Geometric Aspects of Mathematical Optimization and Statistical Theories

Date: December 16–17, 2025 (Japan Standard Time)

Venue: Academic Extension Center (Osaka Metropolitan University)

Contents: Workshop (Hybrid: physical/virtual)

Organizers: Hideto Nakashima (Tokai Univ), Koichi Tojo (Tokai Univ), Hiroto Inoue (NIT), Yoshihiko Konno (OMU), Hideyuki Ishi (OMU), Kenji Fukumizu (ISM)

Support: Osaka Central Advanced Mathematical Institute: MEXT Joint Usage/Research Center on Mathematics and Theoretical Physics JPMXP0619217849

JSPS KAKENHI Grant Numbers JP24KK0059 (Fund for the Promotion of Joint International Research)

December 16 (Tue)

13:00 – 13:50 **Atsumi Ohara** (Otemon Gakuin University)

Conformal Flattening on the Probability Simplex and Its Applications

14:00 – 14:50 **Naomichi Nakajima** (Shibaura Institute of Technology)

Information geometry of multiple roots in maximum likelihood estimation and singularity theory

15:00 – 15:50 **Kaori Yamaguchi** (Ritsumeikan University)

Characterization of δ-almost sufficient statistics via KL divergence

16:10 – 17:00 **Akifumi Okuno** (Institute of Statistical Mathematics)

Algebraic Approach to Ridge-Regularized Mean Squared Error Minimization in Minimal ReLU Neural Network

17:10 – 18:00 **Bruno F. Lorenço** (The Institute of Statistical Mathematics) (Online)

Some recent results on the geometry of hyperbolicity cones

December 17 (Wed)

10:00 – 10:50 **Takaharu Yaguchi** (Kobe University)

Symplectic Methods for Learning Hamiltonian Systems

11:00 – 11:50 **Masaki Yoshioka** (The University of Osaka)

Geometric bias correction of maximum likelihood estimators in one-sided truncated exponential families

13:30 – 14:20 **Michiko Okudo** (Chiba University)

Bayesian shrinkage prediction and estimation based on geometric structure

14:30 – 14:20 **Hisatoshi Tanaka** (Waseda University)

The second score equation and semiparametric efficiency in regression models

15:40 – 16:30 **Bartosz Kołodziejek** (Warsaw University of Technology)

A new class of colored Gaussian graphical models with explicit normalizing constants

16:40 – 17:30 **Piotr Zwiernik** (Pompeu Fabra University) (Online)

Probabilistic PCA on tensors

Conformal Flattening on the Probability Simplex and Its Applications

Atsumi Ohara

Otemon Gakuin University, Japan

a-ohara@haruka.otemon.ac.jp

Abstract

From a viewpoint of affine differential geometry a certain class of non-flat information geometric structure is known to be conformally equivalent to dually flat one [1]. In this presentation we show a transformation between such structures on the probability simplex and provides its applications.

By restricting affine immersions with certain conditions, the probability simplex of dimension n can be realized to be 1-conformally flat statistical manifolds immersed in $\mathbf{R}^{n+1}$. Using this fact, we introduce a concept of *conformal flattening* for such manifolds in order to obtain the corresponding dually flat statistical (Hessian) ones with conformal divergences, and show explicit forms of potential functions and affine coordinates [3, 4].

Finally, we demonstrate applications of the flattening to nonextensive statistical physics, Voronoi partitions and weighted centroids on the probability simplex with respect to geometric divergences, which are not necessarily of Bregman type [2, 4].

This is partly joint work with Hiroshi Matsuzoe and Shun-ichi Amari.

References

[1] T. Kurose, On the divergences of 1-conformally flat statistical manifolds. *Tohoku Math. J.*, 46, 427-433 (1994).

[2] A. Ohara, H. Matsuzoe and S.-I. Amari,, Conformal geometry of escort probability and its applications, *Modern Physics Letters B*, Vol.26, No.10, 1250063 (14pp) (2012).

[3] A. Ohara, Conformal flattening for deformed information geometries on the probability simplex, *Entropy*, Vol.20, No.3, 186 (13pp.) (2018).

[4] A. Ohara, Conformal flattening on the probability simplex and its applications to Voronoi partitions and centroids, *Geometric Structure of Information* (Frank Nielsen eds.), Chapter 4 (pp.51-68), Springer (2019).

INFORMATION GEOMETRY OF MULTIPLE ROOTS IN MAXIMUM LIKELIHOOD ESTIMATION AND SINGULARITY THEORY

NAOMICHI NAKAJIMA (SHIBAURA INSTITUTE OF TECHNOLOGY)

Abstract

Information geometry provides a unified perspective on various fields such as statistics, information theory, and optimization theory [1]. In information geometry, *the dually flat spaces* introduced by Amari and Nagaoka play a central role, and they have a canonical distance-like function, called *the Bregman divergence*, which determines the dually flat structure. A typical example of a dually flat space is an exponential family endowed with the KL divergence as its Bregman divergence, in which the maximum likelihood estimator for a curved exponential family is characterized as the orthogonal projection with respect to the KL divergence.

