Inférence asymptotique pour des données directionnelles bruitées

Diego Bolon[1], Davy Paindaveine[2] & Thomas Verdebout[3]

[1] *Universidade de Santiago de Compostela, Espagne, diego.bolon.rodriguez@usc.es*
[2] *Université libre de Bruxelles, Belgique, Davy.Paindaveine@ulb.be*
[3] *Université libre de Bruxelles, Belgique, Thomas.Verdebout@ulb.be*

Résumé. Nous introduisons des modèles paramétriques pour des données directionnelles bruitées dans lesquels un bruit radial de magnitude σ^2 fait dévier les observations de leur domaine sphérique théorique, à savoir une hypersphère centrée en θ et de rayon r. Nous considérons les problèmes d'inférence — les tests d'hypothèses, l'estimation ponctuelle et l'estimation par zone de confiance — sur le paramètre de position θ, dans un contexte où r et σ^2 restent non spécifiés. Nous introduisons divers scénarios asymptotiques dans lesquels le rayon de l'hypersphère et, de façon plus importante, la magnitude du bruit peuvent dépendre de la taille d'échantillon n d'une façon essentiellement arbitraire. Ceci nous permet de considérer des situations très diverses dans lesquelles l'information a priori que les données appartiennent à une hypersphère est de plus en plus, ou au contraire de moins en moins, pertinente. Nous basons notre étude sur la théorie asymptotique des expériences asymptotiques de Le Cam et notre objectif est d'obtenir une compréhension complète des expériences limites qui en résultent. Les taux de contiguïté associés, qui caractérisent la difficulté des problèmes d'inférence considérés, révèlent des résultats assez contre-intuitifs dans certains des scénarios traités. Nous construisons des estimateurs et des tests qui sont localement asymptotiquement optimaux de façon adaptative à travers les différents régimes. Nous montrons que, dans les scénarios asymptotiques standards, les procédures classiques qui ignorent l'information d'a priori hypersphérique réalisent le taux de convergence optimal mais n'atteignent pas les bornes d'efficacité, et que, dans des scénarios asymptotiques non-standards, ces procédures classiques n'ont pas le bon taux de convergence. Nous étudions par l'intermédiaires d'exercices de Monte Carlo à quel point nos résultats asymptotiques se matérialisent dans des échantillons de taille finie. Les perspectives de recherche future incluent notamment l'extension au cas non paramétrique dans lequel la loi du bruit n'est pas spécifiée.

Mots-clés. Théorie asymptotique des expériences statistiques de Le Cam, double asymptotique, taux de contiguïté, données presque directionnelles, identifiabilité forte

Abstract. We introduce parametric models for noisy directional data, in which a radial noise with magnitude σ^2 makes the observations deviate from their theoretical hyperspherical sample space, namely a hypersphere centered at θ and with radius r. We consider inference — hypothesis testing, point estimation, and confidence zone estimation — on the location parameter θ, in a framework where both r and σ^2 remain unspecified. We introduce several asymptotic scenarios in which the radius of the hypersphere and, most importantly, the noise magnitude may depend on the sample size n in an essentially arbitrary way. This allows

us to consider very diverse cases, in which the a priori information that the data belong to a hypersphere is more and more, or on the contrary less and less, relevant. We base our investigation on Le Cam's asymptotic theory of statistical experiments and aim at a full understanding of the resulting limiting experiments. The corresponding contiguity rates, that characterize how easy/hard inference on θ is, reveal rather counter-intuitive results in some scenarios. We build locally asymptotically optimal tests and estimators, that turn out to be adaptively optimal across all asymptotic scenarios. We show that, in standard asymptotic scenarios, classical procedures that would ignore the hyperspherical a priori information are rate-consistent but do not achieve efficiency bounds, and that, in non-standard asymptotic scenarios, such classical procedures are not even rate-consistent. We investigate the finite-sample relevance of our results through Monte Carlo exercises. Perspectives for future research in particular include the extension to the nonparametric case in which the distribution of the radial noise remains unspecified.

