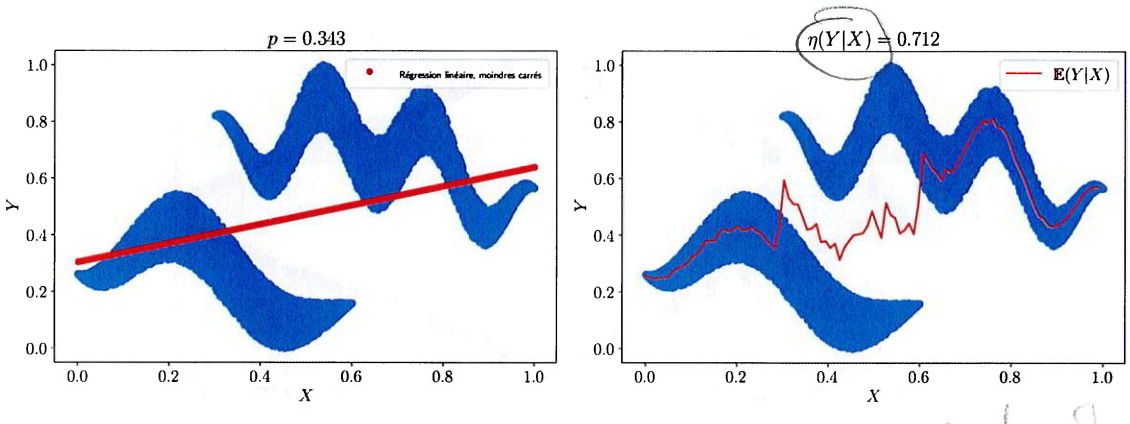

FIGURE 6.1 – Le coefficient de corrélation de Pearsons r, en figure de gauche, mesure une relation linéaire entre les variables X et Y. Le ratio de corrélation, au centre, cherche à déterminer l'existence d'une fonction entre Y et X.

le mieux les valeurs des paires (X, Y) au sens des moindres carrés. Cette fonction est tracée en rouge sur la figure 6.1 Le ratio de corrélation mesure la qualité de cette approximation. Il vaut 1 si Y est une fonction de X et 0 si les variables sont complètement indépendantes, car dans ce cas $Var(Y|X) = Var(Y)$.

Enfin, l'information mutuelle est une grandeur probabiliste. Elle évalue une relation entre les distributions des variables X et Y. Elle vaut 0 si et seulement si les variables sont indépendantes et est maximale lorsque qu'il existe une bijection entre les deux variables aléatoires. Son application à des échantillons statistiques passe par l'estimation des distributions des variables ou de leur entropie.

Nous avons remarqué dans les cas d'exemples présentés au chapitre précédent que l'apprentissage du modèle se traduit par une relation fonctionnelle entre les valeurs de U et la position du BMU Π dans chaque carte, comme rappelé en figure 6.2. Nous nous intéressons à deux méthodes mesurant la qualité de la relation fonctionnelle entre U et Π comme indicateur de l'apprentissage multimodal. L'une s'appuie sur le ratio de corrélation et la seconde sera définie à partir de l'information mutuelle. Ces indicateurs seront adaptés à des architectures de deux et trois cartes, dans lesquelles nous avons constaté cette relation fonctionnelle. Cependant, nous avons noté qu'il n'est pas souhaitable que U soit une fonction de la position du BMU dans toutes les cartes d'une architecture, mais plutôt que la représentation de U soit distribuée entre les cartes, tout en présentant de la redondance en terme d'information. Nous discuterons donc en dernière partie de chapitre des possibilités d'utilisation de l'information mutuelle comme indicateur de l'apprentissage du modèle dans une structure de cartes.

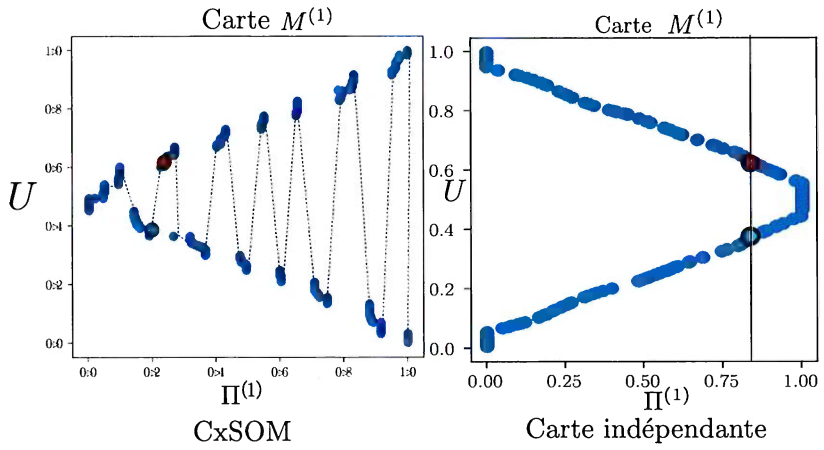

FIGURE 6.2 – Rappel : comparaison de U en fonction de $\Pi^{(1)}$ dans l'expérience exemple à deux cartes, sur des entrées sous forme de cercle. Ce nuage de point fait apparaître une relation semblant bijective entre U et $\Pi^{(1)}$. Nous définiront un indicateur permettant de représenter numériquement cette propriété.

6.2 Utilisation du ratio de corrélation comme mode d'évaluation

Le ratio de corrélation $\eta(\Pi^{(i)}, U)$ permet de mesurer un coefficient d'une relation fonctionnelle non-linéaire entre deux variables. Il atteint la valeur de 1 lorsque U est fonction de la première variable $\Pi^{(i)}$ et est nul lorsque les deux variables sont statistiquement indépendantes.

6.2.1 Définition

La mesure de la dépendance fonctionnelle entre deux variables X et Y peut se décomposer en deux étapes :

1. Trouver une fonction $\varphi(X)$ qui approxime les valeurs de Y
2. Mesurer la qualité de l'approximation.

En considérant deux variables $x \in \Omega_X$, $y \in \Omega_Y$, la fonction approchant le mieux l'ensemble de variables (x, y) au sens des moindres carrés est la fonction :

$$\varphi(x) = \mathbb{E}(Y|X = x), x \in \Omega_X \tag{6.3}$$

Le ratio de corrélation η se définit à partir de φ et mesure la qualité de l'approximation des valeurs de Y par la fonction φ au sens des moindres carrés en calculant l'espérance des variances de la variable $Y|X$ pour chaque valeur possible de X.

