Calcul analytique de courbes COR en imagerie de diffraction X

Fanny MARTICKE[1,2], Guillaume MONTÉMONT[1], Caroline PAULUS[1], Olivier MICHEL[2], Jérôme I. MARS[2], Loïck VERGER[1]

[1]CEA, LETI, Minatec Campus, 17 rue des Martyrs, 38054 Grenoble Cedex 09, France

[2]Univ. Grenoble-Alpes, Gipsa-Lab, 38000 Grenoble, France
CNRS, Gipsa-Lab, 38000 Grenoble, France

fanny.maricke@cea.fr, guillaume.montemont@cea.fr, caroline.paulus@cea.fr
olivier.michel@gipsa-lab.grenoble-inp.fr, jerome.mars@gipsa-lab.grenoble-inp.fr

Résumé – Les caractéristiques opérationnelles de réception (COR) sont utilisées pour étudier la séparabilité entre deux distributions **A** et **B**. Généralement, des courbes COR sont construites en générant de nombreuses réalisations de **A** et **B**. Dans cet article, nous proposons une méthode analytique d'analyse COR basée sur la géométrie d'information. Cette méthode a été validée sur des données simulées de diffraction X.

Abstract – Receiver operation charactéristics (ROC) are used to study the discrimination power of a system or test between two distributions **A** and **B**. In general, ROC curves are obtained by producing numerous observations of **A** and **B**. In this paper, we propose an analytical ROC calculation method based on information geometry. This method has been tested on simulated X-ray diffraction data.

1 Introduction

L'analyse des performances d'un test d'hypothèse binaire en utilisant la caractéristique opérationnelle du récepteur (COR) a été appliquée dans de nombreux domaines et notamment, dans la mise au point de tests diagnostiques en médecine [1]. Dans le cadre de notre étude, le calcul de courbes COR est employé pour étudier la séparabilité de tissus mammaires en diffraction X.

La diffraction est une méthode d'imagerie qui permet d'accéder à la structure moléculaire de l'objet analysé et donc à des informations propres à chaque matériau. Grâce a cette propriété, la diffraction est intéressante pour diverses applications dans le domaine de la sécurité [2] (détection d'explosifs ou de drogue, l'identification de liquides) ou de la santé (cancérologie). Dans la détection du cancer du sein [3], la mammographie basée sur la diffraction X pourrait être plus spécifique que la mammographie classique. Associer une analyse COR à la diffraction X peut donner des renseignements concernant la spécificité et la sensibilité de cette technique d'imagerie en mammographie.

Classiquement, une courbe COR est obtenue à partir de nombreuses observations (simulations ou mesures) des deux lois de probabilité de **A** et de **B** à séparer (ici, distributions pour tissus cancéreux et pour tissus sains) qui sont ensuite attribuées à une des deux classes. Cette méthode, par intégration Monte-Carlo, nécessite un nombre élevé d'observations. Dans cet article, une méthode alternative analytique est proposée en se basant sur la géométrie d'information [4]. En s'inspirant du théorème de Sanov [5], nous proposons une approximation des distributions de probabilité de **A** et **B** relatives à la mesure **x**. Nous montrons qu'avec cette approximation, il est suffisant de ne considérer que les points $\mathbf{x}_0$ de la géodésique liant les distributions **A** et **B** pour couvrir tout l'espace des mesures et de tracer la courbe COR correspondante.

Cette méthode est testée sur des données de diffraction X théoriques et des spectres de diffraction simulés dégradés par le système de mesure.

