2013年統計数理研究所夏期大学院
統計数理研究所　セミナー室 1
9月26日 14時40分～17時50分

ISM Summer School 2013

Information Geometry

Shinto Eguchi

The Institute of Statistical Mathematics, Japan
Email: eguchi@ism.ac.jp, komori@ism.ac.jp
Url: http//www.ism.ac.jp/~eguchi/

Outline

Information geometry

historical remarks

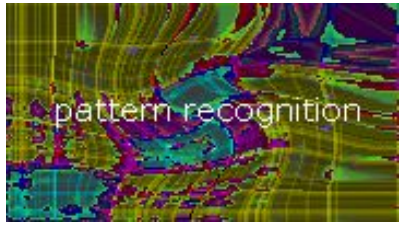
treasure example

Information divergence class

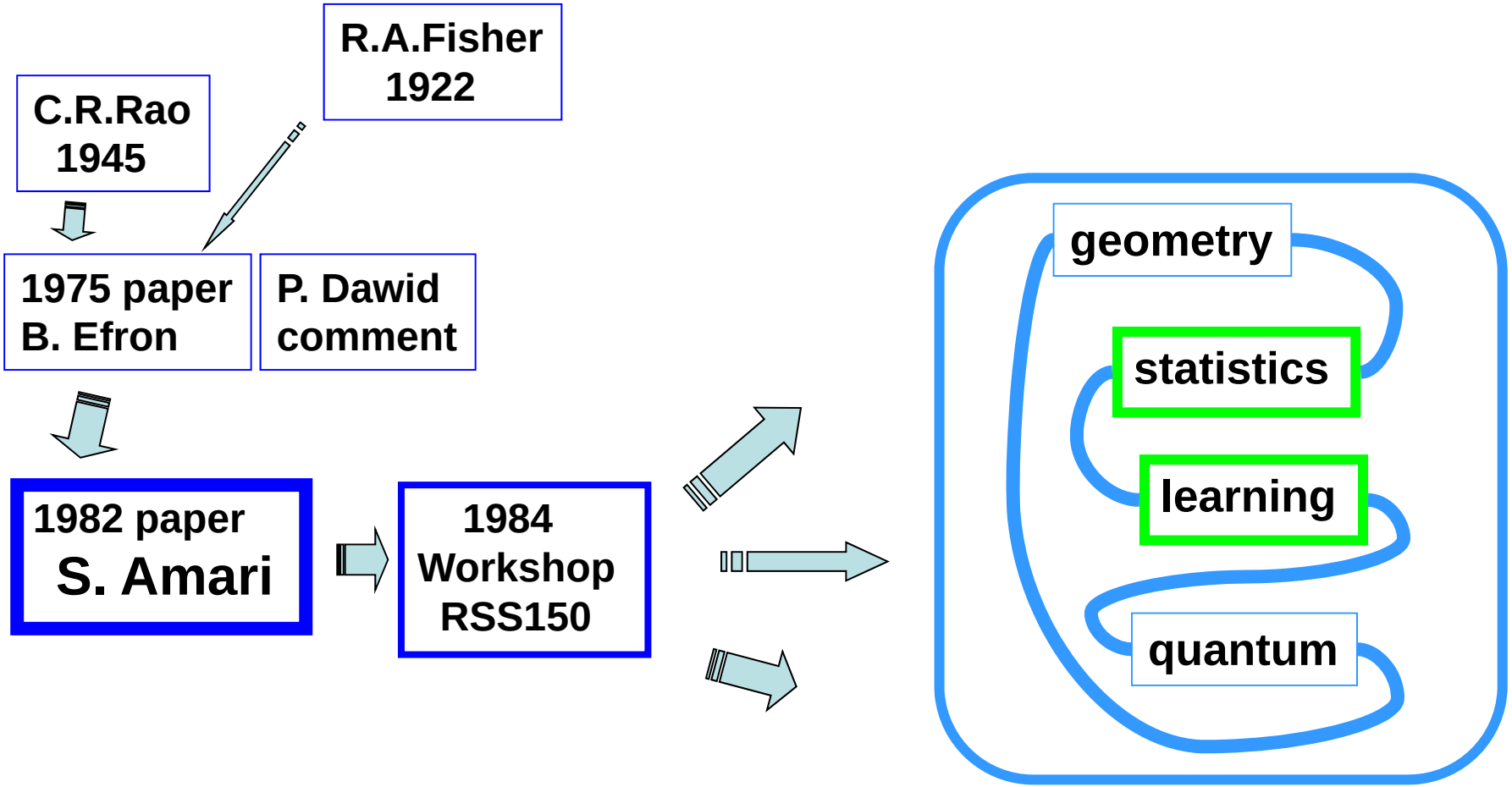
robust statistical methods

robust ICA

robust kernel PCA



Historical comment



Dual Riemannian Geometry

Dual Riemannian geometry gives reformulation for Fisher's foundation

Cf. Einstein field equations (1916)

Information Geometery aims to geometrize

A dualistic structure between modeling and estimation

Cf. Erlangen program (Klein, 1872)

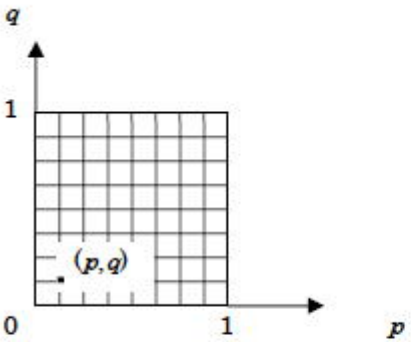
http://en.wikipedia.org/wiki/Erlangen_program

Estimation is projection of data onto model.

"Estimation is an action by projection and model is an object to be projected"

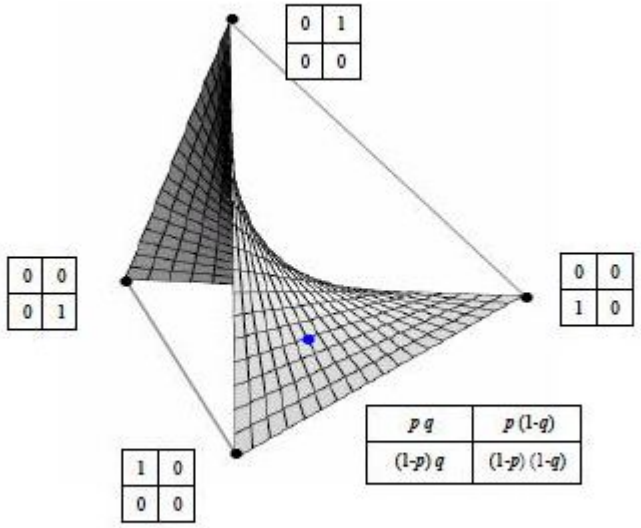
The interplay of action and object is elucidated by a geometric structure.

2 x 2 table



$p\,q$	$p\,(1-q)$
$(1-p)q$	$(1-p)\,(1-q)$

two-by-two table



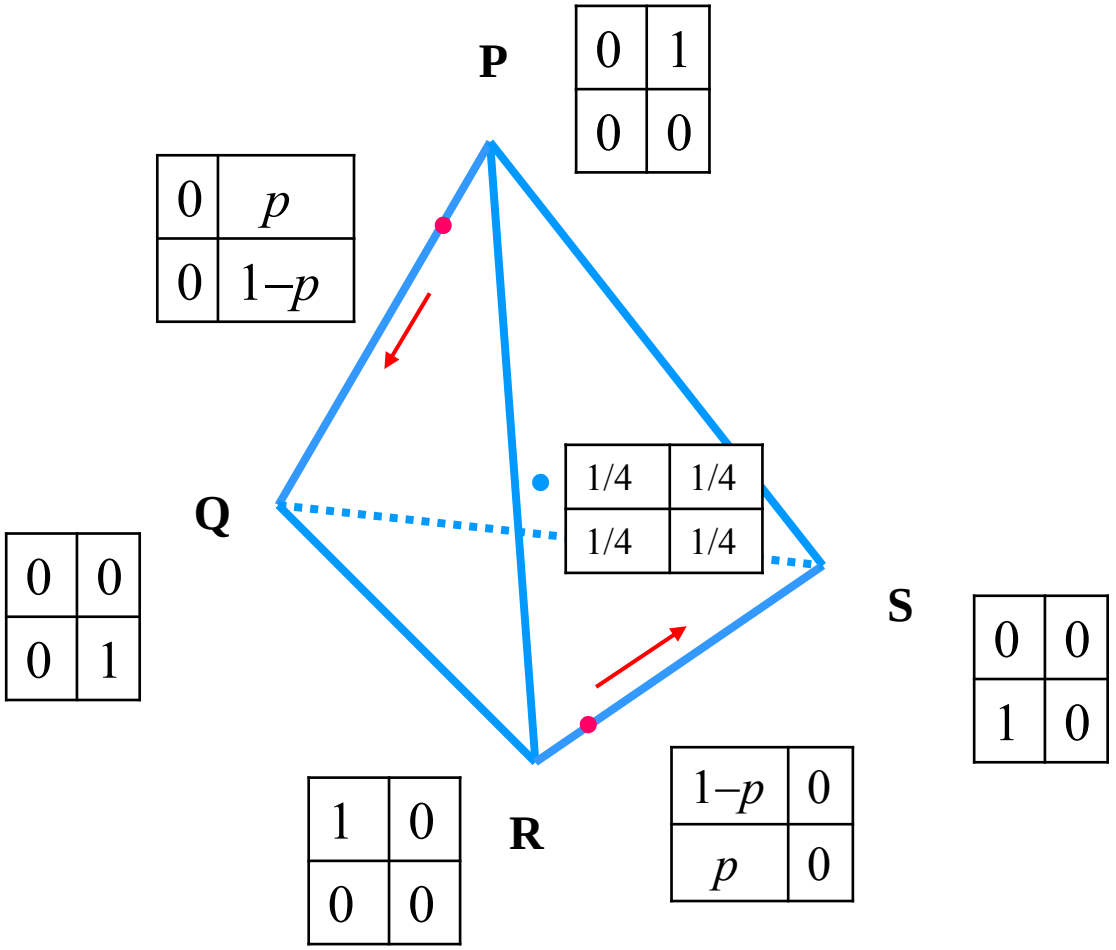
Independent 2×2 table

The space of all 2 × 2 tables associates with a regular tetrahedron

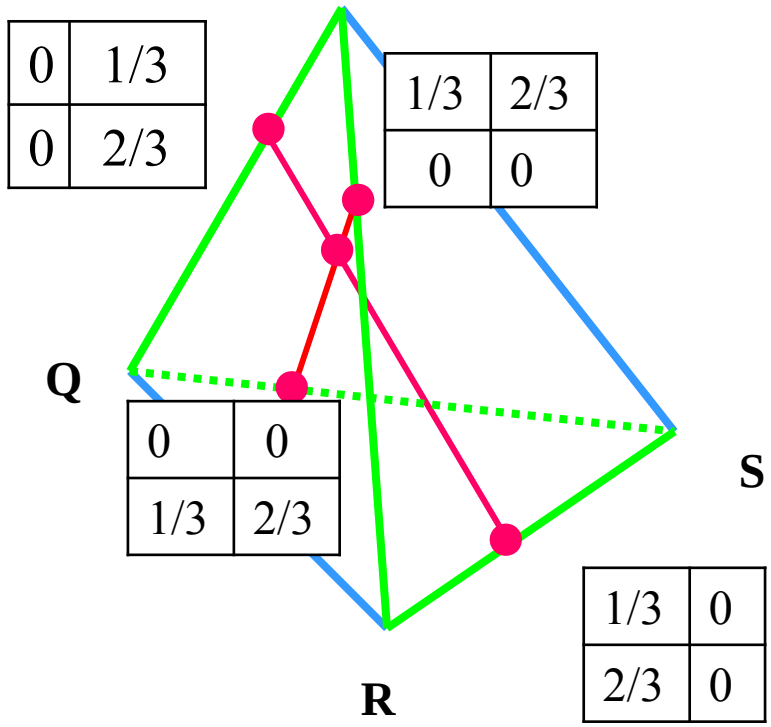
The space of all independent 2 × 2 tables associates with a ruled surface In the regular tetrahedron

We know the ruled surface is exponential geodesic, but does anyone know the minimality of the ruled surface?

regular tetrahedron



regular tetrahedron

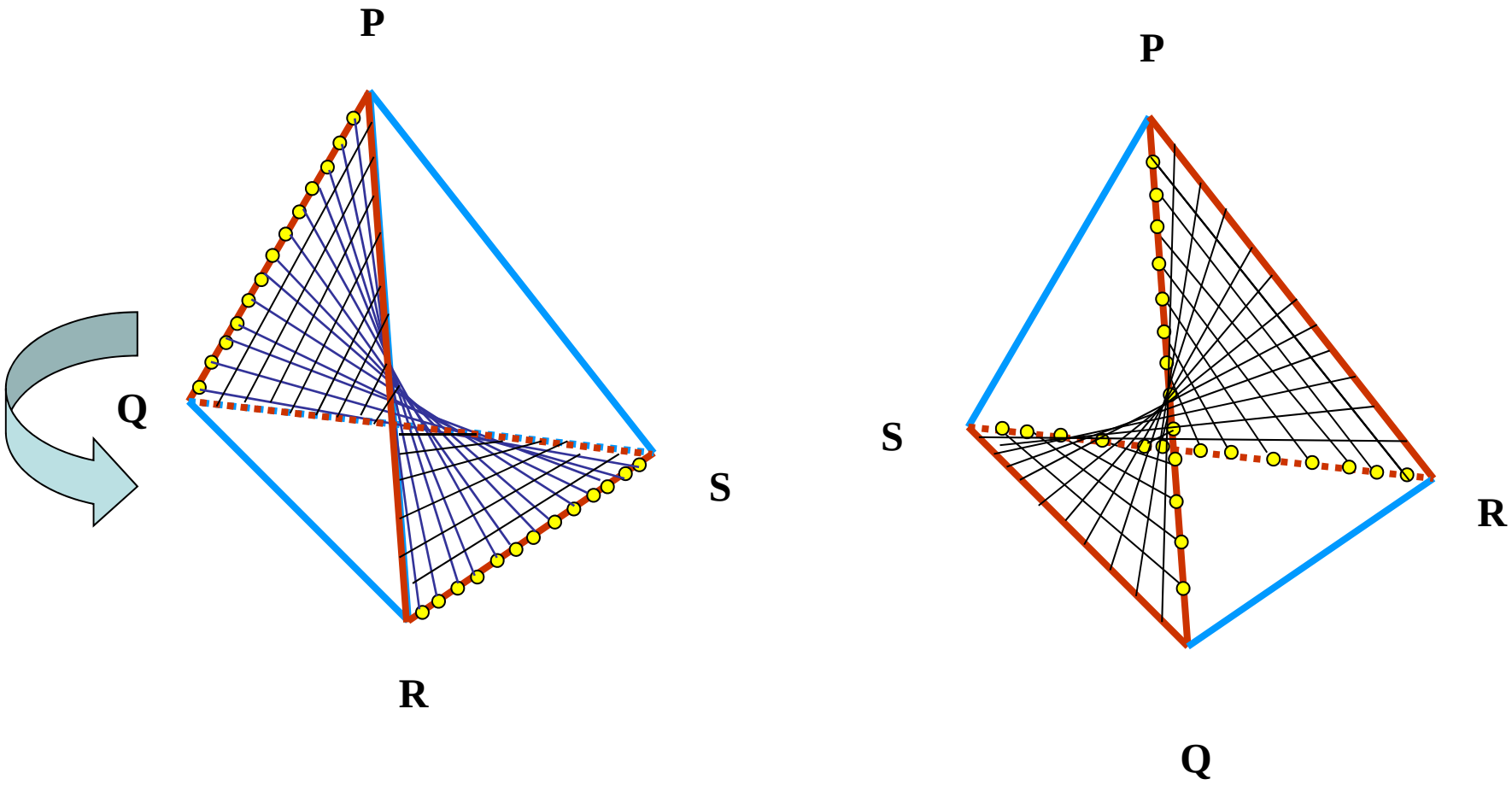


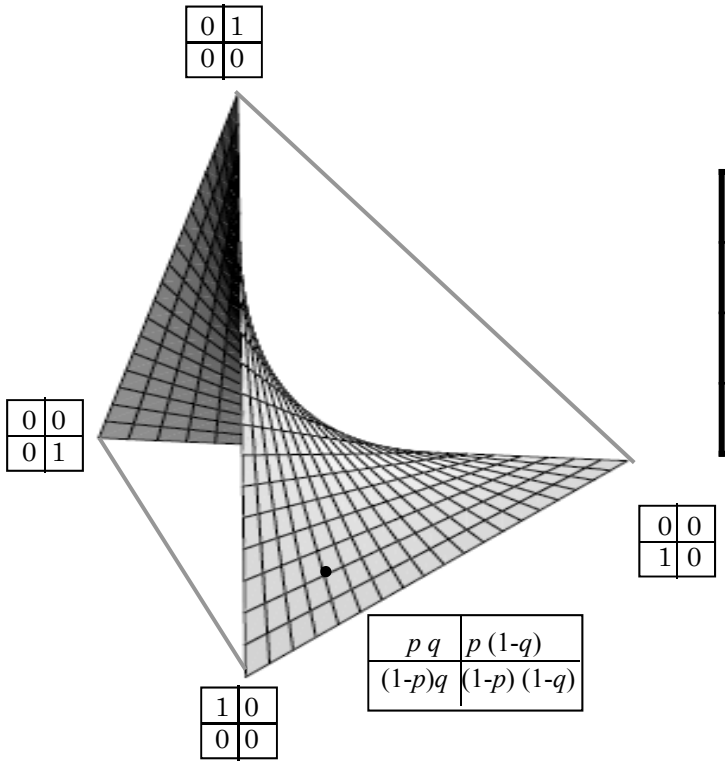
$$\frac{1}{3}\times\begin{array}{|c|c|}\hline 1/3 & 2/3 \\ \hline 0 & 0 \\ \hline\end{array}+\frac{2}{3}\times\begin{array}{|c|c|}\hline 0 & 0 \\ \hline 1/3 & 2/3 \\ \hline\end{array}$$

$$=\frac{1}{3}\times\begin{array}{|c|c|}\hline 1/3 & 0 \\ \hline 2/3 & 0 \\ \hline\end{array}+\frac{2}{3}\times\begin{array}{|c|c|}\hline 0 & 1/3 \\ \hline 0 & 2/3 \\ \hline\end{array}$$

$$=\begin{array}{|c|c|}\hline \frac{1}{3}\times\frac{1}{3} & \frac{1}{3}\times\frac{2}{3} \\ \hline \frac{2}{3}\times\frac{1}{3} & \frac{2}{3}\times\frac{2}{3} \\ \hline\end{array}$$

Ruled surface





	A	B	
C	x	y	e
D	z	w	f
	g	h	n

$$e = x + y,\ f = z + w,$$
$$g = x + z,\ h = y + w$$
$$n = x + y + z + w$$

	A	B	
C	pq	$p(1-q)$	P
D	$(1-p)q$	$(1-p)(1-q)$	$1-p$
	q	$1-q$	1

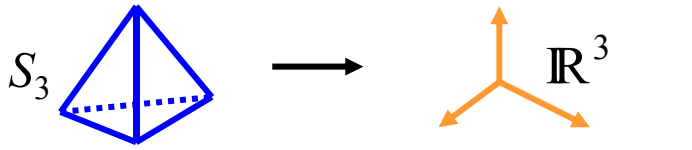
$$\chi^2 = n\frac{(xw - yz)^2}{e\,f\,g\,h}$$