$$\eta(Y; X) = 1 - \frac{\mathbb{E}(Var(Y|X))}{Var(Y)} \tag{6.4}$$

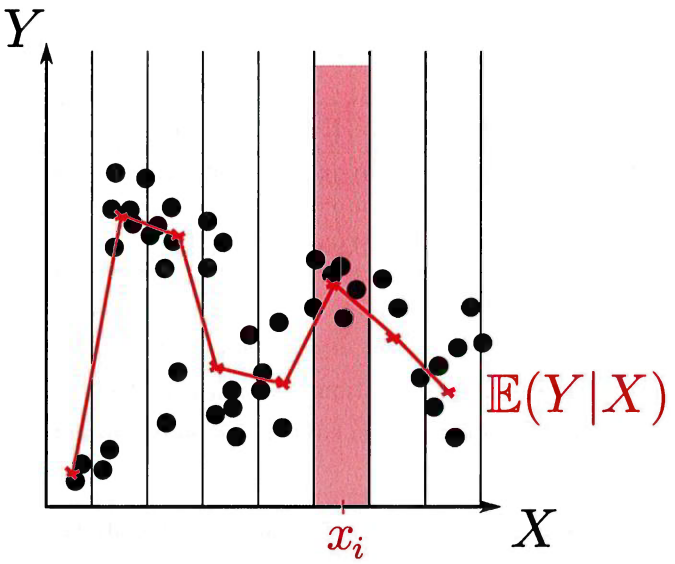

FIGURE 6.3 – Exemple d'approximation non-linéaire de la relation entre X et Y par $\mathbb{E}(Y|X)$. Cette approximation est ici réalisée en discrétisant les valeurs de X. La valeur de la fonction pour chaque x_i est alors la moyenne des valeurs de Y sur l'intervalle considéré.

Une possibilité d'estimation de $\varphi(x)$ est illustré en figure 6.3 : nous discrétiserons les valeurs de X en n valeurs $(x_1, \cdots, x_n)$ et prendrons $\varphi(x_i)$ comme la moyenne des valeurs de Y dans l'intervalle $[x_{i-1}, x_i]$. Le ratio de corrélation n'est pas symétrique. Par le fait qu'il s'appuie sur un rapport, il n'est pas sensible à une transformation linéaire de Y et est sans unité.

6.2.2 Application du ratio de corrélation aux cartes 1D

Nous calculons le ratio de corrélation dans deux cas d'exemple tirés du chapitre précédent sur une architecture de deux cartes, afin de vérifier comment il traduit la relation fonctionnelle entre U et Π que nous avions observé par les tracés :

- Lorsque les entrées sont tirées sur le cercle de centre 0.5 et de rayon 0.5. U correspond à l'angle du point sur le cercle.
- Lorsque les entrées sont tirées sur un anneau, construit en ajoutant un bruit aux points de centre 0.5 et de rayon 0.5. U correspond également à l'angle du point sur le cercle et l'indicateur doit pouvoir refléter l'apprentissage de U malgré le bruit sur les entrées.

Nous comparons également les valeurs du ratio de corrélation à celui obtenu dans le cas de deux cartes indépendantes prenant en entrée l'une $X^{(1)}$ et l'autre $X^{(2)}$ afin de comparer les valeurs du ratio de corrélation.

Les tracés de U en fonction de Π, φ et $\eta(U;\Pi)$ sont représentés en figure 6.4 pour le cercle et figure 6.5 pour l'anneau. Dans les deux cas, la relation entre U et Π est fonctionnelle dans

TABLE 6.1 – Comparaison des valeurs du ratio de corrélation sur plusieurs expériences.

	Entrées		CxSOM		Cartes Simples	
	$\eta(U;X^{(1)})$	$\eta(U;X^{(2)})$	$\eta(U;\Pi^{(1)})$	$\eta(U;\Pi^{(2)})$	$\eta(U;\Pi^{(1)})$	$\eta(U;\Pi^{(2)})$
Cercle	0.45	0.84	0.98	0.99	0.49	0.84
Anneau	0.43	0.83	0.97	0.93	0.44	0.82
Lissajous	0.81	0.80	0.96	0.94		

CxSOM et le ratio de corrélation est proche de 1. Lorsque les entrées sont bruitées, le ratio de corrélation reste élevé, traduisant une relation fonctionnelle. Cet indicateur différencie bien l'organisation des cartes CxSOM des cartes non connectées pour lesquelles le ratio de corrélation est plus faible.

Le tableau 6.1 présente les valeurs de $\eta(U;\Pi)$ sur trois distributions d'entrées : le cercle, l'anneau et la courbe de Lissajous présentée au chapitre 5, que nous comparons aux valeurs obtenues pour les cartes indépendantes. Nous calculons également $\eta(U;X)$, qui est un indicateur sur les entrées, comme point de comparaison. Dans les trois cas, $\eta(U;\Pi)$ est proche de 1 dans CxSOM. Le ratio de corrélation dans les cartes non connectées est similaire à $\eta(U;X)$.

Nous notons que $\eta(U;X^{(2)}) = 0.8$ dans chacune des expériences ; cette valeur est proche de 1 alors que la relation n'est pas « plus fonctionnelle » que pour $\Pi^{(1)}$. Intuitivement, on aurait voulu une valeur similaire dans les deux cas. La valeur seule du ratio de corrélation nous permet donc mal de qualifier la qualité de l'apprentissage de CxSOM ; il faudra la comparer au ratio de corrélation d'entrée $\eta(U;X)$.

Enfin, nous traçons en figure 6.6 l'évolution du ratio de corrélation sur les 200 premiers pas d'apprentissage des cartes. Les mesures sont réalisées sur 10 expériences réalisées sur des distributions d'entrées identiques, puis moyennées sur ces expériences. Nous observons que $\eta(U;\Pi)$ garde une valeur élevée tout au long de l'apprentissage pour CxSOM. Le ratio de corrélation traduit en effet une relation fonctionnelle, mais ne prend pas en compte la proximité des positions. Or, par construction de l'algorithme, une carte, par exemple $M^{(1)}$ définit son BMU en fonction de $X^{(1)}$ et de son entrée contextuelle $\Pi^{(2)}$, représentant directement $X^{(2)}$. U est donc une fonction du BMU dans chaque carte dès le début de l'apprentissage. Le ratio de corrélation ne traduit donc pas l'organisation continue des poids.

6.2.3 Discussion

Le ratio de corrélation $\eta(U;\Pi)$ est une mesure statistique qui exprime par définition la relation fonctionnelle existant entre U et Π, ce que nous cherchions à mesurer. Cette mesure nécessite de discrétiser les valeurs de Π mais pas de U, ce qui est adapté aux cartes auto-organisatrices dans lesquelles Π est en une ou deux dimensions. Il s'agit donc d'une bonne mesure de l'apprentissage

