Mini-symposium PROBAVAR
Méthodes d'estimation de densités de probabilités sur les variétés différentielles et applications

Mini-symposium porté par GSI
≪ Geometric Science of Information ≫
(http ://forum.cs-dc.org/category/72/geometric-science-of-information)

Résumé

En probabilité et en statistique, l'estimation de densité est la construction d'une estimation, basée sur des données observées, d'une fonction de densité de probabilité sous-jacente non observable, considérée comme la densité selon laquelle une grande population est distribuée. Les données sont habituellement considérées comme un échantillon aléatoire de cette population. Les défis actuels sont de généraliser ces estimations, de façon paramétrique ou non-paramétrique, dans des espaces plus généraux (variétés différentielles, espaces métriques, groupes de Lie, ...) avec de nombreuses applications en physique, en analyse de données et en traitement statistique du signal. Nous présentons différentes approches pour aborder ce problème.

Organisateur(s)

1. **Frédéric Barbaresco**, THALES AIR SYSTEMS.

Liste des orateurs

1. **Nicolas Le Bihan**, CNRS / Gipsa-Lab
   Titre : Estimation de densité pour les processus de diffusion multiple sur les hypershpères.
2. **Salem Said**, IMS
   Titre : Lois Gaussiennes dans les espaces symétriques : outils pour l'apprentissage avec les matrices de covariance structurées.
3. **Emmanuel Chevallier**, CMM, Mines ParisTech
   Titre : Densité de matrices Toeplitz ou bloc-Toeplitz hermitiennes définies positives.
4. **Frédéric Barbaresco**, Thales Air Systems
   Titre : Densité de probabilité à maximum d'entropie, définition générale covariante de Jean-Marie Souriau et Jean-Louis Koszul.

Frédéric Barbaresco, Thales Air Systems, Voie Pierre-Gilles de Gennes, 91470 Limours, `frederic.barbaresco@thalesgroup.com`
Nicolas Le Bihan et Florent Chatelain, Gipsa-Lab, 11 Rue des mathématiques, Domaine Universitaire, 38402 Saint Martin d'Hères, `Nicolas.Le-Bihan@gipsa-lab.grenoble-inp.fr`
Salem Said, Université de Bordeaux, Laboratoire IMS, 351 Cours de la libération, 33405 Talence, `salem.said@ims-bordeaux.fr`
Emmanuel Chevalier et Jesus Angluo, Centre de morphologie mathématique, Mines ParisTech, 35 rue Saint-Honoré, 77305 Fontainebleau, `jesus.angulo@mines-paristech.fr`

1 Estimation de densité pour les processus de diffusion multiple sur les hypershpères

On s'intéresse aux processus ponctuels composés sur les hypersphères. Nous présentons tout d'abord plusieurs résultats sur les marches aléatoires isotropes et les processus de diffusion multiple sur les hypersphères. À des fins d'estimation, on s'intéresse ensuite à la décomposition en série de Fourier de la densité de ces processus, ainsi qu'à la conservation du caractère unimodale lors de convolutions multipes sur les hypersphères. En particulier, on introduit les distributions asymptotiques en multi-convolution de distributions de von Mises-Fisher pour les processus composés. Ces résultats permettent de donner les bornes d'estimation pour les paramètres de la distribution asymptotique de processus de Cox composés sur les hypersphères. On présente ensuite une technique d'estimation de type ABC (Approximate Bayesian Computation) pour identifier les distributions des paramètres avec application potentielle en caractérisation/identification de milieux aléatoires.

2 Lois Gaussiennes dans les espaces symétriques : outils pour l'apprentissage avec les matrices de covariance structurées

La notion de loi Gaussienne peut être développée à partir de plusieurs définitions : loi à maximum d'entropie, à minimum d'incertitude (état cohérent), à travers le théorème de la limite centrale, ou la théorie cinétique des gazs. Sur un espace Euclidien, toutes ces définitions mènent à la même forme pour la loi Gaussienne, mais dans des espaces plus généraux, elles donnent lieu à des formes différentes...la présentation propose une définition originale de la notion de loi Gaussienne, valable sur les espaces Riemanniens symétriques de courbure négative. Il s'agit de lois ayant la propriété : le max de vraisemblance est équivalent au barycentre Riemannien. Il n'y a pas de bonne ou de mauvaise définition dans l'absolu, mais celle-ci présente deux avantages, 1) plusieurs espaces de matrices de covariance (réelles, complexes, Toeplitz, Toeplitz par blocs) sont des espaces symétriques de courbure négative, 2) dans ce contexte, elle donne un fondement statistique à l'utilisation du barycentre Riemannien, qui est un outil très populaire pour grand nombre d'applications. Nous allons comparer la définition proposée aux autres définitions possibles, développer les conséquences théoriques de cette définition, et finalement dire comment elle permet de proposer de nouveaux algorithmes d'apprentissage statistique, spécifiquement adaptés au contexte des "big data" et "données en grandes dimensions". Le tout sera illustré par des exemples ... Pour plus de détails, voir les deux papiers [2, 3].

