Osaka Central Advanced Mathematical Institute (OCAMI)
Osaka Metropolitan University
MEXT Joint Usage/Research Center on Mathematics and Theoretical Physics

OCAMI Reports Vol. 8 (2022)
doi: 10.24544/ocu.20221208-007

Mathematical optimization and statistical theories using geometric methods

Organized by
Hideto Nakashima
Yoshihiko Konno
Hideyuki Ishi
Kenji Fukumizu

October 20–21, 2022

Abstract

This workshop was held on October 20–21, 2022 in order to connect researchers in several fields, in particular Statistics, Machine Learning and Mathematics, and to share problems and researches in these fields interdisciplinary.

2020 *Mathematics Subject Classification.*
20G05, 22F30, 43A85, 60E05, 62E10,
62H12, 62J05, 62J07, 62R01

Key words and Phrases.
Algebraic statistics, LASSO, SLOPE, exponential families,
machine learning, geometric analysis

OCAMI. Mathematical optimization and statistical theories using geometric methods. OCAMI Reports. Vol. 8, Osaka Central Advanced Mathematical Institute, Osaka Metropolitan University. 2022, 165 pp.
doi: 10.24544/ocu.20221208-007

Preface

This is a proceedings of the international workshop "Mathematical optimization and statistical theories using geometric methods" held from October 20th to October 21st in 2022. This workshop aimed to connect researchers in several fields, in particular Statistics, Machine Learning and Mathematics, and to share problems and researches in these fields interdisciplinary.

This workshop was supported by Osaka Metropolitan University, Advanced Mathematical Institute MEXT Joint Usage/Research Center on Mathematics and Theoretical Physics, and 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."

This workshop was held in a hybrid format. Domestic speakers are gathered in Academic Extension Center (Osaka Metropolitan University), Foreign speakers participated by Zoom. We had 10 talks, 6 of which were from Japan and the others were from abroad, and 26 people had been registered in this workshop.

Organizers

Hideto Nakashima

Research Center for Statistical Machine Learning, The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan

Email address: `hideto@ism.ac.jp`

Yoshihiko Konno

Department of Mathematics, Osaka Metropolitan University, 1-1, Gakunen-cho, Naka-ku, Sakai-shi, 599-8531

Email address: `konno@omu.ac.jp`

Hideyuki Ishi

Department of Mathematics, Osaka Metropolitan University, 3-3-138, Sugimoto, Sumiyoshi-ku, Osaka, 558-8585, Japan

Email address: `hideyuki-ishi@omu.ac.jp`

Kenji Fukumizu

Research Center for Statistical Machine Learning, The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan

Email address: `fukumizu@ism.ac.jp`

Contents

[1]Slides are not included in this report.

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.

Invariance Learning based on Label Hierarchy

Shoji Toyota
The Graduate University for Advanced Studies
(Joint work with Prof. Kenji Fukumizu)

OCAMI workshop, 20 ~ 21, October, 2022

※ The presentation is based on https://arxiv.org/abs/2203.15549. To appear in Neurips 2022.

Agenda

- Background
- Mathemathical Formulation
- Method
- Theory
- Experiment

Agenda

Background

Mathemathical Formulation

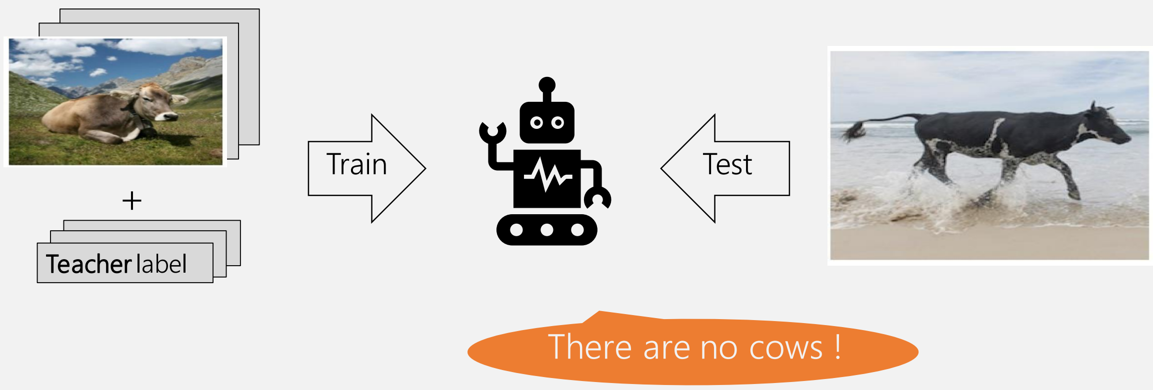
Method

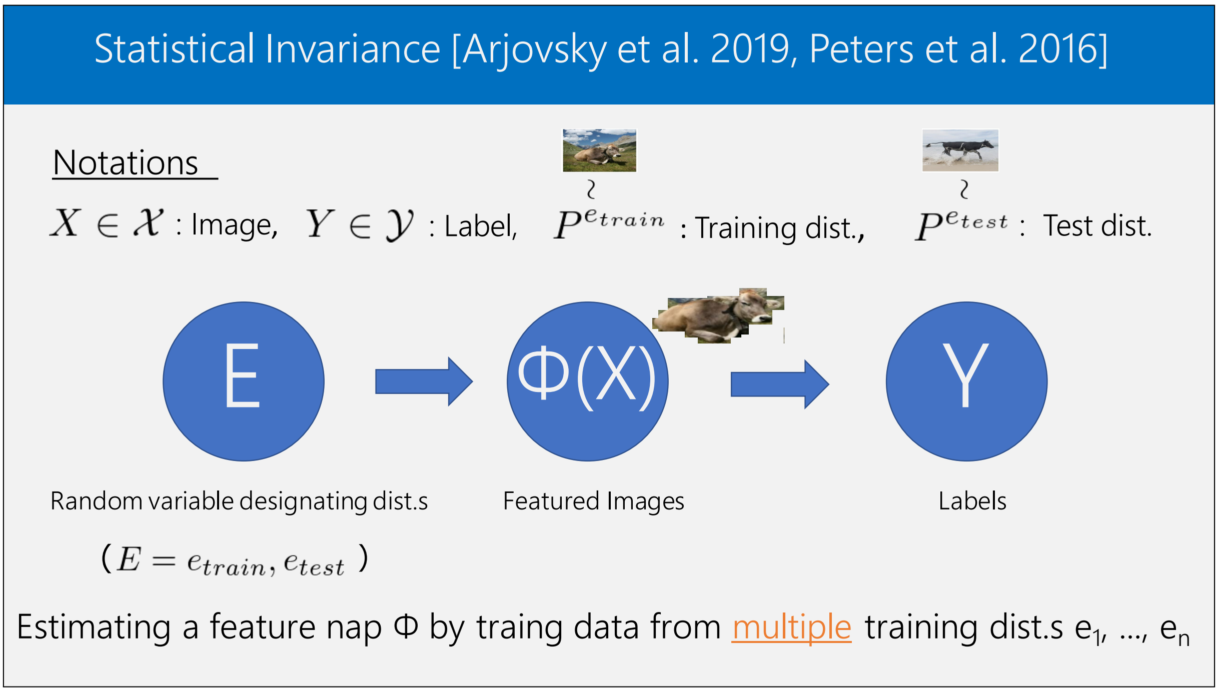
Theory

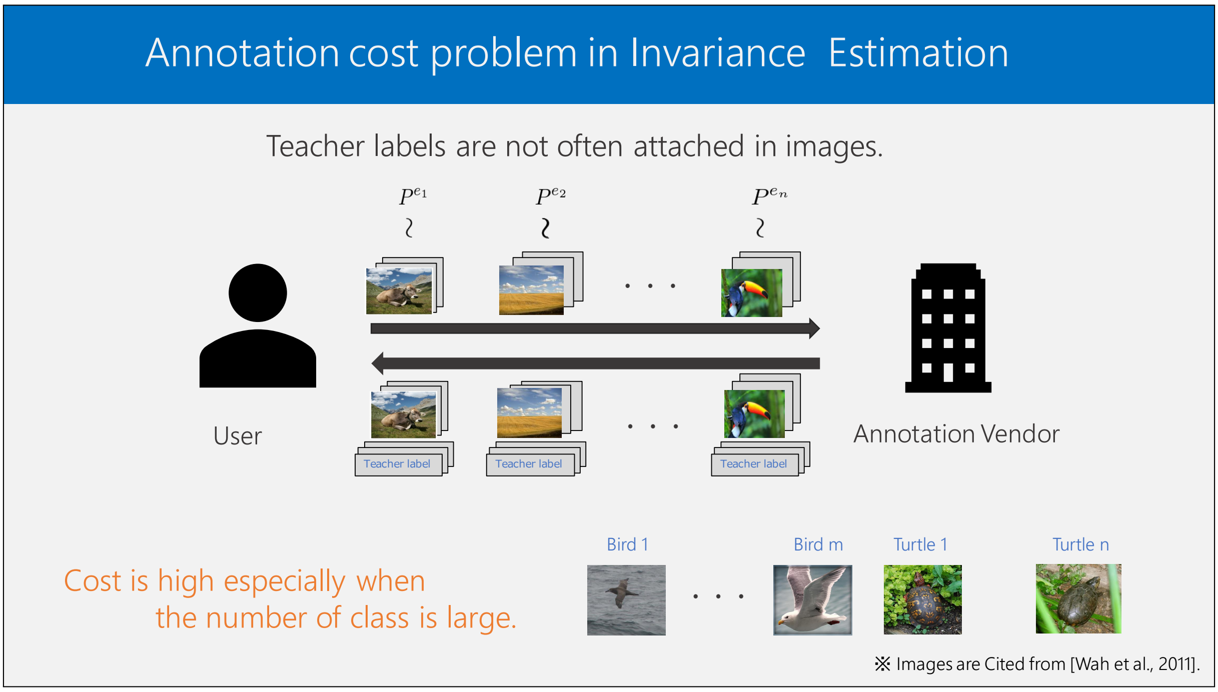
Experiment

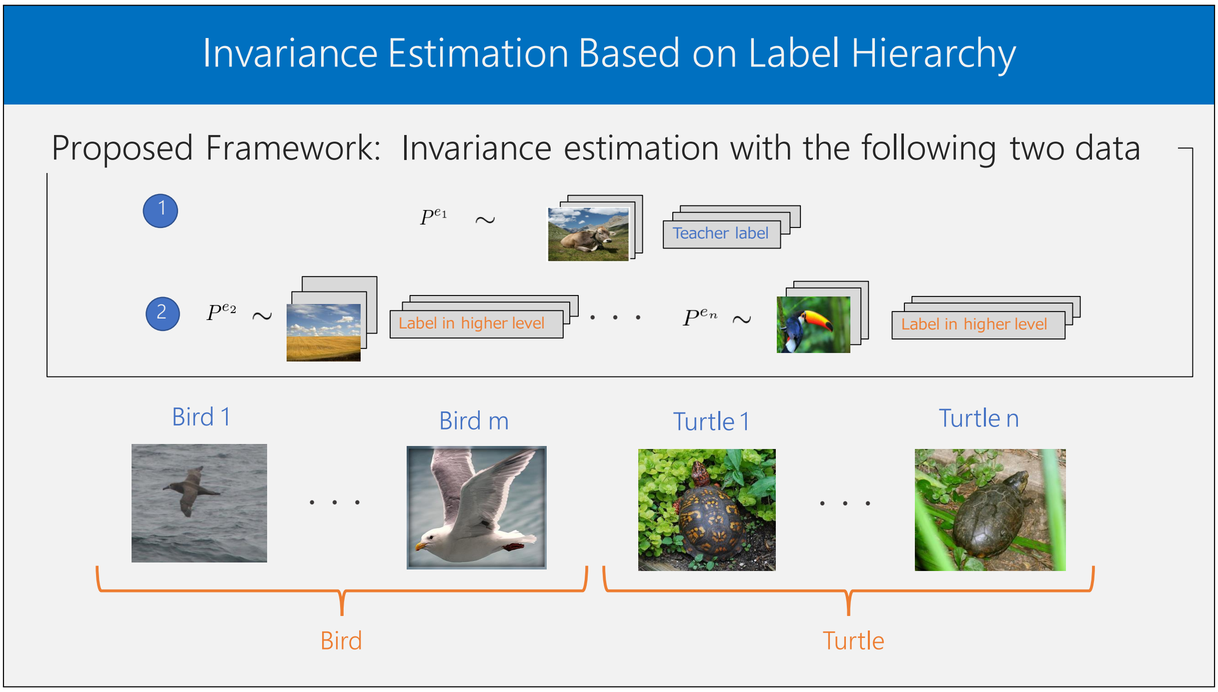
Recent Problem in Machine Learning :
Estimators inherit spurious correlation in training data

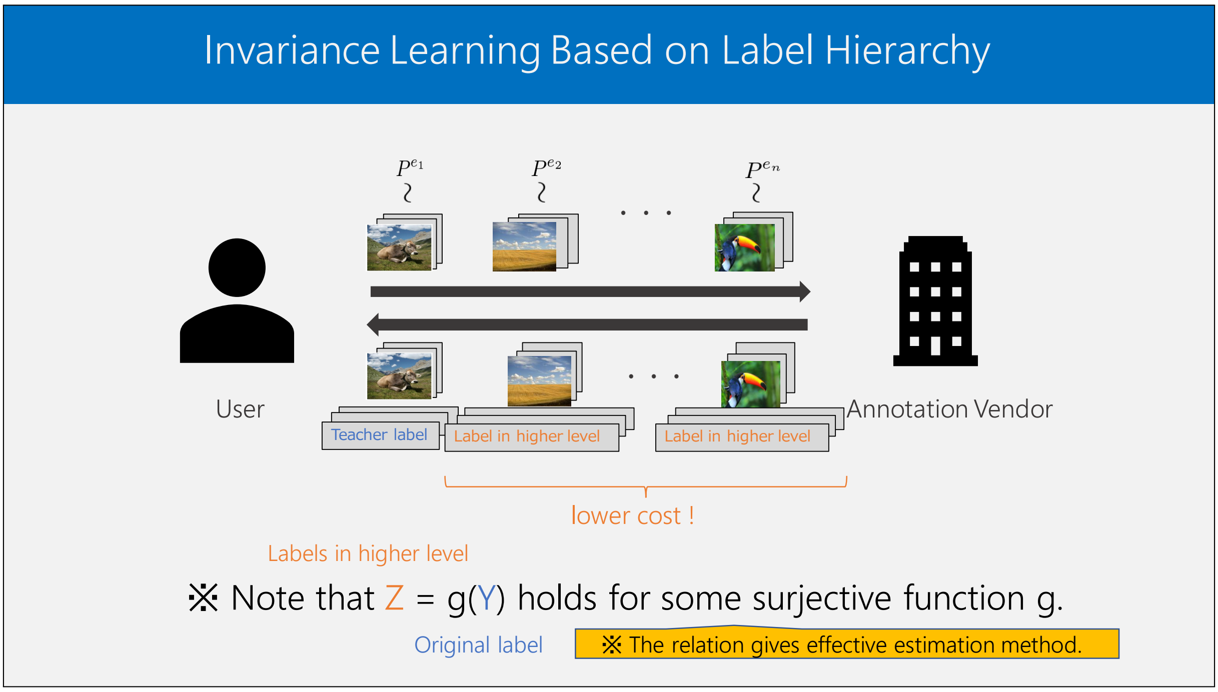
[S. Beery et al. 2018]




Statistical Invariance [Arjovsky et al. 2019, Peters et al. 2016]
Notations
$X \in \mathcal{X}$: Image, $Y \in \mathcal{Y}$: Label, $P^{e_{train}}$: Training dist., $P^{e_{test}}$: Test dist.
E
Φ(X)
Y
Random variable designating dist.s
$(E = e_{train}, e_{test})$
Featured Images
Labels
Estimating a feature nap Φ by traing data from multiple training dist.s $e_1, \ldots, e_n$


Annotation cost problem in Invariance Estimation
Teacher labels are not often attached in images.
P^{e_1}
P^{e_2}
P^{e_n}
User
Teacher label
Teacher label
Teacher label
Annotation Vendor
Cost is high especially when the number of class is large.
Bird 1
Bird m
Turtle 1
Turtle n
※ Images are Cited from [Wah et al., 2011].


Invariance Estimation Based on Label Hierarchy
Proposed Framework: Invariance estimation with the following two data
1
P^{e_1} ∼
Teacher label
2
P^{e_2} ∼
Label in higher level
P^{e_n} ∼
Label in higher level
Bird 1
Bird m
Turtle 1
Turtle n
Bird
Turtle


Invariance Learning Based on Label Hierarchy
P^{e_1}
P^{e_2}
P^{e_n}
User
Annotation Vendor
Teacher label
Label in higher level
Label in higher level
lower cost !
Labels in higher level
※ Note that Z = g(Y) holds for some surjective function g.
Original label
※ The relation gives effective estimation method.

Agenda

- Background
- Mathemathical Formulation
- Method
- Theory
- Experiment

Mathemathical Formulation

$e \in \mathcal{E}$: Index designating a dist. $(X^e, Y^e) \in \mathcal{X} \times \mathcal{Y}$: Image and label on e. $(X^e, Y^e) \sim P^e$

Available samples

1. $\mathcal{D}^{e_1} := \{(X_i^{e_1}, Y_i^{e_1})\}_i \sim P^{e_1}$ for $e_1 \in \mathcal{E}$
2. $\mathcal{D}^{e} := \{(X_i^{e}, Z_i^{e})\}_i \sim P^{e}$ for $\forall e \in \mathcal{E}_{high} (\subset \mathcal{E})$

(=g(Y$_i^e$)): labels in higher hierarchy

Goal: out-of-distribution (o.o.d.) risk Minimization

$$f^{o.o.d.} := \mathrm{argmin}_{f:\mathcal{X}\to\mathcal{Y}} \underbrace{\max_{e\in\mathcal{E}} \mathcal{R}^e(f)}_{\text{Out-of-distribution (o.o.d.) risk}}$$

$\mathcal{R}^e(f) := \int l(Y, f(X)) dP^e(X,Y)$: Risk on $e \in \mathcal{E}$

Assumption: Assume that $\exists \Phi : \mathcal{X} \to \mathcal{H}$, s.t. $P(Y^{e_1} | \Phi(X^{e_1})) = P(Y^e | \Phi(X^e)) (\forall e_1, e_2 \in \mathcal{E})$.

Agenda

- Background
- Mathemathical Formulation
- Method
- Theory
- Experiment

Estimation

Esrimaiton object:

1. Feature map Φ which satisfies E → Φ(X) → Y

   ☹ We can not estimate it by data on a single domain...

2. classifier w predicting a label Y from Φ(X)

Estimation

Esrimaiton object: 1. Feature map Φ which satisfies $E \to \Phi(X) \cancel{\to} Y$
Z
Labels in higher level

2. classifier w predicting a label Y from Φ(X)

Method: We estimate Φ and w stemiously,
by minimizing the following objective function.

$$\underbrace{\hat{O}_\lambda(w,\Phi) := \frac{1}{|\mathcal{D}^{e_1}|}\sum_{(x,y)\in\mathcal{D}^{e_1}} l(y, w\circ\Phi(x))}_{\text{estimating } w \text{ : evaluated by original label data}} + \lambda \underbrace{\left(\text{Dependence measure of } E \to \Phi(X) \to Z\right)}_{\text{estimating } \Phi\text{: evaluated by higher label data}}$$

※ second term: $\sum_{e\in\mathcal{E}_{ad}} \|\frac{1}{|\mathcal{D}^e|}\nabla_{\hat{w}=w} l(g(y), \hat{w}\circ\Phi(x))\|^2$ [M. Arjovsky et al. 2019].

Difficulty of Hyperparameter selection

[Galrajani et al. 2021]

If we select λ by a naive CV method using training data,
famous methods result in random guess classifiers....

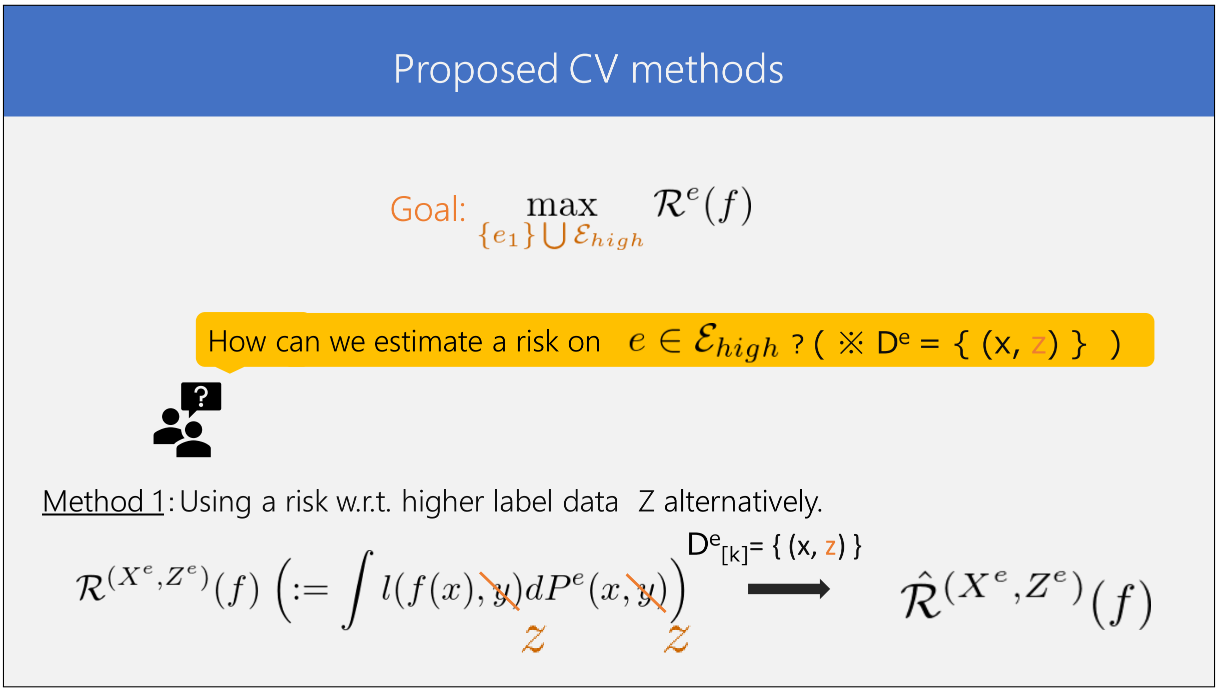
$D_{[-k]}$, $D_{[k]}$: Training and Validation, $f_\lambda \overset{d}{\leftarrow} \hat{f}_\lambda$

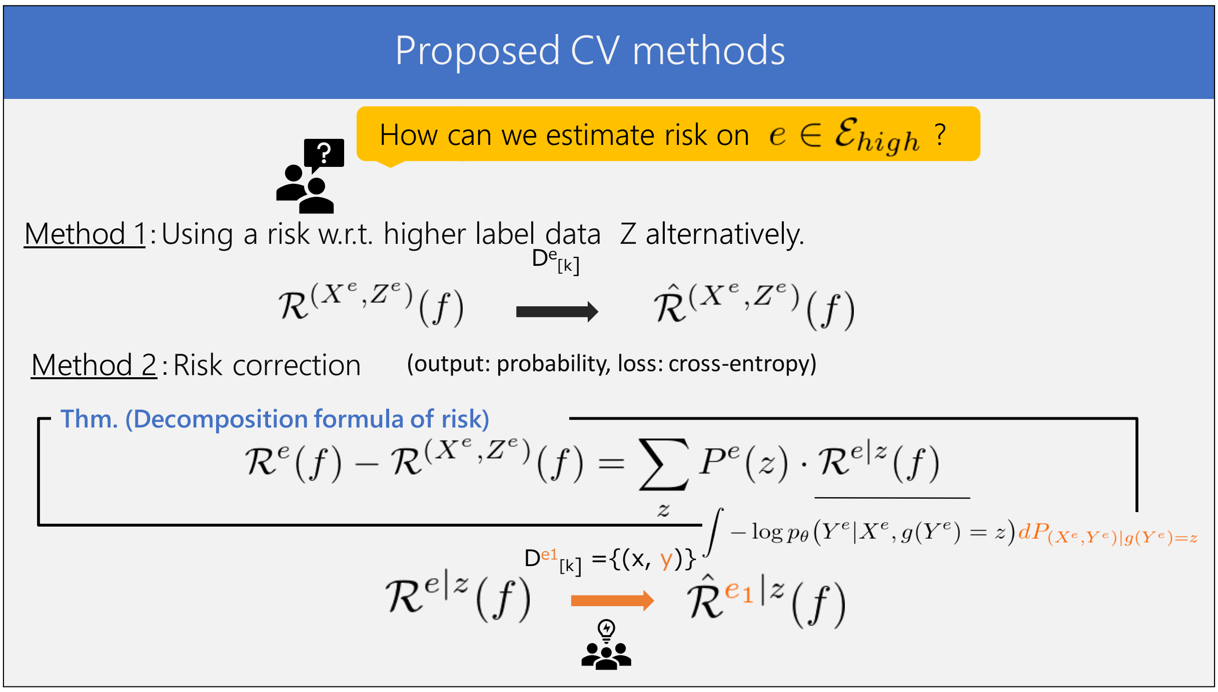
- Cross-Validation (CV) for minimizing an o.o.d. risk $\max_e R(f)$

Goal: $\mathrm{argmin}_\lambda \max_e R^e(f_\lambda)$

$f_\lambda \xrightarrow{D_{[-k]}} \hat{f}_\lambda$ ○ $\max_e R^e(f) \xrightarrow{D_{[k]}} \max_e \hat{R}^e(f)$ ×

Difficulity: o.o.d. risk estimation from validation data


Proposed CV methods
Goal: $\max_{\{e_1\} \bigcup \mathcal{E}_{high}} \mathcal{R}^e(f)$
How can we estimate a risk on $e \in \mathcal{E}_{high}$? (※ D^e = { (x, z) })
Method 1: Using a risk w.r.t. higher label data Z alternatively.
$\mathcal{R}^{(X^e,Z^e)}(f) \left(:= \int l(f(x), \cancel{y}_{z})dP^e(x, \cancel{y}_{z})\right) \xrightarrow{D^e_{[k]} = \{(x, z)\}} \hat{\mathcal{R}}^{(X^e,Z^e)}(f)$


Proposed CV methods
How can we estimate risk on $e \in \mathcal{E}_{high}$?
Method 1: Using a risk w.r.t. higher label data Z alternatively.
$\mathcal{R}^{(X^e,Z^e)}(f) \xrightarrow{D^e_{[k]}} \hat{\mathcal{R}}^{(X^e,Z^e)}(f)$
Method 2: Risk correction (output: probability, loss: cross-entropy)
Thm. (Decomposition formula of risk)
$\mathcal{R}^e(f) - \mathcal{R}^{(X^e,Z^e)}(f) = \sum_z P^e(z) \cdot \mathcal{R}^{e|z}(f)$
$\int -\log p_\theta(Y^e|X^e, g(Y^e) = z)dP_{(X^e,Y^e)|g(Y^e)=z}$
$\mathcal{R}^{e|z}(f) \xrightarrow{D^{e1}_{[k]} = \{(x, y)\}} \hat{\mathcal{R}}^{e_1|z}(f)$

Agenda

- Background
- Mathemathical Formulation
- Method
- Theory
- Experiment

Thoretical analysis of CV methods

$f_\lambda \overset{d}{\leftarrow} \hat{f}_\lambda$ (※ There are some open problems.)

$\mathcal{R}^1(f), \mathcal{R}^2(f)$: Approximations of an o.o.d. risk by Method I and II (ignoring estimation).

($\mathcal{R}^1(f) := \max\{\mathcal{R}^{e_1}(f), \max_{e \in \mathcal{E}_{high}} \mathcal{R}^{(X^e, Z^e)}(f)\}$

$\mathcal{R}^2(f) := \max\Big\{\mathcal{R}^{e_1}(f), \max_{e \in \mathcal{E}_{high}} \{\mathcal{R}^{(X^e, Z^e)}(f) + P^e(z) \cdot \mathcal{R}^{e_1|z}(f)\}\Big\}$)

$$\underbrace{\mathrm{argmin}_\lambda \mathcal{R}^1(f_\lambda)}_{\text{Hyperparameter selected by method I}} \subset \underbrace{\mathrm{argmin}_\lambda \max_{e \in \mathcal{E}} \mathcal{R}^e(f_\lambda)}_{\text{Optimal hyperparameter}}$$

$$\underbrace{\mathrm{argmin}_\lambda \mathcal{R}^2(f_\lambda)}_{\text{Hyperparameter selected by method II}} \subset \underbrace{\mathrm{argmin}_\lambda \max_{e \in \mathcal{E}} \mathcal{R}^e(f_\lambda)}_{\text{Optimal hyperparameter}}$$

The inclusions represent the success of CVs (with ignoring estimations).

Thoretical analysis of CV methods

$\{f_\theta\}_{\theta \in \Theta}$: all m'ble funct.

$\{(X^e, Y^e)\}_{e \in \mathcal{E}} := \{(X, Y) | P_{X_1, Y} = P_{X_1^I, Y^I}\}$: [Rojas-Carulla et al. 2018]

Correctness of Method 1 (Simplified)

(C1) for any λ with $\mathrm{Im}\Phi_2^\lambda \neq \emptyset$, there is $e_\lambda \in \mathcal{E}_{ad}$ such that

$(x, z) \sim P_{X^{e_\lambda}, g(Y^{e_\lambda})}$ satisfies $p^{e^*}(z|\Phi^\lambda(x)) \leq e^{-\beta} - \epsilon$ holds with probability 1.

$$\mathrm{argmin}_\lambda \mathcal{R}^1(f_\lambda) \subset \mathrm{argmin}_\lambda \max_{e \in \mathcal{E}} \mathcal{R}^e(f_\lambda)$$

$\beta := H(Y^{e_1}|X_1^{e_1})$

Correctness of Method 1 (Simplified)

(C2) for any λ with $\mathrm{Im}\Phi_2^\lambda \neq \emptyset$, there is $e_\lambda \in \mathcal{E}_{ad}$ such that

$(x, z) \sim P_{X^{e_\lambda}, g(Y^{e_\lambda})}$ satisfies $p^{e^*}(z|\Phi^\lambda(x)) \leq e^{-\beta_\lambda} - \epsilon$ holds with probability 1.

$$\mathrm{argmin}_\lambda \mathcal{R}^2(f_\lambda) \subset \mathrm{argmin}_\lambda \max_{e \in \mathcal{E}} \mathcal{R}^e(f_\lambda)$$

$\beta_\lambda := H(Y^{e_1}|X_1^{e_1}) - \sum_z P^e(z) \cdot \mathcal{R}^{e_1|z}(f_\lambda)$

$\beta_\lambda \leq \beta$: Method II is more applicable !

Agenda

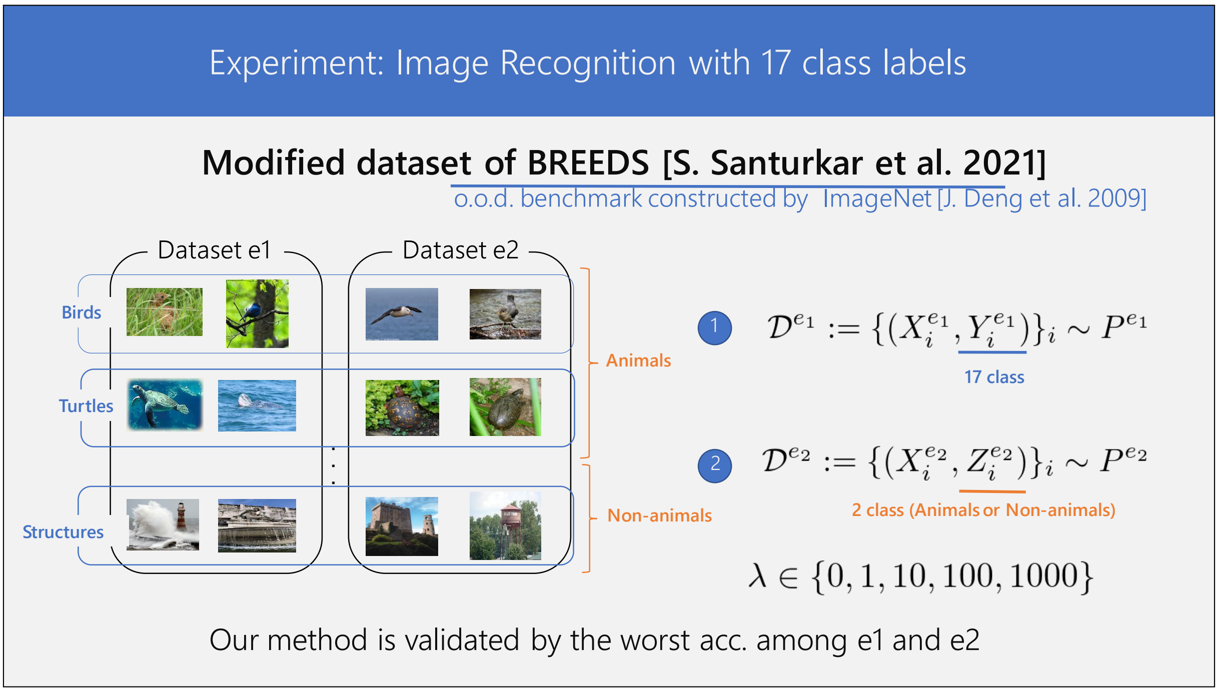
Background

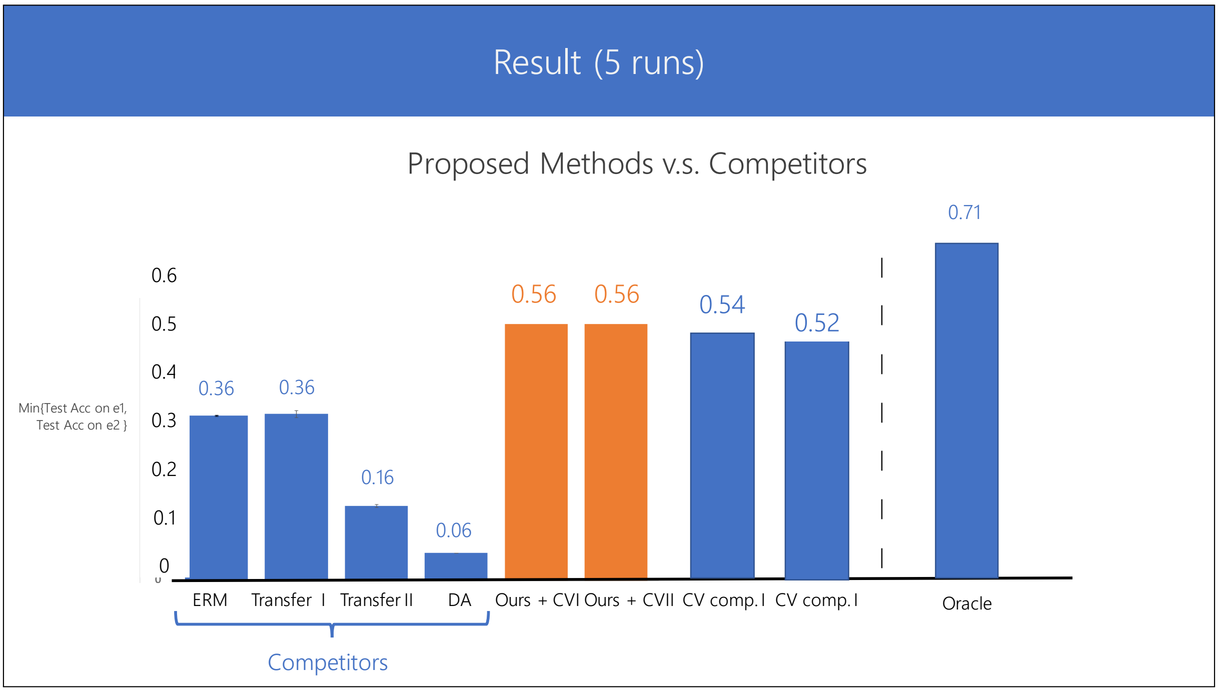
Mathemathical Formulation

Method

Theory

Experiment


Experiment: Image Recognition with 17 class labels
Modified dataset of BREEDS [S. Santurkar et al. 2021]
o.o.d. benchmark constructed by ImageNet [J. Deng et al. 2009]
Dataset e1
Dataset e2
Birds
Turtles
Structures
Animals
Non-animals
1 $\mathcal{D}^{e_1} := \{(X_i^{e_1}, Y_i^{e_1})\}_i \sim P^{e_1}$
17 class
2 $\mathcal{D}^{e_2} := \{(X_i^{e_2}, Z_i^{e_2})\}_i \sim P^{e_2}$
2 class (Animals or Non-animals)
$\lambda \in \{0, 1, 10, 100, 1000\}$
Our method is validated by the worst acc. among e1 and e2


Result (5 runs)
Proposed Methods v.s. Competitors
Min{Test Acc on e1, Test Acc on e2 }
0.6
0.5
0.4
0.3
0.2
0.1
0
0.36
0.36
0.16
0.06
0.56
0.56
0.54
0.52
0.71
ERM
Transfer I
Transfer II
DA
Ours + CVI
Ours + CVII
CV comp. I
CV comp. I
Oracle
Competitors

References

1. S. Beery et al. Recognition in terra in cognita. In CCV, 2018.

2. M. Arjovsky et al., Invariant Risk Minimization, arXiv:1907.02893, 2019.

3. J. Peters et al. Causal inference using invariant prediction: identification and confidence intervals, JRSS-B, 2016.

4. M. Rojas-Carulla et al. Invariant models for causal transfer learning. JMLR, 2018

5. S. Santurkar et al. Breeds: Benchmarks for subpopulation shift. In ICLR, 2021.

6. J. Deng et al. Imagenet: A large-scale hierarchical image database. In CVPR, 2009.

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.

