Mathematical optimization and statistical theories using geometric methods

Date : October 20–21, 2022 (Japan Standard Time)
Venue : Academic Extension Center (Osaka Metropolitan University)
Contents : Workshop (Hybrid: physical/virtual)

- This workshop is held as a part of OCAMI Joint Usage/Research (JPMXP0619217849)
  "MEXT Joint Usage/Research Center on Mathematics and Theoretical Physics"
- This workshop is also supported by Japan Science and Technology Agency, CREST
  "Innovation of Deep Structured Models with Representation of Mathematical Intelligence" in "Creating information utilization platform by integrating mathematical and information sciences, and development to society"

Organizers: Hideto Nakashima (ISM: hideto (at) ism.ac.jp), Yoshihiko Konno (OMU), Hideyuki Ishi (OMU), Kenji Fukumizu (ISM)

Program

• October 20 (Thursday)

13:00–13:50 **Shoji Toyota** (SOKENDAI)
Invariance Learning based on Label Hierarchy

14:00–14:50 **Sho Sonoda** (RIKEN AIP)
Ridgelet Transforms for Neural Networks on Manifolds and Hilbert Spaces

15:00–15:50 **Tomonari Sei** (The University of Tokyo)
Ushio Tanaka (Osaka Metropolitan University)
Stein-type distributions on Riemannian manifolds

16:10–17:00 **Tomasz Skalski** (Wroclaw University of Science and Technology: LAREMA, University of Angers)
On LASSO and SLOPE estimators and their pattern recovery

17:10–18:00 **Carlos Améndola** (Technical University of Berlin)
Likelihood geometry of correlation models

• October 21 (Friday)

9:00– 9:50 **Piotr Zwiernik** (University of Toronto)
Mixed convex exponential families and locally associated graphical models

11:00–11:50 **Koichi Tojo** (RIKEN Center for Advanced Intelligence Project)
Classification problem of invariant q-exponential families on homogeneous spaces

13:50–14:40 **Yoshihiko Konno** (Osaka Metropolitan University)
Adaptive shrinkage of singular values for a low-rank matrix mean when a covariance matrix is unknown

14:50–15:40 **Satoshi Kuriki** (The Institute of Statistical Mathematics)
Expected Euler characteristic heuristic for smooth Gaussian random fields with inhomogeneous marginals

16:00–16:50 **Piotr Graczyk** (LAREMA, University of Angers)
Pattern recovery by SLOPE

Invariance Learning based on Label Hierarchy

Shoji Toyota
The Graduate University for Advanced Studies (SOKENDAI)

Training data used in machine learning may contain features that are spuriously correlated to the labels of data. Deep Neural Networks (DNNs) often learn such biased correlations embedded in training data and hence may fail to predict desired labels of test data generated by a different distribution from one to provide training data. To solve the problem, Invariance Learning (IL) is a rapidly developed approach to overcome the issue of biased correlation, which is caused by some bias in the distribution of a training dataset (e.g., [1]). IL estimates a predictor *invariant* to the change of distributions, aiming at keeping good performance in unseen distributions as well as in the training distributions.

While the IL approach has attracted much attention, requiring training data from multiple distributions may hinder wide applications in practice; preparing training data in many distributions often involves expensive data annotation.

To mitigate the problem of annotation cost, we propose a novel IL framework for the situation where the training data of target classification is given in only *one* distribution, while the task of higher *label hierarchy*, which needs lower annotation cost, has data from multiple distributions. The new IL framework significantly reduces the annotation cost in comparison with previous IL methods; we need exhausting annotation of original classes only for one distribution and just causaer labels for other distributions. Numerical simulations and theoretical analysis verify the effectiveness of our framework.

References

[1] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant Risk Minimization. *arXiv:1907.02893*, 2019.