3 Densité de matrices Toeplitz ou bloc-Toeplitz hermitiennes définies positives

Il existe de nombreuses méthodes non paramétriques d'estimation de densité dans le cas des variables aléatoire à valeur dans $\mathbb{R}^n$. Les trois méthodes les plus classiques sont les histogrammes, la méthode des séries orthogonales (aussi appelé méthode de la fonction caractéristique) et l'estimation par noyaux. Ces méthodes se transposent avec plus ou moins de difficulté au cas des variables aléatoires à valeur dans une variété Riemannienne. Rappelons que dans les variétés Riemanniennes, la métrique induit une mesure de volume, ce qui permet de définir une notion de densité. Nous nous intéressons ici à l'espace de Siegel qui est une variété Riemannienne symétrique à courbure négative pouvant être vu comme une généralisation de l'espace hyperbolique. Le point $x + iy$ du demi plan supérieur de Poincaré devient le point $X + iY$ où X est une matrice symétrique et Y une matrice symétrique définie positive. L'espace hyperbolique et l'espace de Siegel permettent de représenter des matrices symétriques définies positives possédant une structure Toeplitz et Toeplitz par blocs. Ces matrices apparaissent par exemple dans le traitement des signaux radars. Nous analyserons les avantages et les inconvénients des principales méthodes d'estimation de densité dans le cas de l'espace de Siegel.
L'existence de nombreux pavages de l'espace hyperbolique rend envisageable la construction d'histogrammes. Cependant l'absence d'homothétie rend difficile l'adaptation de la taille des cases de l'histogramme à un ensemble de tirages aléatoires donné. L'estimation par séries orthogonales est également envisageable mais la construction de la densité estimée nécessite l'estimation d'une intégrale couteuse à calculer. En exploitant les symétries de l'espace, il est possible d'obtenir des expressions explicites de

noyaux permettant l'estimation d'une densité. La méthode de calcul est basée sur la décomposition de Cartan et la décomposition d'Iwasawa et peut être reproduite sur tous les espaces symétriques. L'estimation par noyaux se trouve être la méthode la plus simple à utiliser sur cet espace. L'estimation de densité est appliquée à un problème de segmentation de signaux radar.