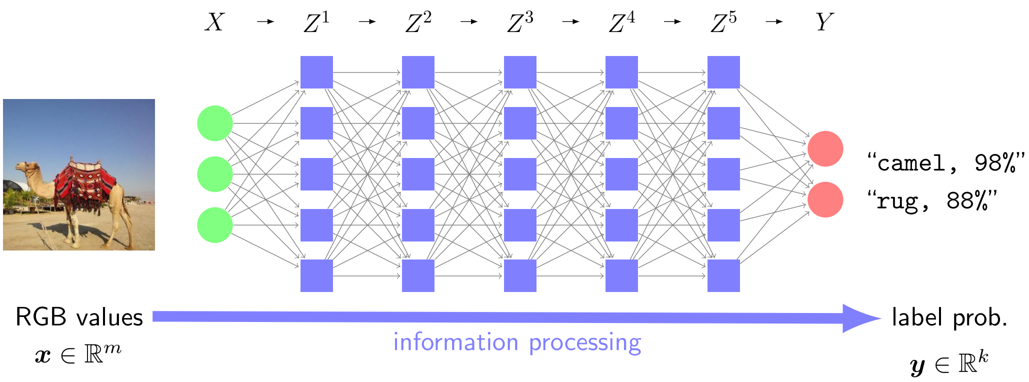
The Ridgelet Transforms of Neural Networks on Manifolds and Hilbert Spaces

Sho Sonoda

Research Scientist
RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan

Mathematical Optimization and Statistical Theories Using Geometric Methods
Osaka Metropolitan University
October 20-21, 2022

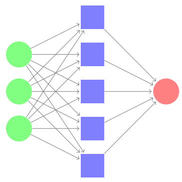
Q. What is a typical solution obtained by deep learning?



- Want to identify what solution is typically acquired via deep learning
- Want to know why (and when) deep learning performs better (than shallow networks)

Reparametrization

Finite-width (Discrete, or "Ordinary") NN

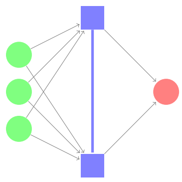


- $\mathtt{SNN}(\boldsymbol{x};\theta_d) = \sum_{i=1}^{d} c_i \sigma(\boldsymbol{a}_i \cdot \boldsymbol{x} - b_i)$
- *nonlinear* parameters: $\theta_d = \{(\boldsymbol{a}_i, b_i, c_i)\}_{i=1}^{d} \in \mathbb{R}^{(m+2)d}$

continuum limit ↓ ↑ discretization

$\gamma_d = \sum_{i=1}^{d} c_i \delta_{(\boldsymbol{a}_i, b_i)}$

Infinite-width (Continuous, or Integral Representation of) NN



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

3 / 20

Definition (Ridgelet Transform)

For any function $f : \mathbb{R}^m \to \mathbb{C}$ and $\rho : \mathbb{R} \to \mathbb{C}$, put

$$R[f;\rho](\boldsymbol{a}, b) = \int_{\mathbb{R}^m} f(\boldsymbol{x})\overline{\rho(\boldsymbol{a} \cdot \boldsymbol{x} - b)}\mathrm{d}\boldsymbol{x}, \quad (\boldsymbol{a}, b) \in \mathbb{R}^m \times \mathbb{R}.$$

Theorem (Reconstruction Formula)

For any $\sigma \in \mathcal{S}'(\mathbb{R}), \rho \in \mathcal{S}(\mathbb{R})$ and $f \in L^2(\mathbb{R}^m)$, we have

$$S[R[f;\rho]](\boldsymbol{x}) = \int R[f;\rho](\boldsymbol{a}, b)\sigma(\boldsymbol{a} \cdot \boldsymbol{x} - b)\mathrm{d}\boldsymbol{a}\mathrm{d}b = (\!(\sigma, \rho)\!) f(\boldsymbol{x}),$$

where $(\!(\sigma, \rho)\!) = (2\pi)^{m-1} \int_{\mathbb{R}} \sigma^\sharp(\omega)\overline{\rho^\sharp(\omega)}|\omega|^{-m}\mathrm{d}\omega$ *and* $\sharp$ *denotes the Fourier transform*

- Meaning 1: Continuous NN is a universal approximator
- Meaning 2: R and S play the same role as Fourier F and inverse Fourier F^{-1} transforms:

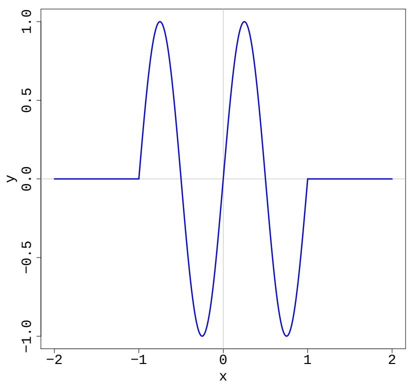
$$F^{-1}[F[f]](\boldsymbol{x}) = (2\pi)^{-m} \int_{\mathbb{R}^m} F[f](\boldsymbol{\xi})e^{i\boldsymbol{x}\cdot\boldsymbol{\xi}}\mathrm{d}\boldsymbol{\xi} = f(\boldsymbol{x})$$

- Independently "discovered" by Murata (1996), Candès (1998), and Rubin (1998)

4 / 20

Numerical example of ridgelet transform $R[f;\rho](a,b)$

- $f(x) = \sin(2\pi x)\,\mathbf{1}_{[-1,1]}(x)$
- $R[f;\rho](a,b) = \int_{\mathbb{R}} f(x)\rho(ax-b)\mathrm{d}x \approx \sum_i \sin(2\pi x_i)\rho(ax_i - b)\Delta x$
- $\sigma(b) = \tanh(b)$
- $\rho(b) = H[\rho_0^{(2)}](b)$ with $\rho_0(b) := \exp(-b^2/2)$, Hilbert transform H



data $f(x)$

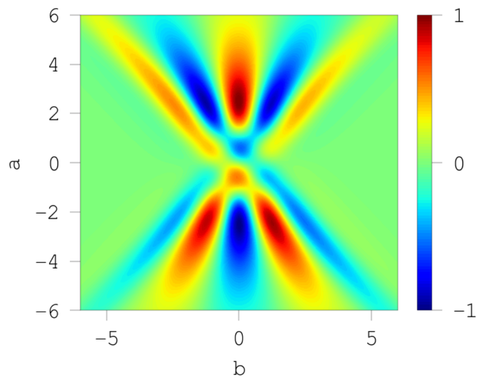
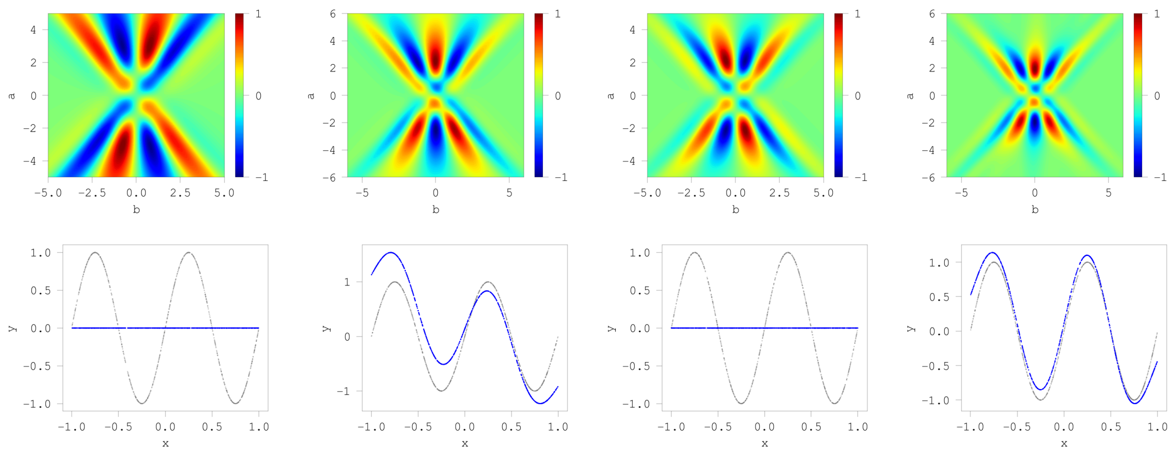


image $R[f;\rho](a,b)$

Visualization results of reconstruction formula $S[R[f;\rho]] = ((\sigma,\rho))f$

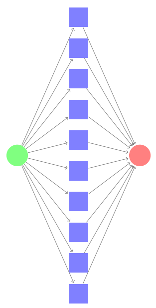
$\rho_1 = H\rho_0^{(1)}$	$\rho_2 = H\rho_0^{(2)}$	$\rho_3 = H\rho_0^{(3)}$	$\rho_4 = H\rho_0^{(4)}$
$((\sigma,\rho_1)) = 0$	$((\sigma,\rho_2)) \neq 0$	$((\sigma,\rho_3)) = 0$	$((\sigma,\rho_4)) \neq 0$

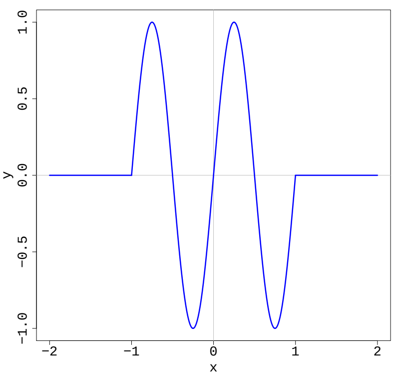


How the parameter distribution looks like?

We will train many ($n = 1,000$) neural networks $\mathtt{SNN}(x;\theta_d) = \sum_{j=1}^{d} c_j \sigma(a_j \cdot x - b_j)$ with $d = 10$ hidden units, and see the distribution of trained parameters (a_j, b_j, c_j).

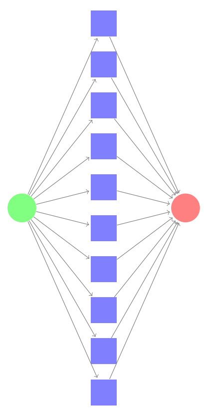
- Data generating function: $f(x) = \sin(2\pi x)\,\mathbf{1}_{[-1,1]}(x)$
- $\sigma(z) = \tanh(z)$
- SGD w. weight decay

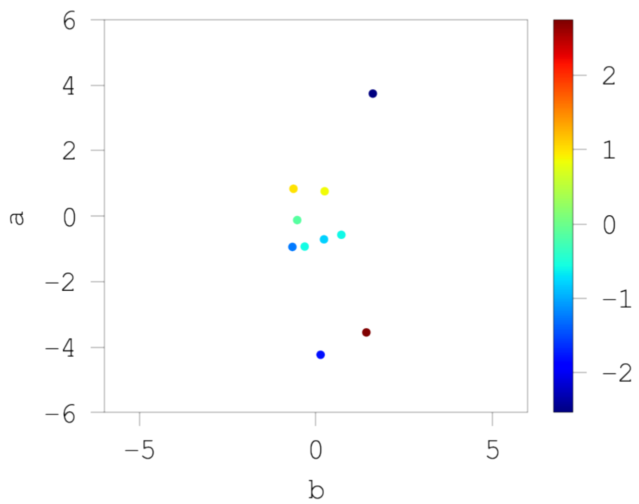




data $f(x)$

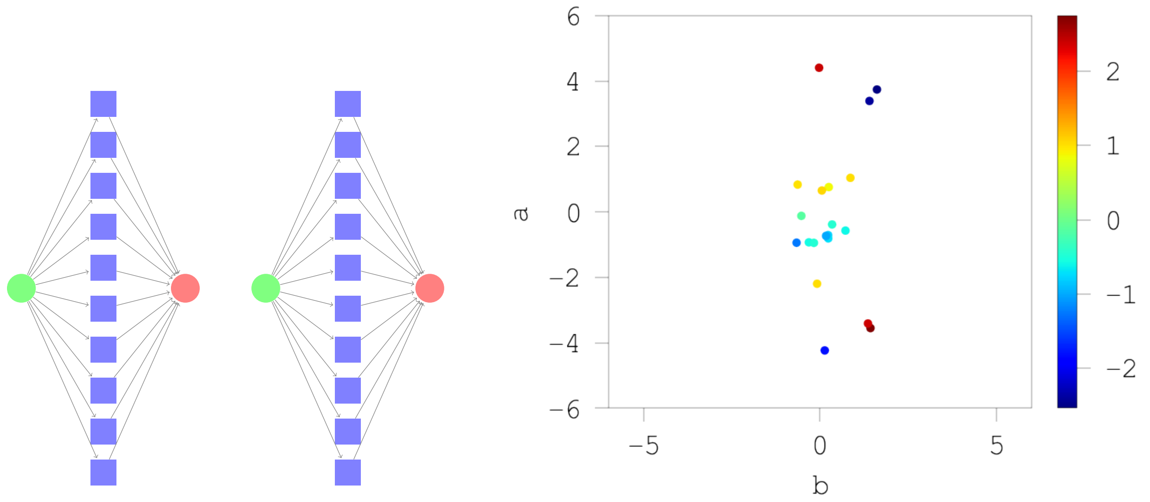
A scatter plot of $d \times n = 10$ hidden parameters (a_j, b_j, c_j) obtained from $n = 1$ neural network $\sum_{j=1}^{d} c_j \sigma(a_j \cdot x - b_j)$ with $d = 10$ hidden units.





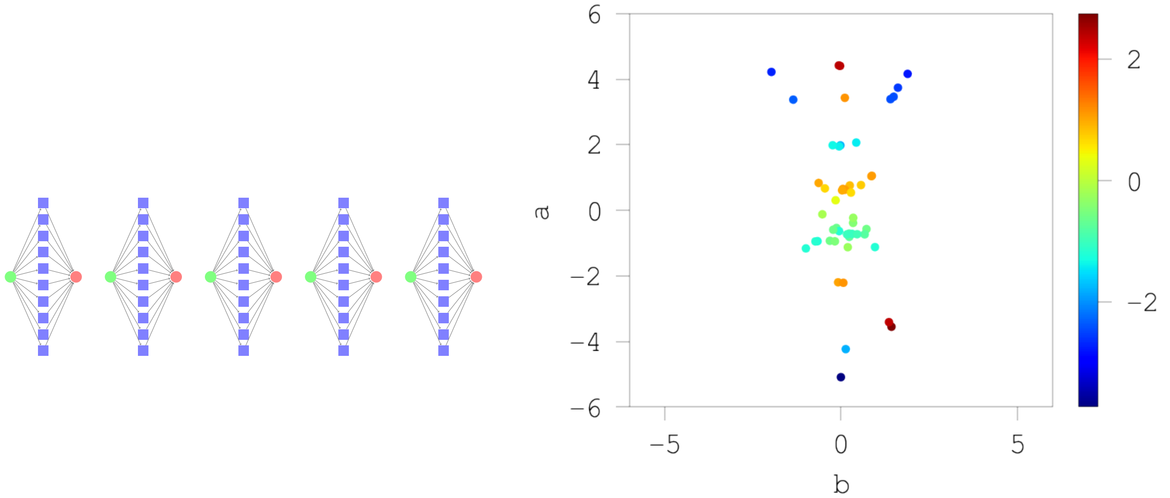
$\arg\min \widehat{L}_n(\theta_d)$

A scatter plot of $d \times n = 20$ hidden parameters (a_j, b_j, c_j) obtained from $n = 2$ neural networks with $d = 10$ hidden units.



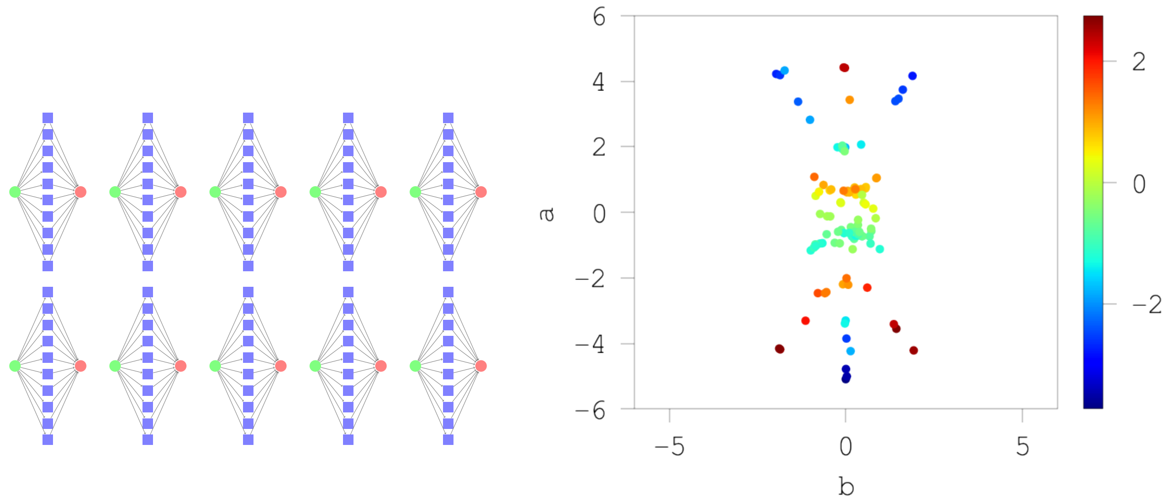
$\arg\min \widehat{L}_n(\theta_d)$

A scatter plot of $d \times n = 50$ hidden parameters (a_j, b_j, c_j) obtained from $n = 5$ neural networks with $d = 10$ hidden units.



$\arg\min \widehat{L}_n(\theta_d)$

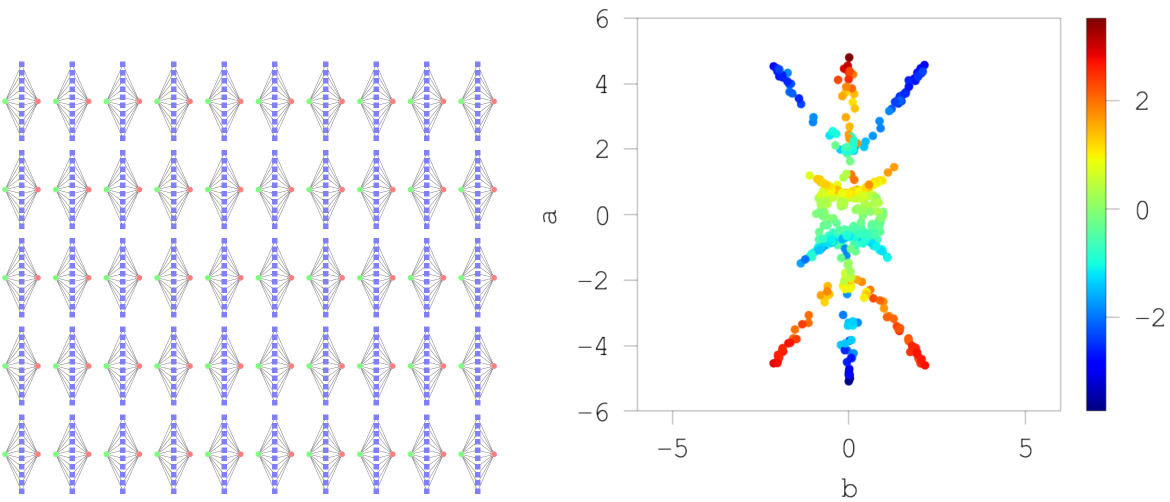
A scatter plot of $d \times n = 100$ hidden parameters (a_j, b_j, c_j) obtained from $n = 10$ neural networks with $d = 10$ hidden units.



$\arg\min \widehat{L}_n(\theta_d)$

7 / 20

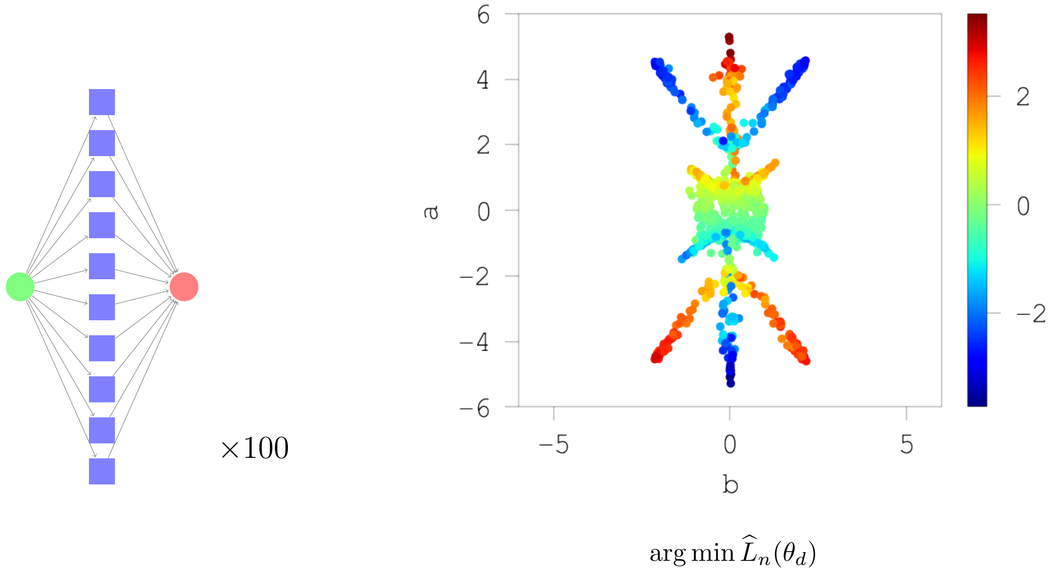
A scatter plot of $d \times n = 500$ hidden parameters (a_j, b_j, c_j) obtained from $n = 50$ neural networks with $d = 10$ hidden units.



$\arg\min \widehat{L}_n(\theta_d)$

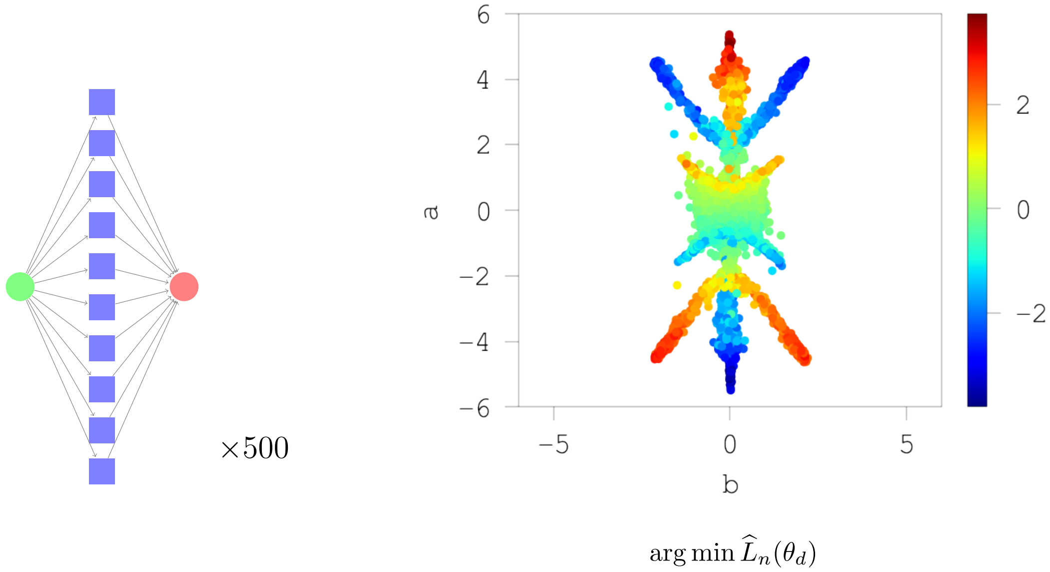
7 / 20

A scatter plot of $d \times n = 1,000$ hidden parameters (a_j, b_j, c_j) obtained from $n = 100$ neural networks with $d = 10$ hidden units.



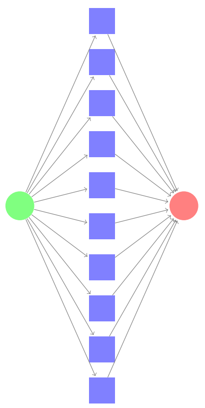
7 / 20

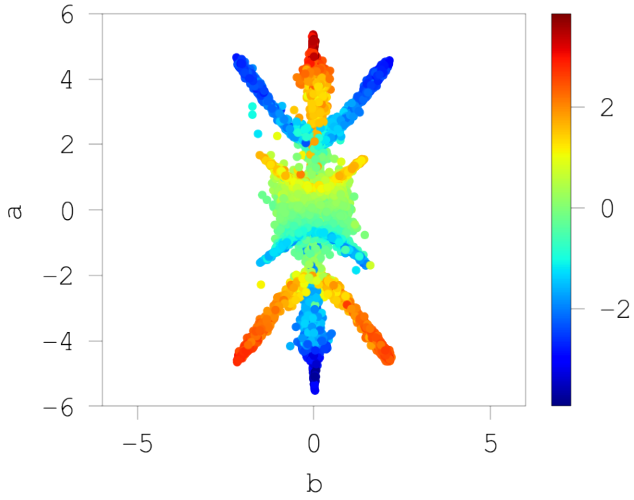
A scatter plot of $d \times n = 5,000$ hidden parameters (a_j, b_j, c_j) obtained from $n = 500$ neural networks with $d = 10$ hidden units.



7 / 20

A scatter plot of $d \times n = 10,000$ hidden parameters (a_j, b_j, c_j) obtained from $n = 1,000$ neural networks with $d = 10$ hidden units.

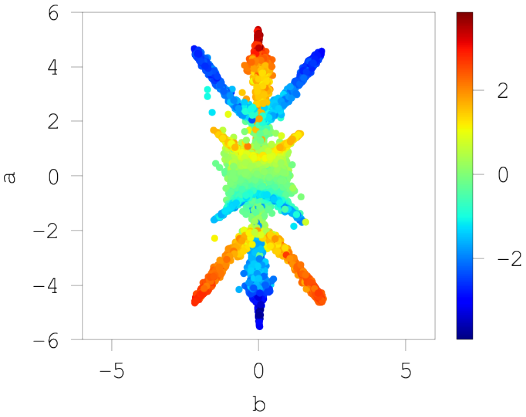




$\arg\min \widehat{L}_n(\theta_d)$

7 / 20

- appears to be the image $R[f;\rho]$ of data f.
- (formal) $\theta_d^{(\infty)} := \mathtt{SGD}(\theta_d^{(0)}, \widehat{L}_n) \sim R[f;\rho]$ (including *sign*!)



scatter plot $\arg\min \widehat{L}_n(\theta_d)$

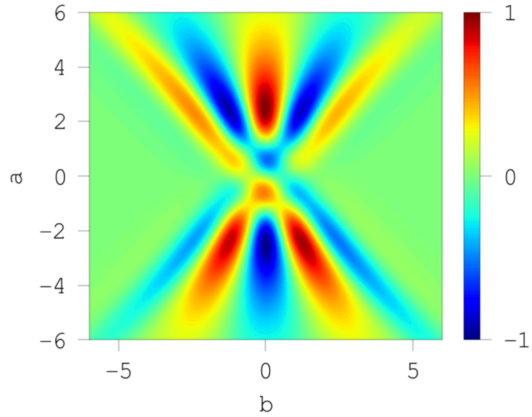


image $R[f](a,b)$

7 / 20

Q. How to Find R?—A. Solve $S[\gamma] = f$

Appendix A.3, in Sonoda-Ishikawa-Ikeda, arXiv:2106.04770

Step 1. Turn the network into a *Fourier expression*

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

$$= \int_{\mathbb{R}^m} \left[\frac{1}{2\pi} \int_{\mathbb{R}} \gamma^\sharp(\boldsymbol{a}, \omega)\sigma^\sharp(\omega) e^{i\omega\boldsymbol{a}\cdot\boldsymbol{x}}\mathrm{d}\omega \right] \mathrm{d}\boldsymbol{a}, \quad \because \tfrac{1}{2\pi}\int_{\mathbb{R}} \gamma^\sharp(\boldsymbol{a},\omega)\sigma^\sharp(\omega)e^{i\omega b}\mathrm{d}\omega = (\gamma(\boldsymbol{a}, \bullet) * \sigma)(b)$$

$$= \frac{1}{2\pi} \int_{\mathbb{R}} \left[\int_{\mathbb{R}^m} \gamma^\sharp(\boldsymbol{\xi}/\omega, \omega) e^{i\boldsymbol{\xi}\cdot\boldsymbol{x}}\mathrm{d}\boldsymbol{\xi} \right] |\omega|^{-m}\sigma^\sharp(\omega)\mathrm{d}\omega, \quad \text{by } (\boldsymbol{a}, \omega) = (\boldsymbol{\xi}/\omega, \omega)$$

where $\cdot^\sharp$ is the Fourier transform in b

Step 2. Assume a *separation-of-variables* form

$$\gamma^\sharp_{f,\rho}(\boldsymbol{\xi}/\omega, \omega) := \widehat{f}(\boldsymbol{\xi})\overline{\rho^\sharp(\omega)}$$

Then, (1) $\gamma_{f,\rho}$ is a particular solution

$$S[\gamma_{f,\rho}] = \frac{1}{2\pi}\left[\int \sigma^\sharp(\omega)\overline{\rho^\sharp(\omega)}|\omega|^{-m}\mathrm{d}\omega \right] \left[\int \widehat{f}(\boldsymbol{\xi}) e^{i\boldsymbol{\xi}\cdot\boldsymbol{x}}\mathrm{d}\boldsymbol{\xi} \right] = (\!(\sigma, \rho)\!) f$$

(2) and $\gamma_{f,\rho}(\boldsymbol{a}, b) = R[f;\rho](\boldsymbol{a}, b)$.

Further Results

Theorem (S-Ishikawa-Ikeda, AISTATS2021)

The empirical regularized least squares parameters in the finite NNs converges to the ridgelet transform:

$$\underset{\gamma_d = \sum_{i=1}^d c_i \delta_{(\boldsymbol{a}_i, b_i)}}{\arg\min} \frac{1}{n}\sum_{i=1}^n |f(\boldsymbol{x}_i) - S[\gamma_d](\boldsymbol{x}_i)|^2 + \beta|\boldsymbol{c}|^2 \xrightarrow{n,d\to\infty,\beta\to+0} S^*[f] = R[f;\sigma_*]$$

- Ridgelet transform can characterize the parameters obtained by learning (loss minimization)

Theorem (S-Ishikawa-Ikeda, arXiv:2106.04770)

The general solution of $S[\gamma] = f$ is given by a sum of ridgelet transforms

$$\gamma = S^*[f] + \sum_{ij} c_{ij} R[e_i; \rho_j]$$

where e_i and ρ_j are ONSs in $L^2(\mathbb{R}^m)$ and $L^2(\mathbb{R}, (\!(\cdot,\cdot)\!))$ resp. satisfying $(\!(\sigma, \rho_j)\!) = 0$

- Ridgelet transform is not only sufficient but also necessary

Extensions to modern network architectures

Based on the *Fourier expression* technique, we have developed new ridgelet transforms for

1. Group convolutional NNs on Hilbert space $\mathcal{H}$
   in S-Ishikawa-Ikeda (NeurIPS2022) and

2. Fully-connected NNs on manifold (noncompact symmetric space) G/K
   in S-Ishikawa-Ikeda (ICML2022)

Group Convolutional NNs on Hilbert Space $\mathcal{H}$[1]

[1]S-Ishikawa-Ikeda, NeurIPS2022

Definition (Group CNN)

Let G be a group, $\mathcal{H}$ be a Hilbert space, and $T : G \to GL(\mathcal{H})$ be a group representation. Let $\mathcal{H}_m \subset \mathcal{H}$ be an m-dimensional subspace equipped with the Lebesgue measure λ. Put

$$S[\gamma](x)(g) := \int_{\mathcal{H}_m \times \mathbb{R}} \gamma(a,b)\sigma((a * x)(g) - b)\mathrm{d}\lambda(a)\mathrm{d}b, \quad x \in \mathcal{H}, g \in G$$

where the (G,T)-convolution is given by

$$(a * x)(g) := \langle T_{g^{-1}}[x], a\rangle_{\mathcal{H}}.$$

Example (Cyclic CNN for multichannel image)

$$\mathtt{CNN}(\boldsymbol{x})(p,q) = \sum_{\ell=1}^{n'} c^\ell \sigma \left(\sum_{k=1}^{n} \sum_{i,j=1}^{m} a_{ij}^{k\ell} x_{i+p,j+q}^{k} - b^\ell \right), \quad \boldsymbol{x} = (x_{ij}^k) \in \mathbb{R}^{m^2 \times n},\ (p,q) \in (\mathbb{Z}/m\mathbb{Z})^2$$

i.e., $G = (\mathbb{Z}/m\mathbb{Z})^2, \mathcal{H} = \mathbb{R}^{m^2 \times n},\ T_{p,q}(\boldsymbol{x}) := (x^{\bullet}_{\bullet-p,\bullet-q})$

In the following, $e \in G$ denotes the identity element.

Definition (Ridgelet Transform)

For any function $f : \mathcal{H}_m \to \mathbb{C}^G$ and $\rho : \mathbb{R} \to \mathbb{C}$, put

$$R[f;\rho](a,b) := \int_{\mathcal{H}_m} f(x)(e)\overline{\rho(\langle a, x\rangle_{\mathcal{H}} - b)}\mathrm{d}\lambda(x).$$

Definition ((G,T)-Equivariance)

A (nonlinear) map $f : \mathcal{H} \to \mathbb{C}^G$ is (G,T)-equivariant when

$$f(T_g[x])(h) = f(x)(g^{-1}h), \quad \forall x \in \mathcal{H}_m, g, h \in G$$

Theorem (Reconstruction Formula)

Suppose that f is (G,T)-equivariant and $f(\bullet)(e) \in L^2(\mathcal{H}_m)$, then $S[R[f;\rho]] = ((\sigma,\rho))f$.

- Meaning: Universality of continuous GCNN
- Corollary: cc-universality of finite GCNNs

Sketch Proof

Step 1. Turn to Fourier expression:

$$S[\gamma](x)(g) = \int_{\mathcal{H}_m \times \mathbb{R}} \gamma(a,b)\sigma(\langle T_{g^{-1}}[x], a\rangle_{\mathcal{H}} - b)\mathrm{d}a\mathrm{d}b$$
$$= \frac{1}{2\pi}\int_{\mathcal{H}_m \times \mathbb{R}} \gamma^\sharp(a,\omega)\sigma^\sharp(\omega)e^{i\omega\langle T_{g^{-1}}[x],a\rangle_{\mathcal{H}}}\mathrm{d}a\mathrm{d}\omega$$
$$= \frac{1}{2\pi}\int_{\mathcal{H}_m \times \mathbb{R}} \gamma^\sharp(\xi/\omega,\omega)\sigma^\sharp(\omega)e^{i\langle T_{g^{-1}}[x],\xi\rangle_{\mathcal{H}}}|\omega|^{-m}\mathrm{d}\xi\mathrm{d}\omega.$$

Step 2. Put separation-of-variables form:

$$\gamma^\sharp_{f,\rho}(\xi/\omega,\omega) := \widehat{f}(\xi)(e)\overline{\rho^\sharp(\omega)}.$$

By the construction it is a particular solution:

$$S[\gamma_{f,\rho}](x)(g) = \frac{1}{2\pi}\int_{\mathcal{H}_m} \widehat{f}(\xi)(e)e^{i\langle T_{g^{-1}}[x],\xi\rangle_{\mathcal{H}}}\mathrm{d}\lambda(\xi)\int_{\mathbb{R}}\sigma^\sharp(\omega)\overline{\rho^\sharp(\omega)}|\omega|^{-m}\mathrm{d}\omega$$
$$= ((\sigma,\rho))f(x)(g).$$

and $\gamma_{f,\rho} = R[f;\rho]$.

14 / 20

Fully-Connected NNs on Noncompact Symmetric Space[2]

[2] S-Ishikawa-Ikeda, ICML2022

15 / 20

Definition (Fully-Connected NNs on Noncompact Symmetric Space G/K)

Let G be a connected semisimple real Lie group, let $G = KAN$ be the Iwasawa decomposition, and let $X := G/K$ be the noncompact symmetric space. Put

$$S[\gamma](x) := \int_{\mathfrak{a}^* \times \partial X \times \mathbb{R}} \gamma(a,u,b)\sigma(a\langle x,u\rangle - b)e^{\varrho\langle x,u\rangle}\mathrm{d}a\mathrm{d}u\mathrm{d}b, \quad x \in X = G/K$$

where $\mathfrak{a}^*$ is the dual of Lie algebra of A, ∂X is the boundary, and $\langle x,u\rangle$ is an X-counter of the Euclidean inner product $\boldsymbol{x}\cdot\boldsymbol{u}$ for $(\boldsymbol{x},\boldsymbol{u}) \in \mathbb{R}^m \times \mathbb{S}^{m-1}$.