The maximum likelihood estimator is usually obtained as a solution of the maximum likelihood equation. However, maximum likelihood equations may have multiple solutions in general, and this phenomenon causes various issues in statistics and numerical analysis, such as root selection [6], lack of model fit [5], interpretation of the maximum likelihood estimator [4], and catastrophe phenomena of MLE. These problems are called *multiple root problems*, which have been investigated for specific models and by ad hoc methods so far [6].

In this talk, we construct a comprehensive framework, not depending on specific models, for studying multiple root problems from the viewpoint of information geometry. Our key idea is to introduce a new geometric notion—*e/m-caustics*, which provide a unified geometric description of the mechanisms behind multiple root problems. They are studied via singularities of the Bregman divergence and are closely related to the extrinsic geometry of statistical models. This connection to extrinsic geometry goes back to Efron's classical work on curved exponential families [2, 3].

References

[1] S. Amari and H. Nagaoka, *Methods of Information Geometry*, Translations of Mathematical Monographs **191**, American Mathematical Society and Oxford University Press, New York (2000).

[2] B. Efron, Defining the Curvature of a Statistical Problem (with Applications to Second Order Efficiency), *Ann. Statist.* **3** (1975), 1189–1242.

[3] B. Efron, The Geometry of Exponential Families, *Ann. Statist.* **6** (1978), 362–376.

[4] B. Efron, Curvature and inference for maximum likelihood estimates, *Ann. Statist.* **46** (2018), 1664–1692.

[5] R. Sundberg, Flat and Multimodal Likelihoods and Model Lack of Fit in Curved Expo- nential Families, *Scand. J. Stat.* **37** (2010), 632–643.

[6] C. G. Small, J. Wang and Z. Yang, Eliminating multiple root problems in estimation, *Statist. Sci.* **15** (2000), 313–341.

This work was partly supported by JSPS KAKENHI Grant Numbers JP25K17299, JP24K21519.

Characterization of δ-almost sufficient statistics via KL divergence

Kaori Yamaguchi (Ritsumeikan University)

Abstract

As a quantitative generalization of sufficient statistics, Yamaguchi and Nozawa proposed the $\boldsymbol{\delta}$-almost sufficient statistic, formulated using the Fisher metric. In this talk, we characterize this $\boldsymbol{\delta}$-almost sufficient statistic using the Kullback-Leibler divergence. We also present a new example of $\boldsymbol{\delta}$-almost sufficient statistics for a model of repeated trials. Furthermore, we analyze the estimation error of the maximum likelihood estimator (MLE) within this framework.

References

[1] Ay, N., Jost, J., Lê, H.V., Schwachhöfer, L.: Information geometry and sufficient statistics. Probab. Theory Relat. Fields 162, 327–364 (2015) https://doi.org/10.1007/s00440-014-0574-8

[2] Kullback, S., Leibler, R.A.: On Information and Sufficiency. Ann. Math. Statist. **22**(1), 79–86 (1951) https://doi.org/10.1214/aoms/1177729694

[3] Absil, P.-A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton, NJ (2007). ISBN 978-0-691-13298-3.

[4] Yamaguchi, K., Nozawa, H. On statistics which are almost sufficient from the viewpoint of the Fisher metrics. Info. Geo. (2024). https://doi.org/10.1007/s41884-024-00160-1

Algebraic Approach to Ridge-Regularized Mean Squared Error Minimization in Minimal ReLU Neural Network

Akifumi Okuno*[1,2,3], Ryoya Fukasaku[4], and Yutaro Kabata[5]

[1]Institute of Statistical Mathematics
[2]The Graduate University for Advanced Studies, SOKENDAI
[3]RIKEN
[4]Kyushu University
[5]Kagoshima University

Keywords: Computational Algebra, ReLU Neural Network

This study investigates a perceptron, a simple neural network model, with ReLU activation and a ridge-regularized mean squared error (RR-MSE). Our approach leverages the fact that the RR-MSE for ReLU perceptron is piecewise polynomial, enabling a systematic analysis using tools from computational algebra. In particular, we develop a Divide-Enumerate-Merge strategy that exhaustively enumerates all local minima of the RR-MSE. By virtue of the algebraic formulation, our approach can identify not only the typical zero-dimensional minima (i.e., isolated points) obtained by numerical optimization, but also higher-dimensional minima (i.e., connected sets such as curves, surfaces, or hypersurfaces). Although computational algebraic methods are computationally very intensive for perceptrons of practical size, as a proof of concept, we apply the proposed approach in practice to minimal perceptrons with a few hidden units. This presentation is based on a preprint of ours (Fukasaku et al., 2025).