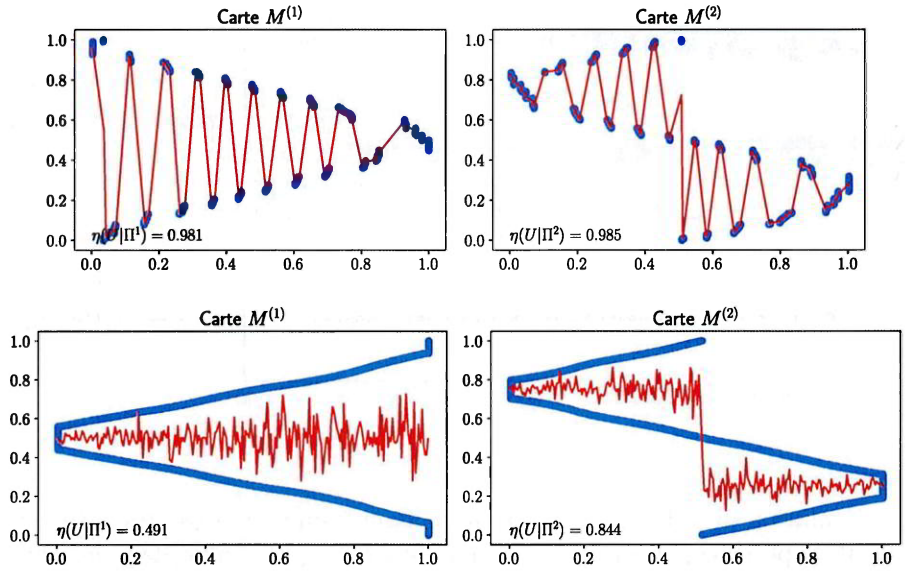

FIGURE 6.4 – Tracé du ratio de corrélation et de φ sur des entrées placées sur un cercle, dans le cas de CxSOM et d'une carte simple. La fonction tracée en rouge est $\mathbb{E}(Y|X = x)$ approximant le nuage de points. Le ratio de corrélation mesure ensuite la variance de Y pour chaque valeur de X.

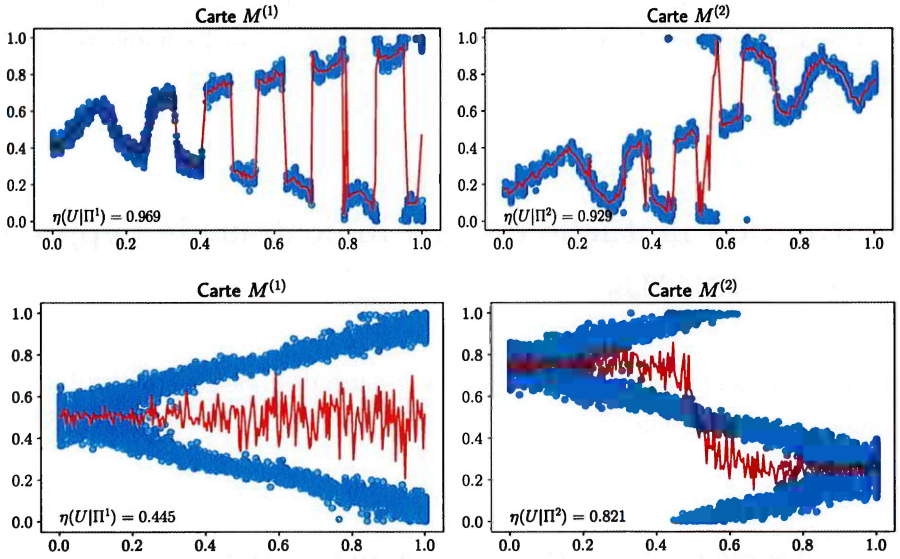

FIGURE 6.5 – Tracé du ratio de corrélation sur cartes CxSOM et cartes simples pour des entrées placées sur un anneau. Les données d'entrées sont bruitées, ce qui conduit à une dispersion plus élevée des réponses des cartes autour de la valeur de U. Le ratio de corrélation traduit toujours bien que la relation entre U et Π s'approche d'une fonction.

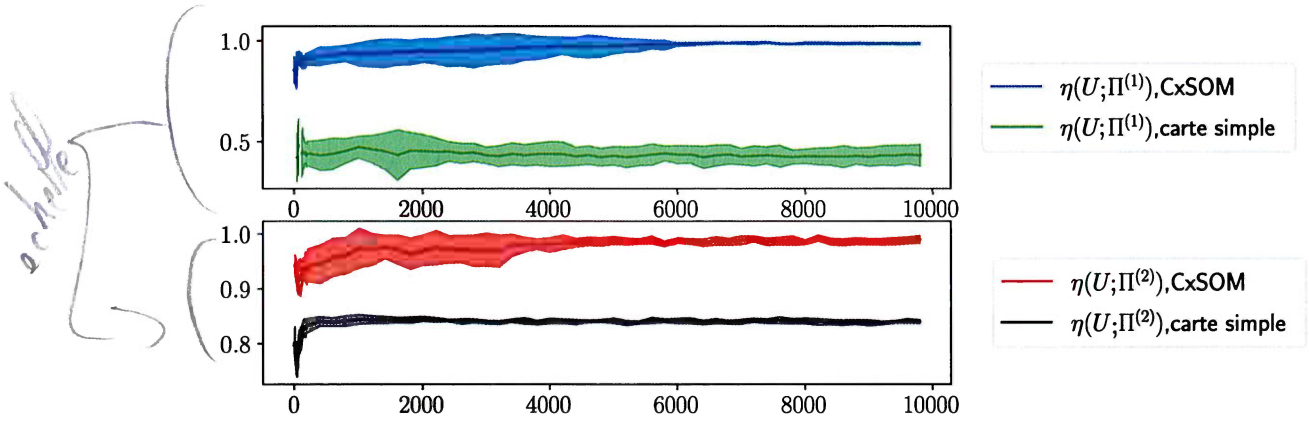

FIGURE 6.6 – Évolution du ratio de corrélation pendant l'apprentissage des cartes, moyenne et écart type sur 10 expériences. Le ratio de corrélation reste plus faible dans les cartes simples que dans les cartes CxSOM. Il garde une valeur élevée tout au long de l'apprentissage : le ratio de corrélation ne traduit pas l'organisation des poids, mais simplement si la carte a un BMU différent pour chaque valeur de U.

de U par une carte et est adaptable pour des cartes en 2D, ainsi que des U en grande dimension. L'utilisation du ratio de corrélation comme indicateur d'un bon apprentissage de U par les BMUs n'est pertinent qu'en le comparant à sa valeur théorique $\eta(U;\Pi)$ ou à la valeur qu'il prend dans une carte auto-organisatrice indépendante. Enfin, il ne traduit pas l'organisation des poids au cours de l'apprentissage.

6.3 L'information mutuelle comme indicateur de l'apprentissage de U par les BMUs

Nous étudions à présent une autre méthode de mesure des relations entre données en s'appuyant sur l'information mutuelle.

6.3.1 Rappel des éléments de théorie de l'information

Les notions d'*entropie* et les valeurs associées, telle que l'*information mutuelle* entre des variables aléatoires, sont des notions fondamentales de la théorie de l'information de Shannon. Ces quantités sont calculées à partir de la distribution des variables aléatoires. L'entropie de Shannon d'une variable aléatoire X à valeurs discrètes dans un ensemble Ω_X, de distribution

P_X, est notée $H(X)$ et définie par la formule :

$$H(X) = -\sum_{x \in \Omega_X} P_X(x)\log(P_X(x)) \tag{6.5}$$

L'entropie de Shannon concerne uniquement des variables discrètes. Une autre version de l'entropie est définie pour des variables continues, l'entropie différentielle :

$$H(X) = -\int_{x \in \Omega_X} p_X(x)\log(p_X(x))dx \tag{6.6}$$

Avec p_X la densité de probabilité de X.