Ridgelet Transforms for Neural Networks on Manifolds and Hilbert Spaces

Sho Sonoda
RIKEN AIP, Tokyo 103–0027 Japan
sho.sonoda@riken.jp

Abstract

To investigate how neural network parameters are organized and arranged, it is easier to study the distribution of parameters than to study the parameters in each neuron. The ridgelet transform is a pseudo-inverse operator (or an analysis operator) that maps a given function f to the parameter distribution γ so that a network

$$S[\gamma](\boldsymbol{x}) := \int_{\mathbb{R}^m \times \mathbb{R}} \gamma(\boldsymbol{a}, b)\sigma(\boldsymbol{a} \cdot \boldsymbol{x} - b)\mathrm{d}\boldsymbol{a}\mathrm{d}b, \quad \boldsymbol{x} \in \mathbb{R}^m$$

represents f, i.e., $S[\gamma] = f$. For depth-2 fully-connected networks on Euclidean space, the ridgelet transform has been discovered up to the closed-form expression, thus we could describe how the parameters are organized. However, for a variety of modern neural network architectures, the closed-form expression has not been known . Recently, our research group has developed a systematic scheme to derive ridgelet transforms for fully-connected layers on manifolds (noncompact symmetric spaces G/K) (Sonoda et al., 2022b) and for group convolution layers on abstract Hilbert spaces $\mathcal{H}$ (Sonoda et al., 2022a). In this talk, the speaker will explain a natural way to derive those ridgelet transforms.

References

S. Sonoda, I. Ishikawa, and M. Ikeda. Universality of Group Convolutional Neural Networks Based on Ridgelet Analysis on Groups. In *Advances in Neural Information Processing Systems 35*, 2022a.

S. Sonoda, I. Ishikawa, and M. Ikeda. Fully-Connected Network on Noncompact Symmetric Space and Ridgelet Transform based on Helgason-Fourier Analysis. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, 2022b.

Stein-type distributions on Riemannian manifolds

Tomonari Sei (The University of Tokyo)[*1]
Ushio Tanaka (Osaka Metropolitan University)[*2]

1. Stein-type distributions on the Euclidean space

Let $\mathcal{P}^2$ be the set of probability distributions μ on $\mathbb{R}^d$ with mean zero and finite second moments such that each marginal distribution μ_i $(i = 1, \dots, d)$ is absolutely continuous with respect to the Lebesgue measure $\mathrm{d}x_i$ on $\mathbb{R}$. We say that a probability distribution $\mu \in \mathcal{P}^2$ is Stein-type if it satisfies

$$\int f(x_i) \left(\sum_{j=1}^{d} x_j \right) \mathrm{d}\mu = \int f'(x_i) \mathrm{d}\mu, \quad i = 1, \dots, d,$$

for any absolutely continuous function $f : \mathbb{R} \to \mathbb{R}$ with bounded derivative f'.

Let $\mathcal{T}_{\mathrm{cw}}$ be the set of coordinate-wise transformations $T(x) = (T_1(x_1), \dots, T_d(x_d))$ such that each T_i is non-decreasing. In [2], it is shown that for any given $\mu_0 \in \mathcal{P}^2$, there exists $T \in \mathcal{T}_{\mathrm{cw}}$ such that $T_\sharp \mu_0$ is Stein-type. The transformation is characterized by a minimizer of a functional

$$F(\mu) = \sum_{i=1}^{d} \int \log \frac{\mathrm{d}\mu_i}{\mathrm{d}x_i} \mathrm{d}\mu_i + \int \frac{1}{2} \left(\sum_{i=1}^{d} x_i \right)^2 \mathrm{d}\mu,$$

over a fiber $\{T_\sharp \mu_0 \mid T \in \mathcal{T}_{\mathrm{cw}}\}$. The fiber is totally geodesic in the L^2-Wasserstein space and F is convex with respect to displacement interpolation. The optimal map T is applied to the problem of determining a general index in [2].

2. Generalization to manifolds

We generalize the Stein-type distributions to those on Riemannian manifolds. The space $\mathbb{R}^d$ is replaced with a product space $M = \prod_{i=1}^{d} M_i$, where each M_i is a Riemannian manifold. The space $\mathcal{P}^2$ of distributions is defined as well. Let $\mathcal{T}_{\mathrm{cw}}$ be the set of coordinate-wise transformations $T(x) = (T_1(x_1), \dots, T_d(x_d))$ such that each $T_i : M_i \to M_i$ is monotone. Here, T_i is said to be monotone if it is written as $T_i(x_i) = \exp_{x_i} \nabla \phi_i(x_i)$ with a cost convex function $\phi_i : M_i \to \mathbb{R}$ (see [1]). The Stein-type distribution is defined by a minimizer of a functional

$$F(\mu) = \sum_{i=1}^{d} \int \log \frac{\mathrm{d}\mu_i}{\mathrm{d}x_i} \mathrm{d}\mu_i + \int V(x) \mathrm{d}\mu,$$

over a fiber $\{T_\sharp \mu_0 \mid T \in \mathcal{T}_{\mathrm{cw}}\}$, where $V : M \to \mathbb{R}$ is a given function.