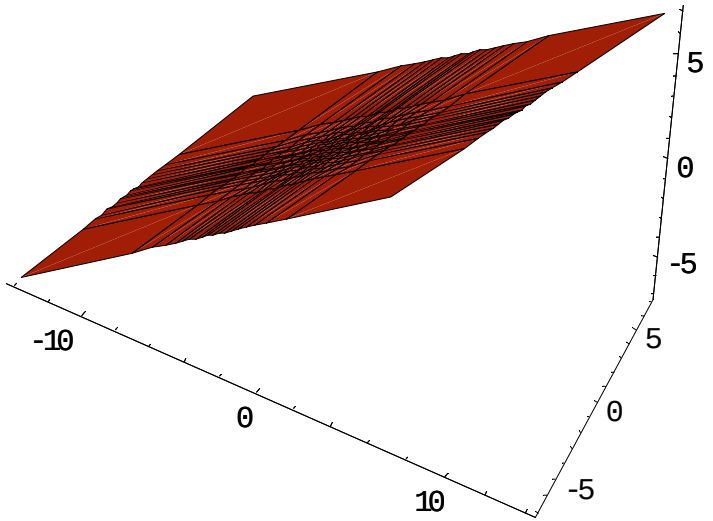
$$= n\frac{\{pq(1-p)(1-q) - p(1-q)(1-p)q\}^2}{p(1-p)q(1-q)}$$

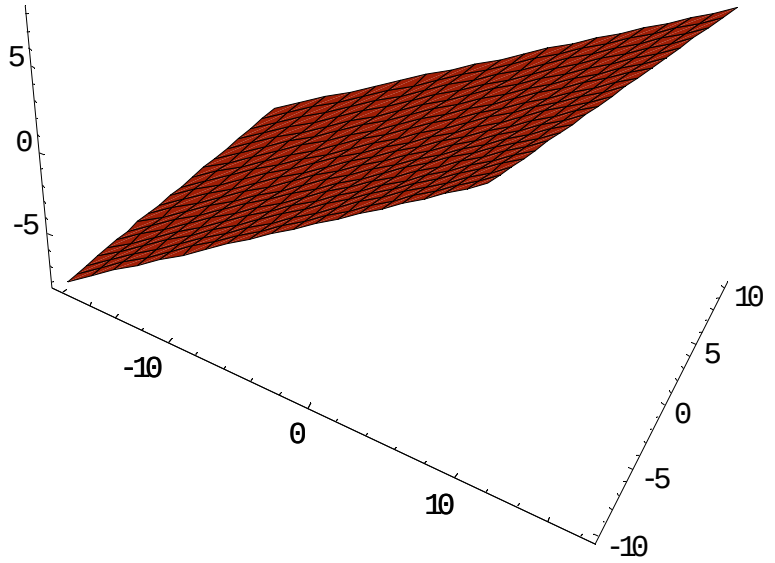
$$= 0$$

e-geodesical

$$(\pi_{11},\pi_{12},\pi_{21}) \longrightarrow \left(\log\frac{\pi_{11}}{\pi_{22}},\log\frac{\pi_{12}}{\pi_{22}},\log\frac{\pi_{21}}{\pi_{22}}\right) \text{ where } \pi_{22}=1-\pi_{11}-\pi_{12}-\pi_{21}$$



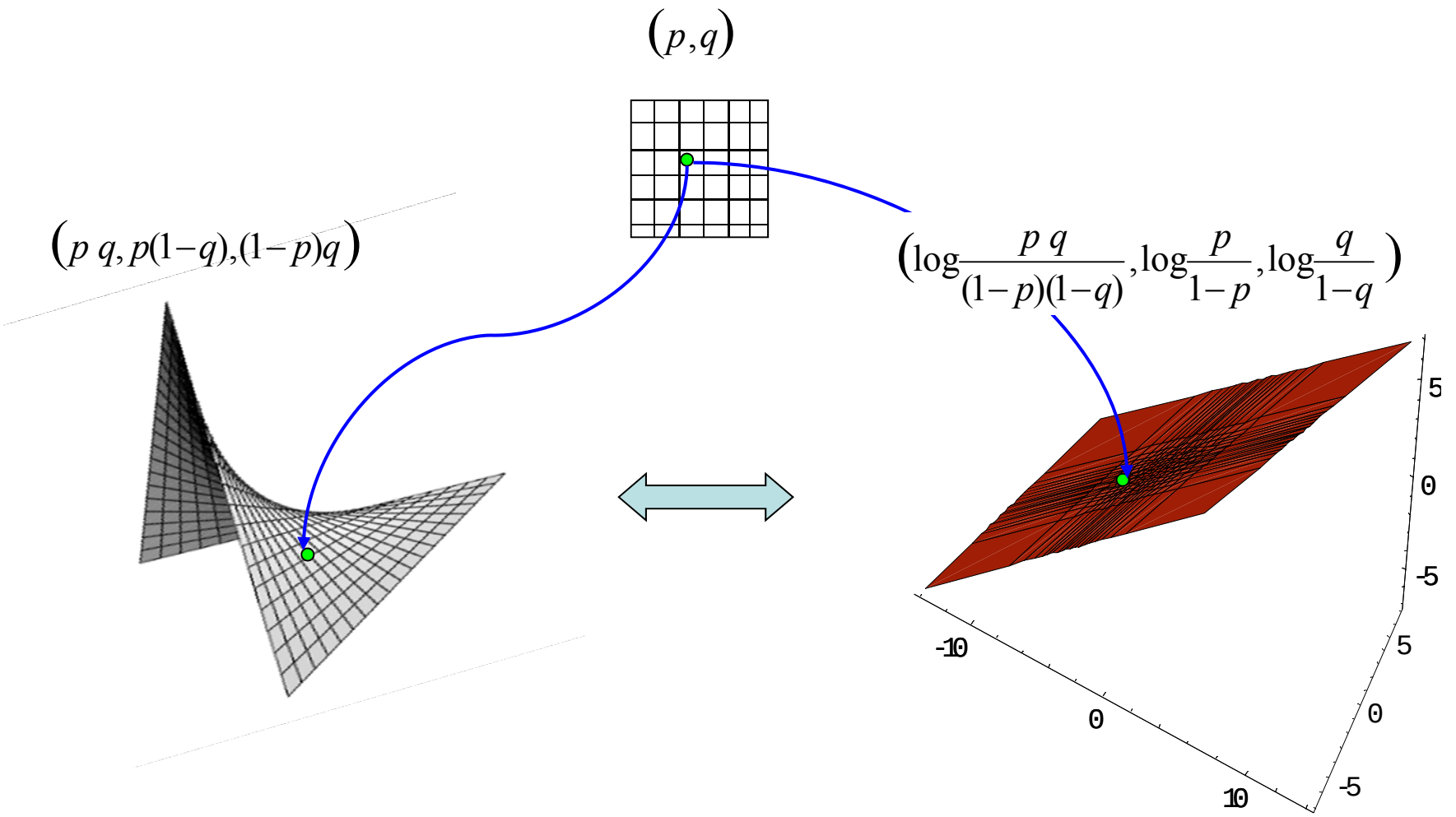




$$\left\{\left(\log\frac{pq}{(1-p)(1-q)},\log\frac{p}{1-p},\log\frac{q}{1-q}\right),\ 0<p<1,\ 0<q<1\right\}$$

$$\{(x+y,x,y),\ x\in\mathbb{R},y\in\mathbb{R}\}$$

Two parametrization

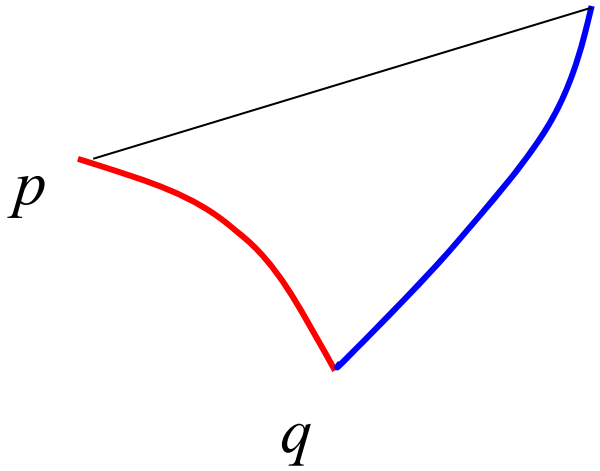


Kullback-Leibler Divergence

Let $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ be probability density functions.

Then **Kullback-Leibler Divergence** is defined by

$$D_{\mathrm{KL}}(p,q) = \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}$$



$$D_{\mathrm{KL}}(p,r) = D_{\mathrm{KL}}(p,q) + D_{\mathrm{KL}}(q,r)$$

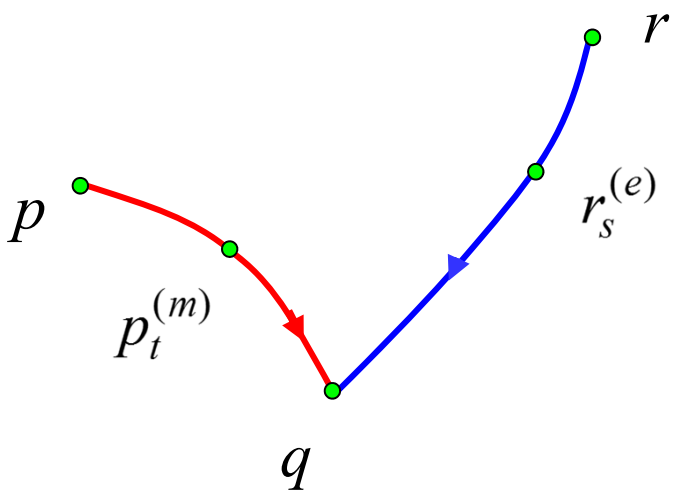
Two one-parameter families

Let $\mathbb{P}$ be the space of all pdfs on a data space.

m-geodesic $$p_t^{(m)}(\boldsymbol{x}) = (1-t)\,p(\boldsymbol{x}) + t q(\boldsymbol{x}), \qquad (p, q \in \mathbb{P})$$

e-geodesic $$r_s^{(e)}(\boldsymbol{x}) = c_s \{r(\boldsymbol{x})\}^{1-s} \{q(\boldsymbol{x})\}^{s}, \qquad (q, r \in \mathbb{P})$$

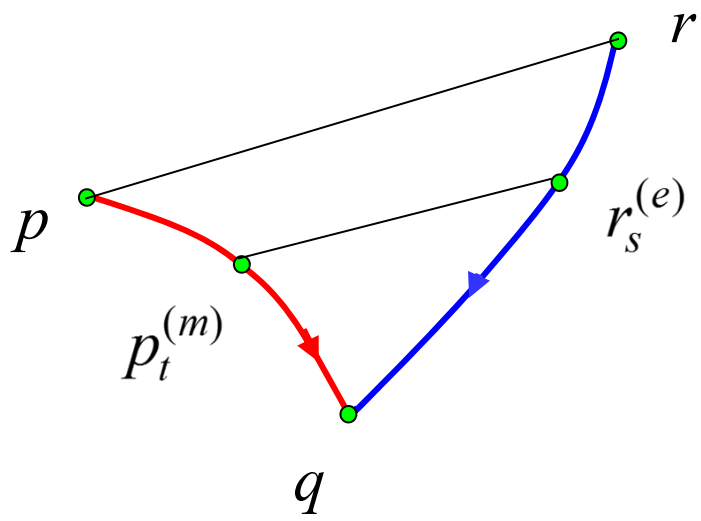
ここで $$c_s = 1 / \int \{r(\boldsymbol{x})\}^{1-s} \{q(\boldsymbol{x})\}^{s} d\boldsymbol{x}$$



Pythagoras theorem Amari-Nagaoka (1982)

$$D(p,r) = D(p,q) + D(q,r) \quad \Rightarrow$$

$$D(p_t^{(m)}, r_s^{(e)}) = D(p_t^{(m)}, q) + D(q, r_s^{(e)}) \qquad (\forall (s,t) \in [0,1] \times [0,1])$$



Proof

$$D_{\mathrm{KL}}(p_t^{(m)}, r_s^{(e)}) - \{D_{\mathrm{KL}}(p_t^{(m)}, q) + D_{\mathrm{KL}}(q, r_s^{(e)})\}$$

$$= \int (p_t^{(m)}(\boldsymbol{x}) - q(\boldsymbol{x}))(\log q(\boldsymbol{x}) - \log r_s^{(e)}(\boldsymbol{x}))d\boldsymbol{x}$$

$$= \int (1-t)(p(\boldsymbol{x}) - q(\boldsymbol{x}))\{(1-s)(\log q(\boldsymbol{x}) - \log r(\boldsymbol{x}) - \log c_s\}d\boldsymbol{x}$$

$$= (1-t)(1-s)\int (p(\boldsymbol{x}) - q(\boldsymbol{x}))(\log q(\boldsymbol{x}) - \log r(\boldsymbol{x}))d\boldsymbol{x}$$

$$-(1-t)\log c_s \int (p(\boldsymbol{x}) - q(\boldsymbol{x}))d\boldsymbol{x}$$

$$= (1-t)(1-s)\{D_{\mathrm{KL}}(p, r) - D_{\mathrm{KL}}(p, q) - D_{\mathrm{KL}}(q, r)\}$$

$$= 0$$

□

ABC in differential geometry

Riemannian metric defines an inner product on any tangent space of a manifold.

$$(M, g) \quad g : \boldsymbol{X}(M) \times \boldsymbol{X}(M) \to \mathbb{R}$$

Geodesic = a minimal arc between x and y

$$\gamma = \underset{c=\{x(t):x(0)=x,x(1)=y\}_{t=0}^{1}}{\operatorname{argmin}} \int_0^1 \sqrt{g(\dot{x}(t), \dot{x}(t))}\, dt$$

Linear connection defines parallelism along a vector field.

$$\nabla : \boldsymbol{X}(M) \times \boldsymbol{X}(M) \to \boldsymbol{X}(M)$$

(1) $\nabla_{fX} Y = f \nabla_X Y,$

(2) $\nabla_X (fY) = f \nabla_X Y + (Xf)Y \quad (\forall f \in \boldsymbol{F}(M), \forall X, Y \in \boldsymbol{X}(M))$

Componetwise $\nabla_{X_i} X_j = \sum_{k=1}^{d} \Gamma_{ij}^{k} X_k$

Cf. slide 41

Geodesic

A one-parameter family (curve) $C = \{\, \boldsymbol{\theta}(t) : -\varepsilon \le t \le \varepsilon \}$ is called geodesic with respect to a linear connection $\{\Gamma_{jk}^{i}(\boldsymbol{\theta}) : 1 \le i, j, k \le p\}$

$$\ddot{\theta}^{i}(t) + \sum_{k=1}^{p}\sum_{j=1}^{p} \Gamma_{jk}^{i}(\boldsymbol{\theta}(t))\,\dot{\theta}^{k}(t)\,\dot{\theta}^{k}(t) = 0 \qquad (\forall i = 1,\ldots,p)$$

speed of acceleration

Remark 2:

If $\Gamma_{jk}^{i}(\boldsymbol{\theta}) = 0 \;\; (1 \le i, j, k \le p)$, any geodesic is a line.

Newton's law of inertia

The property is not invariant with parametrization.

The gedesic C is expressed by a transform rule from parameter $\boldsymbol{\theta}$ to $\boldsymbol{\omega}$

$$\ddot{\omega}^{a}(t) + \sum_{c=1}^{p}\sum_{b=1}^{p} \tilde{\Gamma}_{cb}^{a}(\boldsymbol{\omega}(t))\,\dot{\omega}^{b}(t)\,\dot{\omega}^{c}(t) = 0 \qquad (\forall a = 1,\ldots, p)$$

$$\tilde{\Gamma}_{cb}^{a}(\boldsymbol{\omega}) = \sum_{k=1}^{p}\sum_{j=1}^{p}\sum_{i=1}^{p} \overline{B}_{c}^{k}\,\overline{B}_{b}^{j}\,B_{i}^{a}\,\Gamma_{jk}^{i}(\boldsymbol{\theta}) + \sum_{a=1}^{p} B_{i}^{a}\,\frac{\partial B_{b}^{i}}{\partial \omega^{c}}, \quad B_{i}^{a} = \frac{\partial \omega^{a}}{\partial \theta^{i}},\; \overline{B}_{a}^{i} = \frac{\partial \theta^{i}}{\partial \omega^{a}} \text{ for } \omega^{a} = t^{a}(\boldsymbol{\theta})$$

Change rule on connections

Let us consider a transform ϕ from parameter $\boldsymbol{\theta}$ to $\boldsymbol{\omega}$. Then a geodesic C is written by $\{\boldsymbol{\omega}(t) = \varphi(\boldsymbol{\theta}(t)) : -\varepsilon < t < \varepsilon\}$. Hence we have the following:

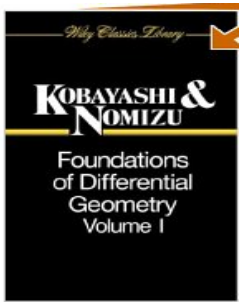
$$\dot{\omega}^a(t) = \sum_{i=1}^{p} B_i^a(\boldsymbol{\theta}(t))\dot{\theta}^i(t) \qquad (\forall a = 1,\ldots, p) \qquad \left(B_i^a = \frac{\partial \varphi^a}{\partial \theta^i} \right)$$

$$\ddot{\omega}^a(t) = \sum_{i=1}^{p} B_i^a(\boldsymbol{\theta}(t))\ddot{\theta}^i(t) + \sum_{j=1}^{p}\sum_{i=1}^{p} \frac{\partial B_i^a(\boldsymbol{\theta}(t))}{\partial \theta^j}\dot{\theta}^i(t)\dot{\theta}^j(t) \quad (\forall a = 1,\ldots, p)$$

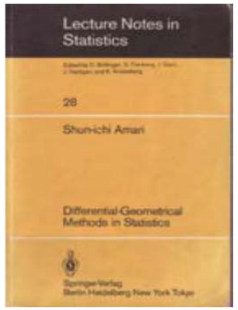
$$\ddot{\omega}^a(t) + \sum_{c=1}^{p}\sum_{b=1}^{p} \widetilde{\Gamma}_{cb}^{\ a}(\boldsymbol{\omega}(t))\dot{\omega}^b(t)\dot{\omega}^c(t) = 0 \qquad (\forall a = 1,\ldots, p)$$

$$\widetilde{\Gamma}_{cb}^{\ a}(\boldsymbol{\omega}) = \sum_{k=1}^{p}\sum_{j=1}^{p}\sum_{i=1}^{p} \overline{B}_c^k \overline{B}_b^j B_i^a \Gamma_{jk}^{\ i}(\boldsymbol{\theta}) + \sum_{a=1}^{p} B_i^a \frac{\partial B_b^i}{\partial \omega^c}, \qquad \overline{B}_a^i = \frac{\partial (\varphi^{-1})^i}{\partial \omega^a} \text{ for } \omega^a = \varphi^a(\boldsymbol{\theta})$$

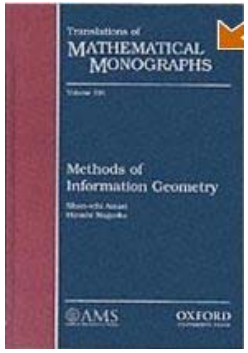
What are reference books?