Cette valeur n'est cependant pas la limite de l'entropie de Shannon calculée par discrétisation de X en N intervalles, $N \to \infty$. L'entropie différentielle et l'entropie de Shannon sont donc deux quantités bien différentes.

L'entropie de Shannon se mesure en *bit/symbole*. Si la distribution de X est concentrée autour d'un point, l'entropie est faible : lors d'une réalisation de X, l'observateur est *plutôt certain* du résultat. En revanche, l'entropie est maximale lorsque X suit une distribution de probabilité uniforme. L'entropie s'interprète également comme la quantité moyenne d'information à fournir, en bits, pour coder une valeur de X. De la même manière, on peut définir l'entropie conjointe de deux variables, qui est l'entropie de leur distribution jointe, et l'entropie conditionnelle, qui est l'entropie de leurs distributions conditionnelles.

Outre les entropies jointes et conditionnelles, l'existence d'une relation statistique entre deux variables aléatoires X, Y à valeurs dans Ω_X, Ω_Y se mesure par *l'information mutuelle*. Elle est définie par :

$$I(X,Y) = \sum_{(x,y) \in \Omega_X, \Omega_Y} P_{XY}(x,y)\log(\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}) \tag{6.7}$$

Avec P_{XY} la distribution de la variable aléatoire jointe (X,Y) Cette valeur mesure la quantité d'information moyenne partagée entre les distributions X et Y : en moyenne, quelle information sur la valeur de Y donne une valeur de X et inversement, quelle information sur la valeur de X donne une valeur de Y.

L'information mutuelle possède les propriétés suivantes :

1. $I(X,Y) = 0 \Leftrightarrow$ X et Y sont indépendantes. L'information mutuelle peut être vue une mesure de la distance entre la distribution jointe de (X,Y), $P(X,Y)$ et la distribution correspondant à l'indépendance des variables, $P(X)P(Y)$.
2. Elle s'exprime à partir de l'entropie de Shannon : $I(X,Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$
3. Elle est symétrique : $I(X,Y) = I(Y,X)$

4. Pour toute fonction f, $I(X,Y) \geq I(X, f(Y))$. L'égalité est atteinte si et seulement si f est *bijective*.

L'information mutuelle se calcule également à partir des densités de probabilité pour des variables à valeur continues de densités de probabilité p_X et p_Y :

$$I(X,Y) = \int_{x \in \Omega_X} \in_{y \in \Omega_Y} p_{XY}(x,y) \log\left(\frac{p_{XY}(x,y)}{p_X(x) p_Y(y)}\right) dx\, dy \tag{6.8}$$

Contrairement à l'entropie, la valeur de l'information mutuelle pour des variables continues correspond bien à une limite des valeurs de l'information mutuelle discrète lorsque le nombre de catégories tend vers l'infini (Cover et Thomas 2005). Cependant, dans le cas continu, les propriétés 2 et 4 ne sont pas vérifiées. L'information mutuelle n'est en effet pas comparable à l'entropie différentielle.

Lors de l'analyse de CxSOM, nous nous intéressons à l'information que portent les positions des BMUs Π d'une carte sur le modèle d'entrée, donc les variables d'entrées $X^{(i)}$ et U.

6.3.2 Méthodes d'estimation de l'information mutuelle

L'information mutuelle et l'entropie sont des grandeurs définies à partir de la distribution des variables aléatoires. Ces distributions, dans notre cas, ne sont pas connues, nous devons donc estimer ces quantités à partir des échantillons de données. Nous considérons ici que les variables que nous étudions sont des variables continues.

Une méthode classique d'estimation de l'information mutuelle et de l'entropie est la méthode dite des *histogrammes*. Cette méthode s'appuie sur une estimation de la distribution des variables U,Π et la distribution de la variable jointe (U, Π) en discrétisant chacune des variables. Cette méthode est représentée en figure 6.7. Les variables U et Π sont discrétisées en *boîtes* de centres x_k et y_k choisis. Une distribution est alors estimée par :

$$P(U = x_i) = \frac{n_{xi}}{N}$$

où n_{xi} est le nombre d'échantillons de U tombant dans la boîte de centre x_i et N le nombre de points. Le même procédé est réalisé pour Π et (U, Π). La précision de l'estimation peut être améliorée en choisissant des tailles de boîtes variables ; nous utilisons ici la méthode simple avec des boîtes de taille fixe. Pour des variables à valeur dans $[0, 1]$, les centres sont définis par $x_k = \frac{k}{M} + \frac{1}{2M}$, avec M le nombre de boîtes. Cette discrétisation permet d'estimer les trois termes d'entropie $\hat{H}(\Pi, U)$, $\hat{H}(U)$ et $\hat{H}(\Pi)$ et d'en tirer l'information mutuelle :

$$\hat{I}(U, \Pi) = \hat{H}(U) + \hat{H}(\Pi) - \hat{H}(U, \Pi) \tag{6.9}$$

La valeur de cet indicateur est très sensible à la résolution choisie pour le calcul des histogrammes. Par ailleurs, plus la taille des boîtes est petite, plus le nombre de points disponibles pour l'estimation doit augmenter. La méthode par histogrammes est difficilement exploitable lorsque la dimension des entrées augmente : le nombre d'échantillons disponibles pour l'estimation doit augmenter exponentiellement avec la dimension des variables pour éviter le phénomène de "boîtes vides". À cause de la dispersion des données, de nombreux intervalles de discrétisation ne contiendront pas de points pour l'estimation alors qu'ils auraient dû en contenir d'après leur distribution théorique, ce qui fausse l'estimation.

Une deuxième méthode régulièrement utilisée pour l'estimation de l'information mutuelle est l'estimateur par KNN (*K-Nearest Neighbors*) de Kraskov (Kraskov et al. 2004). Cet estimateur ne passe pas par l'estimation de la densité de probabilité, contrairement aux histogrammes, mais estime directement l'information mutuelle. Le découpage de l'espace se fait en recherchant, pour N valeurs d'échantillons d'un couple (X, Y), les k plus proches voisins. Une information mutuelle locale est calculée dans cette zone de l'espace, suivant une formule permettant d'approximer les différences de logarithme par la fonction digamma ψ :

$$i_j(X,Y) = \psi(k) - \psi(n_{x_j} + 1) - \psi(n_{y_j} + 1) + \psi(N)$$

Cette information mutuelle locale est ensuite moyennée sur l'ensemble des points :

$$\hat{I}(X,Y) = \psi(k) - \langle \psi(n_{x_j} + 1) + \psi(n_{y_j} + 1) \rangle + \psi(N)$$

L'estimateur de Kraskov est moins sensible aux paramètres choisis pour son estimation qui sont le nombre de voisins considérés (Ross 2014).