Keywords. Le Cam's theory of asymptotic experiments, contiguity rates, double asymptotics, nearly directional data, strong identifiability

1 Contexte

Les problèmes de position multivariés comptent probablement parmi les problèmes les plus étudiés en statistique. Les tests les plus standards pour les problèmes de position à un et deux échantillons sont certainement les tests de Hotelling; voir Hotelling (1931). Ils sont encore beaucoup étudiés aujourd'hui; nous renvoyons par exemple à Chen et al. (2011) et Feng et al. (2017) pour des tests de Hotelling régularisés à un échantillon, et à Li et al. (2020) pour des tests d'une nature similaire pour le problème à deux échantillons. Les tests de Hotelling sont basés sur des moyennes arithmétiques, donc ne sont pas robustes à d'éventuelles observations aberrantes ou à des queues lourdes. Ceci a motivé l'introduction de tests non paramétriques, et en particulier de tests de signes et de rangs signés; voir, parmi beaucoup d'autres, Randles (2000), Hallin et Paindaveine (2002), Larocque, Nevalainen et Oja (2007), Wang, Peng et Li (2015), et Feng, Zou et Wang (2016). Des contributions récentes en inférence pour la position multivariée incluent Agostinelli et Greco (2019) qui a étudié l'estimation par vraisemblance pondérée, Frahm, Nordhausen et Oja (2020), qui a proposé des estimateurs de position pour des données incomplètes et dépendantes, Dürre et Paindaveine (2022), qui a défini des estimateurs de position affine-équivariants fondés sur des simplexes, Chakraborty et Chaudhuri (2017), Kock et Preinerstorfer (2019, 2023), qui ont considéré le cas des grandes dimensions, et Ley et al. (2013), Paindaveine et Verdebout (2017, 2020a,b), qui ont traité les problèmes de position dans le contexte des données directionnelles.

Que ce soit en faible ou en grande dimension, il est de plus en plus courant de supposer que la dimension effective des données, k disons, est plus petite que la dimension d de l'espace ambient; pour ce qui est la grande dimension, nous renvoyons par exemple à Wright et Ma (2022). Sous une hypothèse de linéarité qui veut que les données proviennent d'une version bruitée d'une distribution concentrée sur un hyperplan de dimension k de $\mathbb{R}^d$, la méthodologie

classique dans ce cadre repose alors sur l'analyse en composantes principales, mais bien sûr il est plus prometteur et général de considérer des techniques de réduction de la dimension non linéaires qui permettent aux données de dévier d'une variété k-dimensionnelle courbée. En pratique, tant la dimension k intrinsèque que la variété correspondante restent non spécifiées, et une vaste littérature a considéré le problème de reconstruire ces quantités-clés inconnues; nous renvoyons, par exemple, à Donoho et Grimes (2003), Maggioni, Minsker et Strawn (2016), et aux travaux cités dans ces articles. Des articles récents étudiant l'inférence statistique pour des données déviant de façon modérée d'une variété incluent notamment Shapiro, Xie et Zhang (2021) et Cheng et Xie (2024), qui ont respectivement considéré des tests de goodness-of-fit et des tests à deux échantillons.

Typiquement, les résultats obtenus dans la littérature fournissent des vitesses de convergence et, au mieux, des résultats d'optimalité de type minimax sous des conditions convenables; voir, par exemple, Shapiro, Xie et Zhang (2021) et Cheng et Xie (2024). Pour autant que nous sachions, des résultats d'optimalité/efficacité plus fins, qui établiraient par exemple une optimalité au sens de Le Cam, n'ont pas été obtenus dans ce cadre. Notre travail a pour objectif d'obtenir de tels résultats d'optimalité. Puisqu'il n'y a pas de "free lunch", le prix à payer est de faire des hypothèses plus fortes sur la forme de la variété sous-jacente, par exemple, de supposer que cette variété est une hypersphère ou un hypertore. Dans cet exposé, nous suivons en effet cette route et adoptons un cadre dans lequel l'échantillon à disposition est fait de versions bruitées de vecteurs aléatoires prenant leurs valeurs sur une hypersphère de $\mathbb{R}^d$.

Bibliographie

Agostinelli, C. et Greco, L. (2019), Weighted likelihood estimation of multivariate location et scatter, *Test*, 28, pp. 756-784.