2 Courbes COR

2.1 Définitions et objectifs

A partir d'un ensemble d'observations **x**, on veut décider si elles appartiennent à la distribution de A ou de B (dans la suite, on ne dira plus que **A** ou **B** en désignant les distributions des deux classes A et B). Cette décision est prise par un observateur qui peut être soit humain soit algorithmique. L'observateur classiquement utilisé est le rapport de vraisemblance. Il peut être facilement démontré que toutes stratégies de test se ramènent à un rapport de vraisemblance Λ. Le rapport de vraisemblance utilise toute l'information statistique disponible par rapport à la tâche de discrimination considérée de manière à maximiser sa performance [6]. Les erreurs de décision sont exclusivement dues à la variabilité des deux classes d'objets A et B et au bruit de mesure. Pour pouvoir l'utiliser, il faut connaître les distributions **A** et **B** ce qui peut ne pas être le cas en pratique. Si les distributions sont connues, il est souvent plus commode de considérer le log du rapport de vraisemblance λ :

$$\lambda(\mathbf{x}) = \ln(\Lambda(\mathbf{x})) = \ln\left[\frac{P(\mathbf{x}|\mathbf{A})}{P(\mathbf{x}|\mathbf{B})}\right] \underset{\mathbf{B}}{\overset{\mathbf{A}}{\gtrless}} \lambda_t \qquad (1)$$

La forme de $\lambda(\mathbf{x})$ dépend du type des distributions **A** et **B**. La prise de décision se fait par comparaison de λ à un seuil λ_t.

Chaque point de la courbe COR correspond à un seuil différent. Le coût de détermination d'une courbe ROC est élevé car elle nécessite la génération d'observations des lois $P(\mathbf{x}|\mathbf{A})$ et $P(\mathbf{x}|\mathbf{B})$. Pour répondre à cette alternative, nous proposons une méthode analytique moins coûteuse pour accéder aux courbes COR.

2.2 Calcul analytique

Soient A et B deux éléments (matériaux ou tissus) à séparer et $P(\mathbf{A})$ et $P(\mathbf{B})$ leurs distributions de probabilité associées en diffraction X. En diffraction X, on mesure un nombre de photons par canal. La source de bruit prépondérante est donc le bruit photonique, qui par sa nature physique même est poissonien. Les deux distributions suivent donc une loi de Poisson de paramètres A_i et B_i au canal de mesure i ($1 \leq i \leq k$, $k \in \mathbb{N}$), et leurs supports sont les mêmes. On utilisera la métrique de Fisher, naturelle dans l'espace des lois de probabilité. Le log de leur rapport de vraisemblance est donné par :

$$\lambda(\mathbf{x}) = \sum_{i=1}^{k} x_i \ln(A_i/B_i) - A_i + B_i \tag{2}$$

avec $\sum_{i=1}^{k} A_i = \sum_{i=1}^{k} B_i$. La différentiation de λ par rapport à x_i donne :

$$\nabla_{x_i}\lambda(\mathbf{x}) = \nabla_{x_i} \ln(\Lambda(x)) = \ln(A_i/B_i) \tag{3}$$

Le gradient est donc constant dans tout l'espace. On peut noter que l'on établit aisement que dans le cas présent (lois de Poisson indépendantes par canal), si $x_i = A_i$ alors :

$$\lambda(\mathbf{x} = \mathbf{A}) = D_{KL}(\mathbf{A}||\mathbf{B}) \tag{4}$$

On sait que dans l'espace des lois de probabilités, la forme générale d'une géodésique Γ entre deux distributions P et Q est donnée par [7] :

$$\Gamma : t \in [0;1] \mapsto P^t Q^{1-t} \tag{5}$$

Dans le cas de distributions de Poisson, la géodésique entre deux distributions **B** et **A**, la relation 5 est donnée sur les paramètres des deux processus de Poisson [7] :

$$\Gamma : t \in [0;1] \mapsto A^t B^{1-t} \tag{6}$$

Les points $\mathbf{x}_0$ de cette courbe forment la ligne de moindre séparabilité entre **A** et **B**. Afin de pouvoir étudier la séparabilité entre **A** et **B** dans tout l'espace, nous allons décomposer **A** et **B** sur cette géodésique, et sur ses directions orthogonales. Nous allons donc nous servir de la géométrie de la théorie de l'information au sens de la métrique de Fisher, et approximer les distributions de probabilité d'un point de mesure **x** de l'espace sachant **A** (et de la même manière pour **B**) par :