L'information mutuelle étant une grandeur largement utilisée en théorie de l'information, il existe de nombreuses autres méthodes d'estimation possibles (Doquire et Verleysen 2012).

6.4 Définition d'un indicateur quantifiant la relation fonctionnelle entre U et Π

Nous nous sommes d'abord intéressés à l'utilisation d'une version normalisée de l'information mutuelle entre U et Π comme indicateur de la relation fonctionnelle entre U et Π dans chaque carte. $I(U, \Pi)$ est en effet maximale lorsqu'il existe une bijection entre U et Π. Une version normalisée nous permettrait d'obtenir un indicateur à valeur dans $[0, 1]$ permettant une quantification absolue de l'apprentissage d'une carte.

Nous voulons normaliser l'information mutuelle $I(\Pi, U)$ par la valeur maximale qu'elle pourra prendre dans une carte. Si on considère des variables discrètes, cette valeur maximale est $H(U)$,

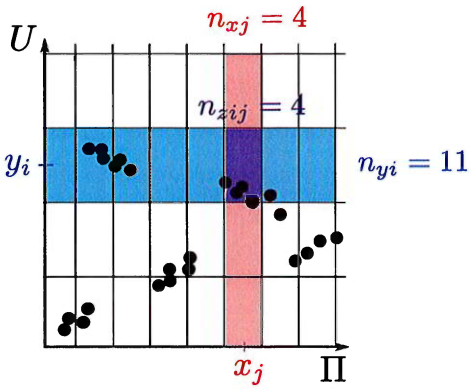

FIGURE 6.7 – Méthode par histogrammes pour estimer les distributions des variables U et Π. Les distributions sont estimées à partir de n_{xj}, n_{yi} et n_{zij}, puis les valeurs de l'entropie H et l'information mutuelle I calculées.

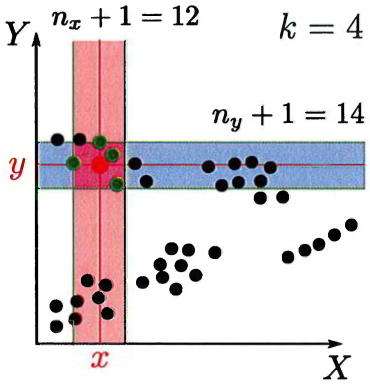

FIGURE 6.8 – Découpage en KNN de Kraskov pour estimer l'entropie et l'information mutuelle des variables X et Y. Les plus proches voisins du point rouge sont trouvés, en vert, et le processus est répété sur tous les points. Les valeurs de n_x et n_y permettent d'estimer directement l'entropie.

atteinte lorsque U est une fonction de Π. En effet, par construction, Π est une fonction de U dans une carte de Kohonen : l'algorithme est déterministe et une sortie est définie pour toute valeur de U. C'est-à-dire, $I(U,\Pi) = I(U, f(U))$. Par propriété de l'information mutuelle, pour toute fonction f et variables X, Y, $I(X, f(Y)) \leq I(X,Y)$. Donc, $I(U,\Pi) \leq I(U,U) = H(U)$. Cette valeur est atteinte si et seulement si U et Π sont en bijection, autrement dit, si et seulement si U est aussi une fonction de Π.

Nous définissons donc un indicateur possible U_c d'une relation fonctionnelle entre U et Π comme :

$$U_c(U|\Pi) = \frac{I(\Pi, U)}{H(U)} \tag{6.10}$$

Ce coefficient n'est pas symétrique et mesure l'information portée par le second terme sur le premier, relativement à la valeur maximale qu'il peut prendre ($H(U)$). On a $U_c(U|\Pi) \in [0,1]$. Cette variante normalisée de l'information mutuelle est s'apparente au *coefficient d'incertitude* entre U et Π et introduit en (Theil et al. 1961). U_c vaut 1 lorsque U est une fonction de Π, et 0 lorsque les deux distributions sont indépendantes.

La normalisation de l'information mutuelle par l'entropie est uniquement valable dans le cas de variables aléatoires discrètes. Pour son utilisation, il est donc nécessaire de considérer Π et U comme des variables discrètes et d'estimer l'information mutuelle et l'entropie par la méthode des histogrammes.

Voyons maintenant ce que cette quantité mesure dans une carte CxSOM. La méthode des

histogrammes est très sensible aux paramètres d'estimation. Pour mieux comprendre ce que représente l'information mutuelle, comparons en figure 6.9 deux exemples de relations entre des variables aléatoires X et Y. Dans le cas de gauche, la relation se rapproche d'une relation fonctionnelle, mais cette relation est bruitée et une même valeur de X correspond à un intervalle de valeurs de Y. Dans le cas de droite, la relation n'est pas une fonction, mais une valeur de X correspond à exactement deux valeurs de U.

L'information mutuelle continue obtenue dans le cas de gauche est faible, de 2.3 bits. En effet, une valeur de Π correspond à tout un intervalle de valeurs pour U. Sur le cas de droite, sa valeur est plus élevée : 4.5 bits. En effet, une valeur de X correspond à deux valeurs de Y. X et Y partagent donc plus d'information que le premier cas de figure. Ce n'est pas ce qu'on veut mesurer dans CxSOM : une fonction bruitée doit être privilégiée par rapport à une relation qui n'est pas fonctionnelle. On cherche en effet à mesurer si une valeur de Π correspond à *un unique* intervalle de U, et non plusieurs intervalles comme dans le cas d'une carte simple, dans laquelle deux valeurs de U éloignées sont codées par une même position de BMU, voir figure 6.2. Pour que U_c traduise cette propriété sur des cartes CxSOM, nous utiliserons un découpage large pour la discrétisation de U. Pour une carte de taille 500, nous avons découpé Π en 500 intervalles et U en 50 intervalles. Ces paramètres d'estimation permettent d'ignorer la dispersion locale sur la valeur de U pour une même position Π.

L'indicateur U_c défini ici doit ainsi être considéré comme un indicateur s'inspirant du coefficient d'incertitude que comme une estimation de sa valeur théorique. C'est cette estimation large qui nous permettra d'évaluer qu'une carte a dissocié les positions de ses BMUs en fonction de U et non seulement de son entrée externe. La valeur de U_c est alors très sensible aux paramètres d'estimation. La taille d'intervalle de discrétisation de U devra par ailleurs être choisie en fonction des données d'entrées.

6.4.1 Application de U_c au cas d'exemple du cercle

Nous traçons l'évolution de $U_c(U|\Pi)$ au cours de l'apprentissage dans un système de deux cartes apprenant sur le cercle en deux dimensions, afin de vérifier que U_c reflète bien la qualité de l'apprentissage du modèle dans une carte. L'organisation finale de U selon Π correspond à celle représentée en figure 6.4.