References

[1] McCann, R. J. (2001). Polar factorization of maps on Riemannian manifolds, *Geometric and Functional Analysis*, **11**, 589–608.

[2] Sei, T. (2022). Coordinate-wise transformation of probability distributions to achieve a Stein-type identity, *Information Geometry*, **5**, 325–354.

[*1] e-mail: sei@mist.i.u-tokyo.ac.jp
[*2] e-mail: utanaka@omu.ac.jp

On LASSO and SLOPE estimators and their pattern recovery

Tomasz Skalski[1,2]

[1]*Wrocław University of Science and Technology, Poland*
[2]*LAREMA, University of Angers, France*

Least Absolute Shrinkage and Selection Operator (LASSO) and Sorted ℓ_1 Penalized Estimator (SLOPE) are the regularization methods used for fitting high-dimensional regression models. They allow to reduce the model dimension by nullifying some of the regression coefficients. Moreover, SLOPE allows the further reduction by equalizing some of nonzero coefficients, which allows to identify situations where some of true regression coefficients are equal.
We shall introduce the notion of the pattern for LASSO and SLOPE and its subdifferential-induced generalization to other convex penalized estimators, which will be illustrated carefully in the case of the orthogonal design matrix. This talk will present new results on the strong consistency of SLOPE estimators and on the strong consistency of pattern recovery by SLOPE when the design matrix is orthogonal. We shall also present the relations of LASSO and SLOPE with root system induced convex hulls.

The research was supported by a French Government Scholarship and by Centre Henri Lebesgue, program ANR-11-LABX-0020-0.

References

[1] M. Bogdan, X. Dupuis, P. Graczyk, B. Kołodziejek, T. Skalski, P. Tardivel, M. Wilczyński. Pattern Recovery by SLOPE. ArXiv 2203.12086.

[2] U. Schneider, P. Tardivel. The geometry of uniqueness, sparsity and clustering in penalized estimation. ArXiv 2004.09106.

[3] T. Skalski, P. Graczyk, B. Kołodziejek, M. Wilczyński. Pattern recovery and signal denoising by SLOPE when the design matrix is orthogonal. ArXiv 2202.08573.

[4] P. Tardivel, T. Skalski, P. Graczyk, U. Schneider. The Geometry of Pattern Recovery by Penalized and Structured Estimators. 2021. hal-03262087.

Likelihood Geometry of Correlation Models

Carlos Améndola

Technical University of Berlin

We present a problem where algebra appears naturally when estimating correlation matrices, that is, standardized covariance matrices. Concretely, we study the geometry of maximum likelihood estimation for correlation matrices, which form an affine space of symmetric matrices defined by setting the diagonal entries to one.

We study the likelihood geometry for this model and linear submodels that encode additional symmetries. We also consider the problem of minimizing two closely related functions of the covariance matrix: the Stein's loss and the symmetrized Stein's loss. Unlike the Gaussian log-likelihood, these two functions are convex and hence admit a unique positive definite optimum.

Studying the critical points in all three settings leads to systems of nonlinear equations, and we compute some of the algebraic degree invariants that measure the algebraic complexity of each optimization problem.

This is joint work with Piotr Zwiernik (University of Toronto, Canada).

Mixed convex exponential families and locally associated graphical models

Piotr Zwiernik (University of Toronto)

Abstract

In statistical exponential families the log-likelihood forms a concave function in the canonical parameters. Therefore, any model given by convex constraints in these canonical parameters admits a unique maximum likelihood estimator (MLE). Such models are called convex exponential families. For models that are convex in the mean parameters (e.g. Gaussian covariance graph models) the maximum likelihood estimation is much more complicated and the likelihood function typically has many local optima. One solution is to replace the MLE with so called dual likelihood estimator, which is uniquely defined and asymptotically has the same distribution as the MLE. In this talk I will consider a much more general setting, where the model is given by convex constraints on some canonical parameters and convex constraints on the remaining mean parameters. We call such models mixed convex exponential families. We propose for these models a 2-step optimization procedure which relies on solving two convex problems. We show that the resulting estimator has asymptotically the same distribution as the MLE. Our work was motivated by locally associated Gaussian graphical models that form a suitable relaxation of Gaussian totally positive distributions.