Example (Continuous Horospherical Hyperbolic NN)

On the *Poincaré ball model* $\mathbb{B}^m := \{\boldsymbol{x} \in \mathbb{R}^m \mid |\boldsymbol{x}| < 1\}$ equipped with the Riemannian metric $\mathfrak{g} = 4(1-|\boldsymbol{x}|)^{-2}\sum_{i=1}^m \mathrm{d}x_i \otimes \mathrm{d}x_i$,

$$S[\gamma](\boldsymbol{x}) := \int_{\mathbb{R}\times\partial\mathbb{B}^m\times\mathbb{R}} \gamma(a,\boldsymbol{u},b)\sigma(a\langle \boldsymbol{x},\boldsymbol{u}\rangle - b)e^{\varrho\langle \boldsymbol{x},\boldsymbol{u}\rangle}\mathrm{d}a\mathrm{d}\boldsymbol{u}\mathrm{d}b, \quad \boldsymbol{x} \in \mathbb{B}^m$$

$\varrho = (m-1)/2$, $\langle \boldsymbol{x},\boldsymbol{u}\rangle = \log\left(\dfrac{1-|\boldsymbol{x}|_E^2}{|\boldsymbol{x}-\boldsymbol{u}|_E^2}\right)$, $\quad (\boldsymbol{x},\boldsymbol{u}) \in \mathbb{B}^m \times \partial\mathbb{B}^m$

Definition (Ridgelet Transform)

For any function $f : X \to \mathbb{C}$ and an auxiliary function $\rho : \mathbb{R} \to \mathbb{C}$, put

$$R[f;\rho](a,u,b) := \int_X \boldsymbol{c}[f](x)\overline{\rho(a\langle x,u\rangle - b)}e^{\varrho\langle x,u\rangle}\mathrm{d}x$$

where $\boldsymbol{c}[f]$ is a Helgason-Fourier multiplier.

Theorem (Reconstruction Formula)

For any $\sigma \in \mathcal{S}'(\mathbb{R}), \rho \in \mathcal{S}(\mathbb{R})$, and $f \in L^2(X)$, we have

$$S[R[f;\rho]] = \int_{\mathfrak{a}^*\times\partial X\times\mathbb{R}} R[f;\rho](a,u,b)\sigma(a\langle x,u\rangle - b)e^{\varrho\langle x,u\rangle}\mathrm{d}a\mathrm{d}u\mathrm{d}b = ((\sigma,\rho))f.$$

where $((\sigma,\rho))$ is a certain scalar product.

- Meaning: Universality of continuous Fully-Connected NN on X
- Corollary: cc-universality of finite Fully-Connected NNs on X

Fourier Analysis on $X = G/K$

Helgason, GGA (1984, Introduction); GASS (2008, Chapter III)

Definition (Helgason-Fourier Transform)

For any function $f : X \to \mathbb{C}$,

$$\widehat{f}(\lambda, u) := \int_X f(x) e^{(-i\lambda+\varrho)\langle x,u\rangle} \mathrm{d}x, \quad (\lambda, u) \in \mathfrak{a}^* \times \partial X$$

with a certain constant vector $\varrho \in \mathfrak{a}^*$.

Theorem (Inversion Formula)

For any $f \in L^2(X)$ (or $f \in C_c^\infty(X)$),

$$f(x) = |W|^{-1} \int_{\mathfrak{a}^* \times \partial X} \widehat{f}(\lambda, u) e^{(i\lambda+\varrho)\langle x,u\rangle} |\boldsymbol{c}(\lambda)|^{-2} \mathrm{d}\lambda \mathrm{d}u, \quad x \in X$$

where $\boldsymbol{c}$ is the Harish-Chandra $\boldsymbol{c}$-function, and $|W|$ is a constant.

This is a "Fourier transform" because $e^{(-i\lambda+\varrho)\langle x,u\rangle}$ is the eigenfunction $e^{(-i\lambda+\varrho)\langle x,u\rangle}$ of the Laplace-Beltrami operator Δ_X on X

Sketch Proof

- Given a function $f : G/K \to \mathbb{C}$, consider solving an integral equation $S[\gamma] = f$ of unknown γ.
- Step 1: Change the frame of $S[\gamma]$ from neurons to a *Fourier expression*:

$$\begin{aligned} S[\gamma](x) &:= \int_{\mathfrak{a}^* \times \partial X \times \mathbb{R}} \gamma(a, u, b) \sigma(a\langle x, u\rangle - b) e^{\varrho\langle x,u\rangle} \mathrm{d}a \mathrm{d}u \mathrm{d}b \\ &= \frac{1}{2\pi} \int_{\mathbb{R}} \left[\int_{\mathfrak{a}^* \times \partial X} \gamma^\sharp(\lambda/\omega, u, \omega) |\boldsymbol{c}(\lambda)|^2 e^{(i\lambda+\varrho)\langle x,u\rangle} \frac{\mathrm{d}\lambda \mathrm{d}u}{|\boldsymbol{c}(\lambda)|^2} \right] |\omega|^{-r} \sigma^\sharp(\omega) \mathrm{d}\omega, \end{aligned}$$

  where $\sharp$ denotes the Euclidean-Fourier transform in b.
- Step 2: Since inside $[\cdots]$ is the *inverse Helgason-Fourier transform*, put a separation-of-variables form:

$$\gamma_{f,\rho}^\sharp(\lambda/\omega, \boldsymbol{u}, \omega) = \widehat{f}(\lambda, \boldsymbol{u}) \overline{\rho^\sharp(\omega)} |\boldsymbol{c}(\lambda)|^{-2}.$$

  Then, by the construction, it is a particular solution:

$$S[\gamma_{f,\rho}] = (\!(\sigma, \rho)\!) f,$$

  where $(\!(\sigma, \rho)\!) := \frac{|W|}{2\pi} \int_{\mathbb{R}} \sigma^\sharp(\omega) \overline{\rho^\sharp(\omega)} |\omega|^{-m} \mathrm{d}\omega$.
- In the end, we can verify that $\gamma_{f,\rho}$ is the ridgelet transform $R[f; \rho]$.

Conclusion

- Ultimate goal:
  - Characterize deep solutions
- We have seen:
  - Shallow solutions are characterized by ridgelet transform
- Take home message:
  - If there is a Fourier transform, then so is the ridgelet transform
- We will see:
  - A ridgelet transform for depth

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 $(i = 1, \ldots, d)$ 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

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

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 $T(x) = (T_1(x_1), \ldots, T_d(x_d))$ 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 $T(x) = (T_1(x_1), \ldots, T_d(x_d))$ 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`

Stein-type distributions on Riemannian manifolds

Tomonari SEI Ushio Tanaka

The University of Tokyo Osaka Metropolitan University

Oct 20 (Thu), OCAMI workshop
"Mathematical optimization and statistical theories using geometric methods"

The Stein identity

We begin with the following fact.

Proposition (Stein identity)

A random variable X follows $N(0,1)$ if and only if

$$E[Xf(X)] = E[f'(X)]$$

for any differentiable function f with bounded f'.

Proof: ($\Rightarrow$) For the density function $\phi(x) = (2\pi)^{-1/2}e^{-x^2/2}$,

$$\int xf(x)\phi(x)dx = \int f(x)\{-\phi(x)\}'dx = \int f'(x)\phi(x)dx.$$

($\Leftarrow$) If $E[Xf(X)] = E[f'(X)]$, it is shown that X has density $p(x)$. Then the identity is equivalent to

$$p'(x) + xp(x) = 0.$$

The unique solution is $p(x) = \phi(x)$. $\square$

Application of the Stein identity

Why is the Stein identity important?

- Stein's unbiased risk estimator (statistics)
- Central limit theorem (probability theory)
- Stein discrepancy (machine learning)

Application: Stein's unbiased risk estimator

- Let $X \sim N_d(\theta, I_d)$, where $\theta \in \mathbb{R}^d$ is unknown parameter.
- Consider an estimator $X + f(X)$ of θ. The risk is

$$\begin{aligned} &E[\|X + f(x) - \theta\|^2] \\ &= E[\|X - \theta\|^2] + 2E[f(X)^\top (X - \theta)] + E[\|f(X)\|^2] \\ &= d + 2E[\nabla^\top f(X)] + E[\|f(X)\|^2] \quad \text{(Stein identity)} \\ &= E[\underbrace{d + 2\nabla^\top f(X) + \|f(X)\|^2}_{\text{risk estimator}}] \end{aligned}$$

Another application of Stein identity

Application: Poincaré inequality (Chernoff 1981, Chen 1982)

- If $X \sim N(0, 1)$, then

$$V[g(X)] \le E[g'(X)^2],$$

  with equality if and only if $g(x) = ax + b$.
- Indeed,

$$\begin{aligned} V[g(X)] &\le E[(g(X) - g(0))^2] \\ &= E[(\textstyle\int_0^X g'(x)dx)^2] \\ &\le E[X \textstyle\int_0^X g'(x)^2 dx] \quad \text{(Cauchy–Schwarz*)} \\ &= E[g'(X)^2] \quad \text{(Stein identity)}. \end{aligned}$$

(* valid even for $X < 0$.)

Outline of this talk

In this talk, we generalize the Stein identity in the following manner.

1. Define Stein-type distributions on $\mathbb{R}^d$ by an identity
$$E[(X_1 + \cdots + X_d)f(X_i)] = E[f'(X_i)].$$
2. Define Stein-type distributions on the direct product of Riemaniann manifolds (on-going work).

We first see the background of the problem in a couple of slides.

Background: Objective general index (OGI)

S. (2016) pointed out that the Stein identity is related to a scaling problem, which is a motivation of this work.

- First, consider d random variables $X_1, \ldots, X_d$.
- For example, X_i is academic score of students on i-th subject.

Proposition (S. 2016)

There exist unique $w_1, \ldots, w_d > 0$ such that
$$\mathrm{Cov}(Y, w_i X_i) = 1 \quad (i = 1, \ldots, d),$$
where $Y = w_1 X_1 + \cdots + w_d X_d$, under a mild condition.

- The proof is based on matrix scaling (Marshall–Olkin 1968).
- We call Y the objective general index (OGI).
- The Stein identity appears in a functional version of this fact.

Illustration

A numerical example

Suppose that the covariance matrix of X_1, X_2, X_3 is

$$(\mathrm{Cov}(X_i, X_j))_{i,j=1}^3 = \begin{pmatrix} 1 & -0.5 & -0.5 \\ -0.5 & 1 & 0 \\ -0.5 & 0 & 1 \end{pmatrix}.$$

In this case,

$$\mathrm{Cov}(X_1 + X_2 + X_3, X_1) = 1 - 0.5 - 0.5 = 0.$$

But, a weight $(w_1, w_2, w_3) = (2.135779, 1.667566, 1.667566)$ gives

$$\mathrm{Cov}(\underbrace{w_1X_1 + w_2X_2 + w_3X_3}_{\text{OGI}}, w_iX_i) = 1, \quad i = 1, 2, 3.$$

Functional OGI

- Next, consider a random variable X with density $p(x)$.
- Define an infinite number of variables by Heaviside function:

$$h_\xi(X) = I_{\{X \geq \xi\}} - E[I_{\{X \geq \xi\}}], \quad \xi \in \mathbb{R}.$$

- What is OGI of $\{h_\xi(X)\}_{\xi \in \mathbb{R}}$?

Proposition (S. 2016)

There exists a unique positive function $w(\xi)$ such that

$$\mathrm{Cov}(Y, w(\xi)h_\xi(X)) = 1 \quad (\xi \in \mathbb{R}), \qquad (*)$$

where $Y = \int_{\mathbb{R}} w(\xi)h_\xi(X)p(\xi)d\xi$. In fact, $Y \sim N(0, 1)$.

- We call Y the functional OGI of X.
- The identity $(*)$ is considered as a version of the Stein identity.
- Let us check it.

Functional OGI and Stein identity

- It is shown that the condition of the functional OGI

$$\mathrm{Cov}(Y, w(\xi)h_\xi(X)) = 1$$

is equivalent to the Stein identity

$$E[Yf_\xi(Y)] = E[f'_\xi(Y)]$$

for $f_\xi(y) = h_\xi(T^{-1}(y))$ and $T(x) = \int_{\mathbb{R}} w(\xi)h_\xi(x)p(\xi)d\xi$.

- In other words, the functional OGI is characterized by an increasing function T that attains the Stein identity.
- The Stein-type distribution we now discuss is a generalization of $N(0,1)$ based on this fact.

Variational characterization

Before proceeding, we recall a variational characterization of $N(0,1)$.

Proposition

$$F(p) = \int_{\mathbb{R}} p(x)\log p(x)dx + \int_{\mathbb{R}} \frac{x^2}{2}p(x)dx$$

has a unique minimizer $p(x) = \phi(x) = (2\pi)^{-1/2}e^{-x^2/2}$.

- Proof 1: $F(p) = \int p(x)\log(p(x)/\phi(x))dx + \mathrm{const}$
- Proof 2: Let p_0 be a minimizer of F. Let $T(x) = x + \varepsilon f(x)$ be an increasing function. Then,

$$F(T_\sharp p_0) - F(p_0) = \varepsilon\left(-\int p_0(x)f'(x)dx + \int f(x)p_0(x)dx\right) + o(\varepsilon).$$

The stationary condition is the Stein identity. So $p_0 = \phi$.

Fiber

Now let us go on to the $\mathbb{R}^d$ case.

- Let $\mathcal{P}^2$ be the set of probability distributions μ on $\mathbb{R}^d$ such that the marginal distribution μ_i satisfies

$$\int_{\mathbb{R}} x_i d\mu_i = 0, \quad \int_{\mathbb{R}} x_i^2 d\mu_i < \infty, \quad \mu_i \ll \text{Leb}.$$

- We call $T : \mathbb{R}^d \to \mathbb{R}^d$ a coordinate-wise transformation if

$$T(\boldsymbol{x}) = (T_1(x_1), \ldots, T_d(x_d)), \quad T_i'(x_i) > 0.$$

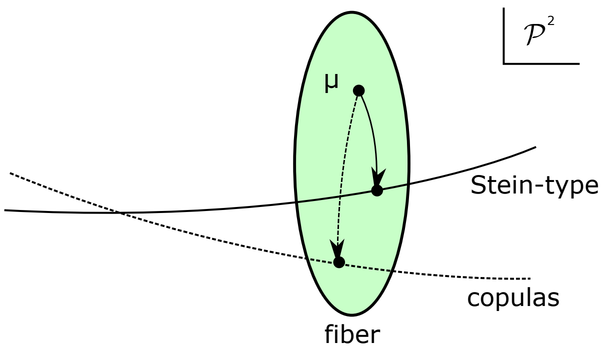
- For each $\mu \in \mathcal{P}^2$, define the μ-fiber

$$\mathcal{F}_\mu = \{T_\sharp \mu \in \mathcal{P}^2 \mid T \text{ is coordinate-wise}\},$$

where $T_\sharp$ denotes the push forward.

Picture

- The space $\mathcal{P}^2$ is decomposed into the set of fibers.
- We define a Stein-type distribution in each fiber.



Remark

- $\mathcal{F}_\mu$ is totally geodesic in the Wasserstein space.
- $\mathcal{F}_\mu$ has a unique copula (Sklar's theorem). A copula refers to a distribution with uniform marginals on $[0, 1]$.

A free-energy functional

- Define a functional $F : \mathcal{P}^2 \to \mathbb{R}$ by

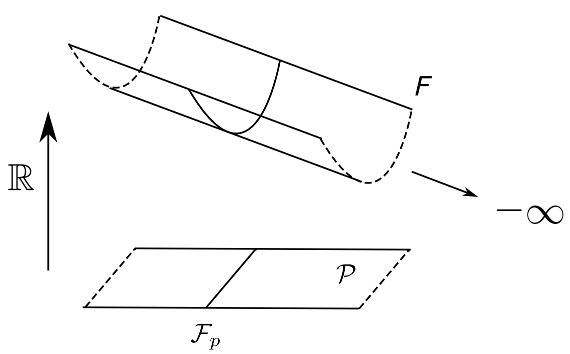
$$F(\mu) = \sum_{i=1}^{d} \int_{\mathbb{R}} \log \frac{d\mu_i}{dx_i} d\mu_i + \frac{1}{2} \int_{\mathbb{R}^d} \left(\sum_{i=1}^{d} X_i \right)^2 d\mu,$$

- We can further consider

$$F(\mu) = \sum_{i=1}^{d} \int_{\mathbb{R}} \log \frac{d\mu_i}{dx_i} d\mu_i + \int_{\mathbb{R}^d} V(\boldsymbol{x}) d\mu,$$

  with some smooth function $V(\boldsymbol{x})$ (S. 2017).
- This appears in the optimal transport theory (McCann 1997) except that the entropy term is replaced with $\int \log \frac{d\mu}{dx} d\mu$.

Minimization over the fiber



- F is not bounded from below on the whole space $\mathcal{P}^2$.
- But F may be bounded from below on each fiber $\mathcal{F}_\mu$.

Stein-type distribution

- To minimize F over the fiber, consider a perturbation of the transformation around the identity:

$$T_\varepsilon = \mathrm{Id} + \varepsilon f, \quad f(\boldsymbol{x}) = (f_1(x_1), \dots, f_d(x_d)), \quad \varepsilon \in \mathbb{R}.$$

- Then we have, as $\varepsilon \to 0$, the first variation

$$F((T_\varepsilon)_\sharp \mu) \simeq F(\mu) + \varepsilon \sum_i \int \{-f_i'(x_i) + f_i(x_i)(x_1 + \cdots + x_d)\} d\mu.$$

Definition (Stein-type distribution)

A distribution μ is called a Stein-type distribution if it satisfies

$$\int_{\mathbb{R}^d} (x_1 + \cdots + x_d) f_i(x_i) d\mu = \int_{\mathbb{R}^d} f_i'(x_i) d\mu, \quad \forall i, \ \forall f_i \in \mathrm{C}^1(\mathbb{R}).$$

Examples

Example 1 (independent case)

If $X_1, \dots, X_d$ are independent and have zero mean, then the equation

$$E[(X_1 + \cdots + X_d) f(X_i)] = E[f'(X_i)]$$

forces

$$E[X_i f(X_i)] = E[f'(X_i)].$$

Thus, only the independent Stein-type distribution is the standard normal distribution.

Examples

Example 2 (Gaussian)

Let $(X_1, \ldots, X_d) \sim N_d(0, S)$. Then the distribution is Stein-type if and only if

$$\sum_{j=1}^{d} S_{ij} = \mathrm{Cov}\left(X_i, \sum_j X_j\right) = 1$$

for $i = 1, \ldots, d$. This is the same as the OGI property.

Example 3 (non-Gaussian)

Let $Z \sim N(0,1)$ and U be any distribution with $E[U] = 0$ and $E[U^2] < \infty$. Then the random vector (X_1, X_2) with

$$X_1 = \frac{Z+U}{\sqrt{2}}, \quad X_2 = \frac{Z-U}{\sqrt{2}}$$

is Stein-type.

Functional OGI (revisited)

We briefly discuss an application of our results.

Problem

- Let $X_1, \ldots, X_d$ be random variables with joint density $p(\boldsymbol{x})$, which represent students' scores on d academic subjects.
- How to define the overall score?

An answer

- Let $Y = \sum_{j=1}^{d} T_j(X_j)$, where $T(X)$ is the Stein-type.
- Then the Heaviside function $f(x_i) = h_\xi(x_i)$ yields

$$E[Y \mid X_i > \xi] \;>\; E[Y \mid X_i < \xi], \quad \forall \xi \in \mathbb{R}, \ \forall i.$$

- Interpretation: students with higher score on each subject i has higher overall score in mean.

Assumption on μ: copositivity

S. (2022) established an existence and uniqueness theorem. We suppose some conditions.

- For each $\mu \in \mathcal{P}^2$, denote the product measure of marginal distributions by

$$\mu^{\perp} = \textstyle\prod_{i=1}^{d} \mu_i.$$

Definition (Copositivity)

We say that μ is copositive if

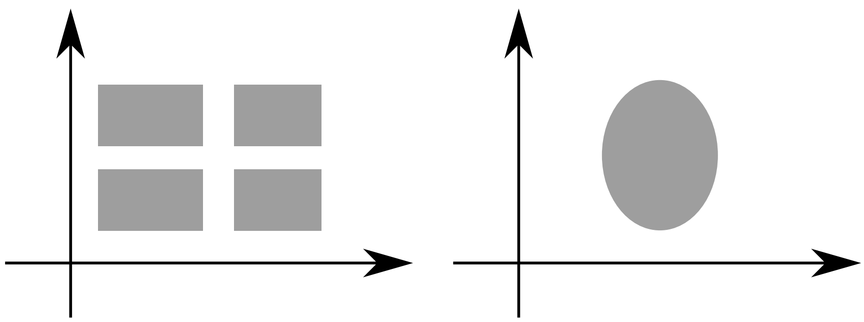
$$\beta(\mu) = \inf_{T:\mathrm{cw}} \frac{\int \{\sum_i T_i(x_i)\}^2 d\mu}{\int \{\sum_i T_i(x_i)\}^2 d\mu^{\perp}} > 0.$$

- Trivially, if μ is independent $(\mu = \mu^{\perp})$, then $\beta(\mu) = 1$.
- Sufficient conditions for copositivity are discussed later.

Assumption on μ: regular support

Definition (Regularity)

We say that μ has a regular support if the support of μ is the direct product of the supports of μ_i's.

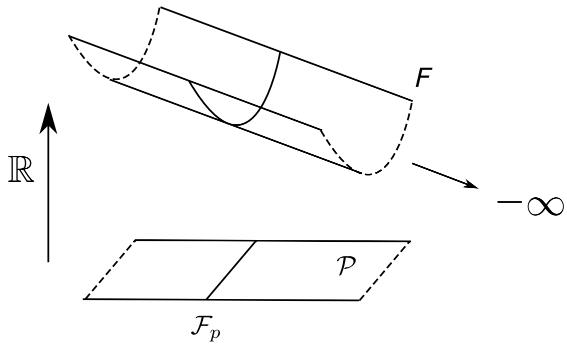


regular non-regular

Existence and uniqueness theorem

Theorem (Existence and uniqueness)

Suppose that μ is copositive and has a regular support. Then there exists a unique Stein-type distribution in the μ-fiber.



Proof sketch.

- Uniqueness follows from the displacement convexity

$$F([(1-\lambda)T_0+\lambda T_1]_\sharp\mu) > (1-\lambda)F((T_0)_\sharp\mu)+\lambda F((T_1)_\sharp\mu),$$

  where strict inequality follows from the regular support condition.
- For existence, we use the copositivity to obtain

$$F(\mu) \geq \int \log\frac{d\mu^\perp}{dx}d\mu^\perp + \frac{\beta}{2}\int\left(\sum_i x_i\right)^2 d\mu^\perp.$$

  Then the problem is essentially reduced to the independent case $\mu=\mu^\perp$.

$\square$

Sufficient conditions for copositivity

- We establish sufficient conditions for copositivity

$$\beta(\mu) = \inf_{T:\mathrm{cw}} \frac{\int \{\sum_i T_i(x_i)\}^2 d\mu}{\int \{\sum_i T_i(x_i)\}^2 d\mu^{\perp}} > 0.$$

- The notion of positive dependence plays a significant role.

Definition (e.g. Rüschendorf 2013)

1. $p(\boldsymbol{x})$ is called MTP$_2$ (multivariate totally positive of order 2) if $p(\boldsymbol{x} \vee \boldsymbol{y})p(\boldsymbol{x} \wedge \boldsymbol{y}) \geq p(\boldsymbol{x})p(\boldsymbol{y})$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$.
2. $p(\boldsymbol{x})$ is said to be associated if $\int \phi\psi p \mathrm{d}\boldsymbol{x} \geq \int \phi p \mathrm{d}\boldsymbol{x} \int \psi p \mathrm{d}\boldsymbol{x}$ for all increasing $\phi, \psi : \mathbb{R}^d \to \mathbb{R}$.
3. $p(\boldsymbol{x})$ is called PSMD (positive super-modular dependent) if $\int \phi(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \geq \int \phi(\boldsymbol{x})p^{\perp}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ for any super-modular function ϕ.

Sufficient conditions

Theorem (FKG 1971, Christofides 2004, S. 2017)

MTP$_2$ $\Rightarrow$ associated $\Rightarrow$ PSMD $\Rightarrow$ copositive.

- MTP$_2$ is relatively easy to confirm.

Open problems

There are some open problems.

Conjectures

1. The marginal support of any Stein-type density is $\mathbb{R}$.
2. Existence implies uniqueness.
3. A Gaussian distribution is copositive if the covariance matrix is strictly copositive.

For the rest of talk, we generalize the Stein-type distributions on $\mathbb{R}^d$ to the direct space of Riemannian maniolds.

Optimal transport on Riemannian manifolds

We recall the optimal transport theory on Riemannian manifolds according to McCann (2001).

- Let (M, g) be a Riemannian manifold that is C^3, compact and connected without boundaries.
- An example in mind is $M = S^1$ (circle).
- Let $d(x, y)$ be the geodesic distance between $x, y \in M$.
- A cost is defined by $c(x, y) = d(x, y)^2/2$.
- A function $\phi : M \to \mathbb{R}$ is called cost-convex if there exists $\phi^* : M \to \mathbb{R}$ such that

$$\phi(x) = \sup_{y \in M} \{-c(x, y) - \phi^*(y)\}.$$

- If ϕ is cost-convex, it is Lipschitz and therefore is differentiable vol-a.e. (Rademacher's differentiability theorem).

McCann's theorem

For cost-convex ϕ, a map $T : M \to M$ defined by

$$T(x) = \exp_x(\nabla\phi(x))$$

is considered as a generalization of increasing functions on $\mathbb{R}$.

Theorem (McCann 2001)

Let $\mu \ll \mathrm{vol}$ and ν be probability measures on M. Then there exists a unique cost-convex function ϕ (up to additive constants) such that $T(x) = \exp_x(\nabla\phi(x))$ pushes μ forward to ν. This map is a unique minimizer of the transportation cost $\int c(x, T(x))d\mu$.

Fiber

- Let $M_1, \ldots, M_d$ be C^3 compact Riemannian manifolds.
- Consider the product space $M = \prod_{i=1}^d M_i$.
- Let $\mathcal{P}$ be the set of probability distributions μ on M such that the marginal distribution μ_i satisfies $\mu_i \ll \mathrm{vol}_i$.
- We call $T : M \to M$ a coordinate-wise transformation if

  $$T(\boldsymbol{x}) = (T_1(x_1), \ldots, T_d(x_d)), \quad T_i = \exp_{x_i}(\nabla\phi_i(x_i)),$$

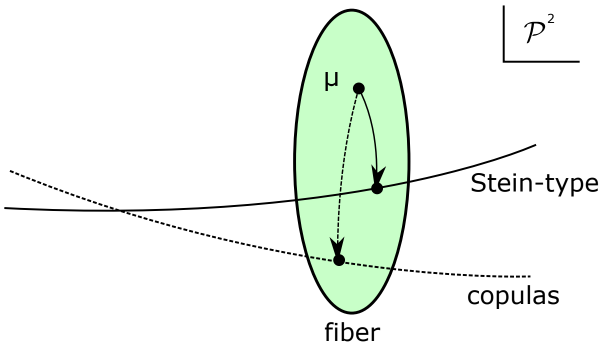
  where ϕ_i is cost-convex.
- For each $\mu \in \mathcal{P}$, define the μ-fiber

  $$\mathcal{F}_\mu = \{T_\sharp\mu \in \mathcal{P} \mid T \text{ is coordinate-wise}\},$$

  where $T_\sharp$ denotes the push forward.

Picture

- The space $\mathcal{P}$ is decomposed into the set of fibers.
- We define a Stein-type distribution in each fiber.



Remark: Sklar's theorem on manifolds

$\mathcal{F}_\mu$ has a unique "copula", which refers to a distribution with uniform marginals on M_i. (cf. circula; Jones et al. (2015))

Stein-type distribution on M

- Let $V(\boldsymbol{x})$ be a smooth function on $M = \prod_{i=1}^{d} M_i$.
- Define a functional $F : \mathcal{P} \to \mathbb{R}$ by

$$F(\mu) = \sum_{i=1}^{d} \int_{M_i} \log \frac{d\mu_i}{dx_i} d\mu_i + \int_M V(\boldsymbol{x}) d\mu.$$

Definition

A Stein-type distribution on M is defined by a minimizer of $F(\mu)$ over a fiber.

Problem: Existence and uniqueness? $\to$ future work..

Stationary condition

- To minimize F over the fiber, consider a perturbation of the transformation around the identity:

$$T_\varepsilon(x) = \exp_x(\varepsilon f(x)), \quad f(\boldsymbol{x}) = (f_1(x_1), \ldots, f_d(x_d)), \quad \varepsilon \in \mathbb{R}.$$

- Then we have the first variation

$$F((T_\varepsilon)_\sharp \mu) \simeq F(\mu) + \varepsilon \sum_i \int \{-\nabla_i f_i(x_i) + f_i(x_i)\nabla_i V(\boldsymbol{x})\} d\mu.$$

Lemma

If μ is Stein-type, then

$$\int_M f_i(x_i)\nabla_i V(\boldsymbol{x}) d\mu = \int_M \nabla_i f_i(x_i) d\mu, \quad \forall i, \ \forall f_i \in \mathrm{C}^1(M_i).$$

Example

Circular case

- Let $M_1 = \cdots = M_d = S^1$.
- We use the coordinate $x_i = (\cos\theta_i, \sin\theta_i) \in M_i$.
- Consider a function

$$V(\boldsymbol{x}) = \frac{1}{2}\{(\textstyle\sum_i \cos\theta_i)^2 + (\sum_i \sin\theta_i)^2\}.$$

The derivative is $\partial_{\theta_i} V(\boldsymbol{x}) = -A(\theta)\sin(\theta_i - \bar{\theta})$, where $A(\theta)$ and $\bar{\theta}$ are defined appropriately.
- Then the Stein-type distribution has to satisfy

$$-\textstyle\int_M f_i(\theta_i) A(\theta)\sin(\theta_i - \bar{\theta}) d\mu = \int_M f_i'(\theta_i) d\mu.$$

- Any application? $\to$ future work...

Summary and future work

Summary

- We defined the Stein-type distributions on Euclidean space and established the existence and uniqueness theorem.
- We generalized it to distributions on Riemannian manifolds.

Future works

- Existence seems OK due to the compactness. Uniqueness may be non-trivial.
- Any analogue of Poincaré inequality?
- We are seeking applications.

Thank you for your attention!

References I

- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*, Springer.
- Chen, L. H. (1982). An inequality for the multivariate normal distribution, *J. Multivariate Anal.*, **12**, 306–315.
- Chen, L. H. Y., Goldstein, L., and Shao, Q. (2011). *Normal Approximation by Stein's Method*, Springer.
- Chernoff, H. (1981). A note on an inequality involving the normal distribution, *Ann. Probab.*, **9** (3), 533–535.
- Christofides, T. C. and Vaggelatou, E. (2004). A connection between supermodular ordering and positive/negative association, *J. Multivariate Anal.*, **88**, 138–151.
- Fallat, S., Lauritzen, S., Sadeghi, K., Uhler, C., Wermuth, N., and Zwiernik, P. (2017). Total positivity in Markov structures, *Ann. Statist.*, **45** (3), 1152–1184.
- Fortuin, C. M., Kasteleyn, P. W., and Ginibre, J. (1971). Correlation inequalities on some partially ordered sets, *Comm. Math. Phys.*, **22**, 89–103.
- Jones, M. C., Pewsey, A., and Kato, S. (2015). On a class of circulas: copulas for circular distributions, *Ann. Inst. Statist. Math.*, **67** (5), 843–862.
- Marshall, A. W., Olkin, I., (1968). Scaling of matrices to achieve specified row and column sums. *Numer. Math.*, 12, 83–90.

References II

- McCann, R. J. (1997). A convexity principle for interacting gases, *Adv. Math.*, **128**, 153–179.
- McCann, R. J. (2001). Polar factorization of maps on Riemannian manifolds, *Geometric and Functional Analysis*, **11**, 589–608.
- Müller, A. and Stoyan, D. (2002). *Comparison Methods for Stochastic Models and Risks*, *Wiley*.
- Nelsen, R. B. (2006). *An Introduction to Copulas*, 2nd ed., Springer.
- Rüschendorf, L. (1981). Characterization of dependence concepts in normal distributions, *Ann. Inst. Statist. Math.*, **33**, 347–359.
- Rüschendorf, L. (2013). *Mathematical Risk Analysis*, Springer.
- Sei, T. (2016). An objective general index for multivariate ordered data, *J. Multivariate Anal.*, 147, 247–264.
- Sei, T. (2017). Coordinate-wise transformation and Stein-type densities, Proceedings of the 3rd Conference on Geometric Science of Information (GSI2017), Nov 7–9, 2017, Mines ParisTech, France.
- Sei, T. (2022). Coordinate-wise transformation of probability distributions to achieve a Stein-type identity, Information Geometry, **5**, 325–354.
- Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables, *Proc. Sixth Berkeley Symp. on Math. Statist. and Prob.*, Vol. 2, 583–602.
- Villani, C. (2003). *Topics in Optimal Transportation*, American Mathematical Society.

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.

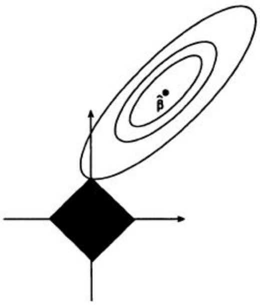
[3] T. Skalski, P. Graczyk, B. Kołodziejek, M. Wilczyński. Pattern recovery and signal denoising by SLOPE when the design matrix is orthogonal. ArXiv 2202.08573.

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

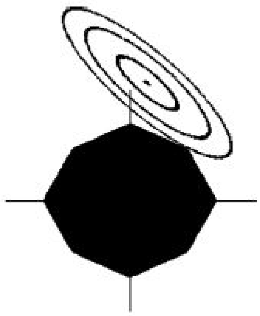
On LASSO and SLOPE estimators and their pattern recovery

Tomasz Skalski

Wrocław University of Science and Technology, Poland
University of Angers, France



Osaka & on-line
2022/10/20



Linear regression model

Linear regression model: $Y = X\beta + \varepsilon$:

- $Y \in \mathbb{R}^n$: response vector
- $X \in \mathbb{R}^{n\times p}$: design matrix
- $\beta \in \mathbb{R}^p$: unknown parameter vector
- $\varepsilon \in \mathbb{R}^n$: random noise term

Noiseless case: $\varepsilon = 0$.
Noisy case: ε has continuous and symmetric distribution.
Goal: to estimate β.

Ordinary Least Squares estimator

- Ordinary Least Squares (Legendre, 1805, Gauss, 1809)
- $\hat{\beta}^{OLS} := \arg\min_{b \in \mathbb{R}^p} \frac{1}{2} \|Y - Xb\|_2^2$
- $\hat{\beta}^{OLS} = (X'X)^{-1}X'Y$

Not defined when $n < p$.
In noisy case: with probability 1 has p pairwise different coordinates.

Penalized estimator

Consider the following penalized estimator

$$\hat{\beta} := \arg\min_{b \in \mathbb{R}^p} \frac{1}{2}\|Y - Xb\|_2^2 + \lambda J(b), \text{ where } J \text{ is a norm.}$$

- $\hat{\beta}$ is well defined when $n \geq p$ as well as when $n < p$.
- The pattern of $\hat{\beta}$ is characterized by its subdifferential ∂_J.
- The dual norm J^* is given by $J^*(x) = \sup\{z'x : J(z) \leq 1\}$.
- $\hat{\beta} = 0$ if and only if $J^*(X'Y) \leq 1$.

Examples of penalized estimators

- Ridge regression (Hoerl & Kennard, 1970)
- $\hat{\beta} := \arg\min_{b\in\mathbb{R}^p} \frac{1}{2}\|Y - Xb\|_2^2 + \lambda\|b\|_2, \quad \lambda > 0$
- LASSO (Chen & Donoho, 1994, Tibshirani, 1996)
- $\hat{\beta}^{LASSO} := \arg\min_{b\in\mathbb{R}^p} \frac{1}{2}\|Y - Xb\|_2^2 + \lambda\|b\|_1, \quad \lambda > 0$
- SLOPE (Bogdan, van den Berg, Sabatti, Su, Candès, 2015)
- $\hat{\beta}^{SLOPE} := \arg\min_{b\in\mathbb{R}^p} \frac{1}{2}\|Y - Xb\|_2^2 + \sum_{i=1}^{p}\lambda_i|b|_{(i)}, \quad \lambda_1 > 0,$
  $\lambda_1 \geq \ldots, \lambda_p \geq 0, \ |b|_{(1)} \geq \ldots \geq |b|_{(p)}$

Least Absolute Shrinkage and Selection Operator (LASSO)

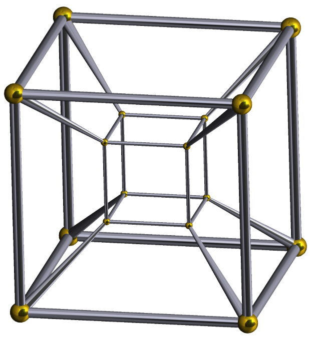
LASSO estimator (Chen & Donoho, 1994, Tibshirani, 1996) minimizes the ℓ^1-penalized Euclidean distance between Y and Xb:

$$\hat{\beta}^{LASSO} := \arg\min_{b\in\mathbb{R}^p} \frac{1}{2}\|Y - Xb\|_2^2 + \lambda\|b\|_1, \quad \lambda > 0.$$

- $\hat{\beta}^{LASSO}$ is well defined both for $n \geq p$ and $n < p$.
- $\partial_{\|\cdot\|_1}(b) = \mathrm{sign}(b)$.