References

Fukasaku, R., Kabata, Y., and Okuno, A. (2025). Algebraic approach to ridge-regularized mean squared error minimization in minimal relu neural network. https://arxiv.org/abs/2508.17783.

*okuno@ism.ac.jp

Some recent results on the geometry of hyperbolicity cones

Bruno F. Lorenço (The Institute of Statistical Mathematics)

Abstract

In this talk, we first overview the basics of hyperbolicity cones and their importance to optimization applications. Then, we present some recent results on the geometry of hyperbolicity cones including a discussion on its facial exposedness properties and automorphism groups of certain types of hyperbolicity cones. This talk is based on joint works with Masaru Ito, Vera Roshchina and James Saunderson.

Symplectic Methods for Learning Hamiltonian Systems

Takaharu Yaguchi

Kobe University/RIKEN AIP

In this talk, I will explain machine learning methods using symplectic geometry to learn the Hamilton equations.

Many physical phenomena are described by classical mechanics. Hamiltonian mechanics is a formalism of classical mechanics, and many phenomena are described by the Hamilton equations. From a geometric point of view, the Hamilton equations are defined as a certain type of gradient flow on a symplectic manifold with respect to a function called the Hamiltonian, which is the total energy of the Hamiltonian system.

To learn the Hamilton equations from observational data of physical phenomena using machine learning, it is necessary to learn the Hamiltonian and the structure of the underlying symplectic manifold from data. In particular, machine learning methods, such as neural networks, generally have universal approximation properties. While this is ideal in the sense that any function can be approximated, it also means that functions that do not preserve the symplectic structure can be obtained. Therefore, machine learning models that naively use such methods may not preserve the symplectic structure. However, since the symplectic structure is a necessary and sufficient condition for the Hamilton equations, if the symplectic structure is lost, the physical properties, including the energy conservation law, are also lost.

In this talk, I will introduce neural networks that can preserve physical laws by preserving the symplectic structure, and present related research.

Geometric bias correction of maximum likelihood estimators in one-sided truncated exponential families

Masaki Yoshioka
Graduate School of Engineering Science, The University of Osaka

In regular statistical models, the asymptotic bias of the maximum likelihood estimator (MLE) is known to be expressed in terms of the connection coefficients of the m-connection, and this provides a systematic geometric method for bias correction (Amari and Kumon, 1983). Such a geometric formulation relies on standard regularity conditions, including the parameter-independent support of the model. When these conditions fail, as in truncated models, the usual information-geometric framework cannot be directly applied.

We consider a one-sided truncated exponential family (oTEF), where the support depends on a truncation parameter γ. This parameter-dependent support makes the model non-regular. Because the support depends on the truncation parameter, the behavior of the MLE bias differs from that in regular models. Existing asymptotic results (Bar-Lev, 1984; Akahira, 2016) do not provide a geometric interpretation of the MLE bias.

The main contribution of the study is to show that the leading term of the asymptotic bias of the MLE in the oTEF coincides with the connection coefficients of the α-connection proposed by Yoshioka and Tanaka (2023). This result provides a clear statistical interpretation for the geometric structure of the oTEF: the structure represents the bias information of the MLE, revealing both the similarities to and distinctions from regular models. Furthermore, this finding enables geometric bias correction for non-regular models, demonstrating that information-geometric bias correction can be extended beyond the regular case.

Bayesian shrinkage prediction and estimation based on geometric structure

Michiko Okudo
Chiba University

The geometric structure of a statistical model often provides valuable insight into the construction of prior distributions in Bayesian analysis. In particular, priors whose mass is concentrated on specific points or subspaces—referred to as shrinkage priors—can improve estimation and prediction when the model exhibits certain geometric properties. Shrinkage estimation is a classical approach in statistics, originating from procedures that shrink the estimator of a multivariate normal mean toward the origin. This talk first surveys existing results on how shrinkage priors have been constructed in Bayesian statistics, with a particular focus on approaches motivated by geometric properties of statistical models.

Building on these ideas, we will present our recent work on shrinkage priors for the mean vector of multivariate normal models and the covariance matrix. When the parameter is a matrix rather than a vector, the notion of shrinkage extends in multiple directions. For covariance matrix estimation in multivariate normal models, shrinking eigenvalues has become a representative strategy. Motivated by the geometric properties of the model—particularly those arising from the Fisher metric— we present priors that introduce shrinkage directions other than eigenvalue modification [2]. For multivariate normal models with unknown means, we develop a prediction method in [1] that serves as an alternative to approximated Bayesian predictive distributions and empirical Bayes approaches. The method constructs estimators on an extended domain larger than the original parameter space and uses them to define predictive distributions; shrinkage priors are then incorporated into this framework to further improve performance.