Foundations of Differential Geometry (Wiley Classics Library)
Shoshichi Kobayashi, Katsumi Nomizu



Differential Geometrical Methods in Statistics
Shun-Ichi Amari 1985 年 Springer



Methods of Information Geometry

Shun-Ichi Amari , Hiroshi Nagaoka

Amer Mathematical Society (2001)

Statistical model and information

Statistical model $M = \{ p_{\boldsymbol{\theta}}(\boldsymbol{x}) = p(\boldsymbol{x}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta \} \quad (\Theta \subset \mathbb{R}^p)$

Score vector $\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{\theta}) = \dfrac{\partial}{\partial \boldsymbol{\theta}} \log p(\boldsymbol{x}, \boldsymbol{\theta})$

Space of score vector $T_{\boldsymbol{\theta}} = \{ \boldsymbol{\alpha}^{\mathrm{T}} s(\cdot, \boldsymbol{\theta}) : \boldsymbol{\alpha} \in \mathbb{R}^p \}$

Fisher information metric $g_{\boldsymbol{\theta}}(u, v) = E_{\boldsymbol{\theta}} \{ uv \} \quad (\forall u, v \in T_{\boldsymbol{\theta}})$

Fisher information matrix $I_{\boldsymbol{\theta}} = E_{\boldsymbol{\theta}} \{ s(\boldsymbol{x}, \boldsymbol{\theta}) s(\boldsymbol{x}, \boldsymbol{\theta})^{\mathrm{T}} \}$

Score vector space

Note 1: $u \in T_{\theta} \Rightarrow E_{\theta}(u) = \int u(\boldsymbol{x},\boldsymbol{\theta})\, p(\boldsymbol{x},\boldsymbol{\theta}) d\boldsymbol{x} = \int \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{s}(\boldsymbol{x},\boldsymbol{\theta})\, p(\boldsymbol{x},\boldsymbol{\theta}) d\boldsymbol{x}$

$= \boldsymbol{\alpha}^{\mathrm{T}} \int \boldsymbol{s}(\boldsymbol{x},\boldsymbol{\theta})\, p(\boldsymbol{x},\boldsymbol{\theta}) d\boldsymbol{x} = 0$

$u, v \in T_{\theta} \Rightarrow g_{\theta}(u,v) = \int u(\boldsymbol{x},\boldsymbol{\theta}) v(\boldsymbol{x},\boldsymbol{\theta})\, p(\boldsymbol{x},\boldsymbol{\theta}) d\boldsymbol{x}$

$= \boldsymbol{\alpha}^{\mathrm{T}} \{ \int \boldsymbol{s}(\boldsymbol{x},\boldsymbol{\theta}) \boldsymbol{s}(\boldsymbol{x},\boldsymbol{\theta})^{\mathrm{T}}\, p(\boldsymbol{x},\boldsymbol{\theta}) d\boldsymbol{x} \} \boldsymbol{\beta} = \boldsymbol{\alpha}^{\mathrm{T}} I_{\theta} \boldsymbol{\beta}$

The space of all random variables with mean 0 and finite variance

$$S_{\theta} = \{ t(\boldsymbol{x},\boldsymbol{\theta}) : E_{\theta}(t(\boldsymbol{x},\boldsymbol{\theta})) = 0,\ V_{\theta}(t(\boldsymbol{x},\boldsymbol{\theta})) < \infty \}$$

Note 2: S_{θ} is an infinite dimensional vector space that includes T_{θ}

For $u(\boldsymbol{x},\boldsymbol{\theta}) = \alpha^{T} \boldsymbol{s}(\boldsymbol{x},\boldsymbol{\theta})$

$$u(\boldsymbol{x},\boldsymbol{\theta})^{k} - E_{\theta}\{u(\boldsymbol{x},\boldsymbol{\theta})^{k}\},\quad \frac{\partial^{k}}{\partial\theta_{i_1} \ldots \partial\theta_{ik}} u(\boldsymbol{x},\boldsymbol{\theta}) - E_{\theta}\{\frac{\partial^{k}}{\partial\theta_{i_1} \ldots \partial\theta_{ik}} u(\boldsymbol{x},\boldsymbol{\theta})\},\quad k = 1,2,\ldots.$$

belong to S_{θ} Cf. Tiatis (2006)

Linear connections

Score vector $\boldsymbol{s}(\boldsymbol{x},\boldsymbol{\theta}) = (s_i(\boldsymbol{x},\boldsymbol{\theta}))_{i=1}^{p}$

m-connection $\overset{\mathrm{m}}{\Gamma}{}_{jk}^{i}(\boldsymbol{\theta}) = \sum_{i'=1}^{p} g^{ii'}\left[E_{\boldsymbol{\theta}}\left\{\frac{\partial s_k}{\partial \theta^j} s_{i'}\right\} + E_{\boldsymbol{\theta}}\left\{s_k s_j s_{i'}\right\}\right] \quad (1 \le i, j, k \le p)$

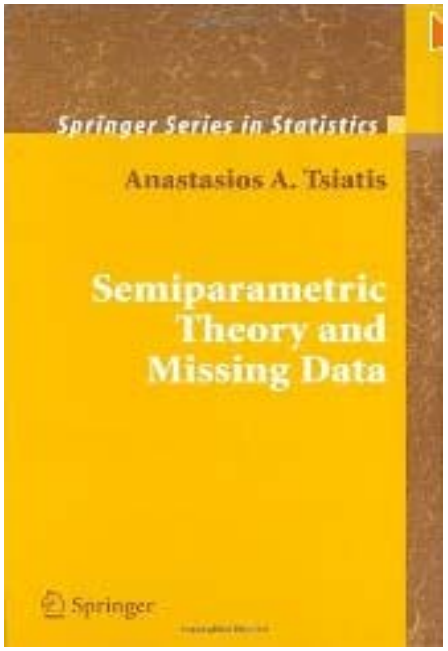
e-connection $\overset{\mathrm{e}}{\Gamma}{}_{jk}^{i}(\boldsymbol{\theta}) = \sum_{i'=1}^{p} g^{ii'} E_{\boldsymbol{\theta}}\left\{\frac{\partial s_k}{\partial \theta^j} s_{i'}\right\} \quad (1 \le i, j, k \le p)$

---

m-geodesic $p_t^{(m)}(\boldsymbol{x}) = (1-t)p(\boldsymbol{x}) + tq(\boldsymbol{x}), \quad (p, q \in \mathbb{P})$

e-geodesic $r_s^{(e)}(\boldsymbol{x}) = c_s\{r(\boldsymbol{x})\}^{1-s}\{q(\boldsymbol{x})\}^{s}, \quad (q, r \in \mathbb{P})$

Semiparamtric thoery



1 Introduction to Semiparametric Models

2 Hilbert Space for Random Vectors

3 The Geometry of Influence Functions

4 Semiparametric Models

5 Other Examples of Semiparametric Models

6 Models and Methods for Missing Data

7 Missing and Coarsening at Random for Semiparametric Models

8 The Nuisance Tangent Space and Its Orthogonal Complement

9 …14.

Semiparametric Theory and Missing Data
(Springer Series in Statistics)
Anastasios Tsiatis

Geodesical models

Definition (i) Statistical model $\boldsymbol{M}$ is said to be **e-geodesical** if

$$p(\boldsymbol{x}),\ q(\boldsymbol{x}) \in M \Rightarrow c_t\, p(\boldsymbol{x})^{1-t} q(\boldsymbol{x})^t \in M \quad (\forall t \in (0,1))$$

$$\text{where} \quad c_t = 1 / \int p(\boldsymbol{x})^{1-t} q(\boldsymbol{x})^t \, d\boldsymbol{x}$$

(ii) Statistical model M is said to be **m-geodesic** if

$$p(\boldsymbol{x}),\ q(\boldsymbol{x}) \in M \Rightarrow (1-t)p(\boldsymbol{x}) + tq(\boldsymbol{x}) \in M \quad (\forall t \in (0,1))$$

Note : Let $\boldsymbol{P}$ be a space of all probability density functions.

By definition $\boldsymbol{P}$ is e-geodesical and m-geodesical.

However the theoretical framework for $\boldsymbol{P}$ is not perfectly complete.

Cf. Pistone and Sempi (1995, AS)

Two type of modeling

Let $p_0(\boldsymbol{x})$ be a pdf and $\boldsymbol{t}(\boldsymbol{x})$ a p-dimensional statistic.

Exponential model $M^{(\mathrm{e})} = \{\, p(\boldsymbol{x},\boldsymbol{\theta}) = p_0(\boldsymbol{x}) \exp\{\boldsymbol{\theta}^T \boldsymbol{t}(\boldsymbol{x}) - \kappa(\boldsymbol{\theta})\};\ \boldsymbol{\theta} \in \Theta\}$

$$\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^p : \kappa(\boldsymbol{\theta}) < \infty\}$$

Matched mean model $M^{(\mathrm{m})} = \{\ p(\boldsymbol{x}) : E_p\{\boldsymbol{t}(\boldsymbol{x})\} = E_{p_0}\{\boldsymbol{t}(\boldsymbol{x})\}\ \}$

Let $\{\, p_i(\boldsymbol{x}) : i = 0, 1, \ldots, I\,\}$ be a set of pdfs.

Exponential model $M^{(\mathrm{e})} = \{\, p(\boldsymbol{x},\boldsymbol{\theta}) = p_0(\boldsymbol{x}) \exp\{\sum_{i=1}^{I} \theta_i \log \frac{p_i(\boldsymbol{x})}{p_0(\boldsymbol{x})} - \kappa(\boldsymbol{\theta})\};\ \boldsymbol{\theta} \in \Theta\}$

$$\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^p : \kappa(\boldsymbol{\theta}) < \infty\}$$

Mixture model $M^{(\mathrm{m})} = \{\ p(\boldsymbol{x},\boldsymbol{\theta}) = (1 - \sum_{i=1}^{I} \theta_i) p_0(\boldsymbol{x}) + \sum_{i=1}^{I} \theta_i p_i(\boldsymbol{x}) : \theta \in \Theta\}$

$$\Theta = \{\ \boldsymbol{\theta} = (\theta_1, \ldots, \theta_I) : 0 < \sum_{i=1}^{I} \theta_i < 1,\ \theta_i > 0\ (\forall i = 1, \ldots, I)\}$$

Statistical functional

A functional $\boldsymbol{f}(p)$ is called a statistical functional if $p(\boldsymbol{x})$ is a pdf. (Hampel, 1974)

A statistical functional $\boldsymbol{f}(p)$ is said to be Fisher-consistent for a model

$M = \{p_{\boldsymbol{\theta}}(\boldsymbol{x}) : \boldsymbol{\theta} \in \Theta\}$ if $\boldsymbol{f}(p)$ satisfies that

(1) $\boldsymbol{f}(p) \in \Theta$,

(2) $\boldsymbol{f}(p_{\boldsymbol{\theta}}) = \boldsymbol{\theta} \quad (\forall \boldsymbol{\theta} \in \Theta)$

Example. For a normal model $M = \{p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}, \ \boldsymbol{\theta} = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+\}$

$$\boldsymbol{f}(p) = \left(\int x p(x)\, dx, \int x^2 dx - \left(\int x p(x)\, dx\right)^2\right)^T$$

is Fisher-consistent for $\boldsymbol{\theta} = (\mu, \sigma^2)$.

Transversality

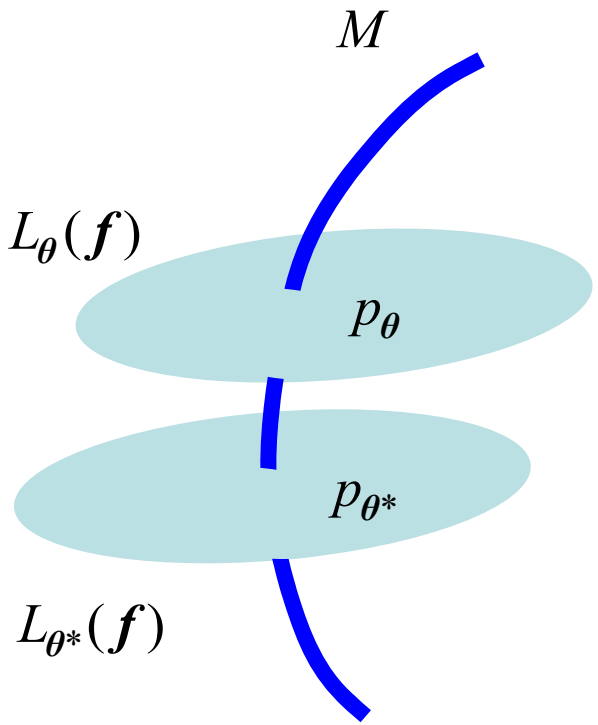
Let $\boldsymbol{f}(p)$ be Fisher consistent functional.

$$L_{\boldsymbol{\theta}}(\boldsymbol{f}) = \{p : \boldsymbol{f}(p) = \boldsymbol{\theta}\},$$

which is called a leaf transverse to M.

(1) $L_{\boldsymbol{\theta}}(\boldsymbol{f}) \cap M = \{p_{\boldsymbol{\theta}}\}$

(2) $\underset{\theta \in \Theta}{\oplus} L_{\boldsymbol{\theta}}(\boldsymbol{f})$ is a local neighborhood.



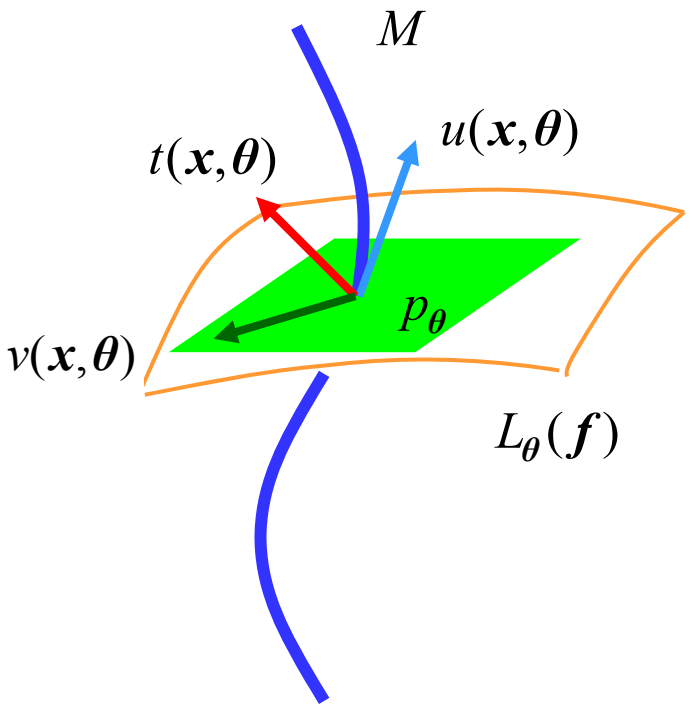
Foliation structure

Statistical model $M = \{ p_\theta(\boldsymbol{x}) = p(\boldsymbol{x}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta \} \quad (\Theta \subset \mathbb{R}^p)$

Foliation $\boldsymbol{P} = \bigoplus_{\theta \in \Theta} L_\theta(\boldsymbol{f})$

Decomposition of tanget spaces

$$T_{p_\theta}(\boldsymbol{P}) = T_{p_\theta}(M) \oplus T_{p_\theta}(L_\theta(\boldsymbol{f}))$$



$$\forall t \in T_{p_\theta}(\boldsymbol{P}),\ \exists u \in T_{p_\theta}(M),\ \exists v \in T_{p_\theta}(L_\theta(\boldsymbol{f}))$$

such that $t(\boldsymbol{x}, \boldsymbol{\theta}) = u(\boldsymbol{x}, \boldsymbol{\theta}) + v(\boldsymbol{x}, \boldsymbol{\theta})$

Transversality for MLE

Let $p_0(\boldsymbol{x})$ be a pdf and $\boldsymbol{t}(\boldsymbol{x})$ a p-dimensional statistic.

Exponential model $M^{(\mathrm{e})} = \{ p(\boldsymbol{x},\boldsymbol{\theta}) = p_0(\boldsymbol{x}) \exp\{\boldsymbol{\theta}^T \boldsymbol{t}(\boldsymbol{x}) - \kappa(\boldsymbol{\theta})\}; \boldsymbol{\theta} \in \Theta \}$

$$\Theta = \{ \boldsymbol{\theta} \in \mathbb{R}^p : \kappa(\boldsymbol{\theta}) < \infty \}$$

For the exponential model $M^{(\mathrm{e})}$ the MLE functional

$$\boldsymbol{f}_{\mathrm{ML}}(p) =: \arg\max_{\boldsymbol{\theta}\in\Theta} E_p\{\log p(\boldsymbol{x},\boldsymbol{\theta})\}$$

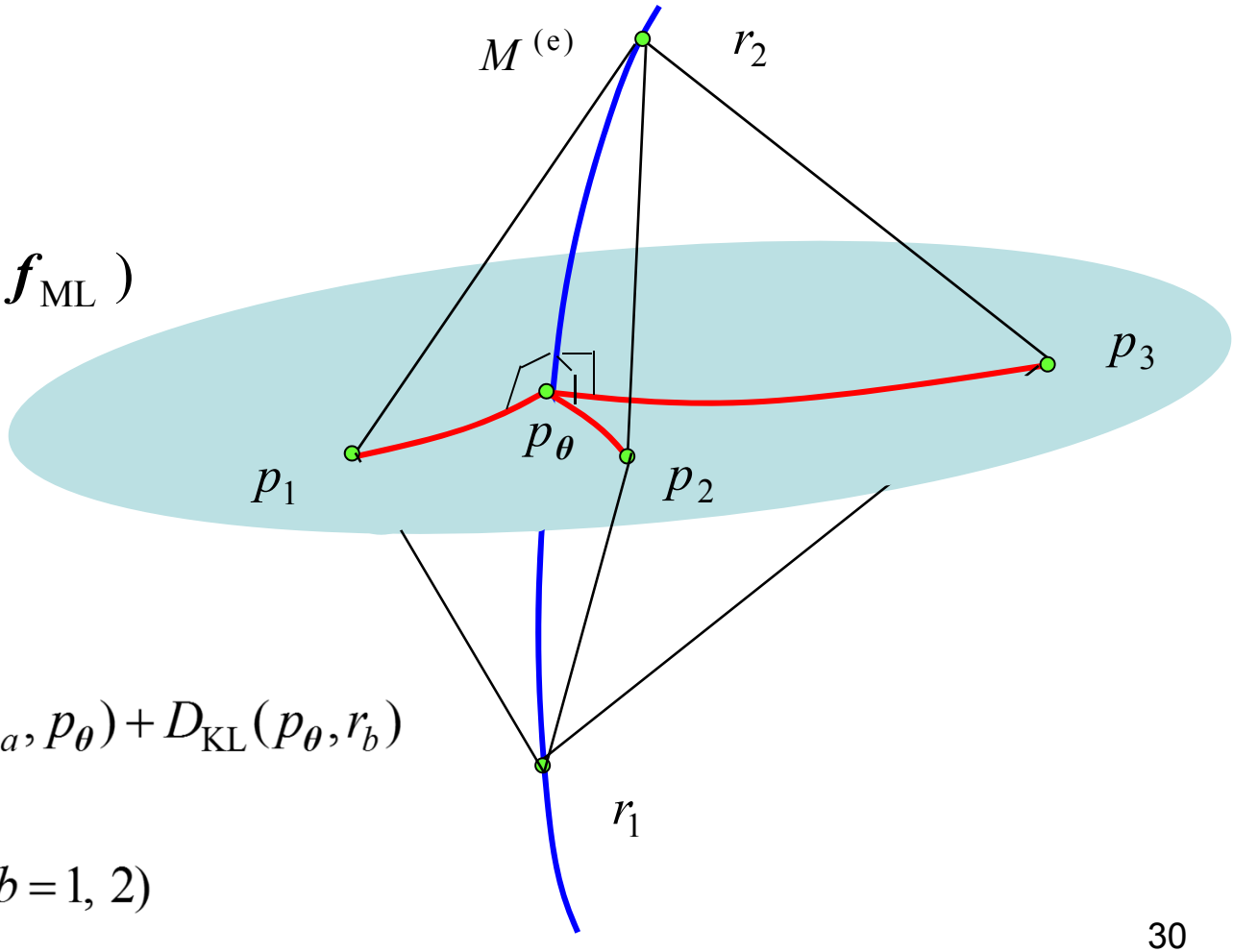
is written by

$$\boldsymbol{f}_{\mathrm{ML}}(p) = \arg_{\boldsymbol{\theta}\in\Theta} \left\{ E_p\{\boldsymbol{t}(\boldsymbol{x})\} = E_{p(\cdot,\boldsymbol{\theta})}\{\boldsymbol{t}(\boldsymbol{x})\} \right\}$$

Hence the foliation associated with $\boldsymbol{f}_{\mathrm{ML}}$ is a matched mean model.

$$L_{\boldsymbol{\theta}}(\boldsymbol{f}_{\mathrm{ML}}) = \{ p : E_p\{\boldsymbol{t}(\boldsymbol{x})\} = E_{p(\cdot,\boldsymbol{\theta})}\{\boldsymbol{t}(\boldsymbol{x})\} \}$$

Maximum likelihood foliation



$$D_{\mathrm{KL}}(p_a, r_b) = D_{\mathrm{KL}}(p_a, p_{\theta}) + D_{\mathrm{KL}}(p_{\theta}, r_b)$$

$$(a = 1,\ 2,\ 3,\ \ b = 1,\ 2)$$

Estimating function

Statistical model $M = \{ p_{\boldsymbol{\theta}}(\boldsymbol{x}) = p(\boldsymbol{x}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta \}$ $(\Theta \subset \mathbb{R}^p)$

p-variable function $\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})$ is unbiased

$\overset{\text{def}}{\Leftrightarrow}$

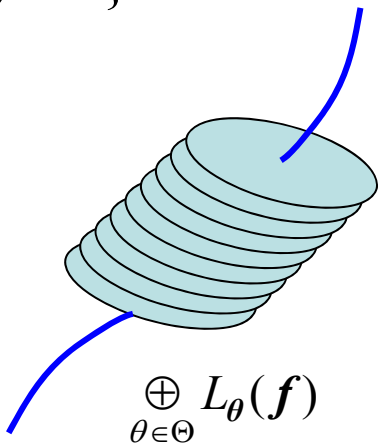
$$E_{p(\cdot,\boldsymbol{\theta})}\{\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})\} = \mathbf{0}, \ \det\left(E_{p(\cdot,\boldsymbol{\theta})}\left\{\frac{\partial \boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\}\right) \neq 0 \quad (\forall \boldsymbol{\theta} \in \Theta)$$

The statistical functional $\boldsymbol{f}(p) = \underset{\boldsymbol{\theta} \in \Theta}{\arg\,\text{solve}} \{ E_p \boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta}) = \mathbf{0} \}$

is Fisher-consitent, and the leaf transverse to M

$$L_{\boldsymbol{\theta}}(\boldsymbol{f}) = \{ p : E_p \boldsymbol{u}(\boldsymbol{x}, \boldsymbol{\theta}) = \mathbf{0} \}$$

is m-geodesical.