LASSO dual ball = hypercube

- $J^*(b) = \|b\|_\infty$
- $B^* = B_\infty(0,\lambda) = [-\lambda,\lambda]$



Sorted ℓ^1 Penalized Estimator (SLOPE)

SLOPE (Bogdan, van den Berg, Sabatti, Su, Candès, 2015) minimizes the sorted ℓ^1 penalized Euclidean distance between Y and Xb:

$$\hat{\beta}^{SLOPE} := \arg\min_{b\in\mathbb{R}^p} \frac{1}{2}\|Y - Xb\|_2^2 + J_\Lambda(b).$$

- Sorted ℓ^1 norm: $J_\Lambda(b) := \sum_{i=1}^{p} \lambda_i |b|_{(i)}$, where $\lambda_1 > 0, \lambda_1 \geq \ldots, \lambda_p \geq 0$ and $|b|_{(1)} \geq \ldots \geq |b|_{(p)}$.
- $\hat{\beta}^{SLOPE}$ is well defined both for $n \geq p$ and for $n < p$.

SLOPE generalizes the previous approaches:

- $\lambda_1 = \ldots = \lambda_p = 0 \Rightarrow \hat{\beta}^{SLOPE} = \hat{\beta}^{OLS}$,
- $\lambda_1 = \ldots = \lambda_p > 0 \Rightarrow \hat{\beta}^{SLOPE} = \hat{\beta}^{LASSO}$.

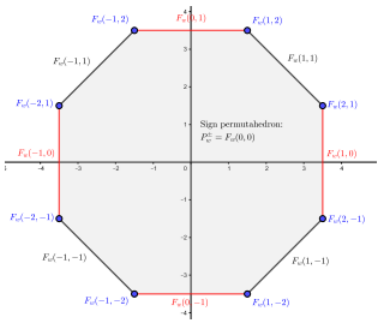
SLOPE dual ball = Signed permutahedron $P^{\pm}(\Lambda)$

The dual of sorted ℓ^1 norm is:

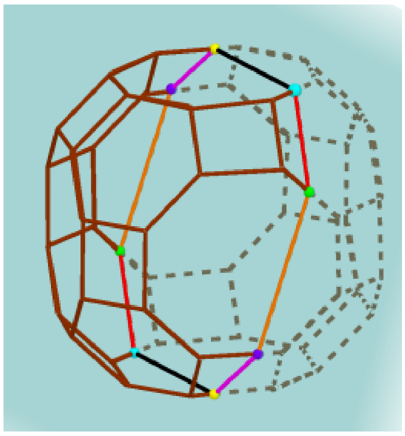
$$J^*_\Lambda(b) = \max\left\{\frac{|b|_{(1)}}{\lambda_1}, \frac{|b|_{(1)}+|b|_{(2)}}{\lambda_1+\lambda_2}, \dots, \frac{|b|_{(1)}+\cdots+|b|_{(p)}}{\lambda_1+\cdots+\lambda_p}\right\}.$$

The unit ball of J^*_Λ is the signed permutahedron $P^{\pm}(\Lambda)$:

$$P^{\pm}(\Lambda) = Conv\{(\pm\lambda_{\pi(1)}, \dots, \pm\lambda_{\pi(p)}) : \ \pi \in \mathcal{S}_p\}.$$

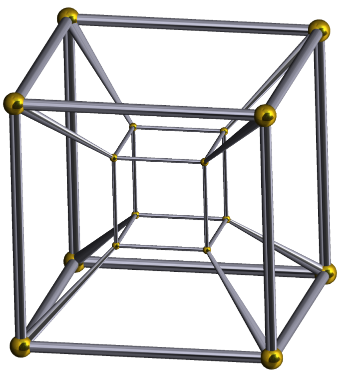


$P^{\pm}(\Lambda)$ in $\mathbb{R}^2$

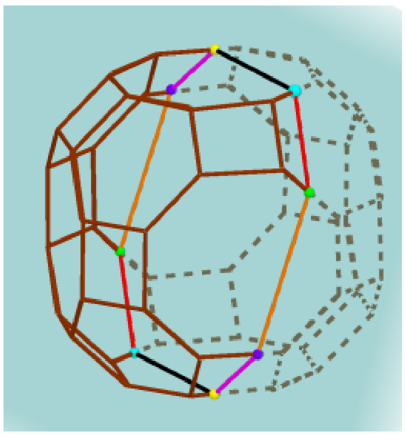


$P^{\pm}(\Lambda)$ in $\mathbb{R}^3$

Root systems and statistics



LASSO: A_1^p



SLOPE: B_p

SLOPE pattern

Definition

- The SLOPE pattern is a function $\mathrm{patt}\colon \mathbb{R}^p \to \mathbb{Z}^p$ defined by

$$\mathrm{patt}(b)_i = \mathrm{sign}(b_i)\mathrm{rank}(|b_i|), \qquad i = 1, \dots, p,$$

  where $\mathrm{rank}(|b_i|) \in \{1, 2, \dots, k\}$ is the rank of $|b_i|$ in a set of nonzero distinct values of $\{|b_1|, \dots, |b_p|\}$ (and $\mathrm{sign}(0) = 0$).
- Properties of $\mathrm{patt}(x)$:
  - $\mathrm{sign}(\mathrm{patt}(x)) = \mathrm{sign}(x)$ (sign preservation),
  - $|x_i| = |x_j| \implies |\mathrm{patt}(x)_i| = |\mathrm{patt}(x)_j|$ (clusters preservation),
  - $|x_i| > |x_j| \implies |\mathrm{patt}(x)_i| > |\mathrm{patt}(x)_j|$ (hierarchy preservation).

Example

$$x = (1.2,\ 1.2,\ 5,\ -5,\ 0,\ 3) \implies \mathrm{patt}(x) = (1,\ 1,\ 3,\ -3,\ 0,\ 2).$$

SLOPE vs. OLS

Theorem (Schneider & Tardivel, 2021)

For $n \geq p$ and $\ker(X) = \{0\}$ we have:
$\hat{\beta}^{OLS} - \hat{\beta}^{SLOPE} = \mathrm{Proj}(\hat{\beta}^{OLS})$ *on* $(X'X)^{-1}P^{\pm}(\Lambda)$.
For $p > n$:
$Y - X(\hat{\beta}^{OLS} - \hat{\beta}^{SLOPE}) = \mathrm{Proj}(\hat{\beta}^{OLS})$ *on* $(X'X)^{-1}\mathrm{row}(X) \cap P^{\pm}(\Lambda)$.

Theorem (Orthogonal design, $n \geq p$)

The orthogonal projection of $\hat{\beta}^{\mathrm{OLS}}$ on $P^{\pm}(\Lambda)$ is equal to $\hat{\beta}^{\mathrm{OLS}} - \hat{\beta}^{\mathrm{SLOPE}}$.

For LASSO: proven by Ewald and Schneider (2018).

SLOPE vs. OLS

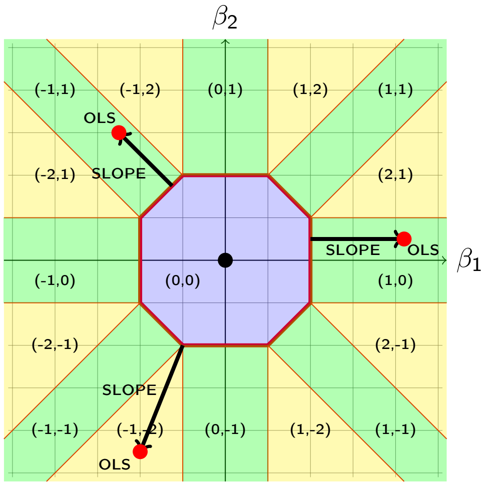


Figure: $\hat{\beta}^{SLOPE}$ and $\hat{\beta}^{OLS}$ in orthogonal design: $X'X = I_p$ for $\Lambda = (2,1)'$.

Simpler expression for SLOPE in orthogonal design: Tardivel, Servien and Concordet (2020).

$Y^{(n)} = X^{(n)}\beta + \varepsilon^{(n)}$

Consider the sequence of regression models: $Y^{(n)} = X^{(n)}\beta + \varepsilon^{(n)}$
with $\varepsilon^{(n)} \sim \mathcal{N}(0, \sigma^2 I_n)$.
No assumptions on relations between $\varepsilon^{(n)}$ and $\varepsilon^{(m)}$ for $n \neq m$.

Theorem

Assume that

$$\lim_n n^{-1}(X^{(n)})'X^{(n)} = C > 0.$$

Let $\widehat{\beta}_n^{SLOPE}$, $n \geq 1$, be the SLOPE estimator corresponding to the tuning vector $\Lambda^{(n)} = (\lambda_1^{(n)}, \lambda_2^{(n)}, \ldots, \lambda_p^{(n)})'$.

- *If $\lim_{n\to\infty} \dfrac{\lambda_1^{(n)}}{n} = 0$, then $\widehat{\beta}_n^{SLOPE} \xrightarrow{a.s.} \beta$.*
- *If $\lambda_0\|\beta\|_\infty > \beta' C\beta/2$ and $\lambda_1^{(n)}/n \to 0$, then $\hat{\beta}^{SLOPE}$ does not converge to β. Hence, $\hat{\beta}^{SLOPE}$ is not strongly consistent for β.*

$Y^{(n)} = X^{(n)}\beta + \varepsilon^{(n)}$

Theorem

Assume that

$$\lim_{n\to\infty} \frac{\lambda_1^{(n)}}{n} = 0$$

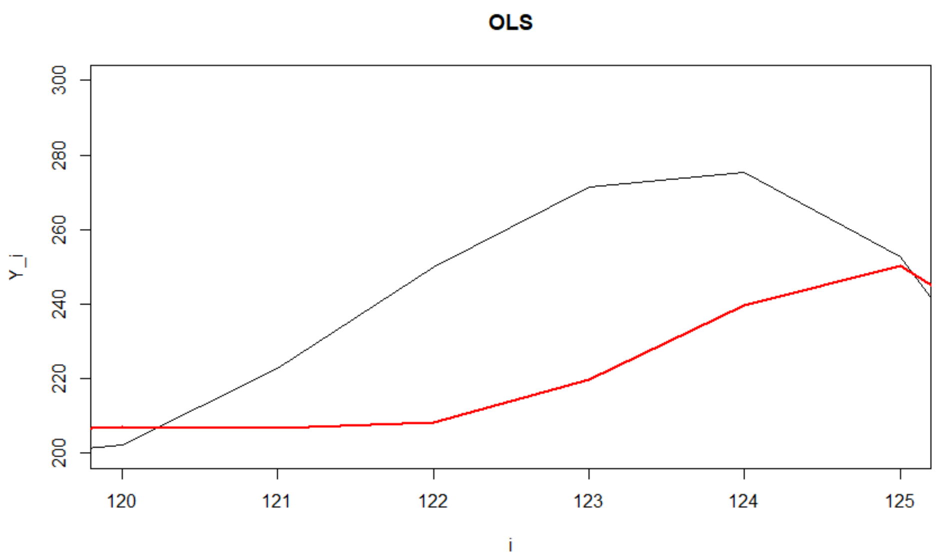
and that there exists $\delta > 0$ *such that*

$$\liminf_{n\to\infty} \frac{\lambda_i^{(n)} - \lambda_{i+1}^{(n)}}{\sqrt{n}\,(\log(n))^{1/2+\delta}} = m > 0 \quad \textit{for} \quad i = 1, \ldots, p-1.$$

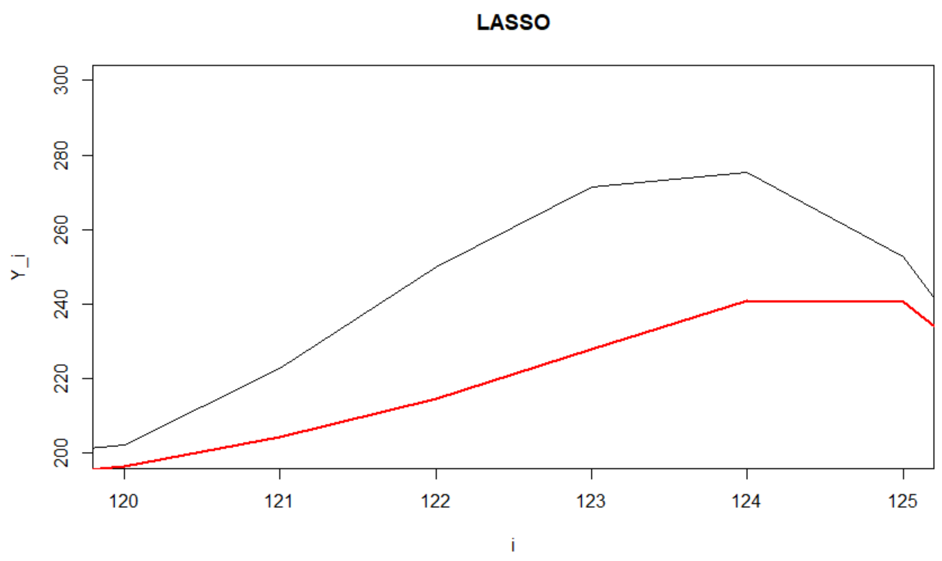
Then we have

$$\mathrm{patt}(\hat{\beta}_n^{SLOPE}) \xrightarrow{a.s.} \mathrm{patt}(\beta).$$

Application of SLOPE: signal denoising

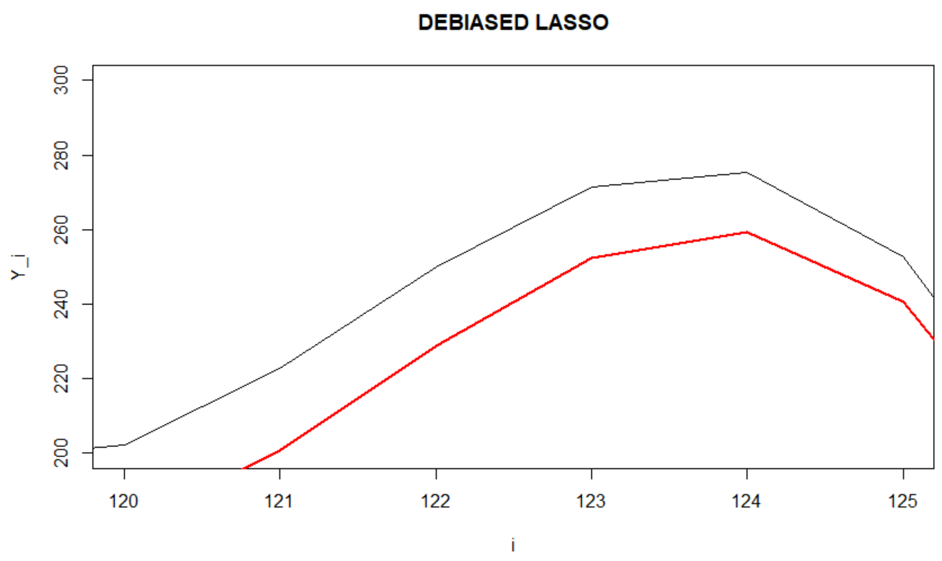


Application of SLOPE: signal denoising



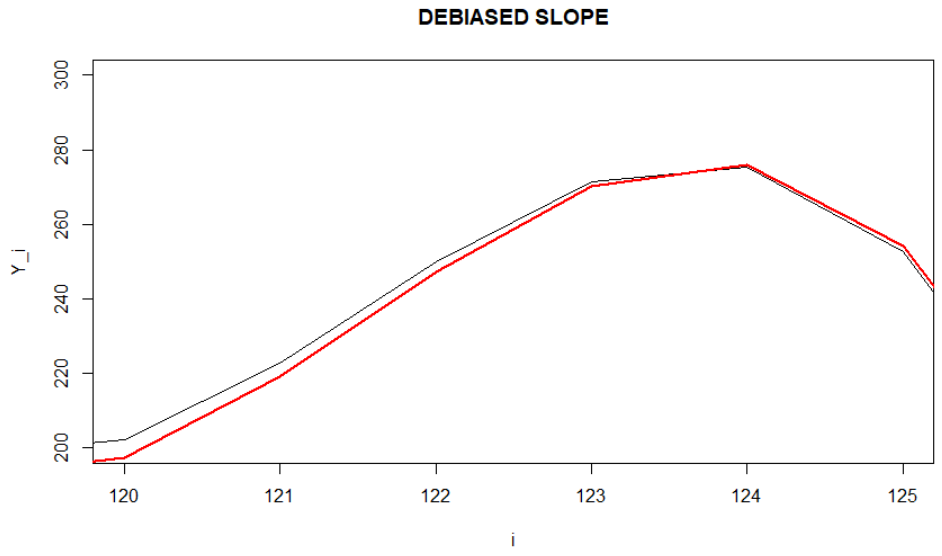
17 / 25

Application of SLOPE: signal denoising

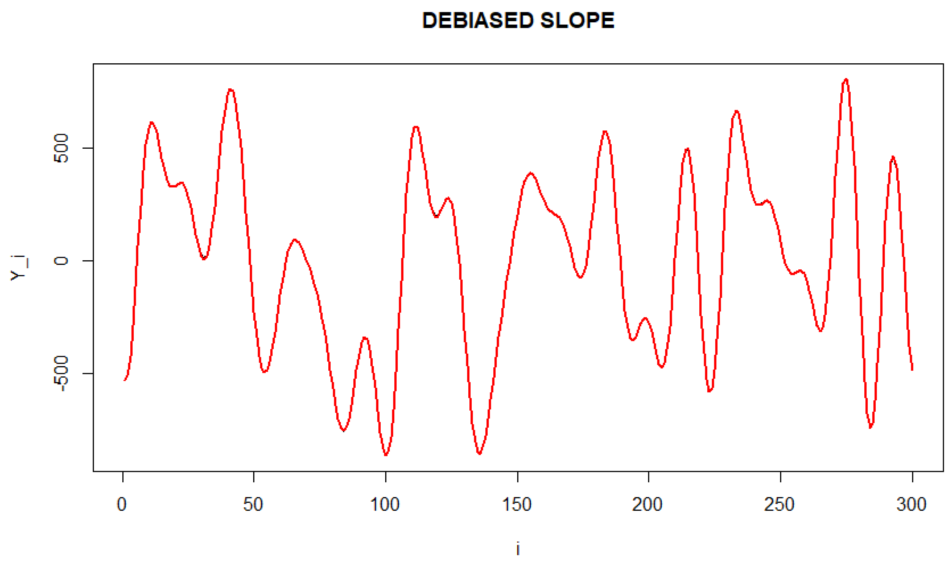


18 / 25

Application of SLOPE: signal denoising



Application of SLOPE: signal denoising

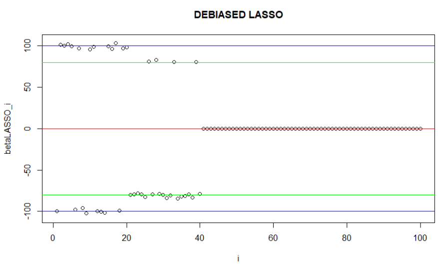


Application of SLOPE: signal denoising

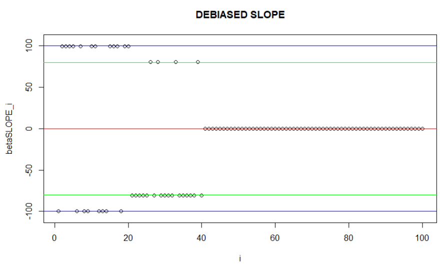
	OLS	LASSO-CV	LASSO-LS	SLOPE-LS
$MSE(,\cdot)$	613.6797	426.3705	171.7957	20.74967

Comparison of MSE between different regression methods

Application of SLOPE: pattern recovery



(a)



(b)

References

- M. Bogdan, E. van den Berg, C. Sabatti, W. Su, E. J. Candès, SLOPE – Adaptive Variable Selection Via Convex Optimization, Annals of Applied Statistics, vol.9, pp. 1103-1140, 2015.
- M. Bogdan, X. Dupuis, P. Graczyk, B. Kołodziejek, T. Skalski, P. Tardivel, M. Wilczyński. Pattern Recovery by SLOPE. ArXiv 2203.12086.
- S. Chen, D. Donoho. Basis pursuit. Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers, 1994, pp. 41-44 vol.1.
- K. Ewald, U. Schneider. Uniformly Valid Confidence Sets Based on the Lasso. EJS, vol. 12, pp. 1358-1387, 2018.
- C. F. Gauss. Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium, 1809.
- A.-M. Legendre. Nouvelles méthodes pour la détermination des orbites des comètes, Paris, Firmin Didot, 1805.
- U. Schneider, P. Tardivel. The geometry of uniqueness, sparsity and clustering in penalized estimation. ArXiv 2004.09106.

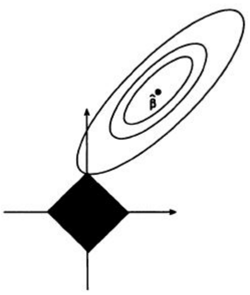
References

- T. Skalski, P. Graczyk, B. Kołodziejek, M. Wilczyński. Pattern recovery and signal denoising by SLOPE when the design matrix is orthogonal. ArXiv 2202.08573.
- P. Tardivel, R. Servien, D. Concordet. Simple expression of the LASSO and SLOPE estimators in low-dimension, Statistics 54, 340-352, 2020.
- P. Tardivel, T. Skalski, P. Graczyk, U. Schneider. The Geometry of Pattern Recovery by Penalized and Structured Estimators. 2021. hal-03262087.
- R. Tibshirani. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), vol. 58, no. 1, [Royal Statistical Society, Wiley], 1996, pp. 267–88.
- R. J. Tibshirani, J. Taylor. The solution path of the generalized lasso. Annals of Statistics, vol. 39, no. 3, pp. 1335-1371, 2011.
- M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). IEEE Trans. Inform. Theory, vol. 55, no. 5, pp. 2183-2202, 2009.
- P. Zhao, B. Yu. On model selection consistency of Lasso. Journal of Machine Learning Research, vol. 7, pp. 2541-2563, 2006.
- H. Zou. The Adaptive Lasso and Its Oracle Properties, Journal ASA 2006.

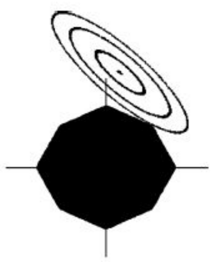
Domo arigato gozaimasu!

Appendix: Pictures from the Title Page

Meeting point of scaled B and scaled unit ball in ℓ^2 of $(Y - Xb)$ is equal to $\hat{\beta}$.



$\mathrm{sign}(\hat{\beta}^{LASSO}) = (0, +)$



$\mathrm{patt}(\hat{\beta}^{SLOPE}) = (1, 1)$

Appendix: Subdifferential

Definition (Subgradient)

Let $f : \mathbb{R}^p \mapsto \mathbb{R}$. Then g is a subgradient of f at b if

$$\forall h \in \mathbb{R}^p \quad f(b+h) \geq f(b) + g'h.$$

Definition (Subdifferential)

The subdifferential $\partial f(b)$ of f at b is the set of all subgradients of f at b.

Appendix: Thresholded estimator

Definition (Thresholded penalized least squares estimator)

Let pen be a penalizer, $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$ and $\lambda > 0$. Given $\hat{\beta} \in S_{X,\lambda\mathrm{pen}}(y)$, we say that $\hat{u}$ is a thresholded estimator of $\hat{\beta}$ if $\partial_{\mathrm{pen}}(\hat{\beta}) \subset \partial_{\mathrm{pen}}(\hat{u})$.

Definition (Thresholded LASSO)

$$\hat{\beta}_i^{\mathrm{LASSO},\tau} = \begin{cases} \hat{\beta}_i^{\mathrm{LASSO}}, & \text{if } |\hat{\beta}_i^{\mathrm{LASSO}}| > \tau, \\ 0, & \text{otherwise.} \end{cases}$$

Appendix: LASSO and SLOPE in orthogonal design

Theorem (Tibshirani, 1996)

Exact formula for LASSO in orthogonal ($X'X = I$) design:

$$\hat{\beta}_i^{LASSO} = \mathrm{sign}(\hat{\beta}_i^{OLS}) \max\{|\hat{\beta}_i^{OLS}| - \lambda, 0\}.$$

Theorem (Tardivel, Servien, Concordet (2020))

Let $|\hat{\beta}^{OLS}|_{(1)} \geq \ldots \geq |\hat{\beta}^{OLS}|_{(p)}$. *Let* $\hat{S}_k := \sum_{i=1}^{k} (|\hat{\beta}^{OLS}|_{(i)} - \lambda_i)$. *Denote a partition* $(k_1, k_2, \ldots, k_s = p)$ *of* $\{1, 2, \ldots, p\}$ *such that* $k_i := \max\{\arg\max_{k > k_{i-1}} \{\frac{\hat{S}_k - \hat{S}_{k-1}}{k - k_{i-1}}\}\}$ *with* $k_0 = \hat{S}_0 = 0$. *Then* $\hat{\beta}_i^{ols} \cdot \hat{\beta}_i^{slope} \geq 0$ *and* $|\hat{\beta}^{slope}|$ *is given by*

$$\left(k_1 \text{ terms} \left(\frac{\hat{S}_{k_1}}{k_1} \right)_+, \ldots, (k_s - k_{s-1}) \text{ terms} \left(\frac{\hat{S}_{k_s} - \hat{S}_{k_{s-1}}}{k_s - k_{s-1}} \right)_+ \right).$$

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).

Likelihood Geometry of Correlation Models

Carlos Enrique Améndola Cerón
(Technical University of Berlin)

OCAMI: Mathematical optimization and statistical theories using geometric methods

October 20, 2022

Setup / Introduction

- $\mathbb{S}^n_+$ real symmetric positive definite $n \times n$ matrices
- *Model*: $M \subseteq \mathbb{S}^n_+$, and *Data*: $S \in \mathbb{S}^n_+$
- What is the 'best' point $\Sigma^* \in M$ that explains S?
- Gaussian *ML* estimation:

$$\hat{\Sigma} = \underset{\Sigma \in M}{\arg\max} \ \log\det(\Sigma^{-1}) - \mathrm{tr}(\Sigma^{-1}S)$$

- Can be seen as minimizing the divergence $\mathcal{I}(S \| \Sigma)$, where

$$\mathcal{I}(\Sigma_1, \Sigma_2) = \mathrm{tr}(\Sigma_1\Sigma_2^{-1}) - \log\det(\Sigma_1\Sigma_2^{-1}) - n$$

- # complex critical points for generic S: *ML degree*
- In this talk: M consists of *correlation* matrices, i.e. $\Sigma_{ii} = 1 \ \forall i$

Motivating Example: Bivariate Correlations

- Let $M \subset \mathbb{S}^2_+$ consist of 2×2 correlation matrices:

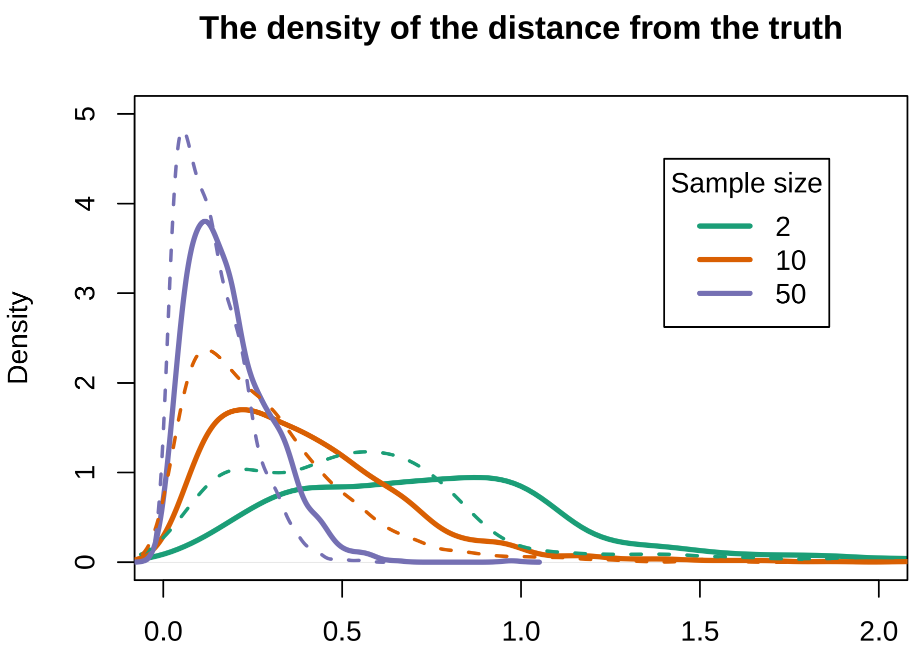
$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \qquad K = \Sigma^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \qquad S = \begin{pmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{pmatrix}$$

  where $-1 < \rho < 1$.
- Finding the MLE corresponding to $\hat{\rho}$ reduces to solving a *cubic* equation [Kendall, Stuart, 1961 "*Advanced Theory of Statistics*"]:

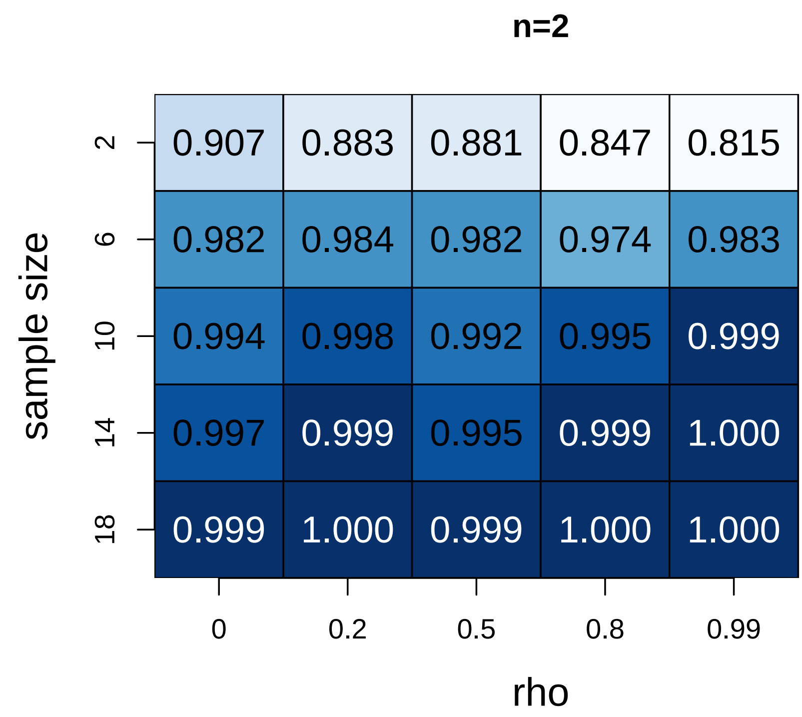
$$\rho^3 - s_{12}\rho^2 + (s_{11} + s_{22} - 1)\rho - s_{12} = 0$$

- ML degree is *3*. There could potentially be three positive definite solutions with a multimodal likelihood function $\ell(\Sigma)$.
- How often does this happen? How bad can it be?

A statistical perspective



Probability of one real critical point



Carlos Améndola | Likelihood Geometry of Correlation Models

Case Study: Bivariate Correlations

- Let $a = \frac{s_{11}+s_{22}}{2}$ and $b = s_{12}$.
- Note that if $S \in \mathbb{S}^2_+$ then $a > 0$ and $|a| > |b|$.
- It holds that $\frac{d}{d\rho}\mathcal{I}(S\|\Sigma) = \frac{2}{(1-\rho^2)^2} f(\rho)$, where

$$f(\rho) = \rho^3 - b\rho^2 - (1-2a)\rho - b.$$

- $f(-1) = -2(a+b) < 0$ and $f(1) = 2(a-b) > 0 \implies$ at least one real root in $(-1,1)$.
- The *discriminant* of f is

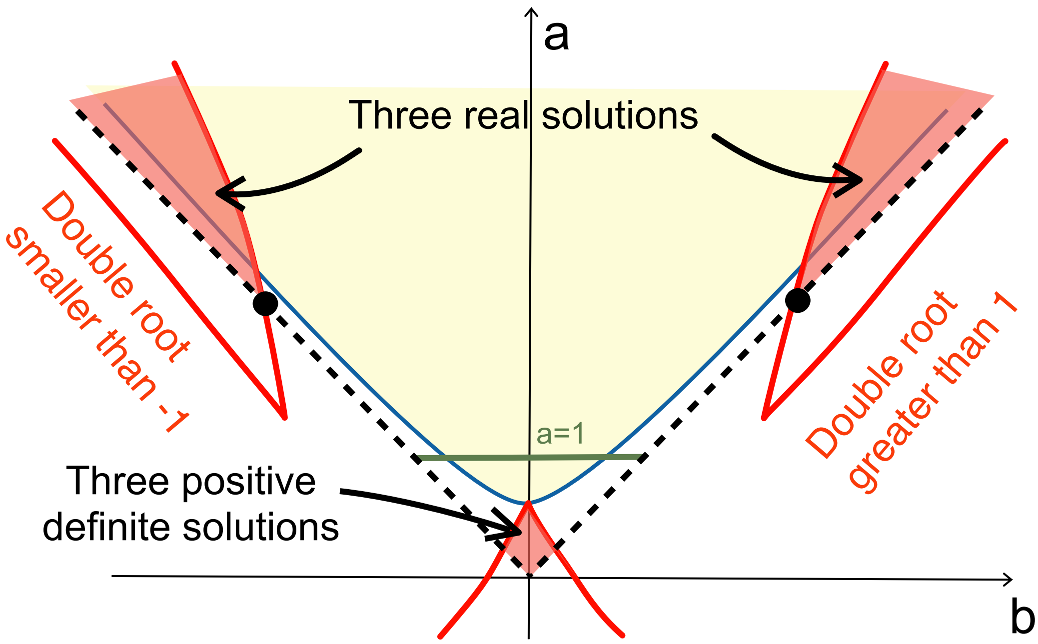
$$\Delta_f(a,b) = -4[b^4 - (a^2+8a-11)b^2 + (2a-1)^3].$$

- f has a single real zero $\iff \Delta_f(a,b) < 0$.
- However, we are more interested in:

  when does f have a single critical point in $(-1,1)$?

Carlos Améndola | Likelihood Geometry of Correlation Models

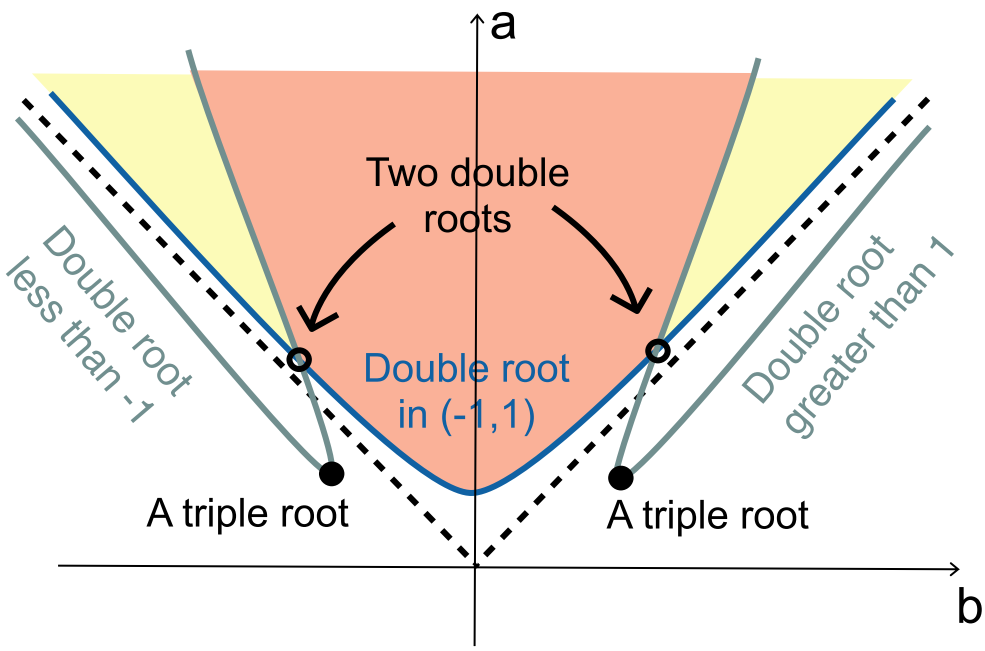
Likelihood Geometry for Bivariate Correlations



Case Study: Bivariate Correlations

- Data matrix
$$S = \begin{pmatrix} a & b \\ b & a \end{pmatrix}$$
where $a > |b| > 0$.
- It holds that $\frac{d^2}{d\rho^2}\mathcal{I}(S||\Sigma) = \frac{2}{(1-\rho^2)^3}g(\rho)$, where
$$g(\rho) = \rho^4 - 2b\rho^3 + 6a\rho^2 - 6b\rho + 2a - 1.$$
- $g(-1) = 8(a+b) > 0$ and $g(1) = 8(a-b) > 0$.
- The *discriminant* of g is
$$\Delta_g = -256\Big(27b^6 - 27(2a^2 + 6a - 5)b^4 + 9(3a^4 + 36a^3 - 32a^2 + 8a + 1)b^2 - (2a-1)(9a^2 - 2a + 1)^2\Big)$$
- If $g(\rho) > 0$ for all $\rho \in \mathbb{R}$ (globally convex) $\Longrightarrow \Delta_g(a,b) \geq 0$.
- However, we are more interested in:
when is g nonnegative in $(-1,1)$?