4 Densité de probabilité à maximum d'entropie, définition générale covariante de Jean-Marie Souriau et Jean-Louis Koszul

Avant de développer la notion de densité à maximum d'entropie (densité de Gibbs), l'exposé rappellera en préambule le cheminement historique de la notion de "fonction caractéristique" introduite en thermodynamique par François Massieu et développée par Williard Gibbs et Pierre Duhem sous la forme de potentiels thermodynamiques. Le travail de Massieu eut une grande influence sur Henri Poincaré, qui introduisit la "fonction caractéristique" en probabilité (un logarithme lie les 2 notions en probabilité et en thermodynamique). Dans le cas des densités à Maximum d'entropie (densité de Gibbs), la métrique Riemannienne associée au hessien de la fonction caractéristique (logarithme de la fonction de partition) est égale à la matrice de Fisher. Les structures de ces géométries hessiennes ont été étudiées en parallèle par le mathématicien Jean-Louis Koszul et son thésard Jacques Vey dans le cadre plus général des cônes convexes saillants (formes de Koszul, fonction caractéristique de Koszul-Vinberg), en cherchant une métrique qui soit invariante par les automorphismes de ces cônes convexes. La notion de densité à maximum d'entropie a été étendue par Jean-Marie Souriau dans les années 60 dans le cadre de la mécanique statistique pour rendre la densité de Gibbs covariante sous l'action des groupes dynamiques de la physique. Il étend la notion d'ensemble canonique de Gibbs à une variété symplectique homogène sur laquelle un groupe agit (groupes dynamiques de la physique ; sous-groupes du groupe affine : groupe de Galilée ou groupe de Poincaré). Lorsque ces groupes sont non commutatifs, l'algèbre de Lie du groupe vérifie des relations de type cohomologique qui brisent la symétrie. Pour rétablir cette symétrie, Souriau introduit une température « géométrique » comme élément de l'algèbre de Lie, et une chaleur « géométrique » (moyenne de l'Energie qui est le moment de l'action hamiltonnienne du groupe) comme élément de son dual, permettant de remettre en dualité, via la transformée de Legendre, l'Entropie « géométrique » et le logarithme de la fonction de partition (fonction caractéristique) définie pour ces nouvelles variables. La densité de Gibbs-Souriau (densité à Maximum d'Entropie) possède alors la propriété d'être covariante pour le groupe qui agit, et l'Entropie de Boltzmann « géométrique » associée est invariante pour tout symplectomorphisme. Si on se restreint dans ce modèle au groupe des translations temporelles, on retrouve la théorie de la thermodynamique classique, et pour un espace euclidien la statistique classique des gaussiennes. Jean-Marie Souriau a appelé cette nouvelle structure élémentaire de la physique statistique « la thermodynamique des groupes de Lie » et précisa que « ces formules sont universelles, en ce sens qu'elles ne mettent pas en jeu la variété symplectique, mais seulement le groupe, son cocycle symplectique et le couple de la température et de la chaleur (géométriques)». Dans le modèle affine de Souriau, comme dans celui de Jean-Louis Koszul, les structures fondamentales sont déduites de la représentation affine des groupes et algèbres de Lie. Les modèles de Souriau et Koszul permettent ainsi d'approfondir et généraliser la notion de densités à maximum d'entropie. L'approche permet de définir des densités sous forme paramétrique, densité de Gibbs à Maximum d'Entropie, dans des espaces très généraux et abstraits.

Références

[1] Le Bihan, N., Chatelain, F. and Manton, J.H., *Isotropic multiple scattering processes on hyperspheres*, IEEE Transactions on Information Theory, Vol. 62, Num. 10, 2016, pp.5740–5752.

[2] Said, S., Hajri, H., Bombrun and L., Vemuri, B.C., *Gaussian distributions on Riemannian symmetric spaces : statistical learning with structured covariance matrices*, Jul. 2016, EEE Trans Inf Theory (under review). Arxiv : https ://arxiv.org/abs/1607.06929.

[3] Said, S., Bombrun, L., Berthoumieu, Y., Manton, J., *Riemannian Gaussian distributions on the space of symmetric positive definite matrices. IEEE Trans Inf Theory (accepted). Arxiv : https ://arxiv.org/abs/1507.01760*,

[4] ref3, *Chevalier, E., Forget, T, Barbaresco and F. Angulo, J*, Kernel density estimation on the Siegel space with applications to radar processing. Entropy, Vol. 18, Num. 396, 2016

[5] Barbaresco, F., *Geometric theory of heat from Souriau Lie groups thermodynamics and Koszul Hessian geometry : applications in information geometry for exponential families*, Entropy, Vol. 18, Num. 386, 2016.

[6] Barbaresco, F., *Koszul information geometry and Souriau Lie group thermodynamics*, AIP Conf. Proc. no. 1641, Proceedings of MaxEnt'14 conference, Amboise, Septembre 2014..

[7] Barbaresco, F., *Koszul information geometry and Souriau geometric temperature/capacity of Lie group thermodynamics*, , Entropy, vol. 16, pp. 4521-4565, in Information, entropy and their geometric structures, MDPI Publisher, Septembre 2015.

[8] Terras, A., *Harmonic Analysis on Symmetric Spaces and Applications II*, Springer : Berlin/Heidelberg, Germany, (2012)..

[9] Pelletier, B., *Kernel density estimation on Riemannian manifolds*, Stat. Probab. Lett. 73, 297–304, (2005)..

[10] Chevallier, E., Forget, T., Barbaresco, F., Angulo, J., *Kernel Density Estimation on the Siegel Space with an Application to Radar Processing*, Entropy, 18(11), 396, (2016)..

[11] Chevallier, E., Kalunga, E., Angulo, J., *Kernel Density Estimation on Spaces of Gaussian Distributions and Symmetric Positive Definite Matrices*, SIAM J. Imaging Sci., 10(1), 191–215, (2017)..