Minimum divergence geometry

$D : M \times M \to \mathbf{R}$ is an **information divergence** on a statistical model M

$\Longleftrightarrow$ (i) $D(p,q) \geqq 0$ with equality if and only if $p = q$

(ii) D is a differentialble on $M \times M$

Let $$D(Z\cdots \mid XY\cdots)(p) = Z_p \cdots X_q Y_q \cdots D(p,q)\mid_{q=p}$$

Then we get a **Riemannian metric** and **dual connections** on M (Eguchi, 1983,1992)

$$g^{(D)}(X,Y) = -D(X \mid Y)$$

$$g^{(D)}(\nabla_X Y, Z) = -D(XY \mid Z) \qquad (\forall Z \in X(M))$$

$$g^{(D)}(\nabla^*_X Y, Z) = -D(Z \mid XY) \qquad (\forall Z \in X(M))$$

Remarks

$g^{(D)}$ is a Riemann metric

$$D(X\,|\,\cdot) = 0$$

$$g^{(D)}(X,Y) - g^{(D)}(Y,X) = D(XY\,|\,\cdot) - D(YX\,|\,\cdot) = D([X,Y]\,|\,\cdot) = 0$$

$$g^{(D)}(X,Y) = -D(X\,|\,Y) = D(XY\,|\,\cdot) = D(\cdot\,|\,XY)$$

∇_X and $\nabla^*{}_X$ are dual connections

$$g^{(D)}(\nabla_{fX}Y, Z) = -f\,D(XY\,|\,Z) = g^{(D)}(f\nabla_X Y, Z) \qquad (\forall Z)$$

$$g^{(D)}(\nabla_X(fY), Z) = -D(X(fY)\,|\,Z) = g^{(D)}(f\nabla_X Y + (Xf)Y, Z) \quad (\forall Z)$$

$$Xg^{(D)}(Y,Z) = g^{(D)}(\nabla_X Y, Z) + g^{(D)}(Y, \nabla^*{}_X Z) = g^{(D)}(\overline{\nabla}_X Y, Z) + g^{(D)}(Y, \overline{\nabla}_X Z)$$

where $\overline{\nabla}_X = \frac{1}{2}(\nabla_X + \nabla^*{}_X)$

Cf. Slide 21

U divergence

Let a triple (U, u, ξ) such that U is a convex function, $u = U'$, $\xi = u^{-1}$

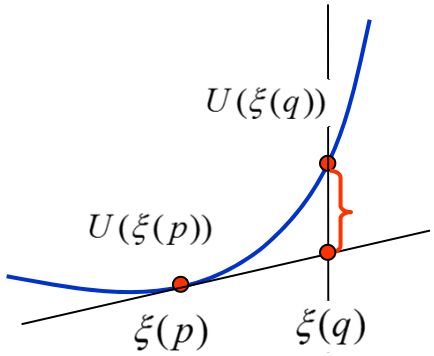
$$\xi(s) = \arg\max_{-\infty < t < \infty} \{ts - U(t)\}$$

$$u(t) = \arg\max_{-\infty < s < \infty} \{st - U^*(s)\}$$

where $U^*(s) = s\xi(s) - U(\xi(s))$

U cross-entropy $L_U(p,q) = \int \{-p\xi(q) + U(\xi(q))\}$

U entropy $H_U(p) = L_U(p,p) = -\int U^*(p)$



U divergence $D_U(p,q) = L_U(p,q) - H_U(p)$

Example of U divergence

$$D_U(p,q) = \int U(\xi(q)) - U(\xi(p)) - p\{\xi(q) - \xi(p)\}$$

KL divergence

$$(U(t), u(t), \xi(s)) = (\exp(t), \exp(t), \log(s))$$

$$D_{\mathrm{KL}}(p,q) = \int \{q - p - p(\log q - \log p)\} = \int p(\log p - \log q)$$

Beta (power) divergence

$$(U_\beta(t), u_\beta(t), \xi_\beta(s)) = \left(\frac{(1+\beta t)^{(1+\beta)/\beta}}{1+\beta}, (1+\beta t)^{1/\beta}, \frac{s^\beta - 1}{\beta} \right)$$

$$D_\beta(p,q) = \int \left\{ \frac{q^{\beta+1} - p^{\beta+1}}{\beta+1} - p\frac{(q^\beta - p^\beta)}{\beta} \right\}$$

Note $\lim_{\beta \to 0} U_\beta(t) = \exp(t)$

$$\lim_{\beta \to 0} D_\beta(p,q) = D_{\mathrm{KL}}(p,q)$$

Geometric formula with D_U

$$(g^{(U)}, \nabla^{(U)}, \nabla^{*(U)}) \quad \text{s.t.} \quad g^{(U)}(X,Y) = -D_U(X \mid Y)$$

$$g^{(U)}(\nabla_X Y, Z) = -D_U(XY \mid Z) \qquad (\forall Z \in X(M))$$

$$g^{(D)}(\nabla^*_X Y, Z) = -D(Z \mid XY) \qquad (\forall Z \in X(M))$$

$$g^{(U)}_{ij}(\theta) = \int \frac{\partial}{\partial \theta^i} q(x,\theta) \frac{\partial}{\partial \theta^i} \xi(q(x,\theta)) dx$$

$$\Gamma^{(U)}_{ij,k}(\theta) = \int \frac{\partial^2}{\partial \theta^i \partial \theta^j} q(x,\theta) \frac{\partial}{\partial \theta^k} \xi(q(x,\theta)) dx$$

$$\Gamma^{*(U)}_{ij,k}(\theta) = \int \frac{\partial}{\partial \theta^k} q(x,\theta) \frac{\partial^2}{\partial \theta^i \partial \theta^j} \xi(q(x,\theta)) dx$$

$$g^{(U)} = g \iff U = \exp \quad \text{(KL divergence)}$$

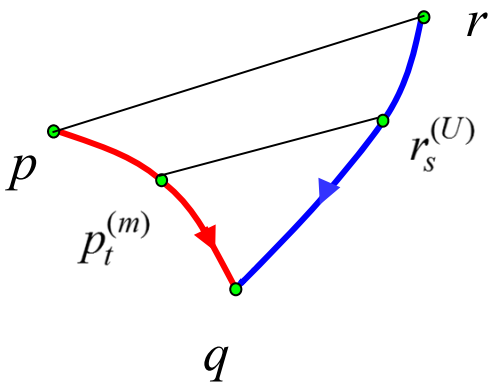
$$\nabla^{(U)} = \overset{\text{m}}{\nabla} \quad (\forall U) \qquad \text{(Eguchi, 2005)}$$

Triangle with D_U

mixture geodesic $\quad \overset{\mathrm{m}}{p}_t(\boldsymbol{x}) = (1-t)p(\boldsymbol{x}) + tq(\boldsymbol{x})$

U geodesic $\quad r_s^{(U)}(\boldsymbol{x}) = u\big((1-s)\xi(r(\boldsymbol{x})) + s\xi(q(\boldsymbol{x})) + \kappa_s\big)$

$$D_U(p_t^{(m)}, r_s^{(U)}) - D_U(p_t^{(m)}, q) - D_U(q, r_s^{(U)})$$
$$= (1-t)(1-s)\{D_U(p,r) - D_U(p,q) - D_U(q,r)\} \quad (\forall (s,t) \in [0,1]\times[0,1])$$



$$\{p_t^{(m)}\} \perp_q \{r_s^{(U)}\} \quad \Rightarrow$$
$$D_U(p_t^{(m)}, r_s^{(U)}) = D_U(p_t^{(m)}, q) + D_U(q, r_s^{(U)})$$

Information divergence class

and

robust statistical methods

Light and shadow of MLE

1. Invariance under data-transformations	1. Non-robust
2. Asymptotic efficiency	2. Over-fitting

Log-likelihood on exponential family **Sufficiency and efficiency**

Likelihood method $\log \rightleftarrows \exp$

U-method $\xi \rightleftarrows u$

Max *U*-entropy distribution

Let us fix a statistics $\boldsymbol{t}(\boldsymbol{x})$.

Equal mean space $\boldsymbol{\Gamma}_{\boldsymbol{\tau}} = \{p : E_P\{\boldsymbol{t}(\boldsymbol{X})\} = \boldsymbol{\tau}\}$

$$p^*(\boldsymbol{x}) = \arg\max_{p \in \boldsymbol{\Gamma}_{\boldsymbol{\tau}}} H_U(p)$$

$$p^*(\boldsymbol{x}) = u(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{x}) - \kappa(\boldsymbol{\theta})) \qquad \textbf{\textit{U}-model}$$

Euler-Lagrange's 1st variation

$$\frac{\partial}{\partial \varepsilon}\{H((1-\varepsilon)p^* + \varepsilon q) - \boldsymbol{\theta}^{\mathrm{T}}(\boldsymbol{t} - \boldsymbol{\tau})((1-\varepsilon)p^* + \varepsilon q) - \lambda((1-\varepsilon)p^* + \varepsilon q - 1)\}|_{\varepsilon=0}$$

$$= \left\langle q\{\xi(p^*) - \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{t} + \kappa\}\right\rangle \qquad (\forall q \in \Gamma_\tau)$$

Hence $\xi(p^*(\boldsymbol{x})) = \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{x}) - \kappa(\boldsymbol{\theta})$

U-estimate

Let $p(\boldsymbol{x})$ be a data density function with statistical model $q_{\boldsymbol{\theta}}(\boldsymbol{x})$

***U*-loss function** $L_U(\boldsymbol{\theta}) = L_U(p, q_{\boldsymbol{\theta}}) = -E_p\{\xi(q_{\boldsymbol{\theta}})\} + \left\langle U(\xi(q_{\boldsymbol{\theta}}))\right\rangle$

***U*-empirical loss function** $L_U^{\mathrm{emp}}(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n}\xi(q_{\boldsymbol{\theta}}(\boldsymbol{x}_i)) + \left\langle U(\xi(q_{\boldsymbol{\theta}}))\right\rangle$

***U*-estimate** $\hat{\boldsymbol{\theta}}_U = \arg\min_{\boldsymbol{\theta}\in\Theta} L_U^{\mathrm{emp}}(\boldsymbol{\theta})$

***U*-estimating function** $\boldsymbol{s}_U(\boldsymbol{x},\boldsymbol{\theta}) = w(\boldsymbol{x},\boldsymbol{\theta})\, s(\boldsymbol{x},\boldsymbol{\theta}) - E_{q_{\boldsymbol{\theta}}}\{w(\boldsymbol{X},\boldsymbol{\theta})\, s(\boldsymbol{X},\boldsymbol{\theta})\}$

where $w(\boldsymbol{x},\boldsymbol{\theta}) = q_{\boldsymbol{\theta}}(\boldsymbol{x})\,\xi'(q_{\boldsymbol{\theta}}(\boldsymbol{x}))$, $\boldsymbol{s}(\boldsymbol{x},\boldsymbol{\theta})$ is score function.

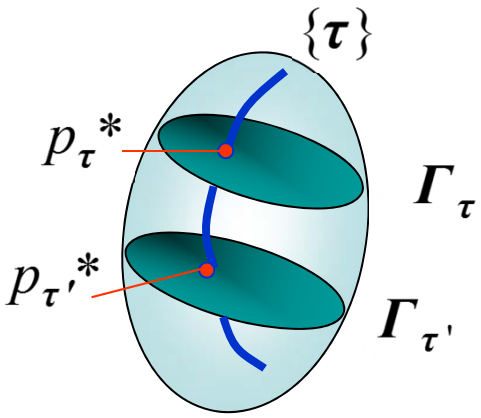
Γ-minimax

Let us fix a statistics $\boldsymbol{t}(\boldsymbol{x})$.

Equal mean space $\boldsymbol{\Gamma}_{\boldsymbol{\tau}} = \{p : E_P\{\boldsymbol{t}(\boldsymbol{X})\} = \boldsymbol{\tau}\}$

$$\max_{p \in \Gamma_{\boldsymbol{\tau}}} \min_{q \in \Gamma_{\boldsymbol{\tau}}} L_U(p,q) = L_U(p_{\boldsymbol{\tau}}^*, p_{\boldsymbol{\tau}}^*) = \min_{q \in \Gamma_{\boldsymbol{\tau}}} \max_{p \in \Gamma_{\boldsymbol{\tau}}} L_U(p,q)$$

$$p_{\boldsymbol{\tau}}^* = \arg\max_{p \in \Gamma_{\boldsymbol{\tau}}} H_U(p)$$



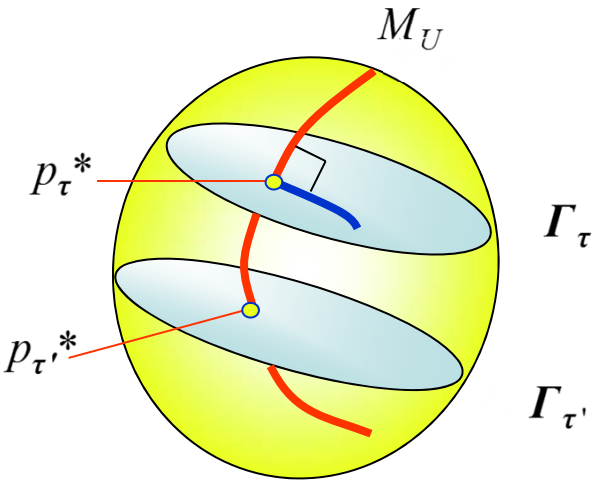
where $p_{\boldsymbol{\tau}}^*(\boldsymbol{x}) = u(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{x}) - \kappa_{\boldsymbol{\theta}})$,

$$\int \boldsymbol{t}(\boldsymbol{x}) u(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{x}) - \kappa_{\boldsymbol{\theta}}) = \boldsymbol{\tau}$$

U-estimator for U-model

U-model $\quad M_U = \{q_\theta(\boldsymbol{x}) = u(\boldsymbol{\theta}^T \boldsymbol{t}(\boldsymbol{x}) - \kappa_\theta) : \boldsymbol{\theta} \in \Theta\}$

U-estimator of mean parameter $\boldsymbol{\tau} = E_{q_\theta}\{\boldsymbol{t}(\boldsymbol{x})\}$ is $\hat{\boldsymbol{\tau}}_U = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{t}(\boldsymbol{x}_i)$



We observe that

$$L_U^{\text{emp}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \xi(q_\theta(\boldsymbol{x}_i)) - \left\langle U(\xi(q_\theta)) \right\rangle$$

$$= \frac{1}{n}\sum_{i=1}^{n} \{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{t}(\boldsymbol{x}_i) - \kappa_\theta\} - \left\langle U(\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{t}(\boldsymbol{x}) - \kappa_\theta) \right\rangle$$

which implies

$$\frac{\partial}{\partial \boldsymbol{\theta}} L_U^{\text{emp}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{t}(\boldsymbol{x}_i) - E_{q_\theta}\{\boldsymbol{t}(\boldsymbol{X})\} = 0$$

Furthermore,

$$L_U^{\text{emp}}(\boldsymbol{\theta}) - L_U^{\text{emp}}(\hat{\boldsymbol{\theta}}_U) = D_U(q_{\hat{\theta}_U}, q_\theta).$$

Influence function

Statistical functional

$$T_U(G) = \arg\min_{\theta\in\Theta}\{-\int \xi(p(x,\theta))dG(x) + \int U(\xi(p(x,\theta))dx\}$$