Pour cette expérience, une phase de test sur 5000 entrées est réalisée à intervalles réguliers lors de l'apprentissage, en utilisant le même jeu d'entrées pour chaque test. Chaque phase de test donne un ensemble d'entrées $(X^{(1)}, X^{(2)}), U$ et un ensemble de réponses des cartes $(\Pi^{(1)}, \Pi^{(2)})$. Nous estimons $U_c(U|\Pi^{(1)})$ et $U_c(U|\Pi^{(2)})$ sur chaque itération considérée et tracer la courbe de l'évolution de l'indicateur au long de l'apprentissage. Ces calculs sont réalisés sur 10 apprentissages séparés, prenant des entrées d'apprentissage aléatoires sur le même cercle. Les cartes sont

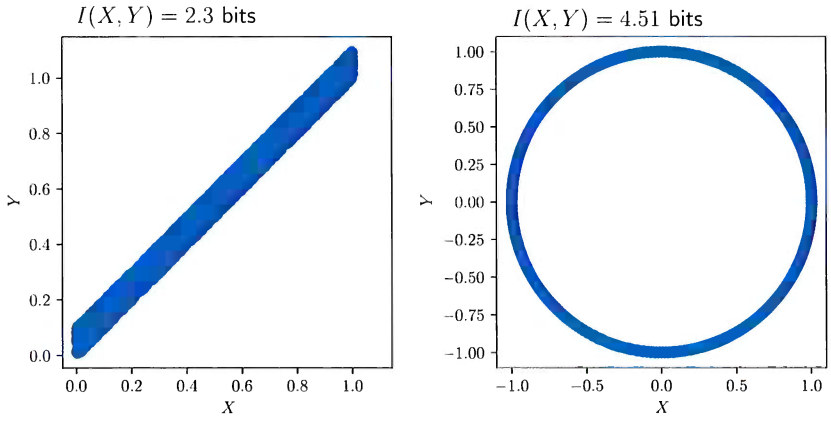

FIGURE 6.9 – Comparaison du calcul de $I(X,Y)$ sur deux distributions. À gauche, la relation entre Y et X se rapproche d'une fonction, mais bruitée. A droite, la relation n'est pas fonctionnelle, mais de telle sorte qu'une valeur de X correspond au maximum à deux valeurs de Y.

initialisées à des poids aléatoires différents au début de chaque apprentissage. Nous comparons les valeurs obtenues pour une carte CxSOM à celles de deux cartes simples apprenant sur les mêmes entrées $X^{(1)}$ et $X^{(2)}$. Les tracés représentent la moyenne, à chaque pas de temps, des indicateurs considérés au pas de temps t.

Pour l'estimation, nous avons discrétisé U en 50 boîtes, et en 500 pour $\Pi^{(i)}$: comme soulevé au paragraphe précédent, il est nécessaire d'utiliser un intervalle plus large pour les valeurs de U, afin de ne pas prendre en compte la dispersion des points au niveau local. L'évolution de U_c est tracée en figure 6.10.

On s'attend à ce que le coefficient d'incertitude soit plus élevée pour la carte au sein de CxSOM que la carte seule. Cela montrera qu'une carte porte de l'information sur son entrée externe mais également sur le modèle global U, donc sur l'autre entrée. On s'attend également à ce que cette valeur atteigne 1, ce qui montrerait qu'une seule carte porte de l'information sur tout le modèle : U est une fonction de Π dans chaque carte.

L'observation du tracé montre que les quantités $U_c(U|\Pi^{(1)})$ et $U_c(U|\Pi^{(2)})$ sont bien toutes deux plus élevées à chaque moment de l'apprentissage que dans le cas ou les cartes sont séparées et leurs valeurs s'approche de 1 à la fin de l'apprentissage. Ces quantités augmentent au cours de l'apprentissage, traduisant bien un gain d'information des cartes sur le modèle au cours de l'apprentissage.

Nous pouvons donc utiliser U_c comme indicateur d'une relation fonctionnelle entre U et Π dans chaque carte, en choisissant bien la taille de discrétisation pour U lors de l'estimation. La taille de l'intervalle doit être assez élevée pour englober le bruit local des données, mais suffisamment faible pour détecter une séparation entre deux intervalles de U codés par une même position de BMU. Cependant, le choix du pas discrétisation de U nécessite une visualisation

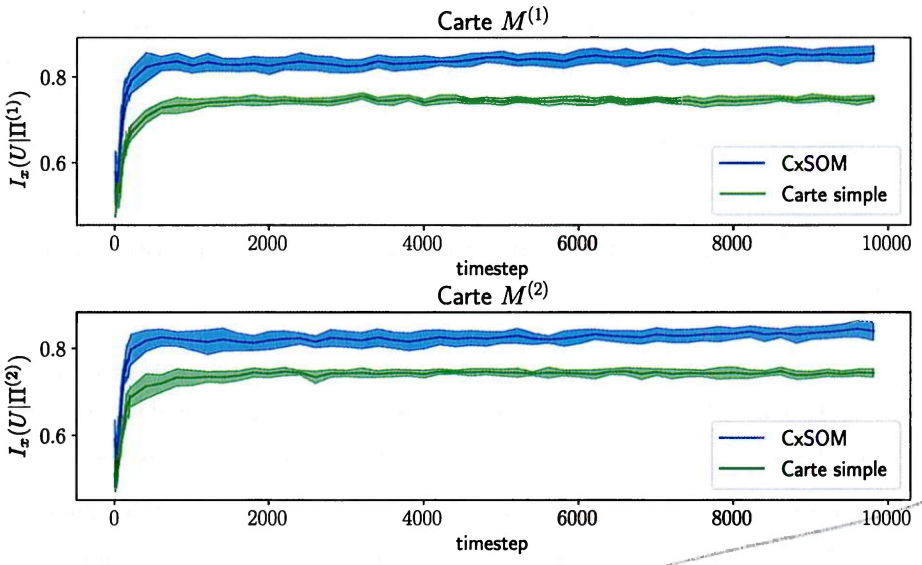

FIGURE 6.10 – Évolution de l'information mutuelle normalisée $I_x(U|\Pi)$ dans chaque carte au long de l'apprentissage. L'intervalle de discrétisation choisi pour U est de 0.02 (50 bins). La courbe bleue correspond à $I_x(U|\Pi)$ dans l'architecture de cartes $M^{(1)}$ et $M^{(2)}$. On compare cette évolution à l'évolution de l'information d'une seule carte apprenant sur les mêmes entrées $X^{(1)}$ ou $X^{(2)}$, sans être connectée.

des valeurs U et Π préalable ; le coefficient d'incertitude sera difficilement exploitable en grande dimension, lorsque les données ne peuvent être visualisées en 2D. Par ailleurs, la discrétisation de U nécessite un très grand échantillon en grande dimension ce qui limite également son utilisation. En conclusion, pour la mesure d'une relation fonctionnelle entre U et Π, il est préférable d'utiliser le ratio de corrélation. Il ne nécessite pas de discrétisation de U et est donc plus exploitable en grande dimension. Cependant, sa valeur reste mal interprétable de manière absolue et il doit être utilisé en comparaison avec les valeurs calculées sur les entrées.