Convexity Analysis



Alternative Loss Functions

From the divergence

$$\mathcal{I}(\Sigma_1, \Sigma_2) = \operatorname{tr}(\Sigma_1 \Sigma_2^{-1}) - \log\det(\Sigma_1 \Sigma_2^{-1}) - n$$

- $\mathcal{I}(\Sigma_1, \Sigma_2) \geq 0$ and is zero if and only if $\Sigma_1 = \Sigma_2$.
- strictly convex in Σ_1 and in Σ_2^{-1}

Fix $S \in \mathbb{S}_+^n$:

1. entropy loss: $\mathcal{I}(S \| \Sigma)$ (minimizer $\hat{\Sigma}$ is *MLE*)
2. Stein's loss: $\mathcal{I}(\Sigma \| S)$ (minimizer $\check{\Sigma}$ is *dual MLE*)
3. symmetrized Stein's loss:

$$L(\Sigma, S) = \frac{1}{2}\left(\mathcal{I}(S \| \Sigma) + \mathcal{I}(\Sigma \| S)\right)$$

(2) and (3) are *strictly convex* in Σ and optimizers are uniquely defined

Full correlation model

Let $M \subset \mathbb{S}^n_+$ be the space of all correlation matrices: $\Sigma_{ii} = 1$ for all $1 \le i \le n$. First order optimality conditions give that the optimum is a correlation matrix $\Sigma = K^{-1}$ satisfying for each $i \neq j$:

1. entropy loss (MLE):
$$K_{ij} = (KSK)_{ij}$$
2. Stein's loss (dual MLE):
$$K_{ij} = (S^{-1})_{ij}$$
3. symmetrized Stein's loss:
$$(KSK)_{ij} = (S^{-1})_{ij}$$

Algebraic Degrees

For the bivariate correlation model $n = 2$, [Brownlees, Llorens-Terrazas (2020)] observed that the dual MLE can be given in closed form (solving a *quadratic* equation!).
From our computations, for $n > 1$ one has

$$\mathrm{dMLdeg}(n) < \mathrm{MLdeg}(n) < \mathrm{SSLdeg}(n)$$

n	1	2	3	4	5	6	7	8	9
SSL degree	1	4	28	292	?	?	?	?	?
ML degree	1	3	15	109	1077	13695	?	?	?
dual ML degree	1	2	5	14	43	144	522	2028	8357

For $n > 4$, computed with the package `LinearCovarianceModels.jl`

how are these numbers growing?

Equicorrelation Model

The model M now consists of all $\Sigma \in \mathbb{S}^n_+$ such that

$$\Sigma_{ii} = 1 \quad \Sigma_{ij} = \rho \text{ for } i \neq j.$$

This means that ρ is restricted to $\frac{-1}{n-1} < \rho < 1$.
Let $W = S^{-1}$. We can exploit the symmetry and set:

$$\overline{W} = \frac{1}{n!} \sum_{P \in \mathcal{S}_n} PWP^T$$

Theorem (Am., Zwiernik (2021))

For the equicorrelation model, the dual ML degree is always 2 for every $n > 1$. The dual MLE $\check{\Sigma}$ admits the explicit form

$$\check{\rho} \;=\; \frac{1 + (n-2)\bar{w} \pm \sqrt{(n\bar{w}+1)^2 - 4\bar{w}}}{2(n-1)\bar{w}}.$$

where $\bar{w}$ is the off-diagonal entry of $\overline{W}$.

Equicorrelation Model

The model M now consists of all $\Sigma \in \mathbb{S}^n_+$ such that

$$\Sigma_{ii} = 1 \quad \Sigma_{ij} = \rho \text{ for } i \neq j.$$

This means that ρ is restricted to $\frac{-1}{n-1} < \rho < 1$.
We can exploit the symmetry and set:

$$\overline{S} = \frac{1}{n!} \sum_{P \in \mathcal{S}_n} PSP^T$$

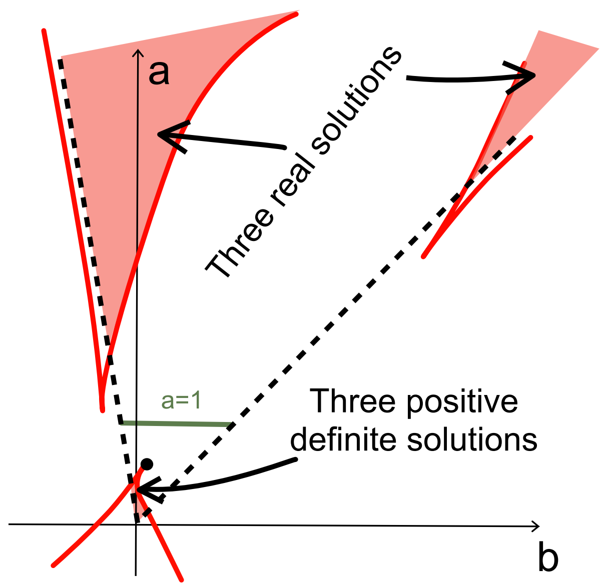
Theorem (Am., Zwiernik (2021))

For the equicorrelation model, the ML degree is always 3 for every $n > 1$. The MLE $\hat{\Sigma}$ satisfies

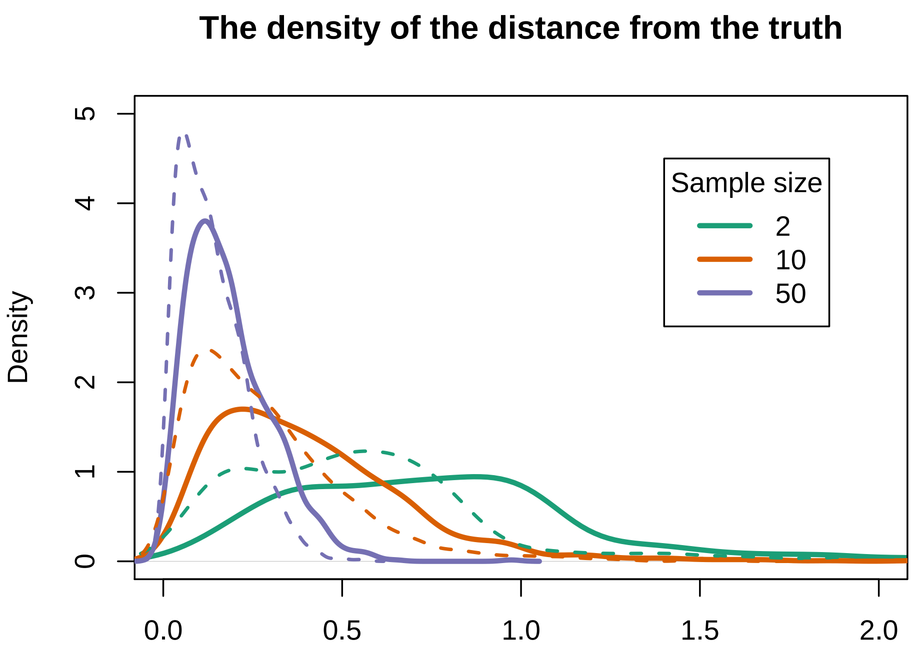
$$(n-1)\rho^3 + ((n-2)(a-1) - (n-1)b)\rho^2 + (2a-1)\rho - b \;=\; 0.$$

where a, b are the diagonal and off-diagonal entries of $\overline{S}$, respectively. The SSL degree is always 4 for every $n > 1$.

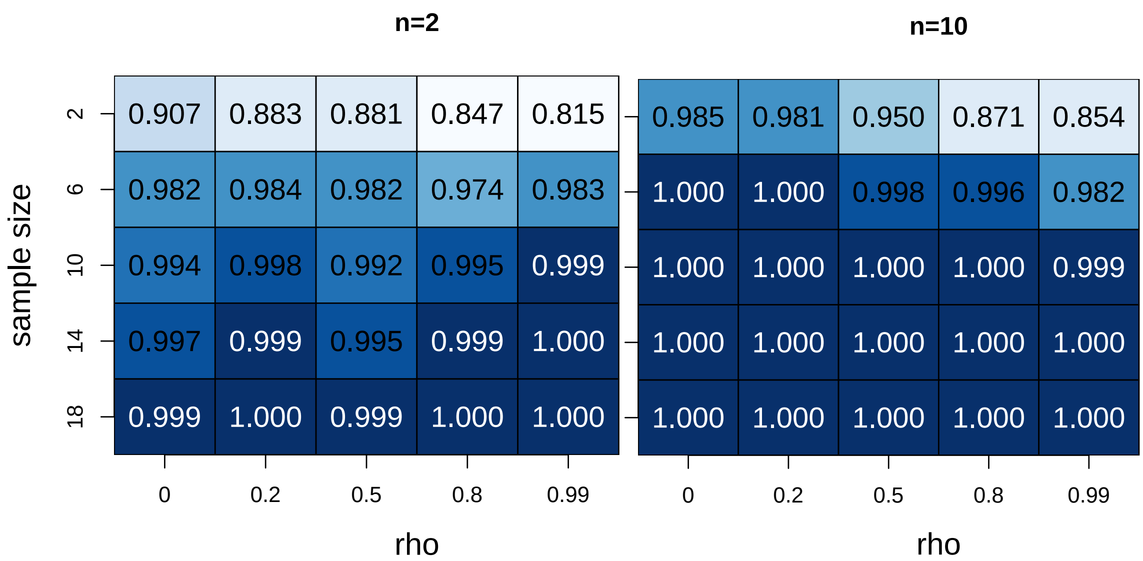
Equicorrelation for $n > 2$



A statistical perspective



A statistical perspective



Summary

- Rich likelihood geometry behind correlation models.
- High ML degree may hint to problematic optimization, but careful analysis shows likelihood function is well-behaved over large regions.
- Introduction of another algebraic complexity measure: *SSL degree*.
- Dual MLE appears to behave best algebraically, how do degrees grow?
- Plenty of relevant submodels (e.g. symmetries) still to be explored.
- Main Reference:
  Améndola, C., & Zwiernik, P., *Likelihood Geometry of Correlation Models*. (2021) *Le Matematiche*, 76(2), pp. 559 - 583.

ありがとうございました!

Carlos Améndola Likelihood Geometry of Correlation Models

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)

Mixed convex exponential families and locally associated graphical models

Piotr Zwiernik

University of Toronto

This story is part of the following paper:

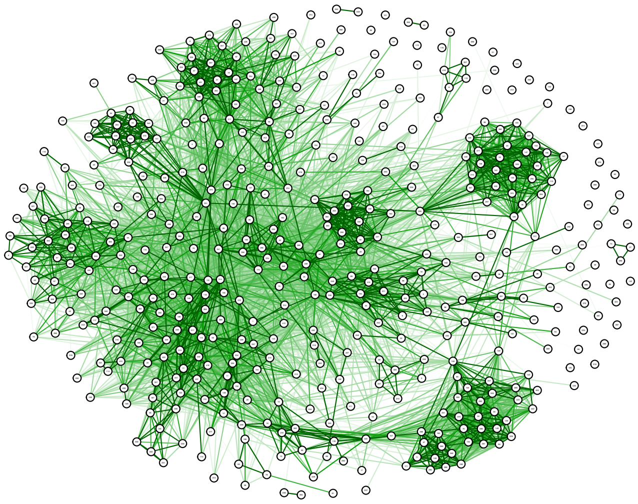
Lauritzen S., & Zwiernik, P., *Locally associated graphical models and mixed convex exponential families*. arXiv:2008.04688.

OCAMI Meeting
21(20) October 2022

Modelling with positive dependence

Example: S&P 500

graphical lasso estimate of the graph:



Note: All edges green (positive partial correlations).

Gaussian totally positive distributions

The zero-mean Gaussian distribution

$$f(x) = \frac{1}{(2\pi)^{d/2}} \sqrt{\det K} \exp(-x^T K x/2)$$

Totally positive: $K = \Sigma^{-1}$ satisfies $K_{ij} \leqslant 0$ for all $i \neq j$. (K is an M-matrix)

- $K_{ij} \leqslant 0$ if and only if $\operatorname{corr}(X_i, X_j | X_{V \setminus \{i,j\}}) \geqslant 0$.

A success story

In some applications it works incredibly well.

Rossell&Zwiernik describe a S&P500 dataset:

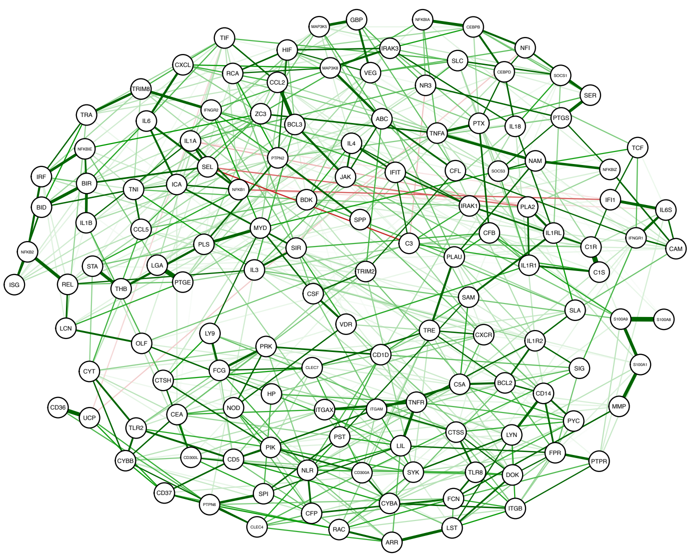
- Our MLE gives a sparser graph and higher likelihood than the best GLASSO estimate!

see also: Agrawal, Roy, Uhler. *Covariance Matrix Estimation under Total Positivity for Portfolio Selection*, 2019.

3

However: Gene expression data

Partial correlations with negative signs additionally penalized.



4

Motivation: Locally associated GGMs

X is associated if $\mathrm{cov}(\mathbf{f}(X), \mathbf{g}(X)) \geqslant \mathbf{0}$ for any $\mathbf{f}, \mathbf{g} : \mathbb{R}^{\mathbf{d}} \to \mathbb{R}$ nondecreasing.

Pitt: A Gaussian X is associated if and only if $\mathbf{\Sigma} \geqslant \mathbf{0}$.

Gaussian graphical model: $\mathbf{X} \sim \mathbf{N_d}(\mathbf{0}, \mathbf{\Sigma})$:

$$\mathbf{M}(\mathbf{G}) = \{\mathbf{\Sigma} \in \mathrm{PD}_{\mathbf{d}} : (\mathbf{\Sigma}^{-1})_{\mathbf{ij}} = \mathbf{0} \text{ for } \mathbf{ij} \notin \mathbf{G}\}.$$

With additional positivity:

$$\mathbf{P}(\mathbf{G}) = \{\mathbf{\Sigma} \in \mathrm{PD}_{\mathbf{d}} : \mathbf{\Sigma}_{\mathbf{ij}} \geqslant \mathbf{0} \text{ for } \mathbf{ij} \in \mathbf{G}\}.$$

Estimation in laGGMs

The log-likelihood ($\mathbf{S}$ sample covariance matrix)

$$\log\det(\mathbf{\Sigma}^{-1}) - \mathrm{tr}(\mathbf{S}\mathbf{\Sigma}^{-1})$$

is concave in $\mathbf{K} = \mathbf{\Sigma}^{-1}$ but not in $\mathbf{\Sigma}$.

Alternative: mixed dual estimate (MDE).

• MDE for mixed convex exponential families is easier to obtain and has the same asymptotics as the MLE.

2-stage estimation procedure

Information divergence (convex in $\boldsymbol{\Sigma}_1$ and in $\mathbf{K}_2$):

$$\mathbf{I}(\boldsymbol{\Sigma}_1 || \mathbf{K}_2) = \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_1\mathbf{K}_2 - \mathbf{I}) - \frac{1}{2}\log\det(\boldsymbol{\Sigma}_1\mathbf{K}_2).$$

$\mathbf{S}$ sample covariance, $S \xrightarrow{1} \widehat{K} \xrightarrow{2} \check{\Sigma}$

1. $\hat{\mathbf{K}}$ minimizer of $\mathbf{I}(\mathbf{S}||\mathbf{K})$ subject to $\mathbf{K} \in \mathbf{M}(\mathbf{G})$.
2. $\check{\boldsymbol{\Sigma}}$ minimizer of $\mathbf{I}(\boldsymbol{\Sigma}||\hat{\mathbf{K}})$ subject to $\boldsymbol{\Sigma} \in \mathbf{P}(\mathbf{G})$.

Note: $\check{\boldsymbol{\Sigma}} \in \mathbf{M}(\mathbf{G})$ and it is a reasonable estimator.

7

Mixed convex exponential families

Regular exponential families

Exponential family $\mathcal{E}$ over $\mathcal{X}$ wrt measure ν

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \exp\{\langle\boldsymbol{\theta}, \boldsymbol{t}(\boldsymbol{x})\rangle - \mathbf{A}(\boldsymbol{\theta})\} \qquad \text{for } \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k,$$

The set of canonical parameters

$$\boldsymbol{\Theta} := \operatorname{int}\left\{\boldsymbol{\theta} \in \mathbb{R}^k : \int_{\mathcal{X}} \exp\{\langle\boldsymbol{\theta}, \boldsymbol{t}(\boldsymbol{x})\rangle\}\, \nu(\mathrm{d}\boldsymbol{x}) < \infty\right\}.$$

In steep exponential families :

- $\boldsymbol{\Theta}$ convex subset of $\mathbb{R}^k$,
- $\mathbf{A}(\boldsymbol{\theta})$ strictly convex, smooth over Θ,
- $\|\nabla\mathbf{A}(\theta)\| \to \infty$ at the boundary.

8

Mixed parametrizations

The split $\boldsymbol{t}(\boldsymbol{x}) = (\boldsymbol{u}(\boldsymbol{x}), \boldsymbol{v}(\boldsymbol{x})) \in \mathbb{R}^{\mathbf{k}}$ induces splits
$\boldsymbol{\theta} = (\boldsymbol{\theta}_{\mathbf{u}}, \boldsymbol{\theta}_{\mathbf{v}}) \in \boldsymbol{\Theta}, \qquad \boldsymbol{\mu} = (\boldsymbol{\mu}_{\mathbf{u}}, \boldsymbol{\mu}_{\mathbf{v}}) \in \mathbf{M}.$
($\boldsymbol{\Theta}$ canonical parameters, $\mathbf{M}$ mean parameters)

$\mathbf{M}_{\mathbf{u}}$ = projection of $\mathbf{M}$ on $\boldsymbol{\mu}_{\mathbf{u}}$
$\boldsymbol{\Theta}_{\mathbf{v}}$ = projection of $\boldsymbol{\Theta}$ on $\boldsymbol{\theta}_{\mathbf{v}}$

Theorem (Barndorff-Nielsen, Mixed Parametrization):

- $(\boldsymbol{\mu}_{\mathbf{u}}, \boldsymbol{\theta}_{\mathbf{v}})$ forms an alternative parametrization.
- $(\boldsymbol{\mu}_{\mathbf{u}}, \boldsymbol{\theta}_{\mathbf{v}}) \in \mathbf{M}_{\mathbf{u}} \times \boldsymbol{\Theta}_{\mathbf{v}}$ (variational independence)

9

Mixed convex exponential family

Fix mixed parametrization $(\boldsymbol{\mu}_{\mathbf{u}}, \boldsymbol{\theta}_{\mathbf{v}}) \in \mathbf{M}_{\mathbf{u}} \times \boldsymbol{\Theta}_{\mathbf{v}}$ of $\mathcal{E}$.

Mixed convex exponential family:

- $\mathbf{M}'_{\mathbf{u}} \times \boldsymbol{\Theta}'_{\mathbf{v}} \subseteq \mathbf{M}_{\mathbf{u}} \times \boldsymbol{\Theta}_{\mathbf{v}}$
- $\mathbf{M}'_{\mathbf{u}} \subseteq \mathbf{M}_{\mathbf{u}}$, $\boldsymbol{\Theta}'_{\mathbf{v}} \subseteq \boldsymbol{\Theta}_{\mathbf{v}}$ rel. closed convex subsets.

Example: Locally associated Gaussian distributions form a mixed convex exponential family.

Example: The Gaussian case

Sufficient statistics: $\boldsymbol{t}(\boldsymbol{x}) = -\frac{1}{2}\boldsymbol{x}\boldsymbol{x}^T$,

Canonical/mean parameters: $\boldsymbol{\theta} = K$, $\boldsymbol{\mu} = -\frac{1}{2}\Sigma$

Gradient map: $\mathbf{A}(\mathbf{K}) = -\frac{1}{2}\log\det K$, $\nabla\mathbf{A}(\mathbf{K}) = -\frac{1}{2}K^{-1}$.

e.g. in locally associated Gaussian graphical models:

- $K_{ij} = \theta_{ij} = (\Sigma^{-1})_{ij} = 0$ for $ij \notin G$, and
- $\Sigma_{ij} = -2\mu_{ij} \geqslant 0$ for $ij \in G$.

So this is a mixed convex exponential family.

see also **Gaussian Double Markovian Distributions** by Boege, Kahle, Kretschmer, Rötger (arXiv:2107.00134)

This leads to an interesting observation:

Fix positive definite $\mathbf{d} \times \mathbf{d}$ matrices $\mathbf{A}, \mathbf{B}$.

For any set $\mathcal{I}$ of indices there **exists** a **unique** positive definite matrix $\mathbf{\Sigma}$ such that:

- $\mathbf{\Sigma_{ij}} = \mathbf{A_{ij}}$ for $(i, j) \in \mathcal{I}$;
- $(\mathbf{\Sigma}^{-1})_{\mathbf{ij}} = \mathbf{B_{ij}}$ for $(i, j) \notin \mathcal{I}$.

12

Kullback-Leibler divergence

Fenchel conjugate: $\mathbf{A}^*(\boldsymbol{\mu}) \;=\; \sup\{\ell(\overline{\boldsymbol{\theta}}; \boldsymbol{\mu}) : \overline{\boldsymbol{\theta}} \in \mathbb{R}^{\mathbf{k}}\}$.

Two distributions in $\mathcal{E}$: one with mean parameter $\boldsymbol{\mu}^{(1)} \in \mathbf{M}$, the other with canonical parameter $\boldsymbol{\theta}^{(2)} \in \mathbf{\Theta}$.

$$K(\boldsymbol{\mu}^{(1)}, \boldsymbol{\theta}^{(2)}) \;=\; -\langle \boldsymbol{\mu}^{(1)}, \boldsymbol{\theta}^{(2)} \rangle + \mathbf{A}^*(\boldsymbol{\mu}^{(1)}) + \mathbf{A}(\boldsymbol{\theta}^{(2)})$$

Note: K is strictly convex both in $\boldsymbol{\mu}^{(1)}$ and in $\boldsymbol{\theta}^{(2)}$.

13

Mixed dual estimator

Mixed exponential family: $(\mu_{\mathbf{u}}, \theta_{\mathbf{v}}) \in \mathbf{M}'_{\mathbf{u}} \times \mathbf{\Theta}'_{\mathbf{v}}$.

Sufficient statistics $t = \frac{1}{n}\sum_{i=1}^{n} t(\mathbf{X}^{(i)}) = (u, v)$.

Two-step procedure:

(S1) $\widehat{\theta} := \arg\min K(t, \theta)$ over θ s.t. $\theta_{\mathbf{v}} \in \mathbf{\Theta}'_{\mathbf{v}}$.

(S2) $\check{\mu} := \arg\min K(\mu, \widehat{\theta})$ over μ s.t. $\mu_{\mathbf{u}} \in \mathbf{M}'_{\mathbf{u}}$.

Some properties:

- **Theorem:** $\check{\mu}$ lies in the mixed convex family.
- $\check{\mu}$ exists if and only if $\widehat{\theta}$ exists,
- if exists, it is unique (convexity),

14

Summary + bibliography + thank you!

We study submodels of exponential families where the model constraints are convex in the mixed parameters.

Our main motivation is in local association.

The likelihood function is not concave so the MLE may be complicated to compute.

We propose a simple and sensible alternative.

This story is part of the following paper:

Lauritzen S., & Zwiernik, P., *Locally associated graphical models and mixed convex exponential families.* To appear in Annals of Statistics.

15

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.

Classification problem of invariant q-exponential families on homogeneous spaces

Koichi Tojo[1], joint work with Taro Yoshino[2]

[1]RIKEN Center for Advanced Intelligence Project, Tokyo, Japan,

[2]Graduate School of Mathematical Science, The University of Tokyo

Octorber 21, 2022

Introduction
Step 1: G/H-method
Step 2: q-deformation
Another topic: natural projection

Contents

Problem

Aim(rough)

We want to know all the "good" families of distributions on important spaces.

Mathematically, let G be a Lie group, H a closed subgroup of G and G/H the homogeneous space of G. Take $q \in \mathbb{R}$.

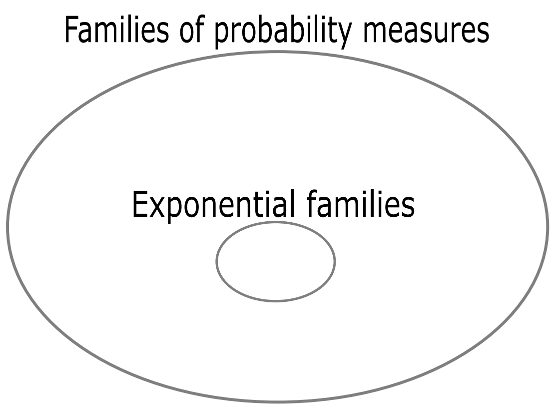
Problem 1.1.

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

A family of probability measures and machine learning

Learning by using a family of probability measures is one of important methods in the field of machine learning.

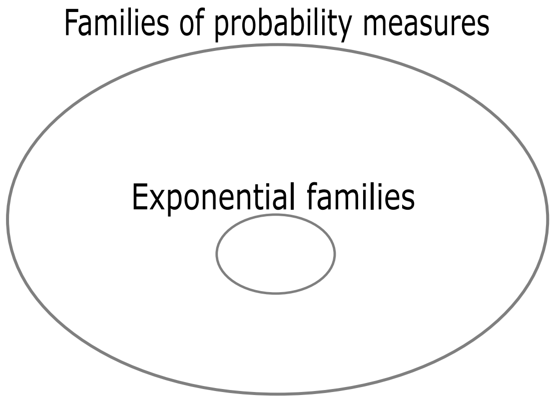
Learning=to optimize the parameters in the family of probability measures



Exponential family

Exponential family

- Exponential families are important subject in the field of information geometry.
- Exponential families are useful for Bayesian inference.
- Exponential families include many widely used families.



Examples (exponential families)

Table: Examples of exponential families

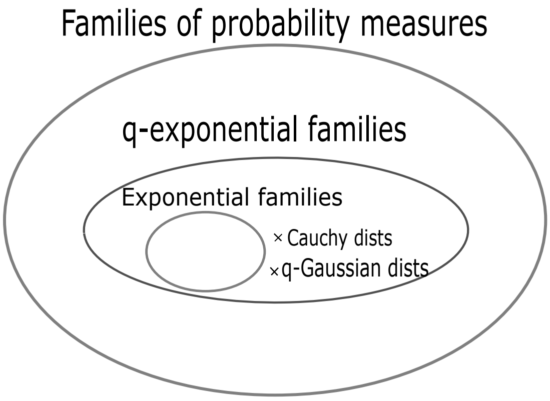
distributions	sample sp. X
Normal	$\mathbb{R}$
Multivariate normal	$\mathbb{R}^n$
Bernoulli	$\{\pm 1\}$
Categorical	$\{1, \cdots, n\}$
Gamma	$\mathbb{R}_{>0}$
Inverse gamma	$\mathbb{R}_{>0}$
Wishart	$\mathrm{Sym}^+(n, \mathbb{R})$
Von Mises	S^1
Poincaré	$\mathcal{H}$

q-exponential family

q-exponential family ($q \in \mathbb{R}$)

q-exponential family

- is a generalization of exponential family ($q = 1$).
- is also important subject in the field of information geometry.
- is useful for Tsallis statistics.



Relation

	exponential family	q-exponential family
Amari's α-family	$\alpha = 1$	$\alpha = 2q - 1$
Entropy	• Shannon entropy • maximization with expected value constraint	• Tsallis entropy • extremization with q-expected value constraint

Definition of q-exponential family

X: manifold, $\mathcal{R}(X)$: the set of all Radon measures on X.

Definition 1.2 (q-exponential family).

$\mathcal{P} \subset \mathcal{R}(X)$ is an q-exponential family on X if there exists a triple (μ, V, T) such that

1. $\mu \in \mathcal{R}(X)$,
2. V is a finite dimensional vector space over $\mathbb{R}$,
3. $T : X \to V$, $x \mapsto T(x)$ is a continuous map,
4. For any $p \in \mathcal{P}$, there exists $\theta \in V^\vee$ such that
$$dp(x) = c_\theta^{-1} \exp_q(-\langle \theta, T(x) \rangle) d\mu(x),$$
where $c_\theta = \int_{x \in X} \exp_q(-\langle \theta, T(x) \rangle) d\mu(x)$ (normalizing constant).

We call the triple (μ, V, T) a *realization* of $\mathcal{P}$.

Definition of $\exp_q$

For $q \in \mathbb{R}$, we put $I_q := \{x \in \mathbb{R} \mid (1-q)x + 1 > 0\}$.

Definition 1.3.

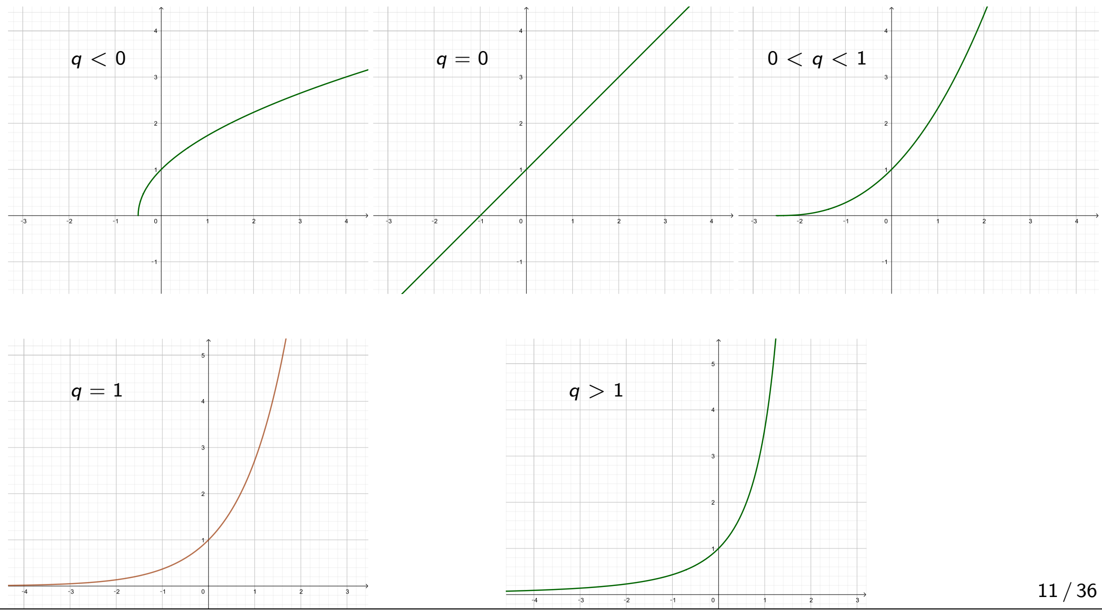
The map $\exp_q \colon I_q \to \mathbb{R}_{>0}$ is defined by
$$\exp_q x := \begin{cases} e^x & (q = 1), \\ ((1-q)x + 1)^{\frac{1}{1-q}} & (q \neq 1). \end{cases}$$

Remark 1.4.

$\exp_q$ is defined as the inverse map of the q-logarithm function $\ln_q \colon \mathbb{R}_{>0} \to \mathbb{R}$
$$\ln_q x := \int_1^x \frac{1}{t^q} dt = \begin{cases} \ln x & (q = 1) \\ \frac{1}{1-q}(x^{1-q} - 1) & (q \neq 1). \end{cases}$$

Graph of $\exp_q$

$$\exp_q x := \begin{cases} e^x & (q = 1), \\ ((1-q)x+1)^{\frac{1}{1-q}} & (q \neq 1). \end{cases}$$



Introduction
Step 1: G/H-method
Step 2: q-deformation
Another topic: natural projection

Problem
Exponential family and q-exponential family
Background

Example: a family of normal distributions

Example 1.5.

The following family of normal distributions is an exponential family on $\mathbb{R}$ $(q = 1)$:

$$\mathcal{P} := \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-m)^2}{2\sigma^2} \right) dx \right\}_{(\sigma,m)\in\mathbb{R}_{>0}\times\mathbb{R}}$$

1. $\mu =$Lebesgue measure,
2. $V = \mathbb{R}^2$,
3. $T\colon X = \mathbb{R} \to \mathbb{R}^2$, $x \mapsto \begin{pmatrix} x^2 \\ x \end{pmatrix}$.

(μ, V, T) is a realization of $\mathcal{P}$.

Example: a family of Cauchy distributions

Example 1.6.

The following family of Cauchy distributions is a 2-exponential family on $\mathbb{R}$:

$$\mathcal{P} := \left\{ \frac{1}{\pi} \frac{\gamma}{(x - x_0)^2 + \gamma^2} \right\}_{(\gamma, x_0) \in \mathbb{R}_{>0} \times \mathbb{R}}$$

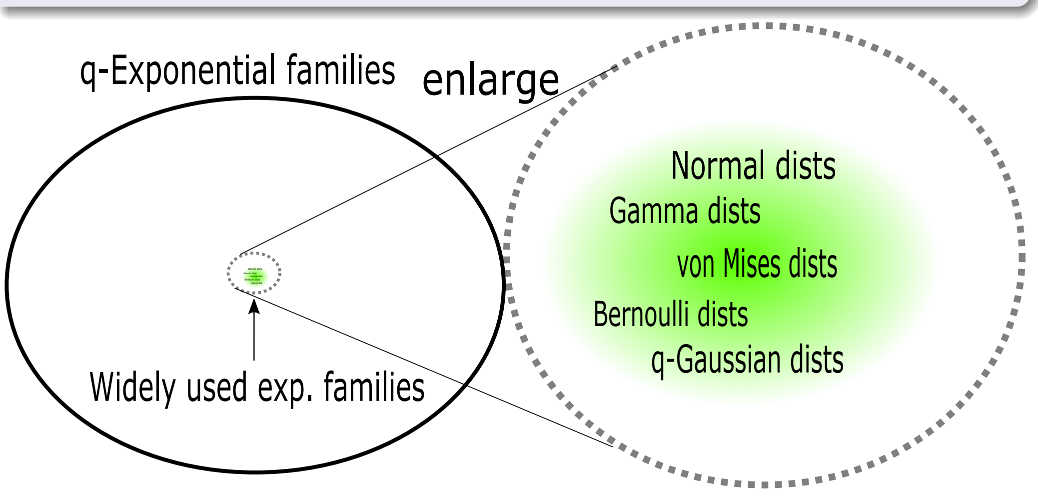
1. $\mu =$Lebesgue measure,
2. $V = \mathbb{R}^2$,
3. $T : X = \mathbb{R} \to \mathbb{R}^2$, $x \mapsto \begin{pmatrix} x^2 \\ x \end{pmatrix}$.

(μ, V, T) is a realization of $\mathcal{P}$.

Background

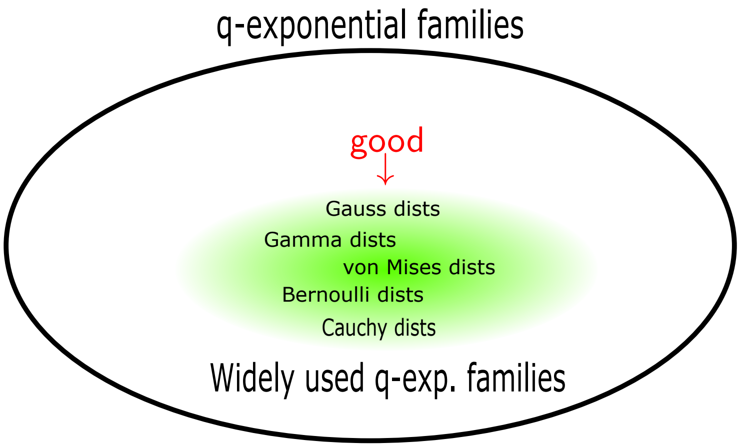
Remark 1.7.