$$P(\mathbf{x}|\mathbf{A}) \propto e^{-D(\mathbf{x}||\mathbf{A})} \tag{7}$$

avec $D(\mathbf{x}||\mathbf{A}) = \sum_{i=1}^{k} x_i \ln\left(\frac{x_i}{A_i}\right) - x_i + A_i$ (divergence de Kullback-Leibler généralisée) et **x** ayant le même support que A. Cette approximation correspond à une version généralisée du théorème de Sanov [5]. L'utilisation de la divergence KL implique plusieurs avantages. D'une part, elle permet de respecter la variation de la quantité d'information lorsque le nombre de canaux de mesure varie. D'autre part, elle permet une décomposition facile des distributions sur la géodésique et les iso-Λ par la relation suivante :

$$\begin{aligned} &D(\mathbf{x}||\mathbf{x}_0(t)) + D(\mathbf{x}_0(t)||\mathbf{A}) = \\ &D(\mathbf{x}||\mathbf{A}) + \sum_{i=1}^{k} [x_i - x_{0,i}(t)] \cdot \ln(A_i/x_{0,i}(t)) \end{aligned} \tag{8}$$

On considère que **x** appartient à une courbe parallèle à la géodésique entre **B** et **A**, paramétrée de la même manière. $\mathbf{x}_0$ est la projection orthogonale au sens de la métrique de Fisher de **x** sur la géodésique entre **B** et **A**, c'est-à-dire la loi la plus proche de l'observation **x** qui se trouve sur la géodésique Γ. Les points **x** et $\mathbf{x}_0$ font donc partie de la même iso-Λ. En remplaçant $x_{0,i}(t)$ par $A_i^t B_i^{1-t}$ (éq. 6), l'équation 8 devient :

$$\begin{aligned} &D(\mathbf{x}||\mathbf{x}_0(t)) + D(\mathbf{x}_0(t)||\mathbf{A}) = \\ &D(\mathbf{x}||\mathbf{A}) + \sum_{i=1}^{k} [x_i - x_{0,i}(t)](1-t) \cdot \ln(A_i/B_i) \end{aligned} \tag{9}$$

Le gradient $\ln(A_i/B_i)$ est orthogonal à l'iso-Λ dont font partie **x** et $\mathbf{x}_0$ et le dernier terme de l'équation 9 est donc nul. On retrouve le théorème de Pythagore généralisé [5]. Chaque probabilité peut alors être décomposée en deux termes :

$$P(\mathbf{x}|\mathbf{A}) \propto e^{-D(\mathbf{x}||\mathbf{x}_0(t))} e^{-D(\mathbf{x}_0(t)||\mathbf{A})} \tag{10}$$

Le deuxième terme est constant pour tous les **x** sur une iso-Λ (même valeur de t) et le premier terme dépend de la valeur de la divergence entre la géodésique et la courbe parallèle à laquelle appartient **x**. La figure 1 montre une illustration de ces considérations pour le cas de deux canaux. En sommant sur toutes ces courbes parallèles de l'espace, on obtiendrait la probabilité pour une iso-Λ. On en déduit qu'il suffit de considérer seuls les points $\mathbf{x}_o$ sur la géodésique entre **B** et **A** en normalisant de la manière suivante :

$$P(\mathbf{x}(t)|\mathbf{A}) = \frac{e^{-D(\mathbf{x}_0(t)||\mathbf{A})}}{\sum_t e^{-D(\mathbf{x}_0(t)||\mathbf{A})}} \tag{11}$$

On calcule les distributions $P(\mathbf{x}(t)|\mathbf{A})$ et $P(\mathbf{x}(t)|\mathbf{B})$ et le λ correspondant pour les points $\mathbf{x}_0(t)$ de la géodésique, pour tracer la courbe ROC. Ensuite, on détermine la probabilité de vrais positifs et de faux positifs pour des seuils λ_t différents. Ces probabilités correspondent aux aires hachurées dans la figure 2. En pratique, il suffit de calculer les fonctions de répartition en fonction de λ, et de tracer $F_{\mathbf{B}}$ en fonction de $F_{\mathbf{A}}$ pour obtenir la courbe ROC.