(Joint work with Steffen Lauritzen, University of Copenhagen)

Classification problem of invariant q-exponential families on homogeneous spaces

Koichi Tojo

RIKEN Center for Advanced Intelligence Project

Abstract

Q-exponential family is a natural generalization of exponential family and is an important subject in the fields of information geometry and statistics. Widely used q-exponential families such as normal distributions and Cauchy distributions have a symmetry. More precisely, the sample space can be regarded as a homogeneous space G/H and the family of distributions on it is G-invariant with respect to the induced G-action by pushforward. Then the following problem naturally arises:

Classify G-invariant q-exponential families on G/H.

I would like to talk about a strategy to solve this problem using "q-deformation" of an exponential family. Moreover, we give a new $SL(2,\mathbb{R})$-invariant q-exponential family on the upper half plane.

This is a joint work with Taro Yoshino.

Adaptive shrinkage of singular values for a low-rank matrix mean when a covariance matrix is unknown

Yoshihiko Konno
Department of Mathematics, Osaka Metropolitan University

Assume that m, n, p are positive integers such that $\min\{m, n\} \geq p$ and that we observe a matrix $\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix}$ which is modeled as $\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Xi} \\ \boldsymbol{0}_{n\times p} \end{bmatrix} + \boldsymbol{E}$ where $\boldsymbol{\Xi}$ is an $m \times p$ non-random matrix(unknown and its rank may be less than $\min\{p, m\}$), $\boldsymbol{E}$ is an $(m + n) \times p$ error matrix(unobservable) whose rows are identically distributed as $N_p(\boldsymbol{0}_p, \boldsymbol{\Sigma})$, a p-variate real normal distribution with zero mean vector and covariance matrix $\boldsymbol{\Sigma}$. We assume that $\boldsymbol{\Sigma}$ is a $p \times p$ positive-definite and unknown matrix.

We consider the problem of estimating $\boldsymbol{\Xi}$ under a low-rank mean matrix condition, i.e.,

$$\operatorname{rank} \boldsymbol{\Xi} = r < p; \quad r \text{ is unknown}$$

under a loss function $\mathsf{L}(\widehat{\boldsymbol{\Xi}}, \boldsymbol{\Xi}|\, \boldsymbol{\Sigma}) = \operatorname{tr}\left\{(\widehat{\boldsymbol{\Xi}} - \boldsymbol{\Xi})^\top (\widehat{\boldsymbol{\Xi}} - \boldsymbol{\Xi})\boldsymbol{\Sigma}^{-1}\right\}$, where $\widehat{\boldsymbol{\Xi}} := \widehat{\boldsymbol{\Xi}}(\boldsymbol{X}, \boldsymbol{Y})$ is an estimator of $\boldsymbol{\Xi}$. Here $\boldsymbol{A}^\top$ and $\operatorname{tr}\boldsymbol{A}$ stand for the transpose and the trace of a square matrix $\boldsymbol{A}$. The risk function of $\mathsf{R}(\widehat{\boldsymbol{\Xi}}, \boldsymbol{\Xi}|\, \boldsymbol{\Sigma})$ is given by the expected value of the loss function where the expectation is taken with respect to the joint distribution of $(\boldsymbol{X}, \boldsymbol{Y})$.