- By definition, there are too many q-exponential families.
- Only a small part of them are widely used.



Motivation

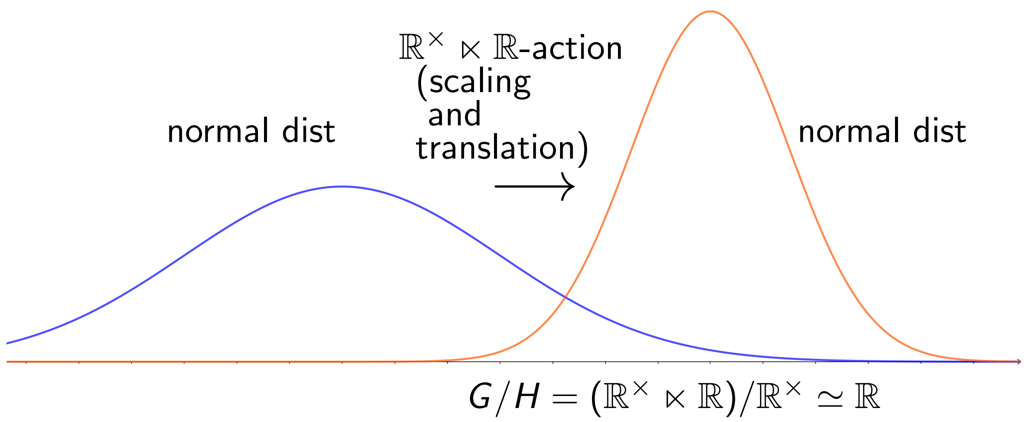
We can expect there exist "good" q-exponential families.
We want a framework to understand "good" q-exponential families systematically.



Observation 1.8.

Useful q-exp. families have the same symmetry as the sample spaces.

- Sample space : homogeneous space G/H
- Family : invariant under the induced G-action



Strategy

Problem 1.1 (again)

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

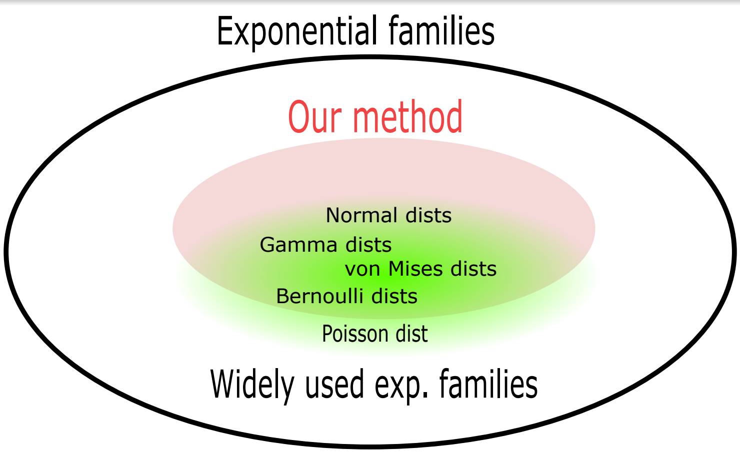
Step 1 Classify G-invariant exponential families on G/H by using G/H-method.

Step 2 Classify G-invariant q-exponential families on G/H by q-deformation of G-invariant exponential families on G/H.

G/H-method

We proposed a method to construct exponential families.

- The method generate many well-known families.
- Families obtained by the method can be classified.



G/H-method: overview

G/H-method = a method to construct a family of probability measures on G/H from

- a finite dim. real representation $\rho : G \to GL(V)$,
- a nonzero H-fixed vector $v_0 \in V$.

See [TY18, TY19, TY20] for the details.

Examples obtained by our method

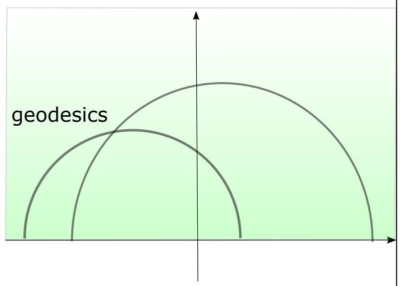
Table: Examples and inputs (G, H, V, v_0) for them

distributions	sample sp. X	G	H	V	v_0
Normal	$\mathbb{R}$	$\mathbb{R}^\times \ltimes \mathbb{R}$	$\mathbb{R}^\times$	$\mathrm{Sym}(2,\mathbb{R})$	E_{22}
Multi. normal	$\mathbb{R}^n$	$GL(n,\mathbb{R}) \ltimes \mathbb{R}^n$	$GL(n,\mathbb{R})$	$\mathrm{Sym}(n+1,\mathbb{R})$	$E_{n+1,n+1}$
Bernoulli	$\{\pm 1\}$	$\{\pm 1\}$	$\{1\}$	$\mathbb{R}_{sgn}$	1
Categorical	$\{1, \cdots, n\}$	$\mathfrak{S}_n$	$\mathfrak{S}_{n-1}$	W	w
Gamma	$\mathbb{R}_{>0}$	$\mathbb{R}_{>0}$	$\{1\}$	$\mathbb{R}$	1
Inverse gamma	$\mathbb{R}_{>0}$	$\mathbb{R}_{>0}$	$\{1\}$	$\mathbb{R}_{-1}$	1
Wishart	$\mathrm{Sym}^+(n,\mathbb{R})$	$GL(n,\mathbb{R})$	$O(n)$	$\mathrm{Sym}(n,\mathbb{R})$	I_n
Von Mises	S^1	$SO(2)$	$\{I_2\}$	$\mathbb{R}^2$	e_1
Poincaré	$\mathcal{H}$	$SL(2,\mathbb{R})$	$SO(2)$	$\mathrm{Sym}(2,\mathbb{R})$	I_2

Here $W = \{(x_1, \cdots, x_n) \in \mathbb{R}^n \mid \sum_{i=1}^n x_i = 0\}$,
$w = (-(n-1), 1, \cdots, 1) \in W$.

Example: Poincaré dists on the upper half plane

Upper half plane $\mathcal{H} := \{z = x + iy \in \mathbb{C} \mid y > 0\}$ admits the linear fractional transformation of $SL(2,\mathbb{R})$.
$\rightsquigarrow G = SL(2,\mathbb{R}),\ H = SO(2),\ X := G/H \simeq \mathcal{H}$.



- Low dimensional representation:
  $\rho : SL(2,\mathbb{R}) \to GL(\mathrm{Sym}(2,\mathbb{R}))$,
  $\rho(g)S = gS{}^tg \quad (S \in \mathrm{Sym}(2,\mathbb{R}))$.
  $v_0 := I_2$.

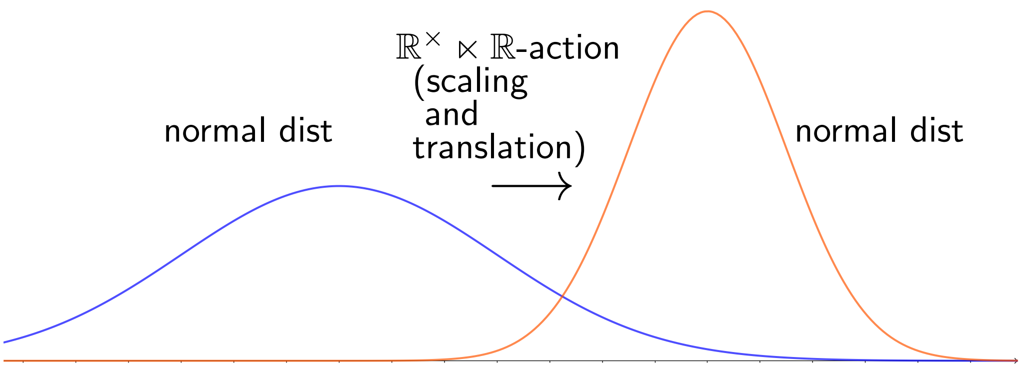
  $$\rightsquigarrow \left\{ \frac{De^{2D}}{\pi} \exp\left(-\frac{a(x^2+y^2)+2bx+c}{y}\right) \frac{dxdy}{y^2} \right\}_{\begin{pmatrix} a & b \\ b & c \end{pmatrix} \in \mathrm{Sym}^+(2,\mathbb{R})}$$

  Here $D = \sqrt{ac - b^2}$.
- Higher dimensional cases:
  We obtain new families by G/H-method.

$\mathcal{P}$ is a G-invariant exponential family

Theorem 2.1.

Any family obtained by our method is a G-invariant exponential family on G/H.



We obtain a family with the symmetry of G/H !

Question

Conversely,

Question 2.2.
Are any G-invariant exponential families on G/H obtained by our method?

$\rightsquigarrow$ Yes, under a mild assumption.
$\rightsquigarrow$ Roughly speaking,

$$\{G\text{-invariant exponential family on } G/H\}$$
$$\text{``}=\text{''}\{\text{family on } G/H \text{ obtained by } G/H\text{-method}\}$$

Introduction
Step 1: G/H-method
Step 2: q-deformation
Another topic: natural projection
Method to construct families
G-invariance of our family
Classification of G-invariant families

Answer to the question

Setting 2.3.
$\mathcal{P} := \{p_\theta\}_{\theta\in\Theta}$ is a G-invariant exponential family on G/H. Here Θ is the parameter space.

Theorem 2.4.
Assume

1. *G/H admits a nonzero relatively G-invariant measure,*
2. *Θ is open.*

Then, $\mathcal{P}$ is a subfamily of a certain family obtained by G/H-method.

For the details, see our paper [TY20].

Classification of G-invariant exponential families

Let us consider an important homogeneous space G/H such as a sphere and a hyperbolic space, more generally symmetric spaces.

Step 1

Classify G-invariant exponential families on G/H.

By Theorem 2.4, this problem above is reduced to the following:

Question 2.5.

Classify families obtained by G/H-method on G/H.

q-deformation of exponential family

q-deformation is a method to construct a q-exponential family from an exponential family with its realization.

Definition 3.1.

Let $\mathcal{P}$ be an exponential family on X and (μ, V, T) realization of $\mathcal{P}$. Put

$$d\tilde{q}_\theta(x) := \exp_q(-\langle\theta, T(x)\rangle)d\mu(x) \quad (\theta \in V^\vee, x \in X)$$

$$\Theta := \{\theta \in V^\vee \mid -\langle\theta, T(x)\rangle \in I_q \text{ for any } x \in X, \int_X d\tilde{q}_\theta < \infty\}$$

$$q_\theta := c_\theta^{-1}\tilde{q}_\theta, \quad c_\theta := \int_X d\tilde{q}_\theta \quad (\theta \in \Theta)$$

$$\mathcal{P}_q := \{q_\theta\}_{\theta\in\Theta}$$

Then $\mathcal{P}_q$ is a q-exponential family on X. We call $\mathcal{P}_q$ a q-deformation of exponential family $(\mathcal{P}, (\mu, V, T))$.

Example

Example 3.2.
The family of Cauchy distributions is obtained by 2-deformation of the family of normal distributions.

- $\mathcal{P}$: family of normal distribution
- μ: Lebesgue measure,
- $V = \mathbb{R}^2$,
- $T : X = \mathbb{R} \to \mathbb{R}^2, x \mapsto \begin{pmatrix} x^2 \\ x \end{pmatrix}$.

$\rightsquigarrow \mathcal{P}_2$: the family of Cauchy distributions.

Property of q-deformation

Let $X := G/H$ be a homogeneous space admitting nonzero relatively G-invariant measure and $q \in \mathbb{R}$.

Proposition 3.3.
Let $\mathcal{P}$ be a G-invariant exponential family on X. Then, there exists a realization (μ, V, T) of $\mathcal{P}$ such that μ is a relatively G-invariant measure on X. Moreover, If $q > 1$ and $\mathcal{P}$ is full , then the q-deformation $\mathcal{P}_q$ of $(\mathcal{P}, (\mu, V, T))$ is G-invariant q-exponential family on X.

Question

Conversely,

Question 3.4.

Are any G-invariant q-exponential families on G/H obtained by q-deformation of some exponential family?

$\rightsquigarrow$ Yes if $q > 1$ under a mild assumption.
$\rightsquigarrow$ Roughly speaking,

$$\begin{aligned} &\{G\text{-invariant } q\text{-exponential family on } G/H\} \\ \text{“}{=}\text{”}\, &\{q\text{-deformation of } G\text{-invariant exponential family on } G/H\} \end{aligned}$$

Answer to the question

Setting 3.5.

$\mathcal{P}_q = \{p_\theta\}_{\theta\in\Theta}$ is a G-invariant q-exponential family on G/H $(q > 1)$.

Theorem 3.6.

Assume

1. *G/H admits a nonzero relatively G-invariant measure,*
2. *Θ is open.*

Then, $\mathcal{P}_q$ is a subfamily of a q-deformation of a certain G-invariant exponential family with a relatively G-invariant base measure.

Classification of G-invariant q-exponential families

G/H: important space

Step 2

Classify G-invariant q-exponential families on G/H by using q-deformation.

By Theorem 3.6, this problem above is reduce to the following:

Question 3.7.

Classify families obtained by q-deformation of G-invariant exponential families.

New family of distributions on the upper half plane

Theorem 3.8 (q-deformation of the family of Poincaré distributions).

Let $q \in [1, 2)$. The following family of distributions is $SL(2, \mathbb{R})$-invariant q-exponential family on the upper half plane.

$$\left\{ c_\theta^{-1} \exp_q \left(- \frac{a(x^2 + y^2) + 2bx + c}{y} \right) \frac{dxdy}{y^2} \right\}_{\theta := \begin{pmatrix} a & b \\ b & c \end{pmatrix} \in \mathrm{Sym}^+(2,\mathbb{R})}$$

$$c_\theta := \frac{\pi(\exp_q(-2D))^{2-q}}{(2-q)D}, D := \sqrt{ac - b^2}.$$

Family on a group and the natural projection

G: Lie group, H: closed subgroup of G,
$\pi\colon G \to G/H$, $g \mapsto gH$ natural projection
$\pi_*\colon \mathcal{P}(G) \to \mathcal{P}(G/H)$ pushforward

Question 4.1.

What kind of families can we obtain by the pushforward π_* of G-invariant exponential family on G?

- Is pushforward of exponential family also exponential family?
- Is pushforward of G-invariant family also G-invariant?

Introduction
Step 1: G/H-method
Step 2: q-deformation
Another topic: natural projection
Questions
Example

Family on a group and the natural projection

G: Lie group, H: closed subgroup of G,
$\pi\colon G \to G/H$, $g \mapsto gH$ natural projection
$\pi_*\colon \mathcal{P}(G) \to \mathcal{P}(G/H)$ pushforward

Question 4.1.

What kind of families can we obtain by the pushforward π_* of G-invariant exponential family on G?

- Is pushforward of exponential family also exponential family?
  $\rightsquigarrow$No, in general.
- Is pushforward of G-invariant family also G-invariant?
  $\rightsquigarrow$Yes.

Example: exp. family on affine transformation group

$G := \mathbb{R}_{>0} \ltimes \mathbb{R}$, $\alpha \in \mathbb{R}$,

$$\rho_\alpha \colon G \to GL(3,\mathbb{R}),\ \rho_\alpha(a,b) := \begin{pmatrix} 1 & 0 & 0 \\ b & 1 & 0 \\ \frac{b^2}{2} & b & 1 \end{pmatrix} \mathrm{diag}(a^\alpha, a^{\alpha+1}, a^{\alpha+2}),$$

$v_0 := {}^t(1,0,0)$.

Proposition 4.2.

If $\alpha \neq 0$, by applying G/H-method to (ρ_α, v_0), we get a $\mathbb{R}_{>0} \ltimes \mathbb{R}$-invariant exponential family on $\mathbb{R}_{>0} \ltimes \mathbb{R}$ as follows:

$$\left\{ \frac{|\alpha|\sqrt{u}}{\sqrt{\pi}\Gamma(r)} \left(\frac{\det D}{u}\right)^r \exp(-a^\alpha(s+tb+ub^2))a^{\alpha(r+\frac{1}{2})-1} da db \right\}_{(r,S)\in\Theta}$$

Here, $D := \begin{pmatrix} s & \frac{t}{2} \\ \frac{t}{2} & u \end{pmatrix}$, $(a,b) \in G$ and $\Theta := \mathbb{R}_{>0} \times \mathrm{Sym}^+(2,\mathbb{R})$.

Pushforward of the obtained family

$$\mathcal{P} := \left\{ \frac{|\alpha|\sqrt{u}}{\sqrt{\pi}\Gamma(r)} \left(\frac{\det D}{u}\right)^r \exp(-a^\alpha(s+tb+ub^2))a^{\alpha(r+\frac{1}{2})-1} da db \right\}_{(r,S)\in\Theta}$$

$G := \mathbb{R}_{>0} \ltimes \mathbb{R}$, $H := \mathbb{R}_{>0}$, $\pi \colon G \to G/H \simeq \mathbb{R}$, $\pi_* \colon \mathcal{P}(G) \to \mathcal{P}(\mathbb{R})$.

Proposition 4.3.

The family $\pi_*\mathcal{P}$ on $\mathbb{R}$ is given as follows:

$$\left\{ \frac{\Gamma(\frac{1}{q-1})}{\sqrt{\pi}\Gamma(\frac{3-q}{2(q-1)})} \sqrt{\frac{q-1}{2}} \frac{1}{\sigma} \exp_q\left(-\frac{(b-m)^2}{2\sigma^2}\right) \right\}_{(q,m,\sigma)\in(1,3)\times\mathbb{R}\times\mathbb{R}_{>0}}$$

Remark 4.4.

- Each distribution is a q-Gaussian distribution.
- The family does not depend on $\alpha \in \mathbb{R}^\times$.

References

[TY18] K. Tojo, T. Yoshino, *A method to construct exponential families by representation theory*, arXiv:1811.01394v4, to appear in Info. Geom.

[TY19] K. Tojo, T. Yoshino, *On a method to construct exponential families by representation theory*, GSI2019, Lecture Notes in Computer Science, vol 11712, 147–156 (2019).

[TY20] K. Tojo, T. Yoshino, *Harmonic exponential families on homogeneous spaces*, Info. Geo. (2020). https://doi.org/10.1007/s41884-020-00033-3

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.

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

Yoshihiko KONNO

Osaka Metropolitan Univeristy/JWU

Workshop

Mathematical optimazaiton and statistical theories using geometric methods

20-21 Octorber 2022

1. The reconstruction of a low-rank matrix from its noisy observation is useful in many applications. This problem is reformulated into a constrained nuclear norm minimization problem (regularized problem).
2. An important ingridient of this problem is how to choose a regularization parameter based on data. Usually the data is independently and identically distributed with unknown variance.
3. (1) The discrepacy principle approach, (2) Stein's Unbiased risk estimator(SURE) approach.
4. Inspired by approach(2) we consider the problem of estimating a low-rank matrix mean in MANOVA(Mulitivariate Analysis of Variance) setting [1]when a positive-definite covariance matrix of error is unknown.

[1]We have data for unknown covariance matrix. The distribution of this data is mean-zero.

MANOVA model and its canonical model

Let $\boldsymbol{m}, \boldsymbol{n}, \boldsymbol{p} \in \mathbb{N}$ such that $\mathbf{min}(\boldsymbol{m}, \boldsymbol{p}) \geq \boldsymbol{p}$. Consider a multivariate regression model

$$\underbrace{\boldsymbol{W}}_{(m+n)\times p} = \underbrace{\boldsymbol{A}}_{(m+n)\times m} \underbrace{\boldsymbol{B}}_{m\times p} + \underbrace{\mathbf{Err}}_{(m+n)\times p},$$

where $\boldsymbol{A}$ is a known design matrix of full rank, $\boldsymbol{B}$ is an unknown regression matrix of rank $\boldsymbol{r}\,(< \boldsymbol{min}(\boldsymbol{m}, \boldsymbol{p})$ and $\boldsymbol{r}$ is unknown$)$, and **Err** is an unobservable error matrix. Here rows of **Err** are independently and identically distributed as $\boldsymbol{N_p}(\mathbf{0}_p, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a $\boldsymbol{p} \times \boldsymbol{p}$ positive-definite unknown matrix.

Notation

1

$$\mathbf{Err} = \begin{bmatrix} e_1^\top \\ e_2^\top \\ \vdots \\ e_{m+n}^\top \end{bmatrix} : (m+n) \times p, \quad \mathrm{vec}(\mathbf{Err}) := \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{m+n} \end{bmatrix}$$

where e_j's are independently and identically distributed as $N_p(0_p, \Sigma)$ $(j = 1, 2, \ldots, (m+n))$.

2 Write

$$\begin{aligned} \mathbb{COV}(\mathbf{Err}) &= \mathbb{E}\Big[\big\{\mathrm{vec}(\mathbf{Err} - \mathbb{E}[\mathbf{Err}]\big\}\big\{\mathrm{vec}(\mathbf{Err} - \mathbb{E}[\mathbf{Err}]\big\}^\top\Big] \\ &= I_{m+n} \otimes \Sigma, \\ \mathbf{Err} &\sim N_{(m+n)\times p}(0_{(m+n)\times p}, I_{m+n} \otimes \Sigma). \end{aligned}$$

MANOVA model and its canonical mode
Problem set-up
Mean matrix estimation when a covariance is known
Mean matrix estimation when a covarianc matrix is unknown
Concluding remarks

1 Let

$$P = (A^\top A)^{-1/2} A^\top : m \times (m+n)$$

and take $P^\perp : n \times (m+n)$ s.t.

$$P(P^\perp)^\top = 0_{m\times n} \quad \text{and} \quad P^\perp (P^\perp)^\top = I_n,$$

Note that

$$\begin{bmatrix} P \\ P^\perp \end{bmatrix} [P^\top, (P^\perp)^\top] = I_{m+n}.$$

2 Put $\Xi := \left(A^\top A\right)^{1/2} B$ and

$$\begin{bmatrix} X \\ Y \end{bmatrix} := \begin{bmatrix} P \\ P^\perp \end{bmatrix} W \sim N_{(m+n)\times p}\left(\begin{bmatrix} \Xi \\ 0_{n\times p} \end{bmatrix}, I_{m+n} \otimes \Sigma \right).$$

Problem set-up

Assume that $\min\{m, n\} \geq p$ and that

$$\begin{array}{cc} & p \\ \begin{array}{c} m \\ n \end{array} & \begin{pmatrix} X \\ Y \end{pmatrix} \end{array} = \begin{pmatrix} \Xi \\ 0 \end{pmatrix} + E; \qquad E = \begin{array}{c} p \\ \begin{pmatrix} \longleftrightarrow \\ \longleftrightarrow \\ \vdots \\ \longleftrightarrow \\ \longleftrightarrow \end{pmatrix} \end{array}$$

where $\begin{bmatrix} X \\ Y \end{bmatrix}$ is observation and Ξ is an $m \times p$ non-random matrix(unknown) of rank $r < p$, E is an $(m+n) \times p$ error matrix(unobservable) whose rows are identically distributed as $N_p(0, \Sigma)$. Here Σ is a $p \times p$ positive-definite and unknown matrix.

MANOVA model and its canonical mode
Problem set-up
Mean matrix estimation when a covariance is known
Mean matrix estimation when a covarianc matrix is unknown
Concluding remarks

We consider the problem of estimating Ξ under a low-rank mean matrix condition, i.e.,

$$\mathrm{rank}\,\Xi = r < p; \qquad r \text{ is unknown}$$

under a loss fucntion and its risk

$$\mathrm{L}_\Sigma(\widehat{\Xi}, \Xi) = \mathrm{tr}\left\{(\widehat{\Xi} - \Xi)\Sigma^{-1}(\widehat{\Xi} - \Xi)^\top\right\} =: \|\widehat{\Xi} - \Xi\|^2_{F,\Sigma}$$

and

$$\mathrm{R}_\Sigma(\widehat{\Xi}, \Xi) = \mathbb{E}[\mathrm{L}_\Sigma(\widehat{\Xi}, \Xi)]$$

where $\widehat{\Xi}$ is an estimator based on (X, S). Here $S = Y^\top Y \sim W_p(\Sigma, n)$, which is the Wishart distribution with the degree of freedom n and the scale matrix Σ.

Mean matrix estimation when a covariance is known

- Assume that m, p are postive integers s.t. $m \geq p$.
- Let

$$Z = \begin{pmatrix} z_1^\top \\ z_2^\top \\ \vdots \\ z_m^T \end{pmatrix}$$

be an $m \times p$ data matrix whose row vectors are independently distributed as

$$z_i : \ p \times 1 \sim N(\widetilde{\xi}_i, \sigma^2 I_p), \qquad (i = 1, 2, \ldots, m)$$

Here $\widetilde{\Xi}^\top := (\widetilde{\xi}_1, \ldots, \widetilde{\xi}_m)$ is unknown but $\sigma > 0$ are known.

MANOVA model and its canonical mode
Problem set-up
Mean matrix estimation when a covariance is known
Mean matrix estimation when a covarianc matrix is unknown
Concluding remarks

- We assume that low-rank mean matrix condition, i.e.,

$$\mathrm{rank}\left(\widetilde{\Xi}\right) = r < p; \quad r \text{ is unknown.}$$

- Consider the problem of estimating $\widetilde{\Xi}$ under a loss fucntion and its risk

$$L_1(\widehat{\widetilde{\Xi}}, \widetilde{\Xi}) = \mathrm{tr}\,\{(\widehat{\widetilde{\Xi}} - \widetilde{\Xi})(\widehat{\widetilde{\Xi}} - \widetilde{\Xi})^\top\} =: ||\widehat{\widetilde{\Xi}} - \widetilde{\Xi}||_F^2$$

and

$$R_1(\widehat{\widetilde{\Xi}}, \widetilde{\Xi}) = \mathbb{E}[L_1(\widehat{\widetilde{\Xi}}, \widetilde{\Xi})].$$

- Here $\widehat{\widetilde{\Xi}}$ is an estimator based on Z.
- $\mathrm{tr}\, A$ and $A^\top$ stand for the traace and the transpose of a matrix A, respectively.
- $||A||_F := \sqrt{\mathrm{tr}\left(A^\top A\right)}$, the Frobenius norm of a matrix A.

Eckart-Young approximation theorem

- Singular Value Decomposition: We can assume that $m \geq p$ without loss of generality. Decompose Z as

$$\begin{aligned} Z &= ULV^{\top}; \quad U = (u_1, \ldots, u_p),\ V = (v_1, \ldots, v_p) \\ L &= \operatorname{diag}(\ell_1, \ell_2, \ldots, \ell_p) \quad \text{with } \ell_1 \geq \ell_2 \geq \cdots \geq \ell_p \geq 0 \end{aligned}$$

where $u_i \in \mathbb{R}^m$, $v_i \in \mathbb{R}^p$ $(i = 1, \ldots, p)$ s.t. $U^{\top} \underbrace{U}_{m \times p} = V^{\top} V = I_p$.

- The total least squres (TLS) pseudo estimator is given by

$$\widehat{\Xi}_{TLS} = \sum_{i=1}^{r} \ell_i u_i v_i^{\top}. \quad \Longleftrightarrow \quad \widehat{\Xi}_{TLS} \in \operatorname*{argmin}_{\Xi:\ \operatorname{rank}(\Xi) \leq r} \|\Xi - Z\|_F^2.$$

Notaton: $\sigma_j(A) > 0$ $(j = 1, 2, \ldots, r)$ are non-zero singular values of a matrix A with $r = \operatorname{rank}(A)$.

Regularization approach

- We consider an estimator which minimizes the penalized least squares criterion

$$\operatorname{Mat}(m, p; \mathbb{R}) \ni \Xi \mapsto \frac{1}{2}\|Z - \Xi\|_F^2 + \operatorname{pen}_{\lambda}(\Xi) \in [0, \infty)$$

where $\operatorname{pen}_{\lambda}(\cdot)$ (≥ 0) is a penality function of Ξ and λ (≥ 0) is a tuning parameter.

- Examples of penalities: For a positive $\lambda > 0$,

  ★ $\mathrm{pen}_\lambda(\Xi) = \lambda \mathrm{rank}\,(\Xi)$

$\Longrightarrow$ a hard-theshholding rule, i.e.,$\mathrm{SVHT}_\lambda(Z) = \sum_{j=1}^{p} \ell_j \mathbb{1}\{\ell_j \geq \lambda\} u_j v_j^\top$,

where $\mathbb{1}\{\text{event}\} = \begin{cases} 1 & \text{if event is true,} \\ 0 & \text{otherwise} \end{cases}$.

★ $\mathrm{pen}_\lambda(\Xi) = \lambda \|\Xi\|_1 := \lambda \sum_{j=1}^{p} \left|\sigma(\Xi)_j\right| \quad \left(\sigma(\Xi)_j\text{: SV's of } \Xi\right)$

where $\left\{(\sigma(\Xi)_j, u_j, v_j)\right\}_{j=1}^{\min(m,p)}$ is a system of singular values of Ξ

$\Longrightarrow$ a soft-thresholding rule, i.e.,

$$\mathrm{SVST}_\lambda(Z) = \sum_{j=1}^{p} (\ell_j - \lambda) \mathbb{1}\{\ell_j \geq \lambda\} u_j v_j^\top.$$

MANOVA model and its canonical mode
Problem set-up
Mean matrix estimation when a covariance is known
Mean matrix estimation when a covarianc matrix is unknown
Concluding remarks

A hard-shreshholding rule

- Assume that σ^2 is known.
- Solve

$$\mathrm{SVHT}_\lambda(Z) = \underset{\Xi}{\mathrm{argmin}}\left[\frac{1}{2}\|\Xi - Z\|_F^2 + \lambda \,\mathrm{rank}\,(\Xi)\right]$$

  where $\lambda > 0$ is a tuning scalar parameter.
- Then the solution is given by

$$\mathrm{SVHT}_\lambda(Z) = \sum_{j=1}^{p} \ell_j\, \mathbb{1}\{\ell_j \geq \lambda\} u_j v_j^\top;\ \mathbb{1}\{\ell_j \geq \lambda\} = \begin{cases} 1 & \ell_j \geq \lambda \\ 0 & \text{otherwise} \end{cases}$$

- The optimal shreshholding is $\frac{4}{\sqrt{3}}\sqrt{p}\sigma$ when $p = m$.

  (See Donoho and Garvish (2017, IEEE, Trans. Inform Theory).

Steps to obtain an adaptive thresholding esitmator

1. Solve regularized minimizaton problem

$$\widehat{\widehat{\Xi}}_\lambda \in \underset{\Xi \in \mathrm{Mat}(m,p;\mathbb{R})}{\arg\min} \left\{ \|Z - \Xi\|_F^2 + \mathbf{pen}_\lambda(\Xi). \right\}.$$

2. Calculate SURE if possible $\left(\text{a closed form of } \widehat{\widehat{\Xi}}_\lambda\right)$:

$$\mathbf{R}_1\left(\widehat{\widehat{\Xi}}_\lambda, \Xi\right) = \mathbb{E}\left[\mathbf{SURE}\left(\widehat{\widehat{\Xi}}_\lambda\right)\right]$$

Note that $\mathbf{SURE}\left(\widehat{\widehat{\Xi}}_\lambda\right)$ is a function of λ and observable data.

3. Solve minimization problem

$$\widehat{\lambda} \in \mathbf{SURE}\left(\widehat{\widehat{\Xi}}_\lambda\right) \Longrightarrow \widehat{\widehat{\Xi}}_{\widehat{\lambda}}.$$

MANOVA model and its canonical mode
Problem set-up
Mean matrix estimation when a covariance is known
Mean matrix estimation when a covarianc matrix is unknown
Concluding remarks

Remarks

1. This method works for the soft-thresholding rule. See Cándes *el al.* (2013).
2. **SURE** does not work for the hard-thresholding rule since Stein's identity, integration-by-parts formula with respect to multivariate normal distribution, fails for the hard-thresholding rule becuase of discontinuity of estimator.

A soft-thersholding rule

- Cèndes et al. define an adaptive soft-shreshholding rule based on SURE:

$$\mathrm{SVST}_\lambda(Z) = \sum_{j=1}^p (\ell_i - \lambda)\, \mathbb{1}\{\ell_i \geq \lambda\} u_j v_j^\top =: \sum_{j=1}^p (\ell_j - \lambda)_+ u_j v_j^\top \tag{1}$$

which is obtained from

$$\min_Y \left\{ \frac{1}{2} \|Z - Y\|_F^2 + \lambda \sum_{j=1}^p \lambda_j \right\} \qquad Y = \mathrm{SVST}_\lambda(Z).$$

- The parameter λ in (1) is selected by minimizieng SURE, Stein's unbiased risk estimate for (1).

- Gaussian integration-by-parts (=Stein's identity) and a bit of algebraic calculation lead to

$$\begin{aligned}
R_1(\mathrm{SVST}_\lambda, \Xi) &= \mathbb{E}[\mathrm{SURE}(\mathrm{SVST}_\lambda)(Z)], \\
\mathrm{SURE}(\mathrm{SVST}_\lambda)(Z) &= -mp\sigma^2 + \sum_{j=1}^p \min\{\ell_j^2, \lambda^2\} \\
&\quad + 2\sigma^2 \mathrm{div}(\mathrm{SVST}_\lambda(X)), \\
\mathrm{div}(\mathrm{SVST}_\lambda(Z)) &= (m-p) \sum_{j=1}^p \left(1 - \frac{\lambda}{\ell_j}\right)_+ + \sum_{j=1}^p \mathbb{1}\{\ell_j > \lambda\} \\
&\quad + 2 \sum_{j=1}^p \sum_{k \neq i}^p \frac{\ell_j(\ell_j - \lambda)_+}{\ell_j^2 - \ell_k^2}
\end{aligned}$$

whenever $\ell_1 > \ell_2 > \cdots > \ell_p \geq 0$.

- An adaptive estimator is given by

$$\mathrm{SVST}_{\widehat{\lambda}}(Z) = \sum_{j=1}^{p} (\ell_i - \widehat{\lambda})_+ u_j v_i^{\top}, \quad (2)$$

$$\widehat{\lambda} \in \operatorname*{arg\,min}_{\lambda \geq 0} \left[\sum_{i=1}^{p} \min\{\ell_i^2, \lambda^2\} + 2\sigma^2 \mathrm{div}(\mathrm{SVST}_{\lambda}(Z)) \right].$$

- Numerical evaluation of the risk of **(2)** was carried out by Candés et. al.
- But it is not clear if $\mathbf{R}_1(\mathrm{SVST}_{\widehat{\lambda}}(Z), \widetilde{\Xi})$ is close to $\mathbf{R}_1(\widehat{\Xi}_{TLS}(Z), \widetilde{\Xi})$ for $\forall \widetilde{\Xi}$ s.t. $\mathrm{rank}\left(\widetilde{\Xi}\right) \leq r < \min(m, p)$.

Mean matrxi estimation when a covarianc matrix is unknown

- Assume that $\min\{m, n\} \geq p$ and that

$$\begin{array}{c} \\ m \\ n \end{array}\!\!\begin{array}{c} p \\ \begin{pmatrix} X \\ Y \end{pmatrix} \end{array} = \begin{pmatrix} \Xi \\ 0 \end{pmatrix} + E; \qquad E = \begin{array}{c} p \\ \begin{pmatrix} \longleftrightarrow \\ \longleftrightarrow \\ \vdots \\ \longleftrightarrow \\ \longleftrightarrow \end{pmatrix} \end{array}$$

- The $m \times p$ mean matrix Ξ is of rank $r < p$
- The error E is an $(m + n) \times p$ error matrix(unobservable) whose rows are identically distributed as $N_p(0, \Sigma)$.
- The covariance matrix Σ is a $p \times p$ positive-definite and unknown.

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

$$\mathrm{rank}\, \boldsymbol{\Xi} = r < \min(m, p); \quad r \text{ is unknown.}$$

- A loss fucntion and its risk are given by

$$\mathrm{L}_{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Xi}}, \boldsymbol{\Xi}) = \mathrm{tr}\,\{(\widehat{\boldsymbol{\Xi}} - \boldsymbol{\Xi})\boldsymbol{\Sigma}^{-1}(\widehat{\boldsymbol{\Xi}} - \boldsymbol{\Xi})^\top\} =: \|\widehat{\boldsymbol{\Xi}} - \boldsymbol{\Xi}\|^2_{F,\boldsymbol{\Sigma}}$$

and

$$\mathrm{R}_{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Xi}}, \boldsymbol{\Xi}) = \mathbb{E}[\mathrm{L}_{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Xi}}, \boldsymbol{\Xi})]$$

where $\widehat{\boldsymbol{\Xi}}$ is an estimator based on $(\boldsymbol{X}, \boldsymbol{S})$.

- $\boldsymbol{S} = \boldsymbol{Y}^\top \boldsymbol{Y} \sim W_p(\boldsymbol{\Sigma}, n)$, which is the Wishart distribution with the degree of freedom n and the scale matrix $\boldsymbol{\Sigma}$.