References

[1] M. Okudo and F. Komaki. Predictive densities for multivariate normal models based on extended models and shrinkage Bayes methods. Electronic Journal of Statistics, vol. 18, 3310–3326, 2024.

[2] M. Okudo and T. Sei. Shrinkage priors for circulant correlation structure models. Information Geometry, accepted.

The Second Score Equation and Semiparametric Efficiency in Regression Models

Hisatoshi Tanaka (Waseda University)

In the classical maximum likelihood estimation theory, the following two properties of the log-likelihood function $\ell_\theta := \log p_\theta$ play a crucial role, where p_θ is a probability density indexed with $\theta \in \Theta \subset \mathbb{R}^k$. The first property $E\dot{\ell}_\theta = 0$ is often referred to as the first score equation. It guarantees the asymptotic normality of the maximum likelihood estimator. The second property $E\dot{\ell}_\theta\dot{\ell}_\theta^\top = -E\ddot{\ell}_\theta$ is called the second score equation and ensures that the maximum likelihood estimator achieves the Cramér-Rao lower bound.

These implications of the score equations are not limited to maximum likelihood estimation. This study generalizes the second score equation as an equivalence between the Riemannian metric defined by the objective function of M-estimation and that derived from the asymptotic variance of the corresponding M-estimator. This generalization allows us to interpret the fact that regression methods frequently used in econometrics, such as Generalized Least Squares (GLS) and Generalized Method of Moments (GMM), achieve semiparametric efficiency in terms of the validity of the second score equation. Consequently, it provides a geometric intuition for the efficiency of these regression methods.

References

[1] Chamberlain, G.: Asymptotic efficiency in estimation with conditional moment restrictions. Journal of econometrics **34**(3), 305–334 (1987)

[2] Eguchi, S.: Geometry of minimum contrast. Hiroshima Mathematical Journal **22**, 631–647 (1992)

[3] Eguchi, S., Komori, O.: Minimum Divergence Methods in Statistical Machine Learning. Springer (2022)

[4] Hansen, L.P.: Large sample properties of generalized method of moments estimators. Econometrica **50**(4), 1029–1054 (1982)

[5] Lee, M.J.: Micro-Econometrics: Methods of Moments and Limited Dependent Variables. Second Edition. Springer (2010)

[6] Nagaoka, H.: The Fisher metric as a metric on the cotangent bundle. Information Geometry **7**(1), 651–677 (2024)

A new class of colored Gaussian graphical models with explicit normalizing constants

Bartosz Kołodziejek (Warsaw University of Technology)

Abstract

We study Bayesian structure learning in colored Gaussian graphical models (CGGMs), which fuse conditional-independence sparsity with symmetry constraints encoded by vertex- and edge-colored graphs. While CGGMs can substantially improve estimation and interpretability in high-dimensional regimes, Bayesian model selection has been hindered by intractable normalizing constants of the colored Diaconis–Ylvisaker (colored G-Wishart) priors.

We introduce a new subclass of CGGMs for which these gamma-type integrals admit explicit formula. Our approach begins with algebraic conditions on the space of colored precision matrices, leading to Block-Cholesky spaces, and proceeds to a combinatorial characterization of the corresponding colored graphs via 2-path regularity, yielding Color Elimination-Regular graphs. The resulting class strictly generalizes tractable decomposable models from the uncolored setting, encompasses all RCOP models built on decomposable graphs, and - under a single vertex color - reveals a close connection to Bose–Mesner algebras from association-scheme theory. We further develop an efficient methodology to compute the required structure constants, enabling closed-form normalizing constants within general Block-Cholesky-spaces. Collectively, these results deliver scalable posterior computation and maximum-a-posteriori selection for a broad family of CGGMs, expanding the range of high-dimensional problems in which symmetry can be exploited for sample-efficient and interpretable graphical modeling.

Probabilistic PCA on tensors

Piotr Zwiernik (Pompeu Fabra University)

Abstract

In probabilistic principal component analysis (PPCA), an observed vector is modeled as a linear transformation of a low-dimensional Gaussian factor plus isotropic noise. We generalize PPCA to tensors by constraining the loading operator to have Tucker structure, yielding a probabilistic multilinear PCA model that enables uncertainty quantification and naturally accommodates multiple, possibly heterogeneous, tensor observations. We develop the associated theory: we establish identifiability of the loadings and noise variance and show that-unlike in matrix PPCA-the maximum likelihood estimator (MLE) exists even from a single tensor sample. We then study two estimators. First, we consider the MLE and propose an expectation maximization (EM) algorithm to compute it. Second, exploiting that Tucker maps correspond to rank-one elements after a Kronecker lifting, we design a computationally efficient estimator for which we provide provable finite-sample guarantees. Together, these results provide a coherent probabilistic framework and practical algorithms for learning from tensor-valued data.

(Joint work with Yaoming Zhen)