Il faut noter que l'approximation de Sanov n'est en général valable que lorsque le nombre d'événements est élevé. Cependant, pour les distributions considérées (carcinome et tissus fibroglandulaires, voir figure 4), la convergence est assez rapide. Nous l'avons vérifié en générant un ensemble de réalisations (à 5, 20 et 100 photons) des distributions **A** (carcinome) et **B**

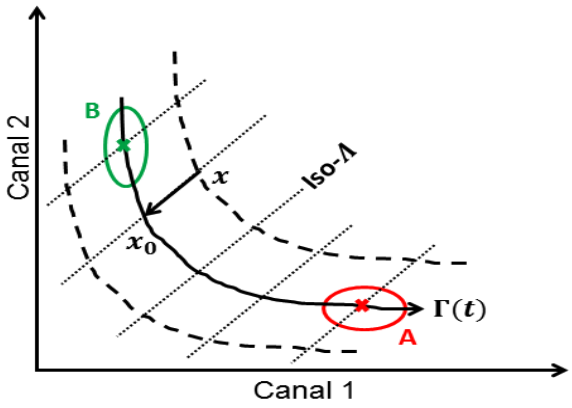


FIGURE 1 – Exemple de géodésique et des iso-Λ correspondantes pour deux canaux.

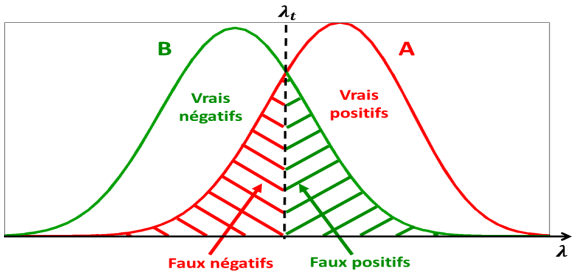


FIGURE 2 – Le seuil λ_t fixe le nombre de vrais et faux positifs.

(fibroglandulaire) par méthode de Monte-Carlo (MC). Ensuite, nous avons calculé les distributions associées à **A** et **B** en fonction de λ. La figure 3a montre ces distributions en fonction de λ superposées à celles obtenues avec notre approximation à 20 photons. Les distributions sont déjà assez proches. La comparaison des courbes COR à 5, 20 et 100 confirme que l'erreur devient négligeable à 20 photons (figure 3b).

3 Application à la diffraction X

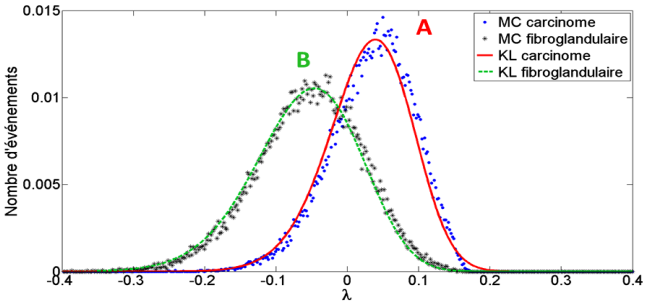
Les signatures de diffraction X peuvent s'identifier soit à des pics discrets appelés pics de Bragg pour des matériaux cristallins soit à une fonction continue appelée fonction d'interférence moléculaire pour les matériaux amorphes. Dans le cadre de la mammographie, les matériaux considérés sont des tissus ayant une structure amorphe. Dans la suite, on ne s'intéressera donc qu'à ce type de spectres.