6.5 Comment utiliser l'information mutuelle continue comme indicateur d'un apprentissage ?

Le ratio de corrélation et le coefficient d'incertitude, présentés ci-dessus, mesurent de deux manières différentes le fait que U est une fonction du BMU dans chaque carte. Bien que le coefficient d'incertitude s'appuie sur l'information mutuelle, l'indicateur que nous avons présenté se détache de la valeur théorique de l'information mutuelle par la discrétisation à gros grains de U.

Sur des architectures à plus de trois cartes, il n'est pas certain ni même souhaitable que U soit une fonction de la position du BMU dans toutes les cartes d'une architecture, mais plutôt que la

représentation de U soit distribuée entre les cartes, tout en présentant de la redondance en terme d'information. Nous envisageons donc dans cette partie des perspectives d'utilisation de l'information mutuelle entre U et les positions des BMUs $(\Pi^{(1)}, \Pi^{(2)})$ pour analyser l'apprentissage dans une architecture de cartes.

6.5.1 Évolution de l'information mutuelle entre U et Π au cours d'un apprentissage

En figure 6.11, nous traçons l'évolution de l'information mutuelle dans les deux cartes, estimée par la méthode de Kraskov. Comme dans les paragraphes précédents, sa valeur est moyennée sur 10 apprentissages. Le jeu de données d'entrée utilisé pour ce calcul est toujours le cercle en 2D.

Nous observons que l'information mutuelle entre U et Π converge vers une valeur plus élevée dans une carte isolée que dans une architecture CxSOM. Ce résultat est étonnant : cela signifie donc que chaque carte au sein de CxSOM n'a pas plus d'information sur le modèle d'entrées qu'une carte isolée, qui ne reçoit pourtant qu'une partie des entrées. Ce résultat s'interprète par le fait que l'information apprise sur le modèle par une carte n'est pas répartie de la même façon dans les deux expériences. Dans une carte indépendante, le niveau de quantification vectorielle sur X est très précis : lorsqu'on présente une entrée X à la carte, le poids du BMU est très proche de cette valeur X. Or, la connaissance de la valeur X donne beaucoup d'information sur le modèle U. Dans CxSOM, on perd ce niveau de quantification sur X, ce qu'on a observé en figure 4.6. On perd donc de l'information sur X.

Le fait que l'information mutuelle soit plus élevée dans une carte indépendante dans les deux expériences traduit ainsi une perte d'information sur l'entrée X dans CxSOM par rapport à une carte indépendante, avec la perte de précision. Cette valeur comprend à la fois un gain d'information qui existe sur $X^{(2)}$ et donc U et une perte d'information sur X ; cette perte domine, d'où la perte globale d'information. Les cartes effectuent donc un compromis : chacune gagne de l'information sur le modèle U, au détriment de l'information apprise sur l'entrée externe. Le seul calcul de l'information mutuelle ne suffit donc pas à analyser l'apprentissage du modèle par les cartes. Des méthodes permettant de séparer l'information entre variables existent dans la littérature. Elles nous permettraient de mesurer le gain d'information sur U dans une ou plusieurs cartes sans s'intéresser à l'information apprise sur l'entrée externe X. Nous avons déjà utilisé une méthode de séparation lors de l'estimation du coefficient d'incertitude U_c : le fait de discrétiser grossièrement la distribution de U a permis de mesurer le gain d'information sur U, sans prendre en compte l'affaiblissement de la précision de la quantification de l'entrée externe. Cette perte d'information pose néanmoins une question concernant la création d'architectures contenant de nombreuses cartes et U de grande dimension : jusqu'à quel point une carte peut-elle se permettre de perdre de l'information sur l'entrée externe pour gagner de l'information sur le modèle ? Cela motive l'idée qu'un apprentissage de U dans une grande architecture devra être

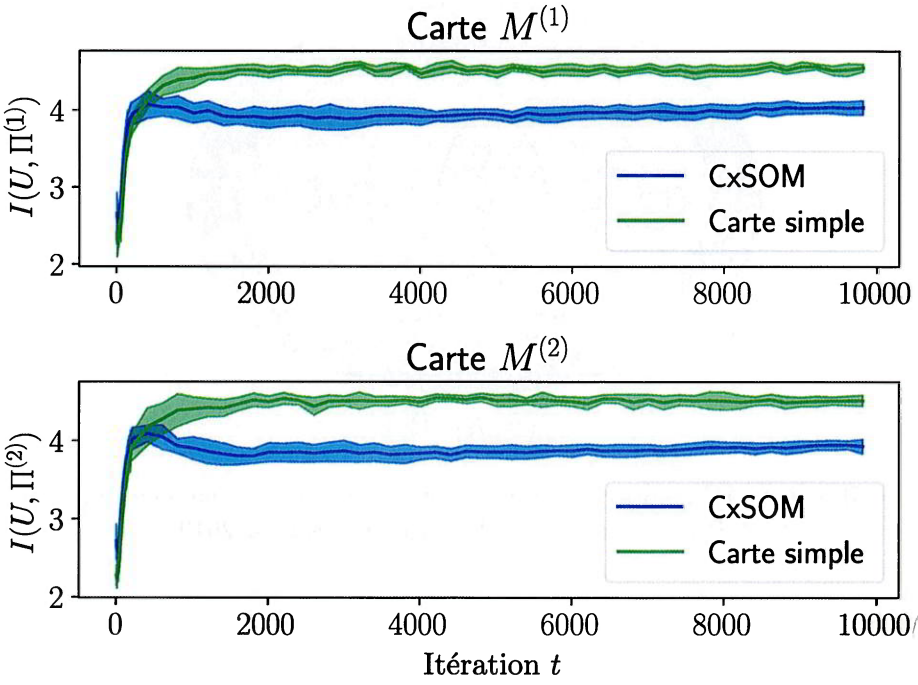

FIGURE 6.11 – Évolution de $I(U, \Pi)$ dans chaque carte au long de l'apprentissage, estimé par la méthode de Kraskov. Cette valeur est moyennée sur 10 expériences. Nous comparons les valeurs obtenue dans une architecture CxSOM, en bleu, au cas d'une carte apprenant indépendamment sur les mêmes entrées $X^{(1)}$ et $X^{(2)}$. Le même échantillon U est utilisé pour chaque phase de test. Nous pouvons voir que dans ces expériences, les positions des BMUs d'une carte indépendante partagent plus d'information avec U que dans le cas de CxSOM.

distribué entre les cartes et ne peut être appris indépendamment dans chaque carte. D'après les seules observations réalisées sur deux et trois cartes dans cette thèse, nous ne pouvons pas affirmer ou non si cette propriété sera vérifiée. Cela constitue une perspective de travaux futurs pour le développement du modèle.