Influence function

$$\mathrm{IF}(T,x) \equiv \left[\frac{\partial}{\partial\varepsilon}T\,(G_\varepsilon)\right]_{\varepsilon=0} \qquad G_\varepsilon = (1-\varepsilon)F_\theta + \varepsilon\delta_x$$

$$\mathrm{IF}(T_U,x) = J_U{}^{-1}(\theta)[w(x,\theta)S(x,\theta) - E_\theta\{w(x,\theta)S(x,\theta)\}]$$

$$\mathrm{GES}(T_U) = \sup_x \; \|\mathrm{IF}(T_U,x)\|$$

Efficiency

Asymptotic variance

$$\sqrt{n}(\hat{\theta}_U - \theta) \underset{D}{\Rightarrow} N(0, J_U(\theta)^{-1} H_U(\theta) J_U(\theta)^{-1})$$

where $J_U(\theta) = E(w(X,\theta) S(X,\theta) S(X,\theta)^T),$

$$H_U(\theta) = Var(w(X,\theta) S(X,\theta))$$

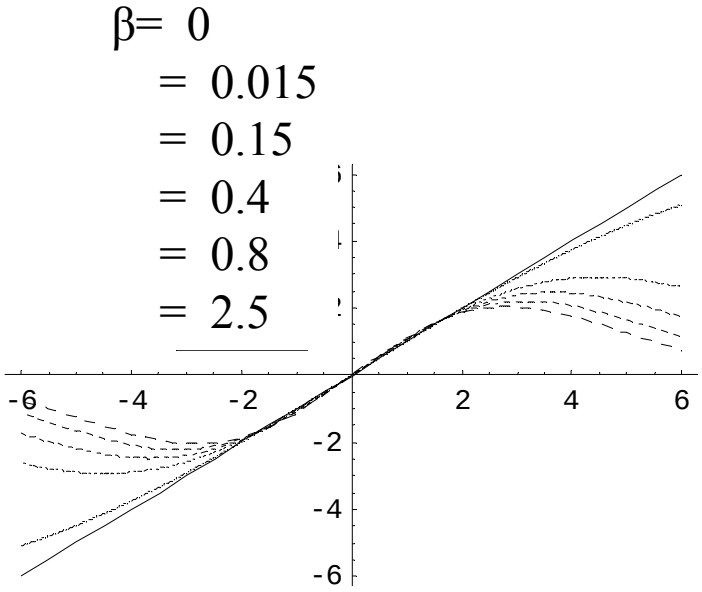
Information inequality

$$I(\theta)^{-1} \leq J_U(\theta)^{-1} H_U(\theta) J_U(\theta)^{-1}$$

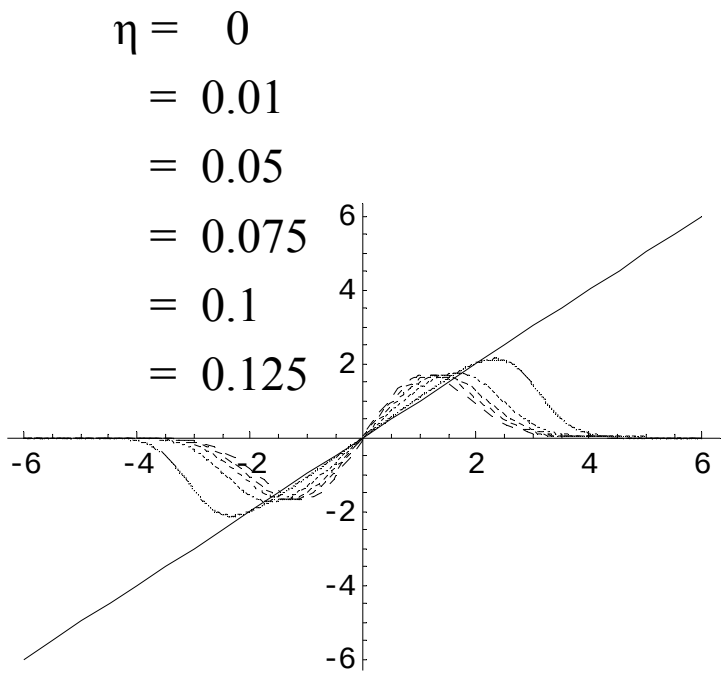
(equality holds iff $U(x) = \exp(x)$)

Normal mean

Influence function



β-power estimates



η-sigmoid estimates

$$U_\eta(t) = \exp(t) + \eta\, t$$

Gross Error Sensitivity

β-power estimates

β	効率	GES
0	1	∞
0.01	0.97	6.16
0.05	0.861	2.92
0.075	0.799	2.47
0.1	0.742	2.21
0.125	0.689	2.04

η-sigmoid estimates

η	効率	GES
0	1	∞
0.015	0.972	1.9
0.15	0.873	1.04
0.4	0.802	0.678
0.8	0.753	0.455
2.5	0.694	0.197

Multivariate normal

pdf is

$$f(y,\mu,\Sigma) = ((2\pi)^p \det\Sigma)^{-\frac{1}{2}} \exp\left\{-\frac{(y-\mu)^T\Sigma^{-1}(y-\mu)}{2}\right\}$$

Estimating equation

$$\begin{cases} \frac{1}{n}\sum w(\boldsymbol{x}_i,\boldsymbol{\theta})(\boldsymbol{x}_i-\boldsymbol{\mu}) = 0 \\ \frac{1}{n}\sum w(\boldsymbol{x},\boldsymbol{\theta})\{(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^T - \Sigma\} = c_U(\theta)\Sigma \end{cases}$$

$$\boldsymbol{\theta} = (\boldsymbol{\mu}, \Sigma)$$

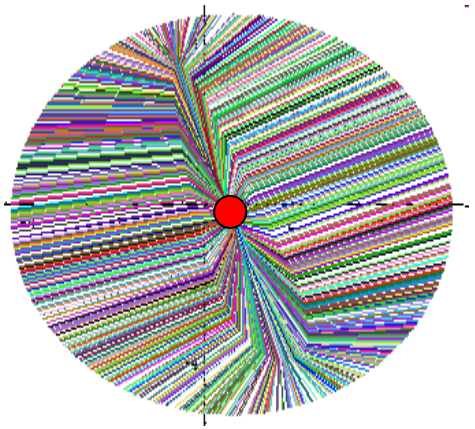
U-algorithm $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \Sigma_k) \rightarrow \boldsymbol{\theta}_{k+1} = (\boldsymbol{\mu}_{k+1}, \Sigma_{k+1})$

繰り返し重み付け平均と分散

$$\mu_{k+1} = \frac{\sum w(x_i, \theta_k)\, x_j}{\sum w(x_i, \theta_k)} \qquad \Sigma_{k+1} = \frac{\sum w(x_i, \theta_k)(x_i - \mu_k)(x_i - \mu_k)^T}{\sum \{ w(x_i, \theta_k) - c_U(\det \Sigma_k) \}},$$

Under a mild condition

$$\mathrm{L}_U(\theta_{k+1}) \geq \mathrm{L}_U(\theta_k) \quad (\forall\, k = 1, \ldots)$$



Simulation

ε-commination

$$G_\varepsilon^{(1)} = (1\text{-}\varepsilon) N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.5\\0.5 & 2\end{pmatrix}\right) + \varepsilon N\left(\begin{pmatrix}5\\-5\end{pmatrix}, \begin{pmatrix}4 & 0\\0 & 1\end{pmatrix}\right)$$

$$G_\varepsilon^{(2)} = (1\text{-}\varepsilon) N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right) + \varepsilon N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}9 & 9\\9 & 9\end{pmatrix}\right)$$

KL error $D_{\mathrm{KL}}(\hat{\theta}, \theta_0)$ with MLE $\hat{\theta}$

$\varepsilon = 0$	$\rightarrow$	$\varepsilon = 0.05$	
3.03	$\rightarrow$	39.24	under $G_\varepsilon^{(1)}$
2.70	$\rightarrow$	16.65	under $G_\varepsilon^{(2)}$

β-power estimates v.s. η-sigmoid estimates

$G_{\varepsilon}^{(1)}$

β	KL error
0	39.24
0.01	35.30
0.05	20.93
0.10	8.91
0.20	12.40
0.30	31.64

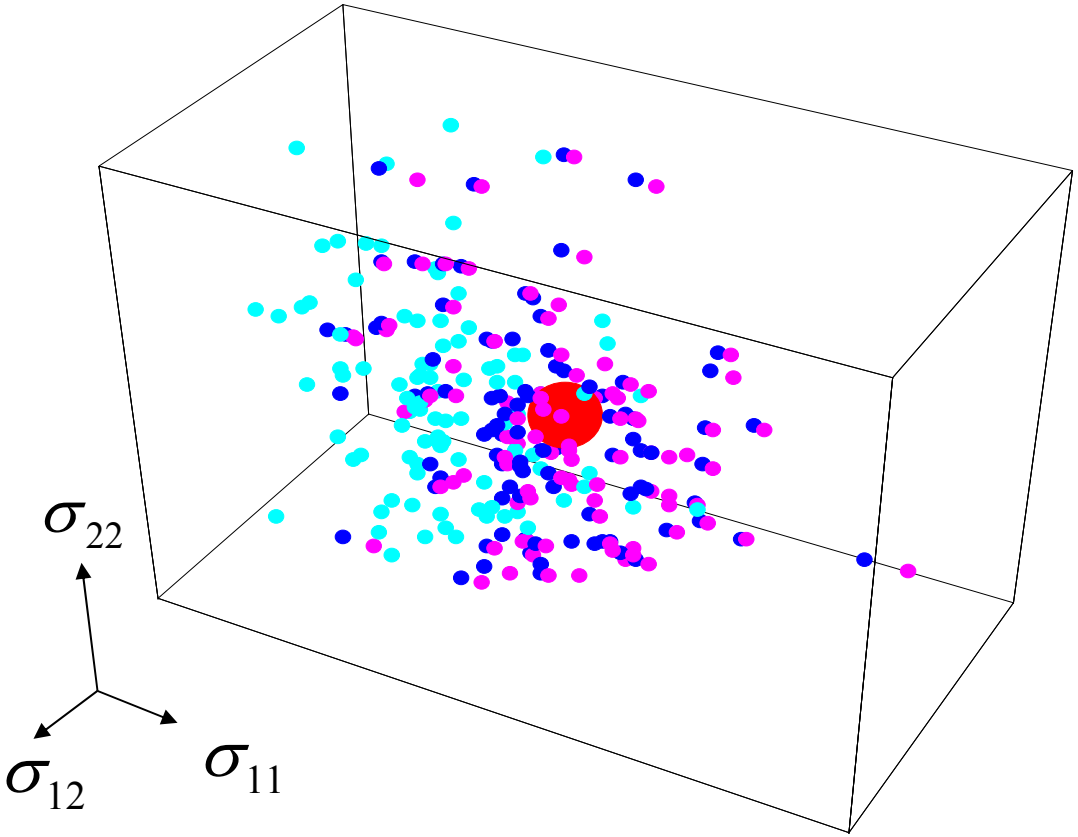
η	KL error
0	39.24
0.0001	23.04
0.00025	6.70
0.0005	4.46
0.00075	4.64
0.001	6.04

$G_{\varepsilon}^{(2)}$

0	16.5
0.01	13.91
0.05	6.65
0.10	5.14
0.20	12.20
0.30	29.68

0	16.5
0.0005	3.36
0.00075	3.19
0.001	3.9
0.002	3.04
0.003	3.07

Plot of MLEs（no outliers）

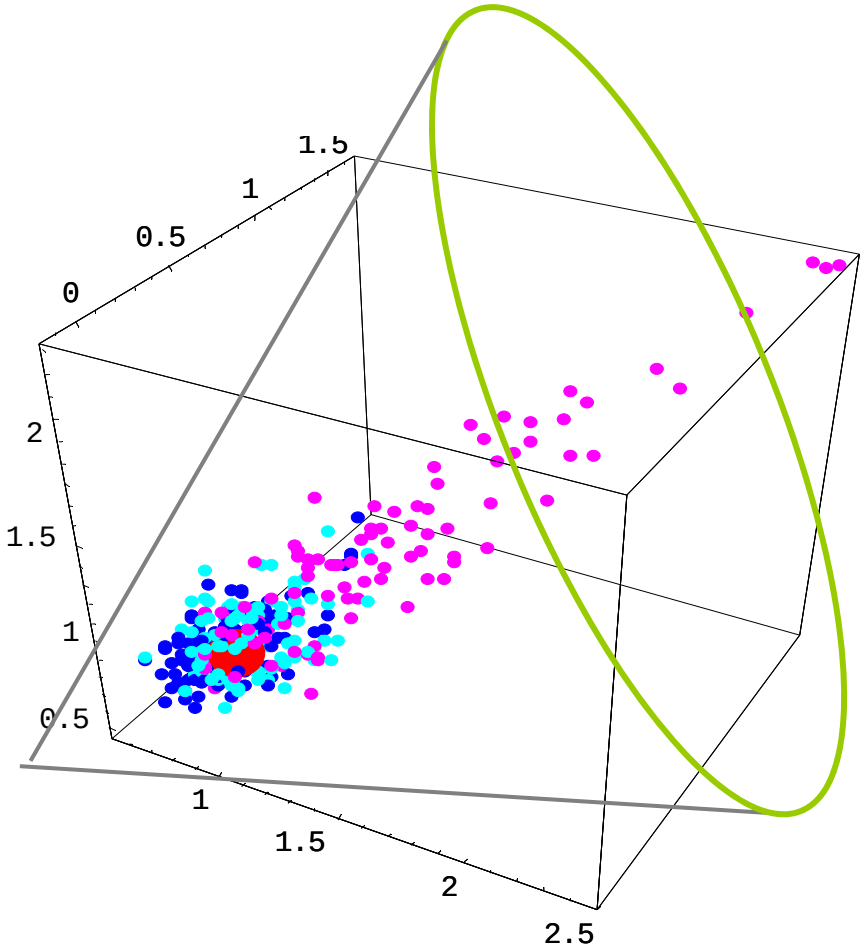


100 replications with
100 size of sample
under Normal dis.

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

- η- Est (η=0.0025)
- β- Est (β=0.1)
- MLE
- true (1,0,1)

Plot of MLEs with ouliers



100 replications with
100 size of sample
under $G_{\varepsilon}^{(2)}$

- η- Est (η=0.0025)
- β- Est (β=0.1)
- MLE
- true (1,0,1)

Selection for tuning parameter

Squared loss function

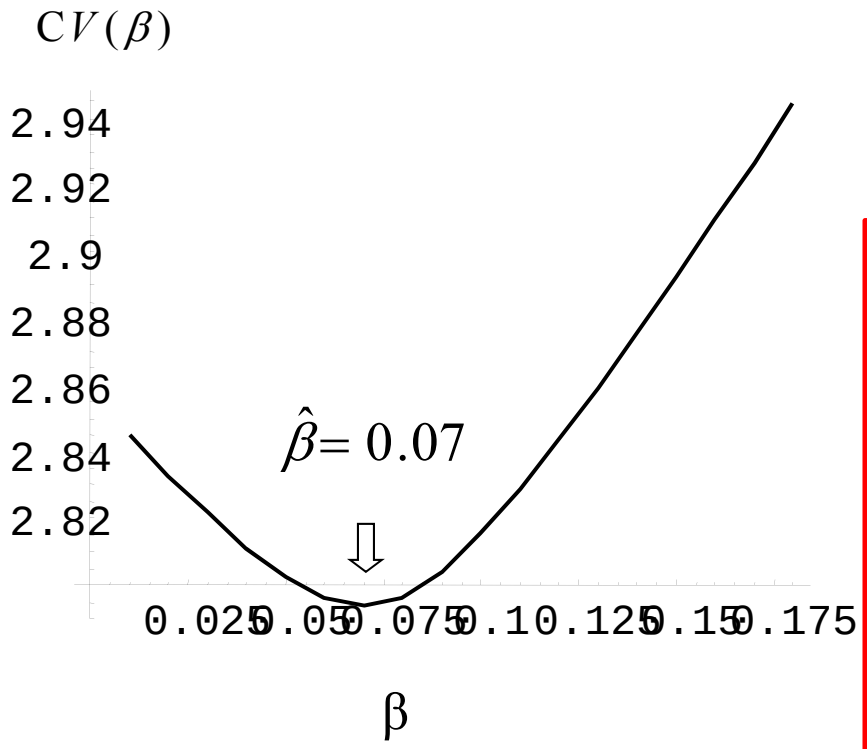
$$\text{Loss}(\hat{\theta}) = \frac{1}{2}\int \{f(y,\hat{\theta}) - g(y)\}^2 dy$$

$$\text{CV}(\beta) = -\frac{1}{n}\sum_{i=1}^{n} f(x_i, \hat{\theta}_\beta^{(-i)}) + \frac{1}{2}\int f(y,\hat{\theta}_\beta)^2 dy$$

$$\hat{\beta} = \arg\min_{\beta} \text{CV}(\beta)$$

Approximate $\hat{\theta}_\beta^{(-i)} \approx \hat{\theta}_\beta + \frac{1}{n-1}\text{IF}(x_i, \hat{\theta}_\beta)$

Normal with mean $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and variance $\begin{pmatrix} .26 & -.1 \\ -.1 & .26 \end{pmatrix}$



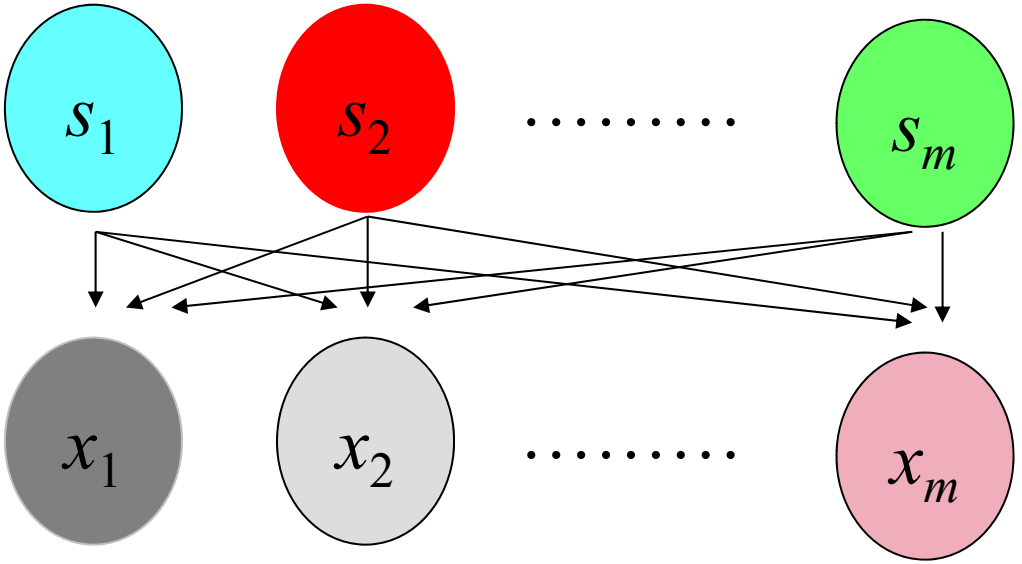
MLE signal	$\begin{pmatrix} 0.054 \\ -0.081 \end{pmatrix}$	$\begin{pmatrix} .228 & -.126 \\ -.126 & .261 \end{pmatrix}$

MLE Contami	$\begin{pmatrix} 0.204 \\ -0.184 \end{pmatrix}$	$\begin{pmatrix} 1.059 & -.263 \\ -.263 & .383 \end{pmatrix}$
β - Est contami	$\begin{pmatrix} 0.086 \\ -0.132 \end{pmatrix}$	$\begin{pmatrix} .293 & -.134 \\ -.134 & .286 \end{pmatrix}$

What is ICA?

Cocktail party effect

Blind source separation



ICA model

(Independent signals) $s = (s_1, \ldots, s_m) \sim p(s) = p_1(s_1) \cdots p_m(s_m)$

$E(S_1) = 0, \cdots, E(S_m) = 0$

(Linear mixture of signals) $W \in \boldsymbol{R}^{m \times m}$, $\mu \in \boldsymbol{R}^m$ s.t. $x = W^{-1}s + \mu$

$$f(\boldsymbol{x}, W, p) = |\det(W)| \, p_1(\boldsymbol{w}_1(\boldsymbol{x} - \boldsymbol{\mu}_1)) \cdots p_m(\boldsymbol{w}_m(\boldsymbol{x} - \boldsymbol{\mu}_m))$$

Aim is to learn W from a dataset $(x_1, \cdots, x_n)$

in which $p(\boldsymbol{s}) = p_1(s_1) \cdots p_m(s_m)$ is unknown.