MANOVA model and its canonical mode
Problem set-up
Mean matrix estimation when a covariance is known
Mean matrix estimation when a covarianc matrix is unknown
Concluding remarks

- To derive a class of estimators, first assume that $\boldsymbol{\Sigma}$ is known.
- Then we have

$$\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2} \sim N_{m\times p}(\widetilde{\boldsymbol{\Xi}}, \boldsymbol{I}_m \otimes \boldsymbol{I}_p), \qquad \widetilde{\boldsymbol{\Xi}} = \boldsymbol{\Xi}\boldsymbol{\Sigma}^{-1/2}$$

which leads to an estimator of $\widetilde{\boldsymbol{\Xi}}$ given by

$$\widehat{\widetilde{\boldsymbol{\Xi}}}_{TLS} \in \mathop{\mathrm{arg\,min}}_{\mathrm{rank}\,\boldsymbol{\Xi} \le r} \|\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2} - \boldsymbol{\Xi}\|_F{}^2 \implies \widehat{\boldsymbol{\Xi}} = \widehat{\widetilde{\boldsymbol{\Xi}}}_{TLS}\boldsymbol{\Sigma}^{1/2}.$$

- Hence we consider a class of estimators of the form

$$\widehat{\boldsymbol{\Xi}}_H = \left(\sum_{i=1}^{p} h_i(\ell_i)\boldsymbol{u}_i\boldsymbol{v}_i^\top\right)\boldsymbol{S}^{1/2}; \ \boldsymbol{X}\boldsymbol{S}^{-1/2} = \boldsymbol{U}\boldsymbol{L}\boldsymbol{V}^\top$$

where $\boldsymbol{L} = \mathrm{diag}(\ell_1, \ldots, \ell_p)$, $\boldsymbol{H} = \mathrm{diag}(h_1, \ldots, h_p)$, $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p)$ and $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p)$ s.t. $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}_p$.

Regularized minimization problem

- Known $\boldsymbol{\Sigma}$ case: For $\lambda \geq 0$,

$$\mathrm{Mat}(m, p; \mathbb{R}) \ni \Xi\Sigma^{-1/2} \mapsto \|X\Sigma^{-1/2} - \Xi\Sigma^{-1/2}\|_F{}^2 + 2\lambda\|\Xi\Sigma^{-1/2}\|_1$$

- Unknonw $\boldsymbol{\Sigma}$ case: For $\lambda \geq 0$, find a minimizer $\widehat{\widetilde{\Xi}}$ of a regularized minimization problem

$$\mathrm{Mat}(m, p; \mathbb{R}) \ni \widetilde{\Xi} \mapsto \|XS^{-1/2} - \widetilde{\Xi}\|_F{}^2 + 2\lambda\|\widetilde{\Xi}\|_1$$

and

$$\widehat{\Xi} = \widehat{\widetilde{\Xi}}S^{1/2} = \left(\sum_{j=1} \ell_j(\ell_j - \lambda)_+ u_j v_j^\top\right) S^{1/2}$$

where $\{(\ell_j, u_j, v_j)\}$ is a system of singular values of $XS^{-1/2}$

MANOVA model and its canonical mode
Problem set-up
Mean matrix estimation when a covariance is known
Mean matrix estimation when a covarianc matrix is unknown
Concluding remarks

- If

$$h_j(\ell_j) = \ell_j - \frac{c}{\ell_j} \ (j = 1, 2, \ldots, p);$$

c is a known positive constant,

then it results in the Efron-Morris estimator which is given by

$$\begin{aligned}\widehat{\Xi}_H &= XS^{-1/2}\left[I_p - c\{(XS^{-1/2})^\top(XS^{-1/2})\}^{-1}\right] S^{1/2}\\ &= X - cX\{X^\top X\}^{-1}S.\end{aligned}$$

- On the other hand, Tsukuma and Kubokawa (2015) considered estimators of the form

$$\widehat{\Xi}_T = X - UTU^\top X$$

where $T = \mathrm{diag}(t_1(\ell_1^2), \ldots, t_p(\ell_p^2))$ and $XS^{-1/2} = ULV^\top$ with $m \times \min(m, p)$ matrix U s.t. $U^\top U = I_{\min(m,p)}$.

- Recall that

$$XS^{-1/2} = ULV^\top \iff L^{-1}U^\top X = V^\top S^{1/2}.$$

From a simple calculation we get

$$\widehat{\Xi}_H = UHV^\top S^{1/2} = UHL^{-1}U^\top X = UL^{-1}HX.$$

- If we set $I_p - T = L^{-1}H\,(t_j(x) = h_i(\sqrt{x}))$, then we have

$$\widehat{\Xi}_H = \widehat{\Xi}_T.$$

- From this we can see that

$$\mathrm{L}_\Sigma(\widehat{\Xi}_H, \Xi) = \mathrm{L}_\Sigma(\widehat{\Xi}_T, \Xi).$$

MANOVA model and its canonical mode
Problem set-up
Mean matrix estimation when a covariance is known
Mean matrix estimation when a covarianc matrix is unknown
Concluding remarks

- Furthermore, using the result due to Tsukuma and Kubokawa (2015), we have

$$\begin{aligned}
\mathrm{R}_\Sigma(\widehat{\Xi}_T, \Xi) &= \mathbb{E}[\mathrm{SURE}(T)];\\
\mathrm{SURE}(T) &= \sum_{j=1}^{p}\Bigg[m + a\ell_j^2 t_j^2 - 2bt_i - 4\ell_j^2 t_j\widetilde{t}_j - 4\ell_i^2\widetilde{t}_j\\
&\quad -2\sum_{k\neq j}^{p}\frac{\ell_j^4 t_j^2 - \ell_k^4 t_k^2}{\ell_j^2 - \ell_k^2} - 4\sum_{k\neq j}^{p}\frac{\ell_j^2 t_j - \ell_k^2 t_k}{\ell_j^2 - \ell_k^2}\Bigg];\\
t_i &= 1 - \frac{h_i(\ell_j)}{\ell_j}; \qquad t_i' = -\frac{1}{2\ell_i^2}\left(\widetilde{h}_j'(\ell_j) + \frac{h(\ell_j)}{\ell_j}\right),\\
a,\ b &: \ \text{known positive constants.}
\end{aligned}$$

- Then we have an adaptive soft-thresholding rule

$$\widehat{\Xi}_{\widehat{\lambda}} \; =: \mathrm{SVST}_{\widehat{\lambda}}(XS^{-1/2})S^{1/2} = \left(\sum_{j=1}^{p}(\ell_j - \widehat{\lambda})_+ u_j v_j^\top\right) S^{1/2}$$

where $\widehat{\lambda} = \mathrm{argmin}_{\lambda \geq 0}\, \mathrm{SURE}(\mathrm{SVST}_\lambda)(XS^{-1/2})$;

$$\mathrm{SURE}(\mathrm{SVST}_\lambda)(XS^{-1/2}) = \sum_{j=1}^{p}\Big[m + a\ell_j^2 t_j^2 - 2bt_i - 4\ell_j^2 t_j \widetilde{t}_j - 4\ell_j^2 \widetilde{t}_j - 2\sum_{k\neq j}^{p}\frac{\ell_j^4 t_j^2 - \ell_k^4 t_k^2}{\ell_j^2 - \ell_k^2} - 4\sum_{j\neq i}\frac{\ell_i^2 t_i - \ell_i^2 t_i}{\ell_i^2 - \ell_j^2}\Big];$$

$$t_j = 1 - \frac{(\ell_j - \lambda)_+}{\ell_j} \; (j = 1, \ldots, p);$$

$$\widetilde{t}_j = -(2\ell_j)^{-2}\left(\mathbb{1}\{\ell_j > \lambda\} + \frac{(\ell_j - \lambda)_+}{\ell_j}\right).$$

Special case

- $\Sigma = \sigma^2 I_p$ where σ is postive but unknown.
- Let $s^2 = \mathrm{tr}(Y^\top Y)/p$.
- Then an adaptive soft-thresholding rule for this case is given by $\widehat{\Xi}_{\widehat{\lambda}} = \sum_{j=1}^{p}\left(\ell_i - \widehat{\lambda}s^2\right)_+ u_j v_j^\top;$, $\quad X = ULV^\top$, with $\widehat{\lambda} = \underset{\lambda \geq 0}{\mathrm{argmin}}\mathrm{SURE}(SVST_\lambda)(X)$ and

$$\mathrm{SURE}(SVST_\lambda)(X) \;=\; \sum_{j=1}^{p}\Big[ms^2 + a\ell_j^2 t_j^2 - 4\ell_j \widetilde{t}_j - 2\sum_{k\neq j}^{p}\frac{\ell_j^4 t_j^2 - \ell_k^4 t_k^2}{\ell_j^2 - \ell_k^2} + s^2\Big(a\ell_j^2 t_j^2 - 4\ell_j^2 t_j \widetilde{t}_j - 4\sum_{k\neq j}^{p}\frac{\ell_j^2 t_j - \ell_k t_k}{\ell_j^2 - \ell_k^2}\Big)\Big]$$

where

$$t_j = 1 - \frac{(\ell_j - \lambda s^2)_+}{\ell_j}; \; \widetilde{t}_j = -\frac{1}{\ell_j^2}\left(\mathbb{1}\{\ell_j > \lambda s^2\} + \frac{(\ell_j - \lambda s^2)_+}{\ell_j}\right).$$

Concluding remarks

1. Derivation of an adaptive threshholding rule:
   - For $\lambda \geq 0$, solve a regularized minimizaton problem(random one) $\widehat{\widehat{\Xi}} \in \underset{\Xi \in \mathrm{Mat}(m,p;\mathbb{R})}{\arg\min} \in \Xi\left\{\frac{1}{2}\|XS^{-1/2} - \Xi\|_F{}^2 + \lambda\|\Xi\|_1\right\}.$
   - We have $\widehat{\Xi}_\lambda = \widehat{\widehat{\Xi}} S^{1/2} = \left(\sum_{j=1} \ell_j(\ell_j - \lambda)_+ u_j v_j^\top\right) S^{1/2}.$

     wherer $\left\{(\ell_j,\, u_j,\, v_j\right\}_{j=1,2,\dots,m}$ is a system of singular values of $XS^{-1/2}$.
   - Obtain SURE $\quad \mathrm{R}_\Sigma\left(\widehat{\Xi}_\lambda,\, \Xi\right) = \mathbb{E}\left[\mathrm{SURE}\left(\widehat{\Xi}_\lambda\right)\right]$
   - Solve the minimization problem $\widehat{\lambda} \in \underset{\lambda \geq 0}{\arg\min}\, \mathrm{SURE}\left(\widehat{\Xi}_\lambda\right) \Longrightarrow \widehat{\Xi}_{\widehat{\lambda}}.$.
2. It is routine to convert this result to case for complex normal distribution.

References

1. Candés, E.J., Sing-Long, C.A., and Trzasko, J.D. (2013): IEEE on Signal Processing **61** 4643–4657.
2. Chételat, D. and Wells, M.T. (2012): AOS **40** 3137–3160.
3. Efron, B. (2004): JASA **99** 619–642.
4. Hansen, N.R. (2018): SPL **135** 76–88.
5. Josse, J. and Sardy, S. (2016): Stat. Comput **26** 715–724.
6. Li, K., Li, H., Chen, R.H., and Wen, Y.-W. (2022): SIAM JSC **44** A2204-A2225.
7. Mukherjee, A., Chen, K., Wang, N., and Zhu, L. (2015): Biometrika **102** 457–477.
8. Tsukuma, H. and Kubokawa, T. (2015): JMVA **139** 312–328.

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_{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_{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

Pattern Recovery by SLOPE

Piotr Graczyk

LAREMA, Université d'Angers, France

OCAMI Workshop 20 - 21 October 2022
Mathematical optimization and statistical theories using **geometric methods**

Joint work with:
T.Skalski (Angers and Wrocław joint PhD)
M. Bogdan, M. Wilczyński (Wrocław)
B. Kołodziejek (Warsaw)
X. Dupuis, P. Tardivel (Dijon)
U. Schneider (Vienna)

[1] M. Bogdan, X. Dupuis, P. Graczyk, B. Kołodziejek, T. Skalski, P. Tardivel, M. Wilczyński, Pattern recovery by SLOPE (2022), arXiv:2203.12086.
This paper is purely analytical, even if some intuitions and notions are geometrical.

[2] P. Tardivel, T. Skalski, U. Schneider, P. Graczyk,
The **Geometry** of Model Recovery by Penalized and Thresholded Estimators (2022), HAL preprint hal-03262087.

....

A geometrical approach to SLOPE was initiated in
[S-T] U. Schneider, P. Tardivel(2020). The Geometry of Uniqueness, Sparsity and Clustering in Penalized Estimation. arXiv preprint arXiv:2004.09106, to appear in 2022.

Linear regression model

We dispose of n observations of p explicative variables (predictors) $X_1, \dots, X_p$ and a response variable Y:
$Y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i,\ i = 1, \dots, n.$

- $X = (x_{ij})_{1 \le i \le n, 1 \le j \le p}$ is the design $n \times p$ matrix.
- The columns of X correspond to p variables
- $\beta = (\beta_1, \dots, \beta_p) \in \mathbb{R}^p$ unknown regression coefficients.
- $\varepsilon = (\varepsilon_1, \dots, \varepsilon_p) \in \mathbb{R}^n$ random noise.

Matrix notation: $Y = X\beta + \varepsilon$

Linear regression $Y = X\beta + \varepsilon$, $X \in \mathbb{R}^{n\times p}$. Estimator of β?

Classical statistics case: $p \leq n$, $\text{rank}X = p$

Ordinary Least Squares estimator:
$\hat{\beta}^{OLS} = \arg\min_{b\in\mathbb{R}^p} \|Y - Xb\|_2^2$

Linear regression $Y = X\beta + \varepsilon$, $X \in \mathbb{R}^{n\times p}$. Estimator of β?

Classical statistics case: $p \leq n$, $\text{rank}X = p$

Ordinary Least Squares estimator:
$\hat{\beta}^{OLS} = \arg\min_{b\in\mathbb{R}^p} \|Y - Xb\|_2^2 = (X'X)^{-1}X'Y$

Chalenging case: $\boxed{p > n}$
$\hat{\beta}^{OLS}$ is not uniquely determined, so no longer useful
Modern statistics resorts to the penalized least squares estimators:

$$\hat{\beta} = \arg\min_{b\in\mathbb{R}^p} \|Y - Xb\|_2^2 + \mathrm{pen}(b),$$

where pen is the penalty on the model complexity.

Penalized estimators LASSO and SLOPE

LASSO (Tibshirani (1996)): $\text{pen}(b) = \lambda\|b\|_1,\ \lambda > 0$

Penalized estimators LASSO and SLOPE

LASSO (Tibshirani (1996)): $\text{pen}(b) = \lambda\|b\|_1,\ \lambda > 0$

SLOPE (Sorted L One Penalized Estimation)
(Bogdan et al. (2015)), defined as

$$\hat{\beta}^{SLOPE} = \underset{b \in \mathbb{R}^p}{argmin}\ \frac{1}{2}\|Y - Xb\|_2^2 + \underbrace{\sum_{i=1}^{p} \lambda_i |b|_{(i)}}_{\text{sorted } \ell_1 \text{ norm}},$$

where $\lambda_1 > 0$, $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ and $|b|_{(1)} \geq \ldots \geq |b|_{(p)}$.

Penalized estimators LASSO and SLOPE

LASSO (Tibshirani (1996)): pen$(b) = \lambda \|b\|_1$, $\lambda > 0$

SLOPE (Sorted L One Penalized Estimation)
(Bogdan et al. (2015)), defined as

$$\hat{\beta}^{SLOPE} = \mathop{argmin}_{b \in \mathbb{R}^p} \frac{1}{2}\|Y - Xb\|_2^2 + \underbrace{\sum_{i=1}^{p} \lambda_i |b|_{(i)}}_{\text{sorted } \ell_1 \text{ norm}},$$

where $\lambda_1 > 0$, $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ and $|b|_{(1)} \geq \ldots \geq |b|_{(p)}$.

When $\lambda_1 = \ldots = \lambda_p > 0$ then SLOPE coincides with LASSO.
Our results for SLOPE give a new approach to LASSO.

Polyhedral penalties and dimensionality reduction

In case when the **penalty** function *pen* is a **polyhedral norm**

Polyhedral penalties and dimensionality reduction

In case when the **penalty** function *pen* is a **polyhedral norm** (i.e. the unit ball $B_{pen}(0,1) \subset \mathbb{R}^p$ in the *pen* norm is a polyhedron)

Polyhedral penalties and dimensionality reduction

In case when the **penalty** function *pen* is a **polyhedral norm** (i.e. the unit ball $B_{pen}(0,1) \subset \mathbb{R}^p$ in the *pen* norm is a polyhedron) penalized estimators usually possess the **dimensionality reduction** properties.

Polyhedral penalties and dimensionality reduction

In case when the **penalty** function *pen* is a **polyhedral norm** (i.e. the unit ball $B_{pen}(0,1) \subset \mathbb{R}^p$ in the *pen* norm is a polyhedron) penalized estimators usually possess the **dimensionality reduction** properties.

It is well known that LASSO estimator has many null components

$$\hat{\beta}_i^{LASSO} = 0$$

.

Dimensionality reduction property of LASSO consists in **elimination of irrelevant predictors** X_i.

SLOPE: dimensionality reduction also by clustering variables

Another important kind of dimensionality reduction consists in clustering (merging, summing) variables with the same values of regression coefficients:

$$\hat{\beta}_i = \hat{\beta}_j \implies Y = ... + \hat{\beta}_i(X_i + X_j) + ...$$

LASSO does not have this property!

SLOPE: dimensionality reduction also by clustering variables

Another important kind of dimensionality reduction consists in clustering (merging, summing) variables with the same values of regression coefficients:

$$\hat{\beta}_i = \hat{\beta}_j \implies Y = ... + \hat{\beta}_i(X_i + X_j) + ...$$

LASSO does not have this property!

Statisticians working with SLOPE observed that many coefficient regression estimates of SLOPE can be:

SLOPE: dimensionality reduction also by clustering variables

Another important kind of dimensionality reduction consists in clustering (merging, summing) variables with the same values of regression coefficients:

$$\hat{\beta}_i = \hat{\beta}_j \implies Y = ... + \hat{\beta}_i(X_i + X_j) + ...$$

LASSO does not have this property!

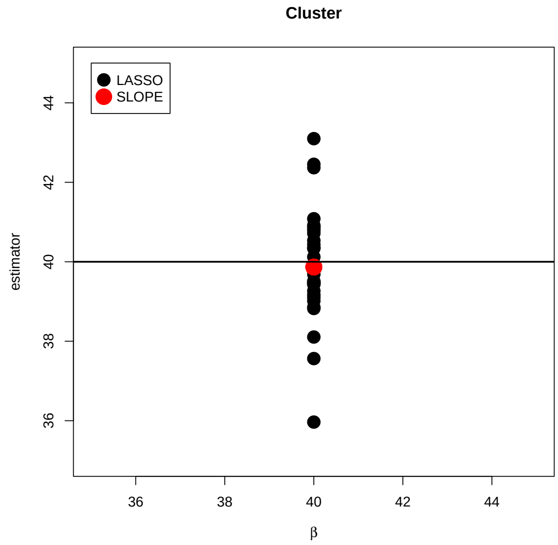
Statisticians working with SLOPE observed that many coefficient regression estimates of SLOPE can be:

- equal $\implies$ clustering predictors
- null $\implies$ eliminating irrelevant predictors like LASSO

Simulations: $n = 100, p = 200$, LASSO and SLOPE on R

We simulated $Y = X\beta + \varepsilon$ where ε has iid $N(0, 5^2)$ entries and
$\beta_1 = \ldots = \beta_{30} = 40, \quad \beta_{31} = \ldots = \beta_{200} = 0.$
The rows of the design matrix X are generated as independent binary Markov chains, with
$\mathbb{P}(X_{i1} = 1) = \mathbb{P}(X_{i1} = -1) = 0.5$ and $\mathbb{P}(X_{i(j+1)} \neq X_{ij}) = 1 - \mathbb{P}(X_{i(j+1)} = X_{ij}) = 0.0476$.

Both LASSO and SLOPE properly estimate at 0 null components of β (not drawn)



Main objective of our research

Why/when does SLOPE recover the clusters and zeros (" SLOPE pattern") of β?

Main objective of our research

Why/when does SLOPE recover the clusters and zeros (" SLOPE pattern") of β?
Explain this phenomenon strictly mathematically.

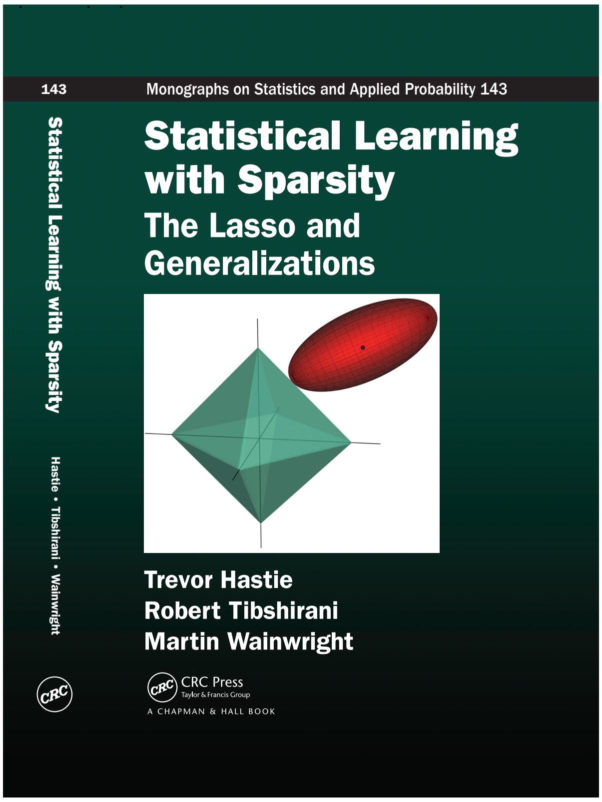
Main objective of our research

Why/when does SLOPE recover the clusters and zeros (" SLOPE pattern") of β?
Explain this phenomenon strictly mathematically.
Give sufficient and necessary conditions for SLOPE pattern recovery.

Main objective of our research

Why/when does SLOPE recover the clusters and zeros (" SLOPE pattern") of β?
Explain this phenomenon strictly mathematically.
Give sufficient and necessary conditions for SLOPE pattern recovery.

A by-product: a new and simple mathematical approach to these questions for LASSO (huge literature on LASSO is very technical)



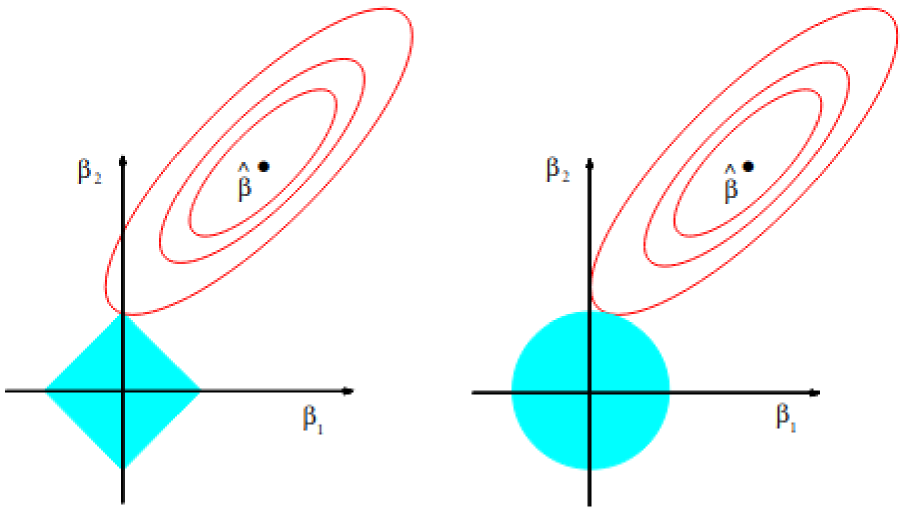


Figure 2.2 *Estimation picture for the lasso (left) and ridge regression (right). T solid blue areas are the constraint regions $|\beta_1|+|\beta_2| \le t$ and $\beta_1^2+\beta_2^2 \le t^2$, respective while the red ellipses are the contours of the residual-sum-of-squares function. T point $\widehat{\beta}$ depicts the usual (unconstrained) least-squares estimate.*

Dimensionality reduction by SLOPE

Some coefficient regression estimates of SLOPE can be null or can be equal in absolute value.

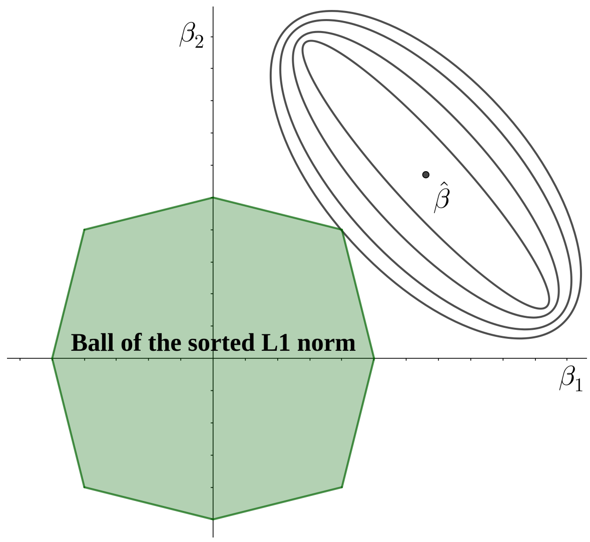


Figure: This figure intuitively illustrates that SLOPE can have some null components or some components equal in absolute value.

Dual penalty norm and dual ball

Suppose that *pen* is a polyhedral norm on $\mathbb{R}^p$.

Dual penalty norm and dual ball

Suppose that *pen* is a polyhedral norm on $\mathbb{R}^p$.

Our results show that the **dual unit ball** B^* plays a crucial role in studying penalized estimators rather than B itself.

Dual penalty norm and dual ball

Suppose that *pen* is a polyhedral norm on $\mathbb{R}^p$.

Our results show that the **dual unit ball** B^* plays a crucial role in studying penalized estimators rather than B itself.

Given a norm $\|\cdot\|$ on $\mathbb{R}^p$, recall that the dual norm $\|\cdot\|^*$ is defined by

$$\|b\|^* = \max\{v'b\colon \|v\| \leq 1\} = \|b^*\|,$$

i.e. it is the norm of b considered as a linear functional b^*.

Dual SLOPE norm and dual ball

Let $\Lambda = (\lambda_1, \ldots, \lambda_p)'$ where $\lambda_1 > 0$ and $\lambda_1 \geq \ldots \geq \lambda_p > 0$.

- The sorted ℓ_1 norm is denoted

$$J_\Lambda(b) = \sum_{i=1}^{p} \lambda_i |b|_{(i)} \text{ where } |b|_{(1)} \geq \ldots \geq |b|_{(p)}.$$

- The dual sorted ℓ_1 norm is equal to

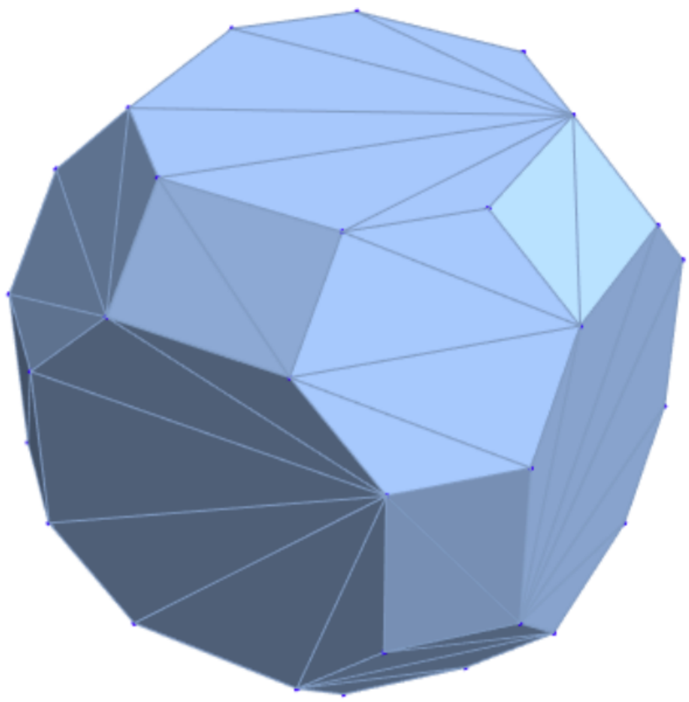
$$J^*_\Lambda(b) = \max\left\{\frac{|b|_{(1)}}{\lambda_1}, \frac{|b|_{(1)} + |b|_{(2)}}{\lambda_1 + \lambda_2}, \ldots, \frac{|b|_{(1)} + \ldots + |b|_{(p)}}{\lambda_1 + \ldots + \lambda_p}\right\}.$$

- The dual SLOPE ball is defined by

$$B^* = \{v \in \mathbb{R}^p |\ J^*_\Lambda(v) \leq 1\}.$$

B^* is a signed permutahedron in $\mathbb{R}^p$: its vertices are signed permutations of Λ.

$p = 3$, $B^* =$ signed permutahedron



Approach of minimization by subdifferential

Let $f: \mathbb{R}^p \to \mathbb{R}$ be a convex function.
The subdifferential ∂f is defined by

$$\partial f(b) = \{v \in \mathbb{R}^p : f(z) \geq f(b) + v'(z - b)\ \forall z \in \mathbb{R}^p\}$$

Evidently, f attains its minimum at a point b if and only if

$$0 \in \partial f(b)$$

.

Approach of minimization by subdifferential

Recall the SLOPE minimization problem:

minimize $b \to f(b) = \frac{1}{2}\|Y - Xb\|_2^2 + J_\Lambda(b)$.
It is a particular case of *pen*–minimization problem
minimize $b \to f(b) = \frac{1}{2}\|Y - Xb\|_2^2 + pen(b)$.

Proposition (Solution of pen-min problem)

$\hat{\beta}$ is a solution of the pen minimization problem if and only if

$$X'(Y - X\hat{\beta}) \in \partial(pen)(\hat{\beta}).$$

Proof. f attains its minimum at a point b if and only if $0 \in \partial f(b)$.
We have
$\partial f(b) = \partial \frac{1}{2}\|Y - Xb\|_2^2 + \partial(pen)(b) = \{-X'(Y - Xb)\} + \partial(pen)(b)$.
The condition $0 \in \partial f(b)$ gives the proposition. $\square$.
Thus we need to understand $\partial(pen)$.

Subdifferential of a norm and the dual ball B^*

Proposition (Subdifferential and the dual ball)

(a) The subdifferential of a norm $\|\cdot\|$ is the following subset of B^*:

$$\partial\|\cdot\|(b) = \{v \in \mathbb{R}^p : \|v\|^* \leq 1 \text{ and } v'b = \|b\|\}$$

(b) If the norm $\|\cdot\|$ is polyhedral, then $\partial\|\cdot\|(b)$ is a face of B^* and all faces of B^* are subdifferentials of $\|\cdot\|$.

Proof. (a) is an easy exercice. Both parts are in the book:
HIRIART-URRUTY, J.-B. and LEMARÉCHAL, C. (2004).
Fundamentals of convex analysis. Springer.

Set $S_{X,\Lambda}(Y)$ of SLOPE solutions. Uniqueness.

We denote $S_{X,\Lambda}(Y) \neq \emptyset$ the set of SLOPE solutions. It is easy to see that it is compact. It may be bigger than a singleton.
The unicity has the following geometrical characterization.

Theorem (Uniqueness, [S-T],[2])

The solution of the pen-minimization problem is unique for all $Y \in \mathbb{R}^n$ if and only if $\mathrm{row}(X)$ does not intersect a face of the dual ball B^* whose codimension is greater than $\dim(\mathrm{col}(X))$.

• Cases in which $S_{X,\Lambda}(Y)$ is not a singleton are very rare.
Indeed, the set of matrices $X \in \mathbb{R}^{n\times p}$ for which there exists a $Y \in \mathbb{R}^n$ where $S_{X,\Lambda}(Y)$ is not a singleton has a null Lebesgue measure on $\mathbb{R}^{n\times p}$ ([S-T])
If $\ker(X) = \{0\}$, then $S_{X,\Lambda}(Y)$ consists of one element.

SLOPE pattern and related notions

The SLOPE pattern (introduced by Schneider and Tardivel (2020)) extracts from a given vector:

a) The sign of the components (positive, negative or null),
b) The clusters (indices of components equal in absolute value),
c) The hierarchy between the clusters.

Definition (SLOPE pattern)

Let $b \in \mathbb{R}^p$. The SLOPE pattern of b, $\mathrm{patt}(b) \in \mathbb{Z}^p$, is defined by

$$\mathrm{patt}(b)_i = \mathrm{sign}(b_i)\,\mathrm{rank}(|b|)_i, \quad i \in \{1,\dots,p\}$$

where $\mathrm{rank}(|b|)_i \in \{0,1,\dots,k\}$, k is the number of nonzero distinct values in $\{|b_1|,\dots,|b_p|\}$.

Example

$b = (4.7, -4.7, 0, 1.8, 4.7, -1.8)' \to \mathrm{patt}(b) = (2, -2, 0, 1, 2, -1)'$.

$\mathcal{P}_p^{\mathrm{SLOPE}} = \mathrm{patt}(\mathbb{R}^p)$ denotes the set of SLOPE patterns.

Identification of patterns as subdifferentials

Theorem (SLOPE pattern= subdifferential(SLOPE *pen*))

Let $\Lambda = (\lambda_1, \ldots, \lambda_p)'$ *where* $\lambda_1 > \ldots > \lambda_p > 0$ *and* $a, b \in \mathbb{R}^p$. *We have* $\mathrm{patt}(a) = \mathrm{patt}(b)$ *if and only if* $\partial J_\Lambda(a) = \partial J_\Lambda(b)$.

Proof. A first (involved) proof was given in [S-T]. In [1] we give a simple proof as a corollary from the (coming below) Proposition on affine characterization of $\partial(J_\Lambda)$ for SLOPE.

Identification of patterns as subdifferentials

Consequently,
for any polyhedral norm penalty *pen*, we define in [2]:

Definition (Pattern= subdifferential(*pen*), [2])

For a penalized estimator with *pen* equal to a polyhedral norm, we say that $\mathrm{patt}(a) = \mathrm{patt}(b)$ if a and b have the same subdifferentials: $\partial pen(a) = \partial pen(b)$.

Example. For LASSO, with $pen = \|\cdot\|_1$, we get

$$\mathrm{patt}(a) = \mathrm{sign}(a).$$

Indeed, the subdifferentials of $pen = \|\cdot\|_1$ (=faces of the unit ball in $\|\cdot\|_\infty$) are in bijection with the set $\{-1, 0, 1\}^p$.

Pattern recovery

Definition (Pattern recovery)

We say that SLOPE pattern is recovered by the SLOPE estimator if there exists $\hat{\beta} \in S_{X,\Lambda}(Y)$ with

$$\mathrm{patt}(\hat{\beta}) = \mathrm{patt}(\beta).$$

Example. Let the true $\beta = (5, 5, 2, -5)'$ and the SLOPE estimator $\hat{\beta}_1 = (4, 4, 3, -4)'$.
Then $\mathrm{patt}(\hat{\beta}) = \mathrm{patt}(\beta) = (2, 2, 1, -2)'$ and we have the pattern recovery.
If $\hat{\beta}_2 = (4.01, 3.99, 3, -4)'$, then $\mathrm{patt}(\hat{\beta}) = (4, 2, 1, -3) \neq \mathrm{patt}(\beta)$ and there is no pattern recovery.
However, it is natural to round up (threshold)

$$\hat{\beta}_2 = (4.01, 3.99, 3, -4)' \approx (4, 4, 3, -4)'.$$

The thresholded estimator $\hat{\beta}_2^{thresh}$ recovers the pattern of β.

Accessibility of a pattern

Not all the patterns can be realized by $\hat{\beta}$ when $p > n$.

Definition (Accessible pattern)

Let $X \in \mathbb{R}^{n\times p}$ and *pen* be a polyhedral norm. We say that $\beta \in \mathbb{R}^p$ has an accessible pattern with respect to X and *pen*, if there exists $y \in \mathbb{R}^n$ and $\hat{\beta} \in S_{X,pen}$ such that $\mathrm{patt}(\hat{\beta}) = \mathrm{patt}(\beta)$.

Accessibility of a pattern

Proposition (Geometric characterization of accessible patterns, [2])

The pattern of $\beta \in \mathbb{R}^p$ is accessible with respect to X and pen if and only if

$$\mathrm{row}(X) \cap \partial(pen)(\beta) \neq \emptyset.$$

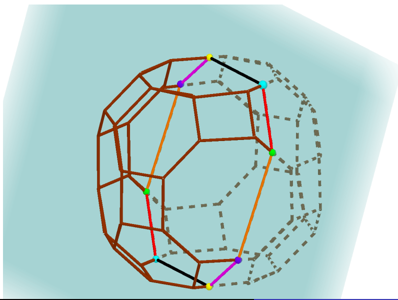
Proof. ($\Longrightarrow$) When the pattern of β is accessible with respect to X and *pen*, there exists $y \in \mathbb{R}^n$ and $\hat{\beta} \in S_{X,pen}(y)$ such that $\partial(pen)(\hat{\beta}) = \partial(pen)(\beta)$. Because $\hat{\beta}$ is a minimizer, $X'(y - X\hat{\beta}) \in \partial(pen)(\hat{\beta}) = \partial(pen)(\beta)$, so that, clearly, $\mathrm{col}(X') = \mathrm{row}(X)$ intersects $\partial(pen)(\beta)$.