6.5.2 Ouvertures possibles

Les mesures proposées dans ce chapitre ont permis d'évaluer un apprentissage du modèle indépendamment dans chaque carte. La mesure de l'information mutuelle est cependant bien plus large que le seul calcul de $I(U, \Pi)$; de nombreux aspects nous semblent intéressants à explorer pour une compréhension de l'apprentissage du modèle dans des architectures comportant plus de cartes. Nous pouvons noter que la méthode d'estimation par KNN présentée dans ce chapitre est une méthode classique d'estimation de l'information, mais il existe de nombreuses autres méthodes d'estimation (Doquire et Verleysen 2012). Des méthodes ont également été développées pour la mesure de l'information mutuelle entre variables continues et discrètes (Ross 2014 ; Gao

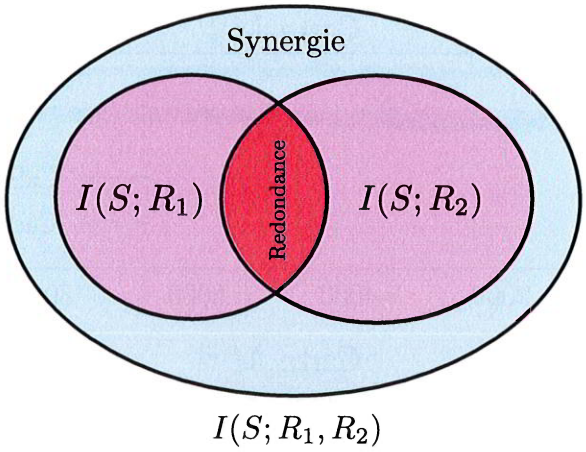

FIGURE 6.12 – Illustration des notions d'information *redondante* et *synergique* entre une variable S et deux variables R_1 et R_2, schéma adapté de (Williams et Beer 2010).

et al. 2017). Enfin, l'information mutuelle a été utilisée pour analyser l'apprentissage dans des réseaux de neurones profonds en (Shwartz-Ziv et Tishby 2017) ou encore directement comme métrique d'apprentissage en (Hjelm et al. 2018). Cette grandeur est ainsi bien documentée et donc pertinente à utiliser dans des travaux futurs. Il sera possible d'appliquer des méthodes d'évaluation utilisées dans d'autres architectures afin de caractériser l'apprentissage associatif des cartes.

Tout d'abord, il est possible de s'intéresser à la notion d'information mutuelle multivariée : étant donné une variable cible S et deux variables R_1 et R_2, $I(S;R_1,R_2)$ désigne l'information mutuelle entre S et la variable jointe (R_1,R_2). Nous pourrons ainsi mesurer, dans une architecture de cartes, $I(U;\Pi^{(1)},\Pi^{(2)})$ par exemple. Il est également possible de décomposer cette information multivariée : (Williams et Beer 2010) définit, en plus de l'information mutuelle, la notion de redondance et de synergie entre variables, illustrée en figure 6.12. La redondance est l'information sur S portée à la fois par R_1 et par R_2, et la synergie l'information portée seulement par la jointure des variables R_1 et R_2. Le calcul de telles grandeurs permettrait par exemple de séparer l'information gagnée sur U et X dans une carte. Le calcul de ces grandeurs entre les entrées, le modèle d'entrée et les BMUs des cartes CxSOM sont une piste d'étude pour une compréhension du stockage d'information dans une architecture de cartes et pour la définition d'un indicateur ciblant spécifiquement le gain d'information sur U lors de l'apprentissage.

Des travaux comme (Lizier et al. 2007 ; Ceguerra et al. 2011) s'intéressent quant à eux à la notion de transfert d'information au sein de systèmes dynamiques complexes. Le calcul d'information entre les éléments des cartes peut ainsi également s'appliquer à la quantification de la dynamique d'apprentissage d'une architecture de cartes.

6.6 Conclusion

Ce chapitre utilise la méthode de représentation des éléments des cartes comme des variables aléatoires proposée au chapitre 4 pour proposer des indicateurs de l'apprentissage multimodal au sein de l'architecture. Les représentations visuelles sont en effet limitées dans des architectures de plus de deux ou trois cartes, et pour des données en plus grande dimension. La définition d'un indicateur permettra également de comparer l'apprentissage d'architecture, autorisant par exemple l'optimisation automatique des paramètres de l'architecture de cartes.

Dans ce chapitre, nous avons introduit deux indicateurs permettant de mesurer que U est une fonction du Π dans chacune des cartes de l'architecture : le ratio de corrélation et le coefficient d'incertitude. Nous avons en effet observé dans les deux chapitres précédents que ce comportement marque l'apprentissage du modèle dans des architectures de deux ou trois cartes en une dimension.

D'une part, nous avons présenté le ratio de corrélation $\eta(U;\Pi)$ qui est une grandeur statistique mesurant directement la relation fonctionnelle entre U et Π. Son calcul passe par une discrétisation des positions Π, mais pas des valeurs de U. Nous avons également étudié l'indicateur $U_c(U|\Pi)$ qui s'appuie sur l'information mutuelle entre U et Π et l'entropie de U. Cet indicateur U_c que nous avons proposé correspond à une estimation d'une version normalisée de l'information mutuelle par la méthode des histogrammes, en discrétisant l'espace des variables U et Π, avec une grande taille d'intervalle pour U. Ce découpage permet de ne pas prendre en compte le fait que les valeurs de U encodées par une position de BMU Π ont une dispersion locale, un bruit. Dans ce cas, l'indicateur permet d'évaluer numériquement si un BMU code pour un seul intervalle de valeur pour U et non plusieurs comme dans le cas d'une carte simple. Il permet de comparer les expériences entre elles, donnant une valeur normalisée entre 0 et 1. Il est cependant limité par la dimension de la variable U : l'estimation par histogrammes, nécessite trop de points si U dépasse la dimension 2 ou 3. Dans un but de mesure de la relation fonctionnelle entre U et Π, le ratio de corrélation sera donc à privilégier, car il est estimable pour des valeurs de U en toute dimension et ne dépend pas de la taille d'intervalle choisie pour U. Sa valeur devra cependant être comparée aux valeurs du ratio de corrélation sur les données d'entrée $\eta(U;X)$.

La relation fonctionnelle entre U et Π n'est toutefois pas une propriété souhaitable dans des plus grandes architectures car elle apporte une forte perte d'information sur l'entrée externe au profit d'un grain d'information sur le modèle. On voudrait plutôt que la représentation de U soit distribuée entre les cartes. Pour étudier l'apprentissage multimodal dans un cadre plus général, nous suggérons aux travaux futurs de s'intéresser à l'information mutuelle et l'information multivariée entre U et les valeurs des BMUs au sein des architectures de cartes.

Finalement, ce chapitre montre que la représentation des éléments d'une carte et des entrées d'un point de vue statistique, que nous avons proposé au chapitre 4 est une méthode pertinente

pour la compréhension des comportements d'apprentissage du modèle CxSOM, mais que leur analyse statistique et le développement d'indicateurs pertinents restent à explorer dans des travaux futurs. Cette approche « comportementale », et non basée sur les poids des cartes, rapproche l'étude des cartes de Kohonen d'autres algorithmes d'apprentissage comportant des entrées et sorties définies. Les perspectives d'études par l'information mutuelle mentionnées ci-dessus sont donc générales à tout type d'architecture d'apprentissage.