Le signal de diffraction dépend du transfert de quantité de mouvement χ qui est défini comme :

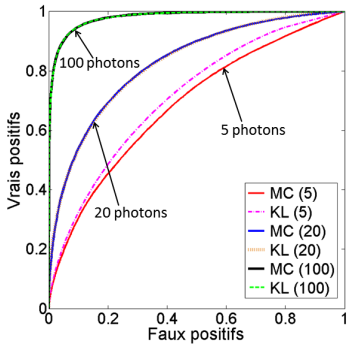
$$\chi = \sin\left(\frac{\theta}{2}\right)\frac{E}{hc} \tag{12}$$

où θ correspond à l'angle de diffraction, E à l'énergie du photon diffusé, h à la constante de Planck et c à la vitesse de la lumière. L'inverse de χ reflète les distances intermoléculaires privilégiées dans le matériau.

Pour tester la méthode de construction de courbes COR proposée, la séparation de tissus fibroglandulaires et de carcinome en diffraction X a été considérée. Les tissus fibroglandulaires sont des tissus très denses du sein. En mammographie classique, il est très difficile de les distinguer de tissus cancéreux. La figure 4 montre les signatures de référence de ces deux types de tissus.



(a) Distributions de probabilité en fonction de λ.



(b) Courbes COR.

FIGURE 3 – Comparaison des résultats obtenus en Monte-Carlo (MC) et avec la méthode proposée (KL).

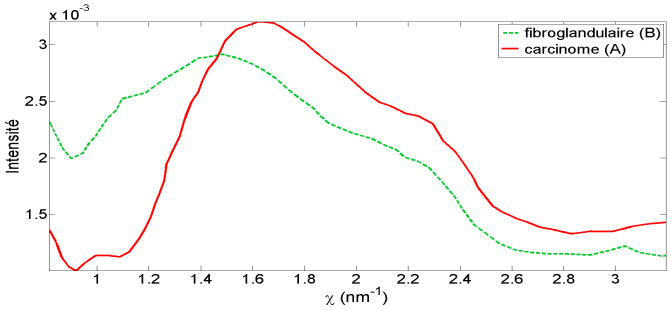
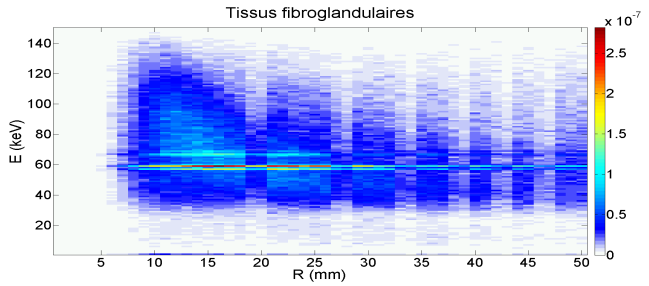


FIGURE 4 – Signatures de diffraction de référence [3] de tissus fibroglandulaires et carcinome (nombre de canaux $k = 500$).

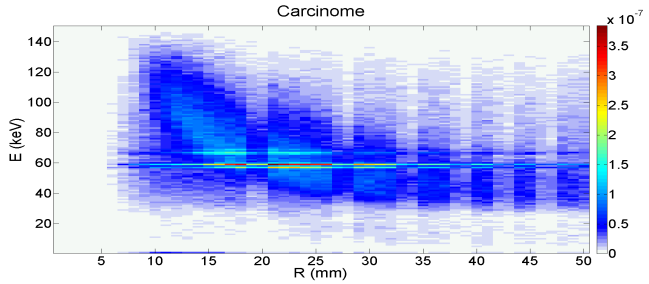
La séparabilité de ces deux tissus a été testée pour un système parfait, c'est-à-dire entre les signatures de référence (figure 4) sans déformation par un système de mesure, et pour des simulations prenant en compte la déformation du spectre liée aux différents facteurs d'un système réel comme la forme du spectre incident, la géométrie du système et la réponse du détecteur [8]. Ces dernières correspondent à des spectres 2D dépendant de l'énergie E et de la position R sur le détecteur.

Chaque R correspond à un angle de diffraction différent. L'espace de représentation des mesures est donc (E, R). La figure 5 montre de tels spectres normalisés par rapport au nombre de photons incidents.

Afin d'étudier l'influence du nombre de photons, les distributions normalisées des deux tissus pour le système parfait et le système réel ont été multipliées par un nombre de photons variable.