Chakraborty, A. et Chaudhuri, P. (2017), Tests for high-dimensional data based on means, spatial signs and spatial ranks, *Annals of Statistics*, 45, pp. 771-799.

Chen, L. S., Paul, D., Prentice, R. L. et Wang, P. (2011), A regularized Hotelling's T^2 test for pathway analysis in proteomic studies, *Journal of American Statistical Association*, 106, pp. 1345-1360.

Cheng, X., et Xie, Y. (2024), Kernel two-sample tests for manifold data, *Bernoulli*. A paraître.

Kock, A. B. et Preinerstorfer, D. (2019), Power in high-dimensional testing problems, *Econometrica*, 87, pp. 1055-1069.

Donoho, D., et Grimes, C. (2003), Hessian eigenmaps: locally linear embedding techniques for high-dimensional data, *Proceedings of the National Academy of Science USA*, 100, pp. 5591-5596.

Dürre, A. et Paindaveine, D. (2022) Affine-equivariant inference for multivariate location under L_p loss functions, *Annals of Statistics*, 50, pp. 2616-2640.

Feng, L., Zou, C., et Wang, Z. (2016), Multivariate-sign-based high-dimensional tests for the two-sample location problem, *Journal of American Statistical Association*, 111, pp. 721-735.

Feng, L., Zou, C., Wang, Z., et Zhu, L. (2017), Composite T^2 test for high-dimensional data, *Statistica Sinica*, 27, pp. 1419-1436.

Frahm, G., Nordhausen, K. et Oja, H. (2020), M-estimation with incomplete and dependent multivariate data, *Journal of Multivariate Analysis*, 176, pp. 104569.

Hallin, M. et Paindaveine, D. (2002), Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks, *Annals of Statistics*, 30, pp. 1103-1133.

Hotelling, H. (1931), The generalization of Student's ratio, *Annals of Mathematical Statistics*, 2, pp. 360-378.

Kock, A. B., et Preinerstorfer, D. (2023), Consistency of p-norm based tests in high dimensions: Characterization, monotonicity, domination, *Bernoulli*, 29, pp. 2544–2573.

Larocque, D., Nevalainen, J. et Oja, H. (2007), A weighted multivariate sign test for cluster-correlated data *Biometrika*, 94, pp. 267-283.

Ley, C., Swan, Y., Thiam, B. et Verdebout, T. (2013), Optimal R-estimation of a spherical location *Statistica Sinica*, 23, pp. 305–333.

Li, H., Aue, A. Paul, D., Peng, J. et Wang, P. (2020), An adaptable generalization of Hotelling's T^2 test in high dimension *Annals of Statistics*, 48, pp. 1815-1847.

Maggioni, M., Minsker, S. et Strawn, N. (2016) Multiscale dictionary learning: non-asymptotic bounds and robustness, *Journal of Machine Learning Research*, 17, pp. 43-93.

Paindaveine, D. et Verdebout, T. (2017), Inference on the mode of weak directional signals: a Le Cam perspective on hypothesis testing near singularities, *Annals of Statistics*, 45, pp. 800-832.

Paindaveine, D. et Verdebout, T. (2020a), Detecting the direction of a signal on high-dimensional spheres: non-null and Le Cam optimality results *Probability Theory and Related Fields*, 176, pp. 1165-1216.

Paindaveine, D. et Verdebout, T. (2020b), Inference for spherical location under high concentration, *Annals of Statistics*, 48, pp. 2982-2998.

Randles, R.H. (2000), A simpler, affine-invariant, multivariate, distribution-free sign test, *Journal of American Statistical Association*, 95, pp. 1263-1268.

Shapiro, A., Xie, Y. et Zhang, R. (2021), Goodness-of-fit tests on manifolds, *IEEE Transactions on Information Theory*, 67, pp. 2539-2553.

Wang, L., Peng, B. et Li, R. (2015), A high-dimensional nonparametric multivariate test for mean vector, *Journal of American Statistical Association*, 110, pp. 1658-1669.

Wright, J. et Ma, Y. (2022), *High-Dimensional Data Analysis with Low-Dimensional Models. Principles, Computation, and Applications*, Cambridge University Press.