ICA likelihood

Log-likelihood function

$$\ell(W) = \sum_{i=1}^{n} \sum_{j=1}^{m} \log p_j(\boldsymbol{w}_j(\boldsymbol{x}_i - \boldsymbol{\mu}_j)) + \log|\det(W)|$$

Estimating equation

$$F(x, W, p) = \frac{\partial \ell(x, W, p)}{\partial W} = (I_m - h(Wx)(Wx)^T)W^{-T}$$

$$h(s) = \left(\frac{\partial \log p_1(s_1)}{\partial s_1}, \ldots, \frac{\partial \log p_m(s_m)}{\partial s_m}\right)$$

Natural gradient algorithm, Amari et al. (1996)

Beta-ICA

β power equation $$\frac{1}{n}\sum_{i=1}^{n} f(x_i, W, \mu)^{\beta} \; F(x_i, W, \mu) = B_{\beta}(W, \mu)$$

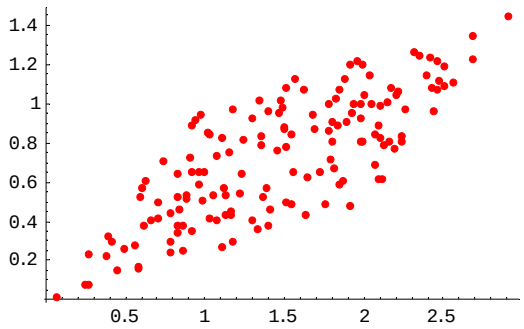
decomposability: $\forall\, s \neq t$

$$\prod_{q \neq s, q \neq t} \mathrm{E}[\{p(w_q X - \mu_q)\}^{\beta}]$$

$$\times \mathrm{E}[\{p_s(w_s X - \mu_s)\}^{\beta} h_s(w_s X - \mu_s)]$$

$$\times \mathrm{E}[\{p_t(w_t X - \mu_t))\}^{\beta} w_t X] = 0$$

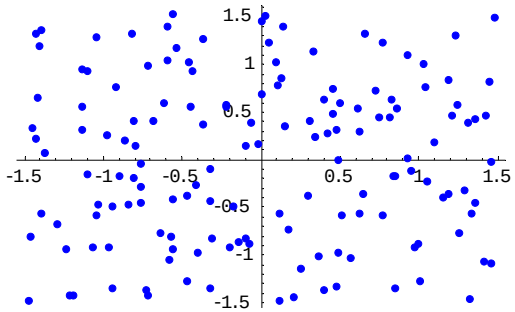
Likelihood **ICA**



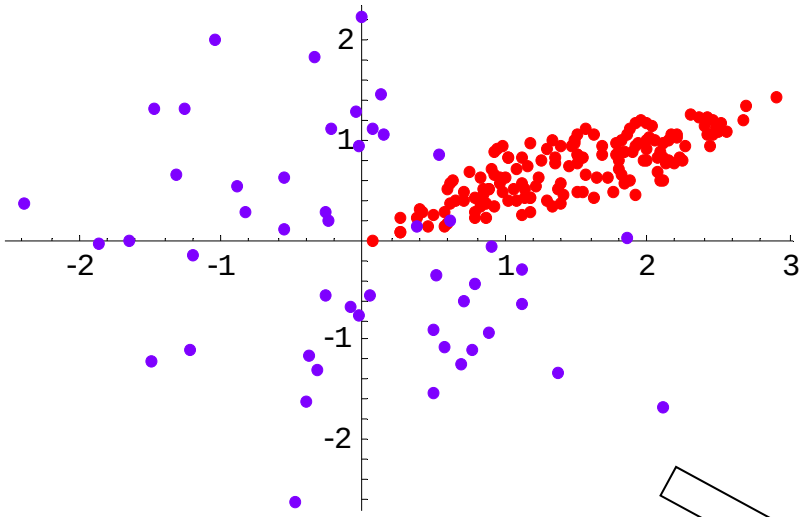
150 signals $\mathrm{U}(0,1) \times \mathrm{U}(0,1)$

Mixture matrix $W^{-1} = \begin{bmatrix} 1 & 2 \\ 1 & 0.5 \end{bmatrix}$

Maximum likelihood

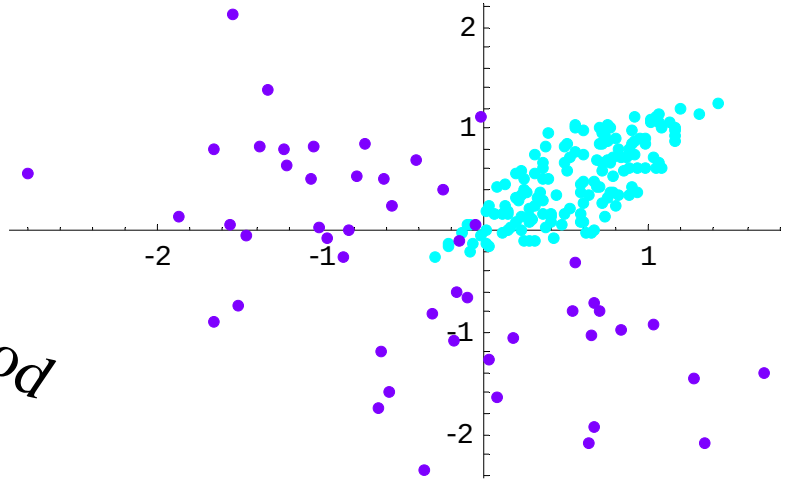


Non-robustness

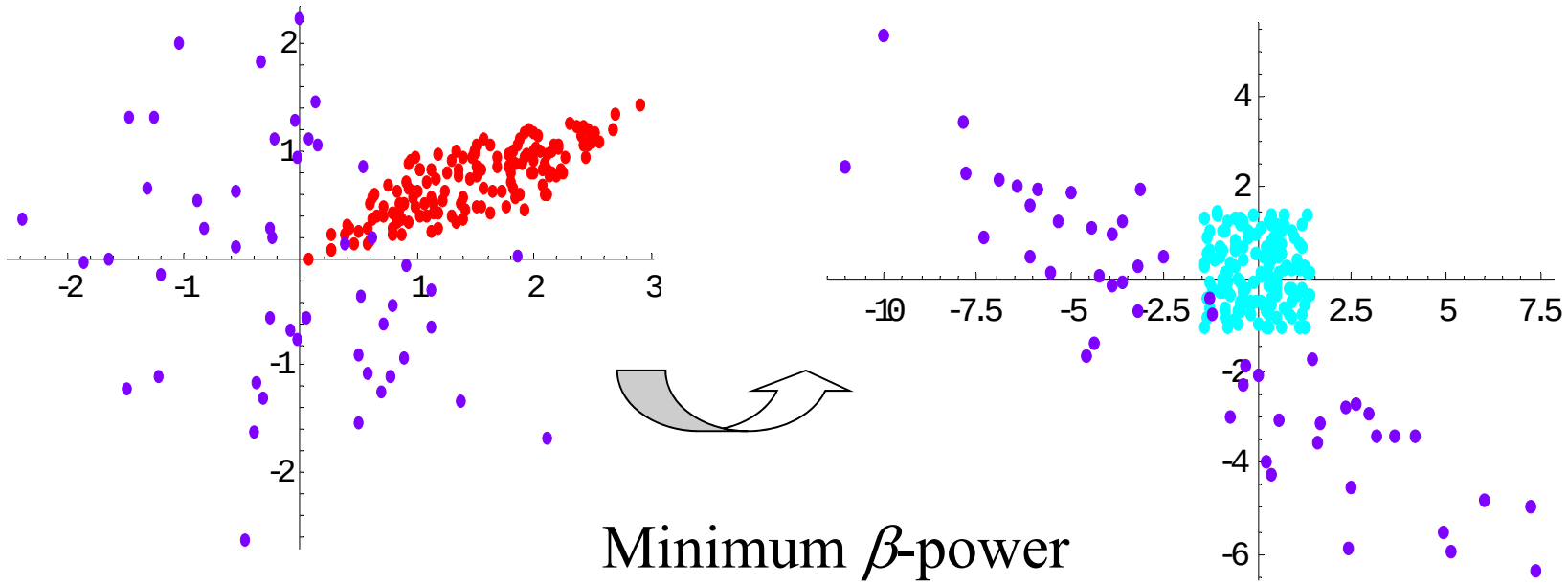


50 Gaussian noise $N(0, 1) \times N(0, 1)$

Maximum likelihood



β power-**ICA** (β=0.2)



Article

Projective Power Entropy and Maximum Tsallis Entropy Distributions

Shinto Eguchi *, Osamu Komori and Shogo Kato

Tsallisべきエントロピーに射影不変性を課したエントロピー、ダイバージェンスを考察した。
平均と分散を一定にする分布族で射影べきエントロピー最大分布モデルが正規分布，
t−分布，ウエグナー分布を含む形で導出された.
そのモデル上での平均と分散の素直な推定法を提案し，推定量が標本平均と標本分散
になることが示された.

Projective γ-power divergence

γ-cross entropy

$$C_\gamma(g,f) = -\mathrm{E}_g\left\{\frac{f(\boldsymbol{X})}{\| f \|_q}\right\}^\gamma$$

where $\| f \|_q = \left(\int f(x)^q \, dx\right)^{\frac{1}{q}}$ with $q = \gamma + 1$

γ-diagonal entropy

$$H_\gamma(f) = C_\gamma(f,f)$$

γ-divergence

$$D_\gamma(g,f) = C_\gamma(g,f) - H_\gamma(g)$$

Cf. Tsallis (1988), Basu et al (1998), Fujisawa-Eguchi (2008), Eguchi-Kato (2010)

Remark. γ = 0 reduces to Boltzmann-Shannon entropy and Kullback-Leibler divergence

Max γ-entropy

Equal moment space

$$\boldsymbol{F}(\mu,\Sigma)=\{f(x):\mathrm{E}_f(X)=\mu,\mathrm{V}_f(X)=\Sigma\}$$

γ-entropy

$$H_\gamma(f)=-\|f\|_q \quad (q=1+\gamma)$$

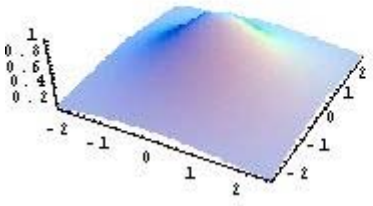
Max γ-entropy

$$H_\gamma(f_\gamma(,\mu,\Sigma)) = \max_{f\in\mathscr{F}_\gamma(\mu,\Sigma)} H_\gamma(f)$$

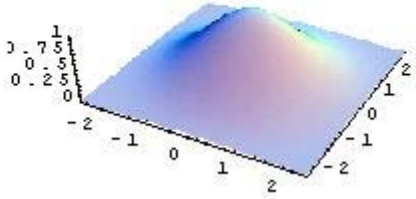
γ-model

$$f_\gamma(x,\mu,\Sigma)=c_\gamma\det(2\pi\Sigma)^{-\frac{1}{2}}\left\{1-\frac{1}{2}\kappa_\gamma(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right\}_+^{\frac{1}{\gamma}}$$

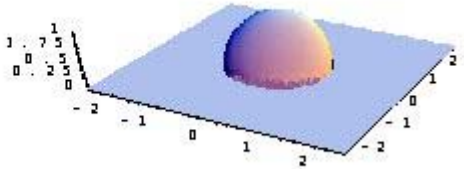
$$\boldsymbol{F}_\gamma(\mu,\Sigma)=\{f\in\boldsymbol{F}(\mu,\Sigma):\ \text{Support}\ (f)\subseteq\{x:(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\le\frac{1}{2}\kappa_\gamma\}$$



$\gamma=-0.3$ (t-distribution)



$\gamma=0$ (Gaussian)



$\gamma=2$ (Wigner)

γ-estimation

Parametric model $M = \{ f(x,\theta) : \theta \in \Theta \}$

γ-Loss function $L_\gamma(\theta) = -\frac{1}{n} z_\gamma(\theta) \sum_{i=1}^{n} f(x_i,\theta)^\gamma \ , \ z_\gamma(\theta) = \left(\int f(x,\theta)^{1+\gamma} dx \right)^{-\frac{\gamma}{1+\gamma}}$

$L_\gamma(\theta) \approx C_\gamma(g, f(\cdot,\theta)) \ \text{ if } \ x_1,\ldots,x_n \sim g(x)$

γ-estimator $\hat{\theta}_\gamma = \arg\min_{\theta \in \Theta} L_\gamma(\theta)$

Gaussian $f_{\mu,\Sigma}(x) = \det(2\pi\Sigma)^{-\frac{1}{2}} \exp\{-\frac{1}{2}(x-\mu)^{\mathrm{T}} \Sigma^{-1} (x-\mu)\}$

γ-estimator $\hat{\theta}_\gamma = (\hat{\mu}_\gamma, \hat{\Sigma}_\gamma)$

$$\hat{\mu}_\gamma = \frac{\sum \{ f_{\hat{\mu}_\gamma,\hat{\Sigma}_\gamma}(x_i) \}^\gamma x_i}{\sum \{ f_{\hat{\mu}_\gamma,\hat{\Sigma}_\gamma}(x_i) \}^\gamma}$$

$$\hat{\Sigma}_\gamma = \frac{\sum \{ f_{\hat{\mu}_\gamma,\hat{\Sigma}_\gamma}(x_i) \}^\gamma (x_i - \hat{\mu}_\gamma)(x_i - \hat{\mu}_\gamma)^{\mathrm{T}}}{(1+\gamma) \sum \{ f_{\hat{\mu}_\gamma,\hat{\Sigma}_\gamma}(x_i) \}^\gamma}$$

Remark $(\hat{\mu}_\gamma, \hat{\Sigma}_\gamma)$ is super-robust. Cf. Fujisawa-Eguchi (2008)

γ-estimation on γ-model

Let $(x_1,\ldots,x_n)$ be a random sample from $f_\gamma(\cdot,\mu,\Sigma)$

γ-loss function
$$L_\gamma(\mu,\Sigma)=\frac{1}{n}\sum_{i=1}^{n}\det(\Sigma)^{-\frac{1}{2}\frac{\gamma}{\gamma+1}}\{1-\frac{1}{2}\kappa_\gamma(x_i-\mu)^{\mathrm{T}}\Sigma^{-1}(x_i-\mu)\}$$
$$=\det(\Xi)^{\frac{\gamma}{p\gamma+2\gamma+2}}-\frac{1}{2}\kappa_\gamma\{(\mu-\bar{x})^{\mathrm{T}}\Xi(\mu-\bar{x})+\mathrm{tr}(S\Xi)\}$$
$$\text{where}\quad \Xi=\det(\Sigma)^{-\frac{1}{2}\frac{\gamma}{1+\gamma}}\Sigma^{-1}$$

Theorem. $$L_\gamma(\bar{x},S)=\min_{(\mu,\Sigma)}L_\gamma(\mu,\Sigma)$$

(Pf). $$L_\gamma(\mu,\Sigma)-L_\gamma(\bar{x},S)=\det(\Xi)^{\frac{\gamma}{p\gamma+2\gamma+2}}-\frac{1}{2}\kappa_\gamma\{(\mu-\bar{x})^{\mathrm{T}}\Xi(\mu-\bar{x})+\mathrm{tr}(S\Xi)\}$$
$$-\det(\hat{\Xi})^{\frac{\gamma}{p\gamma+2\gamma+2}}+\frac{1}{2}\kappa_\gamma\mathrm{tr}(S\hat{\Xi})\}$$
$$=C_\gamma(f_{\bar{x},S},f_{\mu,\Sigma})-C_\gamma(f_{\bar{x},S},f_{\bar{x},S})=D_\gamma(f_{\bar{x},S},f_{\mu,\Sigma})\geq 0$$

(γ-estimator , γ'-model)

loss function \ model	normal-model	γ-model
0-loss (– log-likelihood)	$(\bar{x}, S)$ MLE	**Robust model** MLE
γ-loss	**Robust γ-estimator**	$(\bar{x}, S)$ Moment estimator

Gamma-ICA

Mixture of independent signals $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S} + \boldsymbol{\mu},$

Whitening $\boldsymbol{Z} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})$

Preprocessed form $\boldsymbol{Z} = \widetilde{\boldsymbol{A}}\boldsymbol{S}, \quad \widetilde{\boldsymbol{A}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{A}$

Statistical model $f_{\boldsymbol{Z}}(z; \boldsymbol{W}) = |\det(\boldsymbol{W})| \cdot \prod_{j=1}^{p} f_j(\boldsymbol{w}_j^\top z) = f(\boldsymbol{W}^\top z).$

γ-ICA

$$\begin{cases} (\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) = \underset{\boldsymbol{\mu}, \boldsymbol{\Sigma}}{\operatorname{argmin}}\, \mathcal{D}_\gamma(\widehat{g}_{\boldsymbol{X}}, \xi_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}) \\ \widehat{\boldsymbol{W}} = \underset{\boldsymbol{W} \in \mathcal{SO}_p}{\operatorname{argmin}}\, \mathcal{D}_\gamma(\widehat{g}_{\boldsymbol{Z}}, f_{\boldsymbol{Z}}(\cdot; \boldsymbol{W})) \end{cases}$$

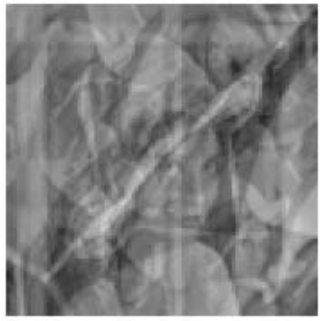
original image

original image

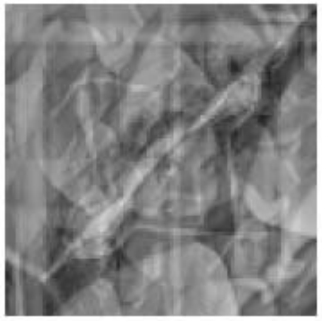
original image

original image

mixed image

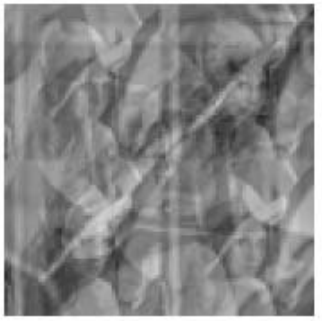


mixed image

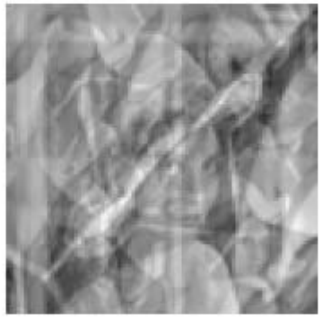


mixed image

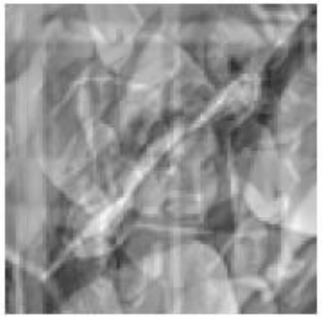
mixed image



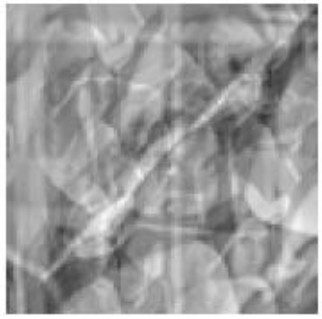
filtered image



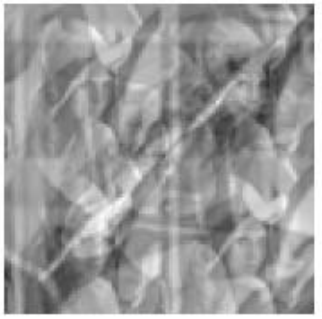
filtered image



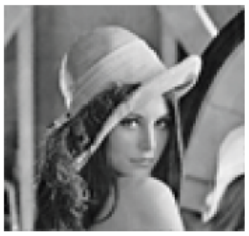
filtered image



filtered image



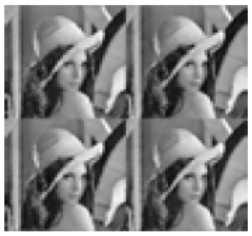
γ–ICA



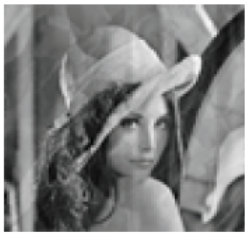
γ–ICA

γ–ICA

γ–ICA



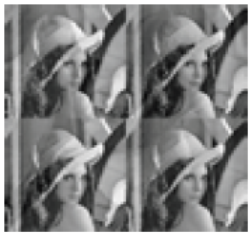
γ–ICA (filtered)



γ–ICA (filtered)

γ–ICA (filtered)

γ–ICA (filtered)



MLE–ICA

MLE–ICA

MLE–ICA

MLE–ICA

MLE–ICA (filtered)