($\Longleftarrow$) If $\mathrm{row}(X)$ intersects the face $\partial(pen)(\beta)$, then there exists $z \in \mathbb{R}^n$ such that $X'z \in \partial(pen)(\beta)$. For $y = X\beta + z$, we have $X'(y - X\beta) = X'z$, so that $\beta \in S_{X,pen}(y)$ and $\mathrm{patt}(\beta)$ is accessible with respect to X and *pen*. $\square$

$n = 2, p = 3$: typically, 17 patterns accessible from 147 The Figure is from [S-T]

colour	type	intersection $\neq \varnothing$	face intersected isometric to	SLOPE models
orange	segments	$\mathrm{row}(X) \cap F_w(\pm(1,0,0))$	$\{5.5\} \times P^{\pm}_{(3.5,1.5)}$	$\pm(1,0,0)$
red	segments	$\mathrm{row}(X) \cap F_w(\pm(1,1,1))$	$P_{(5.5,3.5,1.5)}$	$\pm(1,1,1)$
black	segments	$\mathrm{row}(X) \cap F_w(\pm(0,0,1))$	$\{5.5\} \times P^{\pm}_{(3.5,1.5)}$	$\pm(0,0,1)$
pink	segments	$\mathrm{row}(X) \cap F_w(\pm(-1,0,1))$	$P_{(5.5,3.5)} \times [-1.5, 1.5]$	$\pm(-1,0,1)$
purple	points	$\mathrm{row}(X) \cap F_w(\pm(2,0,-1))$	$\{5.5\} \times \{3.5\} \times [-1.5, 1.5]$	$\pm(2,0,-1)$
green	points	$\mathrm{row}(X) \cap F_w(\pm(2,1,1))$	$\{5.5\} \times P_{(3.5,1.5)}$	$\pm(2,1,1)$
blue	points	$\mathrm{row}(X) \cap F_w(\pm(1,1,2))$	$\{5.5\} \times P_{(3.5,1.5)}$	$\pm(1,1,2)$
yellow	points	$\mathrm{row}(X) \cap F_w(\pm(-1,0,2))$	$\{5.5\} \times \{3.5\} \times [-1.5, 1.5]$	$\pm(-1,0,2)$

Table 1: Accessible SLOPE models with respect to $X = \left(\begin{smallmatrix} 8 & 5 & 8 \\ 10 & 1.25 & -6 \end{smallmatrix}\right)$ and $w = (5.5, 3.5, 1.5)'$.



SLOPE pattern matrix

In order to characterize the SLOPE pattern recovery, we will need some more notions related to a pattern M.

Definition

Let $0 \neq M = (M_1, \ldots, M_p)' \in \mathcal{P}_p^{\mathrm{SLOPE}}$ with $k = \|M\|_\infty$.

Pattern matrix: $U_M \in \mathbb{R}^{p \times k}$ is defined as follows

$$(U_M)_{ij} = \mathrm{sign}(M_i)\mathbf{1}_{(|M_i| = k+1-j)},\ i \in \{1, \ldots, p\}, j \in \{1, \ldots, k\}.$$

Example. Let $M = (1, 2, -2, 0, -1)'$. Then $|M|_\downarrow = (2, 2, 1, 1, 0)'$

$$U_M = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ -1 & 0 \\ 0 & 0 \\ 0 & -1 \end{pmatrix} \qquad U_{|M|_\downarrow} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}$$

$U_M\mathbb{R}^{k+}$ gives all vectors with pattern M

For $k \geq 1$ we denote by $\mathbb{R}^{k+} = \{\kappa \in \mathbb{R}^k \colon \kappa_1 > \ldots > \kappa_k > 0\}$. Definition of U_M implies that for $0 \neq M \in \mathcal{P}_p^{\mathrm{SLOPE}}$ and $k = \|M\|_\infty$, for $b \in \mathbb{R}^p$ we have

$$\mathrm{patt}(b) = M \iff \text{there exists } \kappa \in \mathbb{R}^{k+} \text{ such that } b = U_M\kappa.$$

Example. Let $M = (1, 2, -2, 0, -1)'$ and $\kappa = (\kappa_1, \kappa_2)'$. Then

$$U_M\kappa = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ -1 & 0 \\ 0 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix} = \begin{pmatrix} \kappa_2 \\ \kappa_1 \\ -\kappa_1 \\ 0 \\ -\kappa_2 \end{pmatrix}$$

Clustered matrix $\tilde{X}_M$ and clustered parameter $\tilde{\Lambda}_M$

Definition (Clustered matrix and $\Lambda-$ parameter)

Let $X \in \mathbb{R}^{n\times p}$, $\Lambda = (\lambda_1, \ldots, \lambda_p)$ where $\lambda_1 > \cdots > \lambda_p > 0$.

Clustered matrix: $\tilde{X}_M = XU_M$.

Clustered parameter: $\tilde{\Lambda}_M = (U_{|M|_\downarrow})'\Lambda$.

Example. Let $X = (X_1|X_2|X_3|X_4|X_5)$, $M = (1, 2, -2, 0, -1)'$ and $\Lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)'$ where $\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4 > \lambda_5 > 0$.

$$\tilde{X}_M = (X_2 - X_3|X_1 - X_5) \text{ and } \tilde{\Lambda}_M = \begin{pmatrix} \lambda_1 + \lambda_2 \\ \lambda_3 + \lambda_4 \end{pmatrix}.$$

The clustered design matrix $\tilde{X}_M$ has only $k = 2$ columns instead of $p = 5$.

If $\mathrm{patt}(\beta) = M$, then $X\beta = XU_M\kappa = \tilde{X}_M\kappa$ for $\kappa \in \mathbb{R}^{k+}$. In particular,

1. null components $M_i = 0$ lead to discard the column X_i from the design matrix X,
2. a cluster $K \subset \{1, \dots, p\}$ of M (component of M equal in absolute value) leads to replace the columns $(X_i)_{i \in K}$ by one column equal to the signed sum: $\sum_{i \in K} sign(M_i)X_i$.

New characterization of $\partial(J_\Lambda)$ for SLOPE

The next Proposition provides a new and useful formula for the subdifferential of the sorted ℓ_1 norm, via an optimal system of affine equations. This representation is crucial for the paper [1].

Proposition (Affine characterization of $\partial(J_\Lambda)$ for SLOPE)

Let $b \in \mathbb{R}^p$ and $M = \mathrm{patt}(b)$. Then we have the following formula:

$$\partial J_\Lambda(b) = \left\{ v \in \mathbb{R}^p : J^*_\Lambda(v) \leq 1 \text{ and } U'_M v = \tilde{\Lambda}_M \right\}.$$

Moreover, the affine space generated by $\partial J_\Lambda(b)$ equals $\left\{ v \in \mathbb{R}^p \mid U'_M v = \tilde{\Lambda}_M \right\}$.

Example. For $M = (1, 2, -2, 0, -1)'$ the condition $U'_M v = \tilde{\Lambda}_M$ means

$$v_2 - v_3 = \lambda_1 + \lambda_2, \qquad v_1 - v_5 = \lambda_3 + \lambda_4.$$

This description is much more performant than the hyperplane equation $v'M = J_\Lambda(M)$ that we saw before!

Proof.

Let us prove the inclusion

$$\partial J_\Lambda(b) \supset \left\{ v \in \mathbb{R}^p : J_\Lambda^*(v) \leq 1 \text{ and } U_M' v = \tilde{\Lambda}_M \right\}$$

Assume that $v \in \mathbb{R}^p$ satisfies $J_\Lambda^*(v) \leq 1$ and $U_M' v = \tilde{\Lambda}_M$.
To prove that $v \in \partial J_\Lambda(b)$ it remains to establish that $b'v = J_\Lambda(b)$.
Since $b = U_M s$, where $s \in \mathbb{R}^{k+}$, we have

$$b'v = (U_M s)'v = s'U_M' v = s'\tilde{\Lambda}_M = J_\Lambda(b).$$

The proof of the other inclusion is also elementary but longer, we omit it.

Characterization of pattern recovery by SLOPE

The characterization of pattern recovery by SLOPE given in the next Theorem is the main mathematical result of article.
The main statistical results of paper [1] are based thoroughly on this characterization Theorem.

Given a SLOPE minimizer $\hat{\beta} \in S_{X,\gamma J_\lambda}(Y)$ for which $\mathrm{patt}(\hat{\beta}) = M \neq 0$, we observe that the following two simple properties occur:

Dual ball condition: for $\pi = X'(y - X\hat{\beta})$, we have $\boxed{J^*_\Lambda(\pi) \leq 1.}$

(Actually, we know more: $\pi \in \partial(J_\Lambda)(M)$)

Positivity condition: Consider the vector

$\tilde{X}'_M X\hat{\beta} = \tilde{X}'_M X U_M s = \tilde{X}'_M \tilde{X}_M s$, where $s \in \mathbb{R}^{k+}$.

Thus we have $\boxed{\exists s \in \mathbb{R}^{k+} \quad \tilde{X}'_M X\hat{\beta} = \tilde{X}'_M \tilde{X}_M s.}$

Getting rid of $\hat{\beta}$ in the $\boxed{\textbf{two conditions}}$ by some simple algebraic operations, including:

- the Moore-Penrose pseudo-inverse A^+ of A
- $\tilde{P}_M = (\tilde{X}'_M)^+ \tilde{X}'_M = \tilde{X}_M \tilde{X}^+_M$, the projector onto the space $\mathrm{col}(\tilde{X}_M)$

we derive the **necessity** of two conditions of the next Theorem.

It is next easy to show that these $\boxed{\text{two conditions}}$ are also **sufficient** for the recovery of the pattern M.

Theorem (Characterization of SLOPE pattern recovery by positivity and dual ball conditions)

Let $X \in \mathbb{R}^{n\times p}$, $0 \neq \beta \in \mathbb{R}^p$, $Y = X\beta + \varepsilon$ for $\varepsilon \in \mathbb{R}^n$, $\Lambda \in \mathbb{R}^{p+}$. Let $M = \mathrm{patt}(\beta) \in \mathcal{P}^{\mathrm{SLOPE}}_p$ and $k = \|M\|_\infty$. Define

$$\pi = X'(\tilde{X}'_M)^+ \tilde{\Lambda}_M + X'(I_n - \tilde{P}_M)Y.$$

There exists $\hat{\beta} \in S_{X,\Lambda}(Y)$ with $\mathrm{patt}(\hat{\beta}) = \mathrm{patt}(\beta)$ if and only if the two conditions below hold true:

$$\begin{cases} \text{there exists } s \in \mathbb{R}^{k+} \text{ such} & \text{that } \quad \tilde{X}'_M Y - \tilde{\Lambda}_M = \tilde{X}'_M \tilde{X}_M s \\ & (\textbf{positivity condition}), \\ J^*_\Lambda(\pi) \leq 1 & (\textbf{dual ball condition}). \end{cases}$$

If the positivity and ball conditions are satisfied, then $\hat{\beta} = U_M s \in S_{X,\Lambda}(Y)$ and $\pi = X'(Y - X\hat{\beta})$.

Proof of necessity of two conditions for model recovery.

Let $\hat\beta \in S_{X,J_\Lambda}(Y)$ with $\mathrm{patt}(\hat\beta) = M$, i.e. $\hat\beta = U_M s$, $s \in \mathbb{R}^{k+}$.
We have $X'(Y - X\hat\beta) \in \partial J_\Lambda(M)$. We want to deduce $\tilde X'_M X\hat\beta$ from this inclusion.
Multiplying it by U'_M, by the affine characterization of subdifferential, we get
$\tilde X'_M(Y - X\hat\beta) = \tilde\Lambda_M$ and $\tilde X'_M X\hat\beta = \tilde X'_M Y - \tilde\Lambda_M$.
The positivity condition is proven.
Apply $(\tilde X'_M)^+$ to the last equality $\tilde X'_M X\hat\beta = \tilde X'_M Y - \tilde\Lambda_M$ and use the fact that $\tilde{P_M} = (\tilde X'_M)^+\tilde X'_M$ is the projector onto $col(\tilde X_M)$. We have $X\hat\beta = \tilde X_M s \in \mathrm{col}(\tilde X_M)$ so that $\tilde P_M X\hat\beta = X\hat\beta$. We get

$$(\tilde X'_M)^+\tilde X'_M X\hat\beta = \tilde P_M Y - (\tilde X'_M)^+\tilde\Lambda_M \Rightarrow X\hat\beta = \tilde P_M Y - (\tilde X'_M)^+\tilde\Lambda_M$$

We insert this formula for $X\hat\beta$ in
$B^* \ni X'(Y - X\hat\beta) = X'(\tilde X'_M)^+\tilde\Lambda_M + X'(I - \tilde P_M)Y$.
We proved the dual ball condition.

Necessary condition for model recovery: $\tilde\Lambda_M \in \mathrm{col}(\tilde X'_M)$

Observe that the positivity condition:

$$\boxed{\text{there exists } s \in \mathbb{R}^{k+} \text{ such that } \quad \tilde X'_M Y - \tilde\Lambda_M = \tilde X'_M \tilde X_M s}$$

implies that the property

$$\tilde\Lambda_M \in \mathrm{col}(\tilde X'_M)$$

(or equivalently, the projector $\tilde X'_M(\tilde X'_M)^+\tilde\Lambda_M = \tilde\Lambda_M$)
is necessary for the positivity condition.
The condition $\tilde\Lambda_M \in \mathrm{col}(\tilde X'_M)$ automatically holds when $n \geq k$ and $\mathrm{col}(\tilde X'_M) = \mathbb{R}^k$.

Essential term $X'(\tilde{X}'_M)^+\tilde{\Lambda}_M$ in the dual ball condition

The first term $\boxed{X'(\tilde{X}'_M)^+\tilde{\Lambda}_M}$ in the expression
$\pi = X'(\tilde{X}'_M)^+\tilde{\Lambda}_M + X'(I_n - \tilde{P}_M)Y$
is essential for the dual ball condition. Actually, the second term

$$X'(I_n - \tilde{P}_M)Y = X'(I_n - \tilde{P}_M)X\beta + X'(I_n - \tilde{P}_M)\varepsilon = X'(I_n - \tilde{P}_M)\varepsilon$$

will be shown neglectable, under natural conditions on the (strong) signal β or when $n \to \infty$.

Noiseless case

The second term is null in the **noiseless case** $\varepsilon = 0$.

The dual ball condition becomes $\boxed{J^*_\Lambda(X'(\tilde{X}'_M)^+\tilde{\Lambda}_M) \le 1}$

We check that the positivity condition holds for $\alpha\Lambda$ with Λ verifying the $\boxed{\text{necessary condition } \tilde{\Lambda}_M \in \mathrm{col}(\tilde{X}'_M)}$ and $\alpha > 0$ small enough.

We prove the following characterization of SLOPE pattern recovery in the noiseless case.

SLOPE IR $\iff$ noiseless pattern recovery

Corollary (SLOPE IR $\iff$ pattern recovery for $\varepsilon = 0$)

Consider the noiseless case when $\varepsilon = 0$.
There exists $\alpha > 0$ such that SLOPE with tuning parameter $\alpha\Lambda$ recovers the pattern $\mathrm{patt}(\beta) = M$ if and only if

$$J^*_\Lambda(X'(\tilde{X}'_M)^+\tilde{\Lambda}_M) \leq 1 \text{ and } \tilde{\Lambda}_M \in \mathrm{col}(\tilde{X}'_M).$$

Then $\exists\alpha_0$ such that for all $0 < \alpha < \alpha_0$, SLOPE with tuning parameter $\alpha\Lambda$ recovers the pattern of β.

By analogy to LASSO terminology (Zou, Wainwright, de Geer) we say that the **SLOPE Irrepresentability(IR) Condition holds** if

$$\boxed{J^*_\Lambda(X'(\tilde{X}'_M)^+\tilde{\Lambda}_M) \leq 1 \text{ and } \tilde{\Lambda}_M \in \mathrm{col}(\tilde{X}'_M)}$$

(or equivalently $X'(\tilde{X}'_M)^+\tilde{\Lambda}_M \in \partial J_\Lambda(M))$).
When $\ker(\tilde{X}_M) = \{0\}$ then the SLOPE IR condition reads:

$$J^*_\Lambda(X'\tilde{X}_M(\tilde{X}'_M\tilde{X}_M)^{-1}\tilde{\Lambda}_M) \leq 1.$$

Example, $p = 2$, $n \geq 2$

Let $X = (X_1|X_2) \in \mathbb{R}^{n\times 2}$ such that

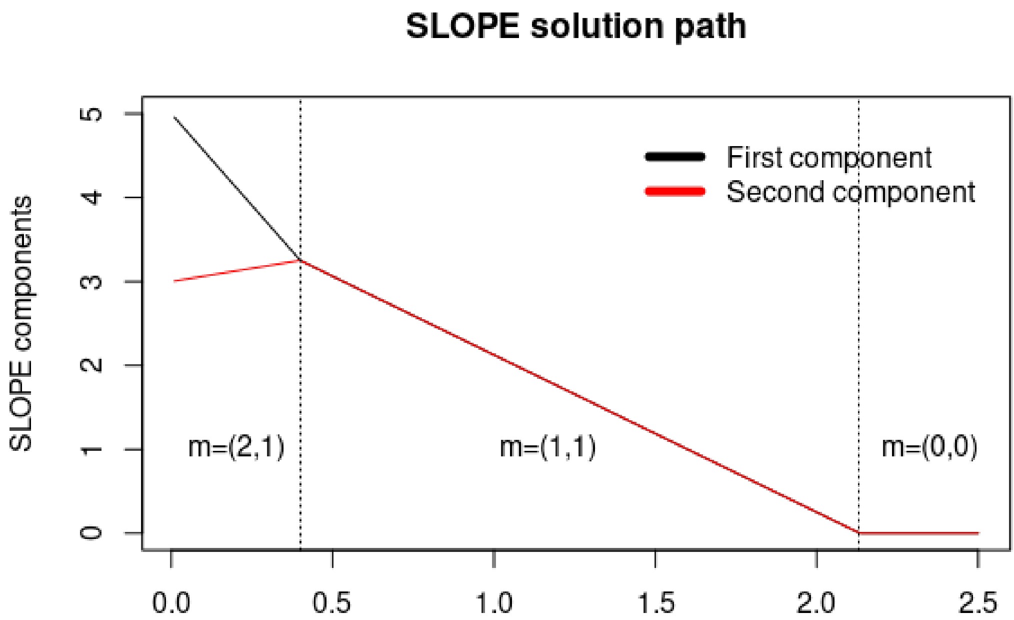
$$X'X = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}.$$

Let $\Lambda = (4,2)'$, $\beta = (5,3)'$, $M = \mathrm{patt}(\beta) = (2,1)'$.
$\tilde{X}_M = X$ and $\tilde{\Lambda}_M = \Lambda$.
$\ker(\tilde{X}_M) = \{0\}$ and

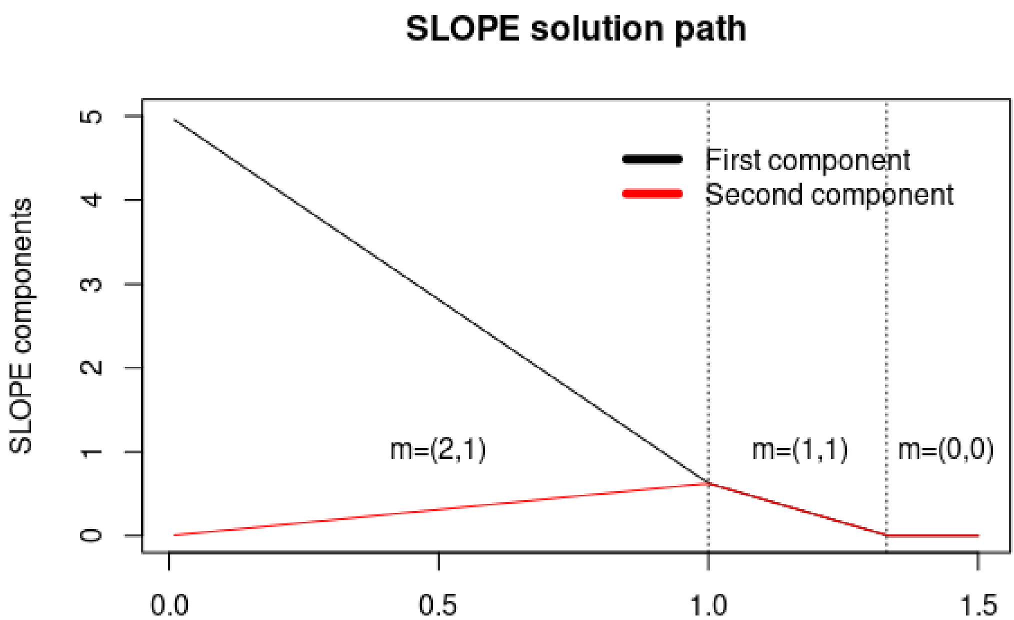
$$J^*_\Lambda(X'(\tilde{X}'_M)^+\tilde{\Lambda}_M) = J^*_\Lambda(X'X(X'X)^{-1}\Lambda) = J^*_\Lambda(\Lambda) = 1 \leq 1.$$

The SLOPE irrepresentability condition holds true, so the noiseless pattern recovery holds for for $\alpha\Lambda$.
Using R, we see that $0 < \alpha < 0.4$ garantees the pattern recovery.

Noiseless pattern recovery holds for $\beta = (5, 3)'$, pattern $= (2, 1)'$



IR does not hold for $\beta = (5, 0)'$, pattern $= (1, 0)'$ $J_{(4,2)}(X'\tilde{X}_M'^{+}\tilde{\Lambda}_M) = 6.4/4 > 1$



Geometrical meaning of $\pi_1 := X'(\tilde{X}'_M)^+\tilde{\Lambda}_M$

Proposition ($\pi_1 := X'(\tilde{X}'_M)^+\tilde{\Lambda}_M$ is a meeting point)

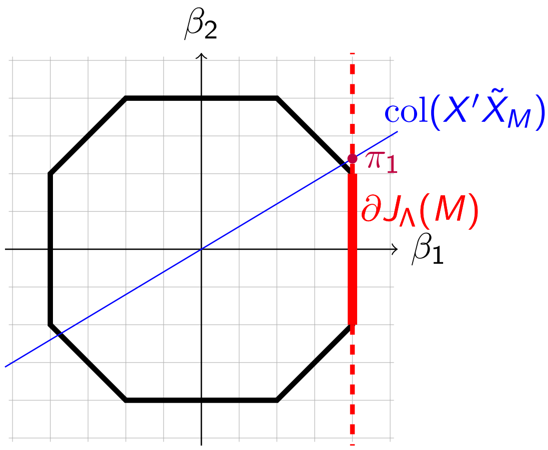
Suppose that $\tilde{\Lambda}_M \in \mathrm{col}(\tilde{X}'_M)$.
Then $\{\pi_1\} = \mathrm{aff}(\partial J_\Lambda(M)) \cap \mathrm{col}(X'\tilde{X}_M)$.

Proof. We use the Proposition on Affine characterization to π_1.
Since $\tilde{X}'_M(\tilde{X}'_M)^+$ is the projection on $\mathrm{col}(\tilde{X}'_M)$ we have

$$U'_M\pi_1 = \tilde{X}'_M(\tilde{X}'_M)^+\tilde{\Lambda}_M = \tilde{\Lambda}_M.$$

Thus $\pi_1 \in \mathrm{aff}(\partial J_\Lambda(M))$.
Moreover, since $\mathrm{col}((\tilde{X}'_M)^+) = \mathrm{col}(\tilde{X}_M)$,
we deduce that $\pi_1 \in \mathrm{col}(X'(\tilde{X}'_M)^+) = \mathrm{col}(X'\tilde{X}_M)$.
We omit the (short) proof of unicity of the meeting point. □

Back to the Example: $\beta = (5,0)'$, pattern $= (1,0)'$ $J^*_{(4,2)}(\pi_1) > 1$,
the meeting point π_1 is not in the pattern face $\partial J_\Lambda(M)$



Meeting point IR for any polyhedral *pen*

For SLOPE, the space $\mathrm{col}(X'\tilde{X}_M) = X'X\mathrm{col}U_M = X'X\mathrm{lin}C_M$ where $C_M = U_M\mathbb{R}^{k+}$ is the "pattern set" of all $x \in \mathbb{R}^p$ with the same pattern as M, i.e.

$$\partial J_\Lambda(x) = \partial J_\Lambda(M)$$

The "pattern set" can be defined for any penalty *pen*.
The meeting point π_1 of aff$\partial pen(x)$ and $XX'\mathrm{lin}C_M$ is well defined for any penalty *pen*.
In [2] we conjecture that the condition $\boxed{\pi_1 \in \partial pen(x)}$ is equivalent to the Noiseless pattern recovery for any polyhedral *pen*.
(proof at finish)

LASSO analogues of our SLOPE characterization Theorem and our SLOPE IR condition

Consider the LASSO sign recovery (*i.e.* existence of estimator $\hat{\beta}^{\mathrm{LASSO}}$ such that $\mathrm{sign}(\hat{\beta}^{\mathrm{LASSO}}) = \mathrm{sign}(\beta) = S \in \{-1,0,1\}^p$)
The LASSO analogue of our characterization Theorem with $\boxed{\text{positivity and dual ball conditions}}$ **is new**. In conclusion we get

Corollary (New LASSO Irrepresentability condition)

Consider the noiseless case when $\varepsilon = 0$.
There exists $\lambda > 0$ such that LASSO with tuning parameter λ recovers $\mathrm{sign}(\beta) = S$ if and only if

$$\|X'(\tilde{X}_S')^+ 1_{\mathbb{R}^k}\|_\infty \le 1 \quad \text{and } 1_{\mathbb{R}^k} \in \mathrm{col}(\tilde{X}_S').$$

Here $\tilde{X}_S'$ is the design matrix X signed and reduced according to S.
Example. If $S = (1,0,-1,0)'$ and $X = (X_1|X_2|X_3|X_4)$, then $\tilde{X}_S' = (X_1, -X_3)$.

New and old LASSO IR condition

The two conditions

$\boxed{\|X'(\tilde{X}'_S)^+ 1_{\mathbb{R}^k}\|_\infty \leq 1 \text{ and } 1_{\mathbb{R}^k} \in \mathrm{col}(\tilde{X}'_S)}$

equivalent to noiseless LASSO sign recovery are new.

When $\ker(\tilde{X}_S) = \{0\}$ then $1_k \in \mathrm{col}(\tilde{X}'_S)$ occurs and $\|X'(\tilde{X}'_S)^+ 1_k\|_\infty \leq 1$ is equivalent to

$$\|X'_{\bar{I}} X_I (X'_I X_I)^{-1} S_I\|_\infty \leq 1$$

where $I = \mathrm{supp}(S)$, $\bar{I} = \{1, \ldots, p\} \setminus I$

(M_I denotes the submatrix of M obtained by keeping columns corresponding to indices in I)

This latter expression is known in literature as the LASSO irrepresentability condition (Fuchs, Zhao, Zou, Wainwright, de Geer).

Symmetric error. Necessity of the SLOPE IR Condition

Corollary

Let $Y = X\beta + \varepsilon$ where ε and $-\varepsilon$ have the same distribution. If $J^*_\Lambda(X'(\tilde{X}'_M)^+ \tilde{\Lambda}_M) > 1$ or $\Lambda_M \notin \mathrm{col}(\tilde{X}'_M)$ then the probability of pattern recovery by SLOPE is smaller than $1/2$.

For LASSO, a similar result when $\ker \tilde{X}_S = \{0\}$, was obtained by Wainwright (2009).

Asymptotic Pattern Recovery (Pattern Consistency) when $\varepsilon \neq 0$. Open IR Condition.

In order to give a sufficient condition for pattern recovery, we must strengthen SLOPE *IR* condition to an Open SLOPE *IR* condition (this also happens with LASSO)

Recall that our SLOPE IR condition is equivalent to

$$X'(\tilde{X}_M')^+\tilde{\Lambda}_M \in \partial J_\Lambda(M)$$

The Open SLOPE *IR* condition is

$$X'(\tilde{X}_M')^+\tilde{\Lambda}_M \in \mathrm{ri}(\partial J_\Lambda(M))$$

where $\mathrm{ri}(F)$ is the relative interior of F.

Open IR Condition is numerically effective

The Open IR Condition $X'(\tilde{X}_M')^+\tilde{\Lambda}_M \in \mathrm{ri}(\partial J_\Lambda(M))$
is equivalent to the following computationally verifiable conditions:

$$\begin{cases} J^*_\Lambda(X'(\tilde{X}_M')^+\tilde{\Lambda}_M) \leq 1 \text{ and } \tilde{\Lambda}_M \in \mathrm{col}(\tilde{X}_M'), \\ \left|\left\{i \in \{1,\dots,p\} \colon \sum_{j=1}^i |X'(\tilde{X}_M')^+\tilde{\Lambda}_M|_{(j)} = \sum_{j=1}^i \lambda_j\right\}\right| = \|M\|_\infty. \end{cases}$$

We count the number of equalities in p inequalities equivalent to $J^*_\Lambda(b) \leq 1$. Recall that

$$J^*_\Lambda(b) = \max\left\{\frac{|b|_{(1)}}{\lambda_1}, \frac{|b|_{(1)} + |b|_{(2)}}{\lambda_1+\lambda_2}, \dots, \frac{|b|_{(1)} + \ldots + |b|_{(p)}}{\lambda_1+\ldots+\lambda_p}\right\}.$$

Asymptotic Pattern Recovery (Pattern Consistency) when $\varepsilon \neq 0$: Open IR, big tuning and strong signal are sufficient

$$S_{X,\alpha\Lambda}(Y) = \underset{b \in \mathbb{R}^p}{argmin} \frac{1}{2}\|Y - Xb\|_2^2 + \alpha J_\Lambda(b).$$

Theorem (Pattern consistency with X fixed)

Let $X \in \mathbb{R}^{n\times p}$, $0 \neq M \in \mathcal{P}_p^{\mathrm{SLOPE}}$, and $\Lambda = (\lambda_1, \ldots, \lambda_p)'$ where $\lambda_1 > \ldots > \lambda_p > 0$. $(\beta^{(r)})_{r\geq 1}$ sequence with pattern M:

- $\beta^{(r)} = U_M s^{(r)}$ *with* $s_1^{(r)} > \ldots > s_k^{(r)} > 0$ *and* $k = \|M\|_\infty$,
- $\Delta_r = \min_{1\leq i<k}\left(s_i^{(r)} - s_{i+1}^{(r)}\right) \xrightarrow{r\to\infty} \infty.$ STRONG SIGNAL

Let $Y^{(r)} = X\beta^{(r)} + \varepsilon$, where ε is an arbitrary vector in $\mathbb{R}^n$. If $\boxed{\alpha_r \to \infty,\ \alpha_r/\Delta_r \to 0 \text{ as } r \to \infty}$ *and*
$X'(\tilde{X}_M')^+\tilde{\Lambda}_M \in \mathrm{ri}(\partial J_\Lambda(M))$, OPEN IR
then $\exists\, r_0 > 0\ \forall r \geq r_0\ \exists \hat{\beta} \in S_{X,\alpha_r\Lambda}(Y^{(r)})$ such that $\mathrm{patt}(\hat{\beta}) = M$.

Pattern consistency with p fixed and $n \to \infty$

We suppose:
$X = X_n$ random, satisfying a natural Lindeberg-Feller condition;
an incremental error $\varepsilon_n = (\epsilon_1, \ldots, \epsilon_n)'$, where $(\epsilon_i)_i$ are i.i.d. centered with finite variance;
$(X_n)_n$ and $(\epsilon_n)_n$ are independent.

Theorem (Pattern consistency with $n \to \infty$)

Let $X \in \mathbb{R}^{n\times p}$ such that $\frac{1}{n}X'X \to C$ almost surely when $n \to \infty$, $0 \neq \beta \in \mathbb{R}^p$ and $M = \mathrm{patt}(\beta)$. If $\lim_{n\to\infty} \frac{\alpha_n}{n} = 0$, $\lim_{n\to\infty} \frac{\alpha_n}{\sqrt{n}} = \infty$ and

$$CU_M(U_M'CU_M)^{-1}\tilde{\Lambda}_M \in \mathrm{ri}(\partial J_\Lambda(M))$$

then

$$\mathrm{patt}(\hat{\beta}_n^{SLOPE}) \xrightarrow{\mathbb{P}} \mathrm{patt}(\beta).$$

Strong Pattern consistency with $n \to \infty$, for all ω

Assume additionnally that the rows of X_n are independent and that each row of X_n has the same law as ξ, where ξ is a random vector whose components are linearly independent a.s. and that $\mathbb{E}[\xi_i^2] < \infty$ for $i = 1, \ldots, p$.

Theorem (Strong Pattern consistency with $n \to \infty$)

Let $X \in \mathbb{R}^{n \times p}$ *such that* $\frac{1}{n} X'X \to C$ *almost surely when* $n \to \infty$, $0 \neq \beta \in \mathbb{R}^p$ *and* $M = \mathrm{patt}(\beta)$. *If* $\lim_{n\to\infty} \frac{\alpha_n}{n} = 0$, $\lim_{n\to\infty} \frac{\alpha_n}{\sqrt{n \log\log(n)}} = \infty$ *and*

$$CU_M(U_M' CU_M)^{-1}\tilde{\Lambda}_M \in \mathrm{ri}(\partial J_\Lambda(M))$$

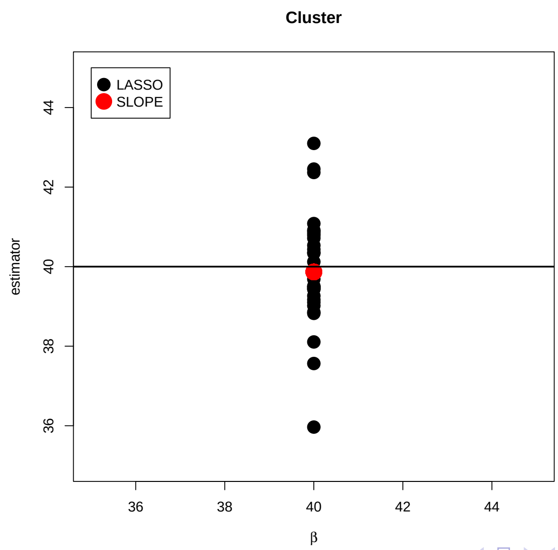
then

$$\mathrm{patt}(\hat{\beta}_n^{SLOPE}) \xrightarrow{\forall \omega} \mathrm{patt}(\beta).$$

Simulation study: example already seen

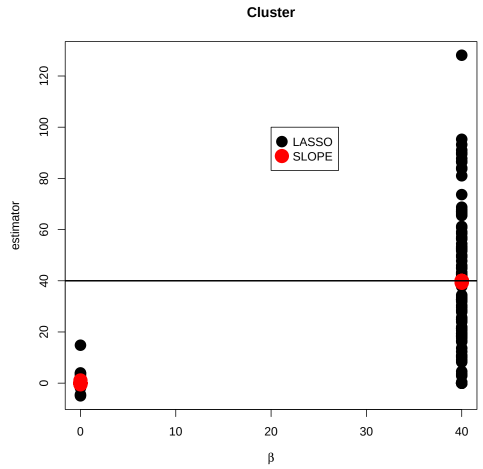
Consider $Y = X\beta + \varepsilon$ where ε has iid $N(0, 5^2)$ entries and

- $\beta_1 = \ldots = \beta_{30} = 40$ and $\beta_{31} = \ldots = \beta_{200} = 0$.
- $X'\tilde{X}_M(\tilde{X}_M'\tilde{X}_M)^{-1}\tilde{\Lambda}_M \in \mathrm{ri}(\partial J_\Lambda(M))$.
- $\|X_{\bar{I}}'X_I(X_I'X_I)^{-1}\mathrm{sign}(\beta_I)\|_\infty \leq 1$.



Consider $Y = X\beta + \varepsilon$ where ε has iid $N(0, 5^2)$ entries and

- $\beta_1 = \ldots = \beta_{100} = 40$ and $\beta_{101} = \ldots = \beta_{200} = 0$.
- $J^*_\Lambda(X'\tilde{X}_M(\tilde{X}'_M\tilde{X}_M)^{-1}\tilde{\Lambda}_M) > 1$.
- $\|X'_{\bar{I}}X_I(X'_IX_I)^{-1}\mathrm{sign}(\beta_I)\|_\infty > 1$.



Theorem [2]. Under the accessibility condition THRESHOLDED SLOPE asymptotically recovers the SLOPE pattern of β.

RESEARCH PROGRAM

RESEARCH PROGRAM (planned with H. Ishi, B. Kołodziejek, H. Nakashima)

Study of Pattern recovery for Graphical SLOPE on Graphical Gaussian Models

Pattern = clusters of equal terms and blocks of 0's $\iff$ Colored Graphical Models

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 (JP-MXP0619217849)
  "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 <u>**Tomonari Sei**</u> (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