We give Steins's unbiased risk estimate for estimators of the form

$$\widehat{\boldsymbol{\Xi}} = \left(\sum_{j=1}^{p} h_j(\ell_j)\boldsymbol{u}_j\boldsymbol{v}_j^\top\right)(\boldsymbol{Y}^\top\boldsymbol{Y})^{1/2}.$$

Here $h_j : [0, \infty) \to [0, \infty)$, $(j = 1, 2, \ldots, p)$ are absolutely continuous functions and $\boldsymbol{U}\boldsymbol{L}\boldsymbol{V}^\top$ is the singular value decomposition of $\boldsymbol{X}(\boldsymbol{Y}^\top\boldsymbol{Y})^{-1/2}$ where $\boldsymbol{U} = (\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_p)$ is an $m \times p$ matrix such that $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_p$ (the $p \times p$ identity matrix), $\boldsymbol{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p)$ is a $p \times p$ orthogonal matrix, and $\boldsymbol{L}$ is a $p \times p$ diagonal matrix whose j-th diagonal element is given by ℓ_j. Note that we may assume that $\ell_1 > \ell_2 > \cdots > \ell_p > 0$ (almost everywhere) with out loss of generality. Based on SURE formula, we propose an adaptive soft-theshholding rule to the singular values $\ell_1, \ell_2, \ldots, \ell_p$. Furthermore, the results above are extended to the complex normal distribution setup.

Expected Euler characteristic heuristic for smooth Gaussian random fields with inhomogeneous marginals

Satoshi Kuriki
The Institute of Statistical Mathematics
10-3 Midoricho, Tachikawa, Tokyo 190-8562, Japan
kuriki@ism.ac.jp

Abstract

Expected Euler characteristic (EC) heuristic is a method for approximating the tail probability of the maximum of a Gaussian random field. In this talk, we provide an expected Euler characteristic formula for the approximate tail probability and its relative approximation error when the index set M is a closed manifold and the mean and variance of the marginal distribution are not necessarily constant. When the variance is constant, [TTA05] proved that the relative approximation error is exponentially small in a general setting where the index set M is a stratified manifold. When the variance is not constant, it is shown that only the subset $M_{\rm supp}$ of M, referred to as the supporting index set, contributes to the maximum tail probability. The proposed tail probability formula is an integral of the Euler characteristic density over $M_{\rm supp}$, and its relative approximation error is proven to be exponentially small as in the case of constant variance. These results are generalizations of [KTT22], who addressed a restricted case of finite Karhunen-Loève expansion by the volume-of-tube method. As an example, the tail probability formula for the largest eigenvalues of noncentral Wishart matrices $\mathcal{W}_p(\nu, \Sigma; \Phi)$ and its relative approximation error are obtained. Numerical experience supports the high accuracy of the expected Euler characteristic formulas regardless of whether the marginals are homogeneous or inhomogeneous.

Keywords: Borel's inequality, Kac-Rice formula, noncentral Wishart distribution, volume-of-tube method, Weyl's tube formula.

References

[KTT22] Satoshi Kuriki, Akimichi Takemura, and Jonathan E. Taylor, *The volume-of-tube method for gaussian random fields with inhomogeneous variance*, Journal of Multivariate Analysis **188** (2022), 104819.

[TTA05] Jonathan E. Taylor, Akimichi Takemura, and Robert J. Adler, *Validity of the expected Euler characteristic heuristic*, Ann. Probab. **33** (2005), no. 4, 1362–1396.

PATTERN RECOVERY BY SLOPE

PIOTR GRACZYK

Abstract

I will present recent results obtained in [1] jointly with M. Bogdan, X. Dupuis, B. Kołodziejek, T. Skalski, P. Tardivel and M. Wilczyński.

SLOPE is a popular method for dimensionality reduction in the high-dimensional regression. Indeed, some regression coefficient estimates of SLOPE can be null (sparsity) or can be equal in absolute value (clustering). Consequently, SLOPE may eliminate irrelevant predictors and may identify groups of predictors having the same influence on the vector of responses.

The notion of SLOPE pattern allows to derive theoretical properties on sparsity and clustering by SLOPE. Specifically, the SLOPE pattern of a vector provides: the sign of its components (positive, negative or null), the clusters (indices of components equal in absolute value) and clusters ranking.

In this research we give a necessary and sufficient condition for SLOPE pattern recovery of an unknown vector of regression coefficients.

References

[1] M. Bogdan, X. Dupuis, P. Graczyk, B. Kołodziejek, T. Skalski, P. Tardivel, M. Wilczyński, *Pattern recovery by SLOPE*(2022), arXiv:2203.12086.

Université d'Angers, CNRS, LAREMA, SFR MATHSTIC, F-49000 Angers, France

Email address: graczyk@univ-angers.fr