MLE–ICA (filtered)

MLE–ICA (filtered)

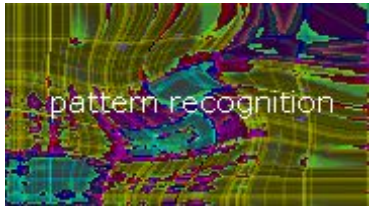
MLE–ICA (filtered)

Outline

Boosting leaning algorithms

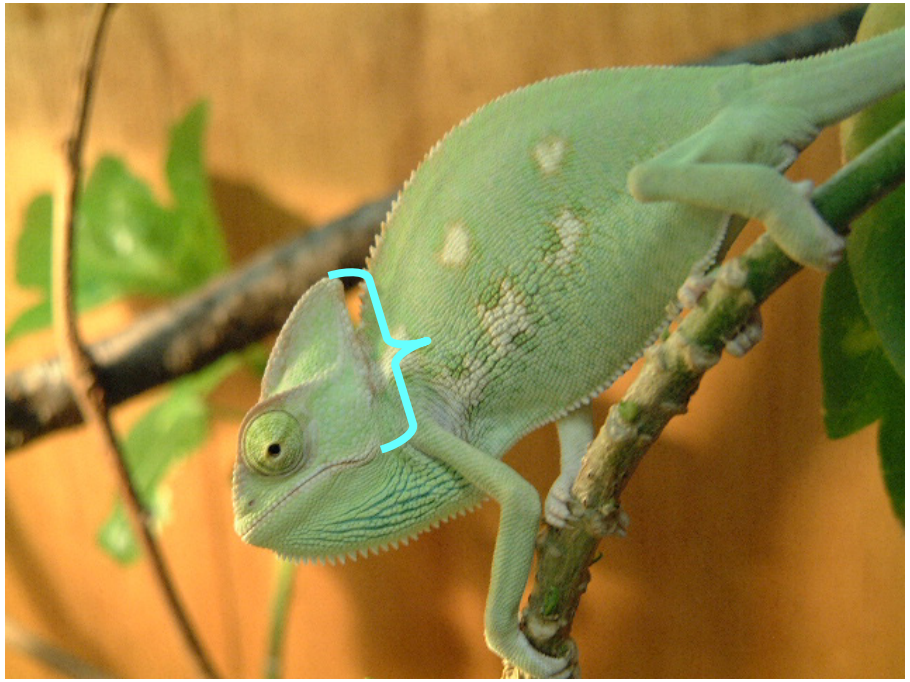
for supervised/unsupervized learning

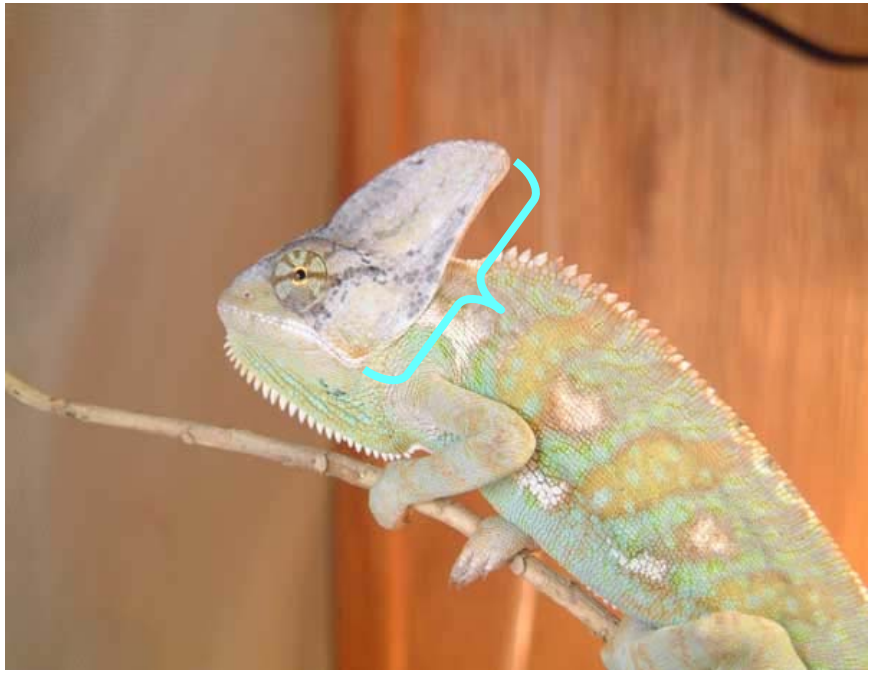
U-loss functions



Which chameleon wins?

Pattern recognition…





What is pattern recognition?

- **There are a lot of examples for pattern recognition.**
- **In principle pattern recognition is a prediction problem for class label which would classify a human interestingness and importance.**
- **Originally human brain wants to label phenomena a few words , for example (good, bad), (yes, no), (dead, alive), (success, failure), (effective, no effect)….**
- **Brain intrinsically predicts the class label from empirical evidence.**

Practice

- Character recognition
- voice recognition
- Image recognition
- face recognition
- finger print recognition
- speaker recognition

- ☆ Credit scoring
- ☆ Medical screening
- ☆ Default prediction
- ☆ Weather forcast
- ☆ Drug response
- ☆ Treatment effect
- ☆ Failure prediction
- ☆ Infectious desease

Binary classification

Feature vector → Class label

$$\boldsymbol{x} = (x_1, \ldots, x_p) \longrightarrow y \in \{-1, +1\}$$

score

$$F : \boldsymbol{x} \to z$$

classifier

$$h_F(\boldsymbol{x}) = \operatorname{sgn}\{F(\boldsymbol{x})\}$$

In a binary class $y = -1, 1$ **(G=2)**

$$F(\boldsymbol{x}) = F(\boldsymbol{x}, +1) - F(\boldsymbol{x}, -1)$$

$$\operatorname{sgn}(F(\boldsymbol{x})) = 1 \quad \Leftrightarrow \quad F(\boldsymbol{x}, +1) > F(\boldsymbol{x}, -1)$$
$$\operatorname{sgn}(F(\boldsymbol{x})) = -1 \quad \Leftrightarrow \quad F(\boldsymbol{x}, +1) < F(\boldsymbol{x}, -1)$$

$$\operatorname{sgn}(F(\boldsymbol{x})) = \arg\max_{y \in \{-1,1\}} F(\boldsymbol{x}, y)$$

Probability distribution

Let $p(\boldsymbol{x}, y)$ be a pdf of random vector $(\boldsymbol{x}, y)$.

$$p(B, C) = \int_B \{ \sum_{y \in C} p(\boldsymbol{x}, y) \, \mathrm{d}\boldsymbol{x} \}$$

Marginal density $p(y) = \int_X p(\boldsymbol{x}, y) \, \mathrm{d}\boldsymbol{x}$ $\qquad$ $p(\boldsymbol{x}) = \sum_{y \in \{1,\ldots,G\}} p(\boldsymbol{x}, y)$

Conditional density $p(y \mid \boldsymbol{x}) = \dfrac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x})}$ $\qquad$ $p(\boldsymbol{x} \mid y) = \dfrac{p(\boldsymbol{x}, y)}{p(y)}$

$$p(\boldsymbol{x}, y) = p(\boldsymbol{x} \mid y) p(y) = p(y \mid \boldsymbol{x}) p(\boldsymbol{x})$$

$$\frac{p(\boldsymbol{x}, y = 1)}{p(\boldsymbol{x}, y = -1)} = \frac{p(y = 1 \mid \boldsymbol{x})}{p(y = -1 \mid \boldsymbol{x})}$$

Error rate

Feature vector $\boldsymbol{x} \in \mathbf{R}^p$, class label $y \in \{1, \cdots, G\}$

Classifier $h_F(\boldsymbol{x})$ has error rate

$$\mathrm{Err}(h_F) = \Pr(h_F(\boldsymbol{x}) \neq y)$$

$$\mathrm{Err}(h_F) = \sum_{i \neq j} \Pr(h_F(\boldsymbol{x}) = i, y = j) = 1 - \sum_{i=1}^{G} \Pr(h_F(\boldsymbol{x}) = i, y = i)$$

Training error

$$\mathrm{Err}^{\text{train}}(h_F) = \frac{\#\{ i : h_F(\boldsymbol{x}_i) \neq y_i \}}{n} \quad \text{for } D_{\text{train}} = \{ (\boldsymbol{x}_i, y_i) : i = 1, \ldots, n \}$$

Test error

$$\mathrm{Err}^{\text{test}}(h_F) = \frac{\#\{ j : h_F(\boldsymbol{x}_j^{\text{test}}) \neq y_j^{\text{test}} \}}{m} \quad \text{for } D_{\text{test}} = \{ (\boldsymbol{x}_j^{\text{test}}, y_j^{\text{test}}) : j = 1, \ldots, m \}$$

False negative/positive

	$y = +1$	$y = -1$
$h_F(\boldsymbol{x}) = +1$	True Positive	False Positive
$h_F(\boldsymbol{x}) = -1$	False Negative	True Negative

FPR $$\mathrm{FP}(h_F) = \mathrm{pr}(h_F(\boldsymbol{x}) = +1 \,|\, y = -1)$$

FNR $$\mathrm{FN}(h_F) = \mathrm{Pr}(h_F(\boldsymbol{x}) = -1 \,|\, y = +1)$$

$$\mathrm{Err}(h_F) = \mathrm{FN}(h_F)\,\mathrm{pr}(y = +1) + \mathrm{FP}(h_F)\,\mathrm{pr}(y = -1)$$

Bayes rule

Let $p(y|\boldsymbol{x})$ be a conditional probability of y given $\boldsymbol{x}$.

Define $F_0(\boldsymbol{x}) = \log \dfrac{p(y=1 \mid \boldsymbol{x})}{p(y=-1 \mid \boldsymbol{x})}$.

The classifier $h_{\text{Bayes}}(\boldsymbol{x}) = \text{sgn}(F_0(\boldsymbol{x}))$ leads to Bayes rule.

Theorem 1

For any classifier h

$$\text{Err}(h_{\text{Bayes}}) \le \text{Err}(h)$$

Note: The optimal classifier is equivalent to the likelihood ratio. However, in practice $p(y|\boldsymbol{x})$ is unknown, so we have to learn $h_{\text{Bayes}}(\boldsymbol{x})$ based on the training data set.

Error rate for Bayes rule

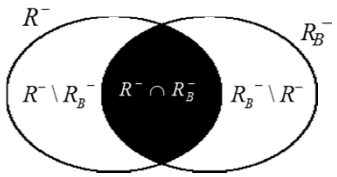
Discriminant space associated with Bayes classifier is given by

$$R_{\mathrm{B}}{}^{+} = \{\boldsymbol{x} \in \boldsymbol{R}^{p} : h_{\mathrm{Bayes}}(\boldsymbol{x}) = +1\} = \left\{\boldsymbol{x} \in \boldsymbol{R}^{p} : p(y=+1 \mid \boldsymbol{x}) \geq p(y=-1 \mid \boldsymbol{x})\right\}$$

$$R_{\mathrm{B}}{}^{-} = \{\boldsymbol{x} \in \boldsymbol{R}^{p} : h_{\mathrm{Bayes}}(\boldsymbol{x}) = -1\} = \left\{\boldsymbol{x} \in \boldsymbol{R}^{p} : p(y=+1 \mid \boldsymbol{x}) < p(y=-1 \mid \boldsymbol{x})\right\}$$

In general, when a classifier h associates with spaces $\{R^{+}, R^{-}\}$

$$\begin{aligned}
\mathrm{Err}(h) - \mathrm{Err}(h_{\mathrm{Bayes}}) &= \left(\int_{R^{-}} - \int_{R_{\mathrm{B}}^{-}}\right) p(\boldsymbol{x}) p(y=+1 \mid \boldsymbol{x}) d\boldsymbol{x} + \left(\int_{R^{+}} - \int_{R_{\mathrm{B}}^{+}}\right) p(\boldsymbol{x}) p(y=-1 \mid \boldsymbol{x}) d\boldsymbol{x} \\
&= \left(\int_{R^{-} \setminus R^{-}{}_{B}} - \int_{R_{B}{}^{-} \setminus R^{-}}\right) p(\boldsymbol{x}) p(y=+1 \mid \boldsymbol{x}) d\boldsymbol{x} + \left(\int_{R^{+} \setminus R_{B}{}^{+}} - \int_{R_{B}{}^{+} \setminus R^{+}}\right) p(\boldsymbol{x}) p(y=-1 \mid \boldsymbol{x}) d\boldsymbol{x} \\
&\geq \left(\int_{R^{-} \setminus R^{-}{}_{B}} - \int_{R_{B}{}^{-} \setminus R^{-}}\right) p(\boldsymbol{x}) p(y=-1 \mid \boldsymbol{x}) d\boldsymbol{x} + \left(\int_{R^{+} \setminus R_{B}{}^{+}} - \int_{R_{B}{}^{+} \setminus R^{+}}\right) p(\boldsymbol{x}) p(y=+1 \mid \boldsymbol{x}) d\boldsymbol{x} \\
&= \left(\int_{R^{-}} - \int_{R_{\mathrm{B}}^{-}}\right) p(\boldsymbol{x}) p(y=-1 \mid \boldsymbol{x}) d\boldsymbol{x} + \left(\int_{R^{+}} - \int_{R_{\mathrm{B}}^{+}}\right) p(\boldsymbol{x}) p(y=+1 \mid \boldsymbol{x}) d\boldsymbol{x} \\
&= \{1 - \mathrm{Err}(h)\} - \{1 - \mathrm{Err}(h_{\mathrm{Bayes}})\}
\end{aligned}$$



$$\mathrm{Err}(h) \geq \mathrm{Err}(h_{\mathrm{Bayes}})$$

Cf. Mclachlan, 2002

Boost learning

Weak learners can be combined into a strong leanrner?

Weak learner (classifier) = error rate is slightly less than .5

Strong learner (classifier) = error rate is slightly less than that of Bayes classifier

Boost by filter (Schapire, 1990)

Bagging, Arching（bootstrap）
(Breiman, Friedman, Hasite)

AdaBoost (Schapire, Freund, Batrlett, Lee)

Set of weak learners

Decision stumps

$$\boldsymbol{F}_{\text{stamp}} = \left\{ f_j(\boldsymbol{x}, a, b) = a\,\text{sgn}(x_j - b) \colon j \in \{1, \cdots, p\}, a \in \{-1, +1\}, b \in \mathbb{R} \right\}$$

Linear classifiers

$$\boldsymbol{F}_{\text{linear}} = \left\{ f(\boldsymbol{x}, \boldsymbol{\beta}) = \text{sgn}\,(\boldsymbol{\beta}_1^{\mathrm{T}} \boldsymbol{x} + \beta_0) \colon \boldsymbol{\beta} = (\boldsymbol{\beta}_1, \beta_0) \in \mathbb{R}^{p+1} \right\}$$

Neural net **SVM** ***k*-nearest neighbur**

Note: **not strong but a variety of characters**

$$\mathsf{F}_{\text{stamp}} \subseteq \mathsf{F}_{\text{linear}}$$

Exponential loss function

Let $D_{\text{train}} = \{ (\boldsymbol{x}_i, y_i) : i = 1, ..., n \}$ be a training (example) set.

Empirical exponential loss function for a score function $F(\boldsymbol{x})$

$$L_{\text{exp}}^{D}(F) = \frac{1}{n} \sum_{i=1}^{n} \exp\{ -y_i F(\boldsymbol{x}_i) \}$$

Expected exponential loss function for a score function $F(\boldsymbol{x})$

$$L_{\text{exp}}^{E}(F) = \int_X \{ \sum_{y \in \{+1,-1\}} \exp\{ -yF(\boldsymbol{x}) \} q(y \mid \boldsymbol{x}) \} q(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$

where $q(y|\boldsymbol{x})$ is the conditional distribution given $\boldsymbol{x}$, q($\boldsymbol{x}$) is the pdf of $\boldsymbol{x}$.

Adaboost

1. Intial : $w_1(i) = \frac{1}{n} \quad (i = 1 \cdots n), \; F_0(\boldsymbol{x}) = 0$

2. For $t = 1, \cdots, T$ $\quad \varepsilon_t(f) = \sum \mathrm{I}(y_i \neq f(\boldsymbol{x}_i)) \frac{w_t(i)}{\sum w_t(i')}$,

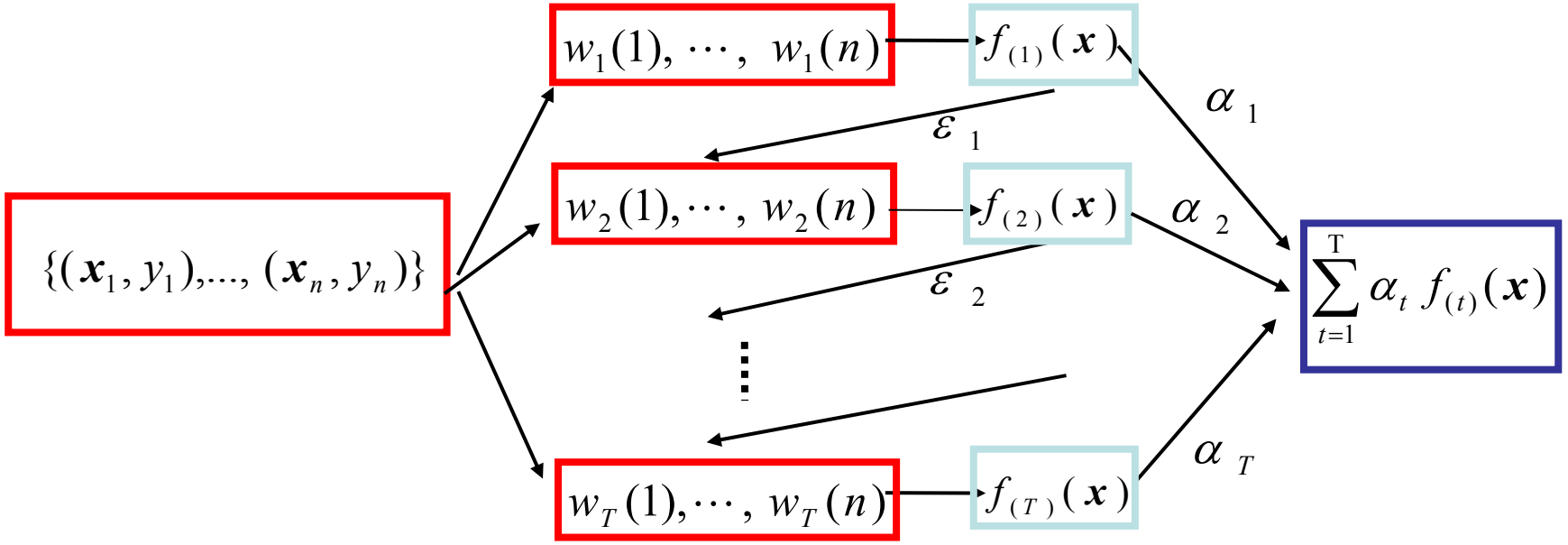
(a) $\quad \varepsilon_t(f_{(t)}) = \min_{f \in \mathsf{F}} \varepsilon_t(f)$

(b) $\quad \alpha_t = \frac{1}{2} \log \frac{1 - \varepsilon_t(f_{(t)})}{\varepsilon_t(f_{(t)})}$

(c) $\quad w_{t+1}(i) = w_t(i) \exp(-\alpha_t f_{(t)}(\boldsymbol{x}_i) y_i)$

3. $\operatorname{sign}\left(F_T(\boldsymbol{x})\right)$, where $F_T(\boldsymbol{x}) = \sum_{t=1}^{T} \alpha_t f_{(t)}(\boldsymbol{x})$

Learning algorithm



Final learner $F_T(\boldsymbol{x}) = \sum_{t=1}^{T} \alpha_t f_{(t)}(\boldsymbol{x})$