(a) Tissus fibroglandulaires.



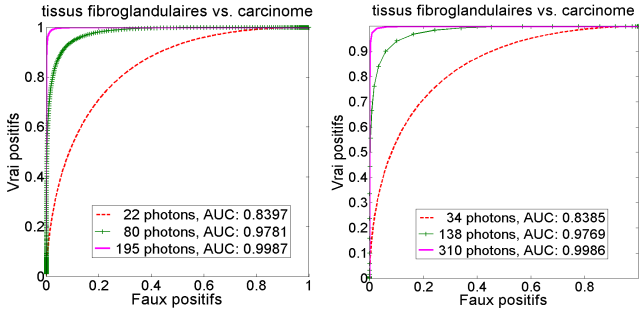
(b) Carcinome.

FIGURE 5 – Réponses du système d'acquisition aux deux tissus (nombre de canaux $k = 150 \times 50$).

4 Résultats et discussion

La figure 6 montre les courbes COR obtenues pour un système idéal (a) et pour un système d'acquisition réel (b) en fonction du nombre de photons détectés. Les différents nombre de photons correspondent à une séparation de 1, 2 et 3σ entre les deux distributions. Pour avoir une séparation presque parfaite entre tissus fibroglandulaires et cancéreux de 3σ, il faut environ 1.6 fois plus de photons avec un système réel qu'avec un système parfait. Cette différence est liée à la perte d'information au cours du processus d'acquisition (déformation par la réponse du système).

La méthode proposée permet donc de calculer des courbes COR pour des spectres sans devoir générer de nombreuses réalisations des deux distributions. Elle permet également d'étudier facilement l'influence du nombre d'événements détectés sur la séparabilité de deux matériaux. Cependant, elle ne permet pas de prendre en compte la variabilité d'objets ce qui sera un point important en mammographie. Dans la suite, il faudra élargir cette méthode afin de pouvoir prendre en compte cette variabilité. Pour d'appliquer cette méthode, il faudra s'assurer que l'approximation de Sanov soit valable. On a observé expérimentalement que la convergence était rapide pour des distributions douces à support large comme c'est le cas pour les matériaux amorphes. Dans la suite, il faudra évaluer la convergence des spectres du type cristallin où on ne trouve que quelques pics de Bragg, intéressants dans le domaine de la sécurité.



(a) Système parfait. (b) Système réel.

FIGURE 6 – Courbes COR obtenues pour différents nombres de photons détectés.

5 Conclusion

Dans cet article, nous avons proposé une méthode qui permet de calculer des courbes COR de manière analytique sans avoir besoin de générer des réalisations des deux distributions à séparer. La méthode a été validée sur des données simulées de diffraction X pour étudier la discrimination de tissus fibroglandulaires et cancéreux avec cette modalité d'imagerie.

Références

[1] L. B. Lusted . Signal detectability and medical decision-making. *Science*, 171:1217–1219, 1971.

[2] G. Harding . X-ray diffraction imaging-A multi-generational perspective. *Applied Radiation and Isotopes*, 67:287–295, 2009.

[3] G Kidane, R D Speller, G J Royle et A M Hanby . X-ray scatter signatures for normal and neoplastic breast tissues. *Physics in Medicine and Biology*, 44:1791–1802, 1999.

[4] S.-I. Amari . Information geometry in optimization, machine learning and statistical inference. *Front. Electr. Electron. Eng. China*, 5:241–260, 2010.

[5] I. Csiszár et P.C. Shields . *Information Theory and Statistics : A Tutorial*. now Publishers Inc., 2004.

[6] H. H. Barrett et K. J. Myers . *Foundations of Image Science*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2004.

[7] A. G. Dabak . *A geometry for detection theory*. Thèse de doctorat, Rice University, 1992.

[8] F. Marticke, C. Paulus, G. Montémont, O. J.J. Michel, J.I. Mars et L. Verger . Multi angle reconstruction of energy dispersive X-ray diffraction spectra. *In WHISPERS IEEE*, 2014.