Update weight

update $w_t(i) \rightarrow w_{t+1}(i)$

$$f_{(t)}(\boldsymbol{x}_i) \neq y_i \Rightarrow \text{Multiply } e^{\alpha_t}$$

$$f_{(t)}(\boldsymbol{x}_i) = y_i \Rightarrow \text{Multiply } e^{-\alpha_t}$$

Weighted error rate $\varepsilon_t(f_{(t)}) \rightarrow \varepsilon_{t+1}(f_{(t)}) \rightarrow \varepsilon_{t+1}(f_{(t+1)})$

$$\varepsilon_{t+1}(f_{(t)}) = \frac{1}{2} \qquad \text{The worst case}$$

$$\varepsilon_{t+1}(f_{(t)}) = \frac{1}{2}$$

$$\varepsilon_{t+1}(f_t) = \sum_{i=1}^{n} I(f_{(t)}(\boldsymbol{x}_i) \neq y_i) \frac{w_{t+1}(i)}{\sum w_{t+1}(i')}$$

$$= \frac{\sum_{i=1}^{n} I(f_{(t)}(\boldsymbol{x}_i) \neq y_i) \exp\{-\alpha_t\, y_i f_{(t)}(\boldsymbol{x}_i)\} w_t(i)}{\sum_{i=1}^{n} \exp\{-\alpha_t\, y_i f_{(t)}(\boldsymbol{x}_i)\} w_t(i)}$$

$$= \frac{\exp\{\alpha_t\} \sum_{i=1}^{n} I(f_{(t)}(\boldsymbol{x}_i) \neq y_i) w_t(i)}{\exp\{\alpha_t\} \sum_{i=1}^{n} I(f_{(t)}(\boldsymbol{x}_i) \neq y_i) w_t(i) + \exp\{-\alpha_t\} \sum_{i=1}^{n} I(f_{(t)}(\boldsymbol{x}_i) = y_i) w_t(i)}$$

$$= \frac{\sqrt{\frac{1-\varepsilon_t(f_{(t)})}{\varepsilon_t(f_{(t)})}}\, \varepsilon_t(f_{(t)})}{\sqrt{\frac{1-\varepsilon_t(f_{(t)})}{\varepsilon_t(f_{(t)})}}\, \varepsilon_t(f_{(t)}) + \sqrt{\frac{\varepsilon_t(f_{(t)})}{1-\varepsilon_t(f_{(t)})}}\, \{1-\varepsilon_t(f_{(t)})\}} = \frac{1}{2} \quad \blacksquare$$

Update in exponential loss

$$L_{\exp}(F) = \frac{1}{n}\sum_{i=1}^{n} \exp\{-y_i F(\mathbf{x}_i)\}$$

Consider $F(\boldsymbol{x}) \rightarrow F(\boldsymbol{x}) + \alpha f(\boldsymbol{x})$

$$L_{\exp}(F+\alpha f) = \frac{1}{n}\sum_{i=1}^{n}\exp\{-y_i F(\boldsymbol{x}_i)\}\exp\{-\alpha y_i f(\boldsymbol{x}_i)\}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\exp\{-y_i F(\boldsymbol{x}_i)\}\left[e^{\alpha}\mathrm{I}(f(\boldsymbol{x}_i)\neq y_i) + e^{-\alpha}\mathrm{I}(f(\boldsymbol{x}_i)= y_i)\right]$$

$$= L_{\exp}(F)\{\, e^{\alpha}\varepsilon(f) + e^{-\alpha}(1-\varepsilon(f))\}$$

where $\varepsilon(f) = \dfrac{\sum_{i=1}^{n} I(f(\mathbf{x}_j)\neq y_i)\exp\{-y_i F(\mathbf{x}_i)\}}{L_{\exp}(F)}$

Sequential optimization

$$L_{\exp}(F+\alpha f) = L_{\exp}(F)\{\varepsilon(f)e^{\alpha} + (1-\varepsilon(f))e^{-\alpha}\}$$

$$\varepsilon(f)e^{\alpha} + (1-\varepsilon(f))e^{-\alpha}$$

$$= \left\{\sqrt{\frac{1-\varepsilon(f)}{e^{\alpha}}} - \sqrt{\varepsilon(f)e^{\alpha}}\right\}^2 + 2\sqrt{\varepsilon(f)\{1-\varepsilon(f)\}}$$

$$\geq 2\sqrt{\varepsilon(f)\{1-\varepsilon(f)\}}$$

Equality holds if and only if $\alpha_{\text{opt}} = \frac{1}{2}\log\frac{1-\varepsilon(f)}{\varepsilon(f)}$

AdaBooost = Seq-min exponential loss

$$\min_{\alpha \in \boldsymbol{R}} L_{\exp}(F_{t-1} + \alpha f_{(t)}) = L_{\exp}(F_{t-1}) \sqrt{\varepsilon(f_{(t)})\{1 - \varepsilon(f_{(t)})\}}$$

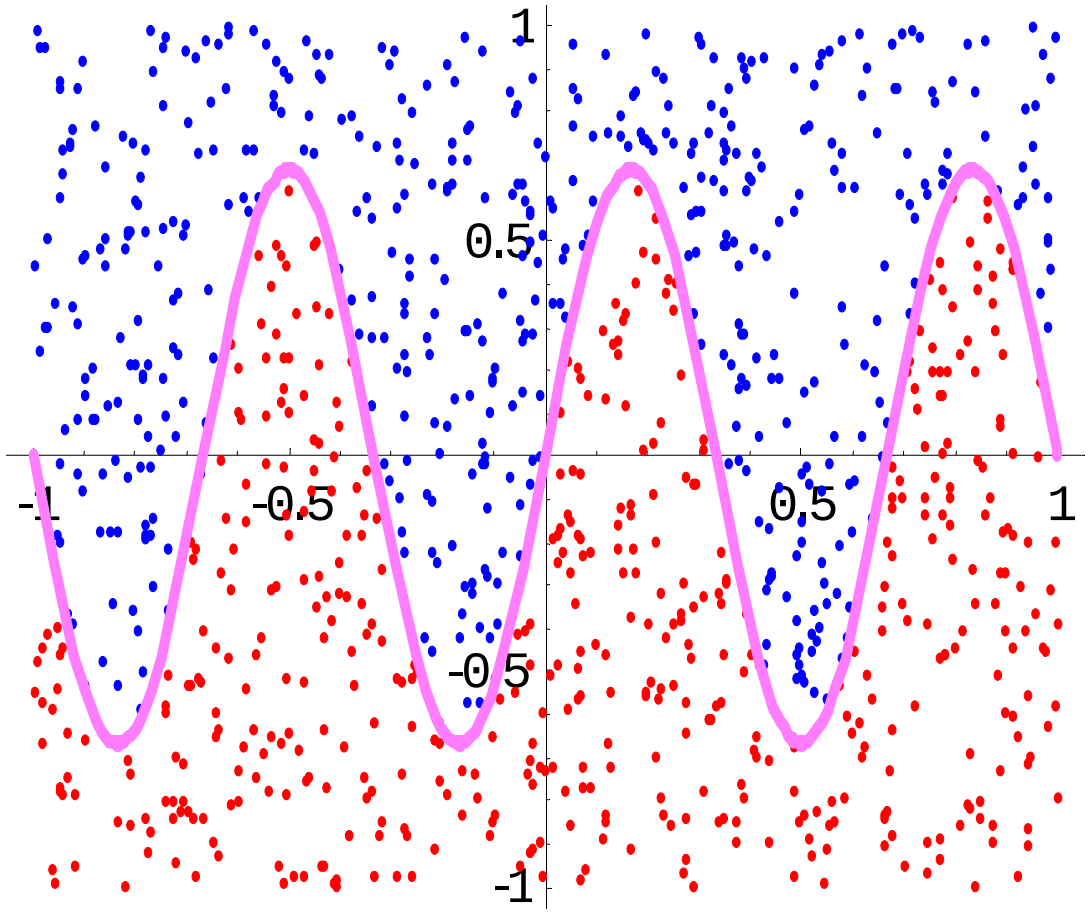
$$\alpha_{\text{opt}} = \frac{1}{2} \log \frac{1 - \varepsilon(f_{(t)})}{\varepsilon(f_{(t)})}$$

(a) $$f_{(t)} = \operatorname*{argmin}_{f \in F} \varepsilon_t(f)$$

(b) $$\alpha_t = \arg\min_{\alpha \in \boldsymbol{R}} L_{\exp}(F_{t-1} + \alpha f_{(t)})$$

(c) $$w_{t+1}(i) \propto w_t(i) \exp\{\alpha_t y_i f_t (x_i)\}$$

Simulation (complete separable)



Feature space

$$[-1,1]\times[-1,1]$$

Decision boundary

$$x_2 = \sin(2\pi x_1)$$

$$\{(\boldsymbol{x}_i, y_i) : i = 1,\cdots,1000\}$$

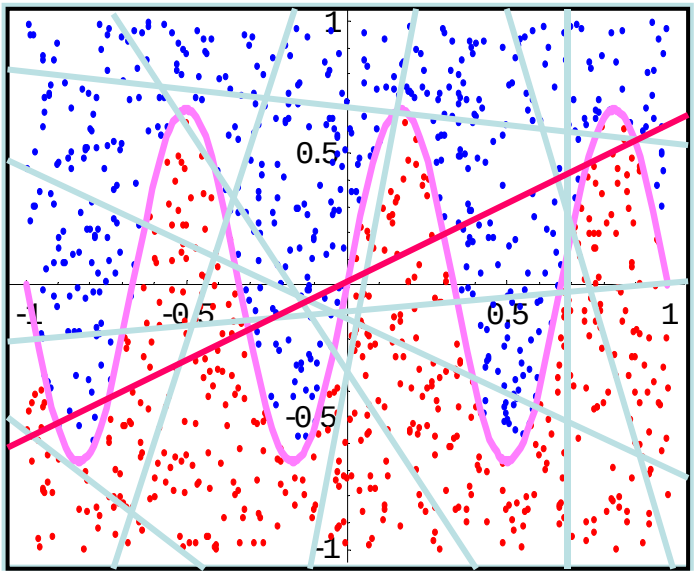
$$\boldsymbol{x}_i \in [-1, 1]\times[-1, 1]$$

$$y_i \in \{-1, +1\}$$

Set of linear classifiers

Linear classification machines

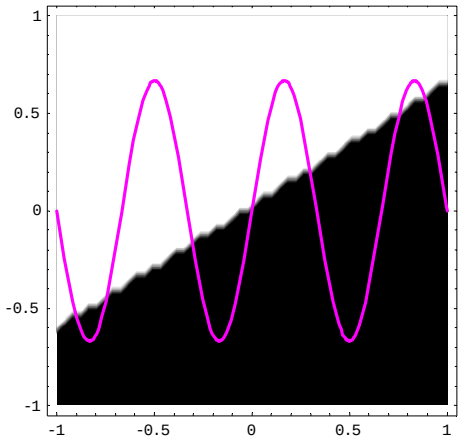
$$f(x_1, x_2) = \operatorname{sgn}(r_1 x_1 + r_2 x_2 + r_3) = \begin{cases} +1 & \text{if } r_1 x_1 + r_2 x_2 + r_3 \geq 0 \\ -1 & \text{if } r_1 x_1 + r_2 x_2 + r_3 < 0 \end{cases}$$



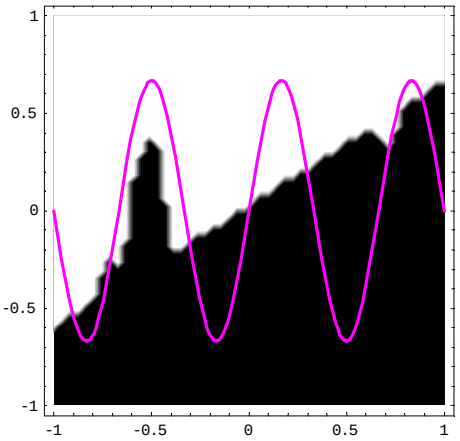
Random generation

$$\{r_1, r_2, r_3\} \sim U(-1, 1)^3$$

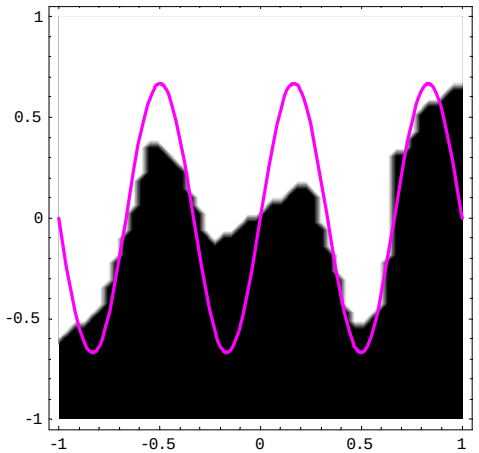
Learning process



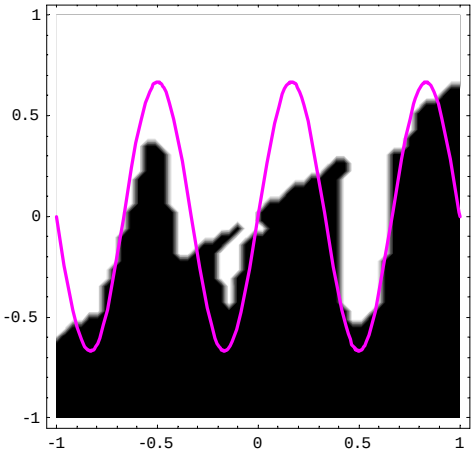
Iter = 1, train err = 0.21



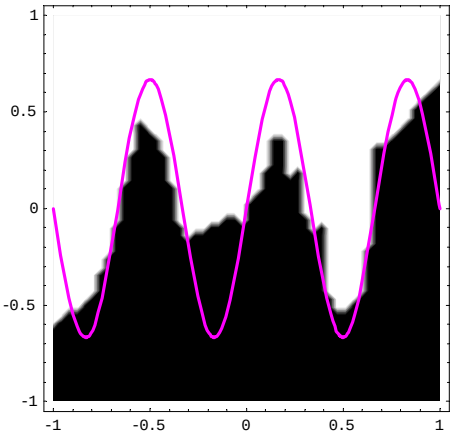
Iter = 13, train err = 0.18



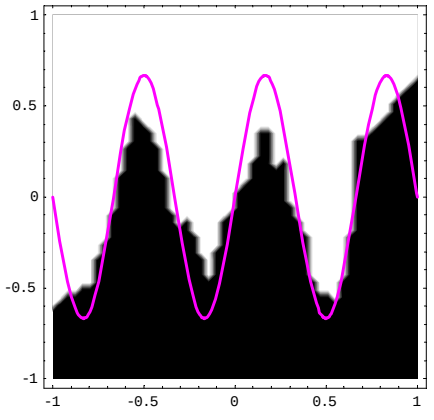
Iter = 17, train err = 0.10



Iter = 23, train err = 0.10

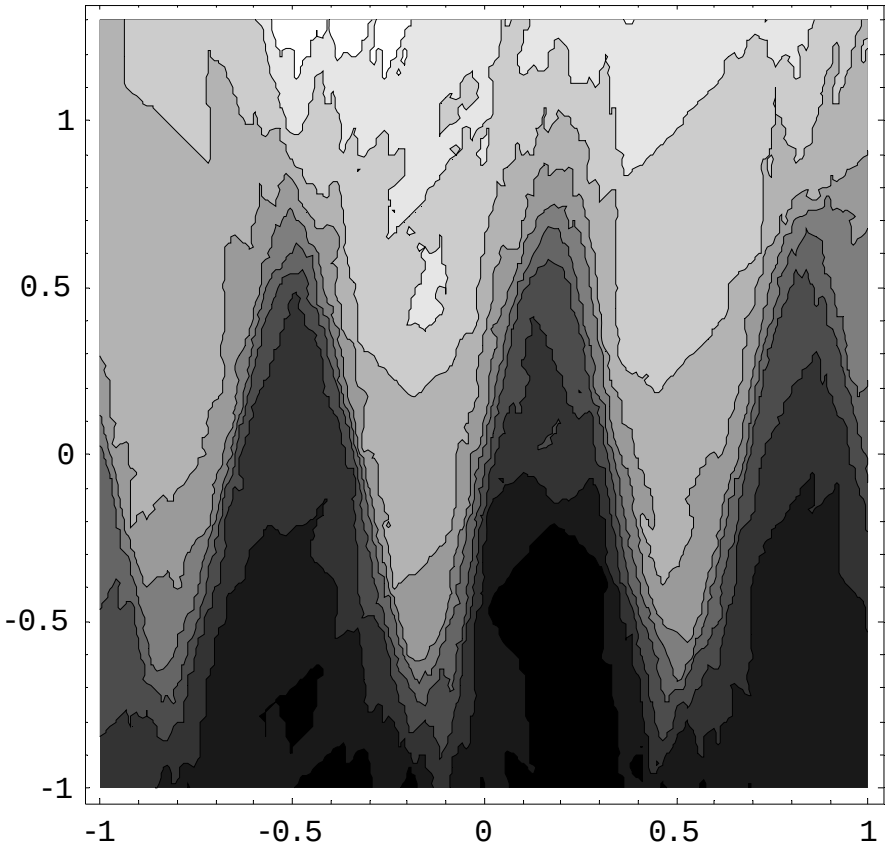


Iter = 31, train err = 0.095

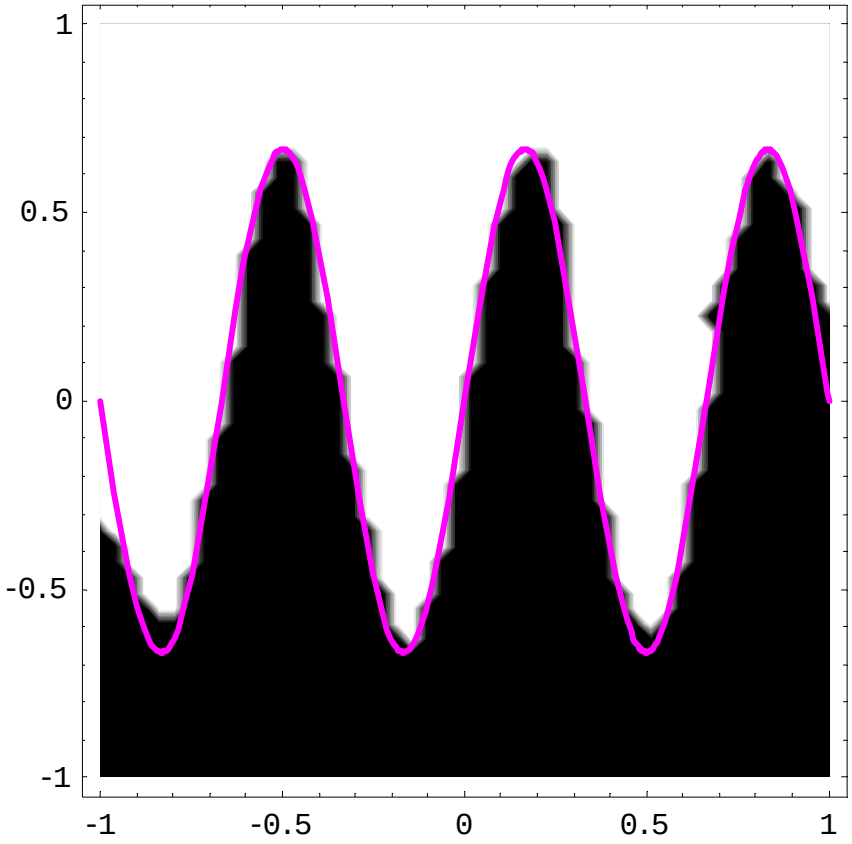


Iter = 47, train err = 0.08

Final decision boundary

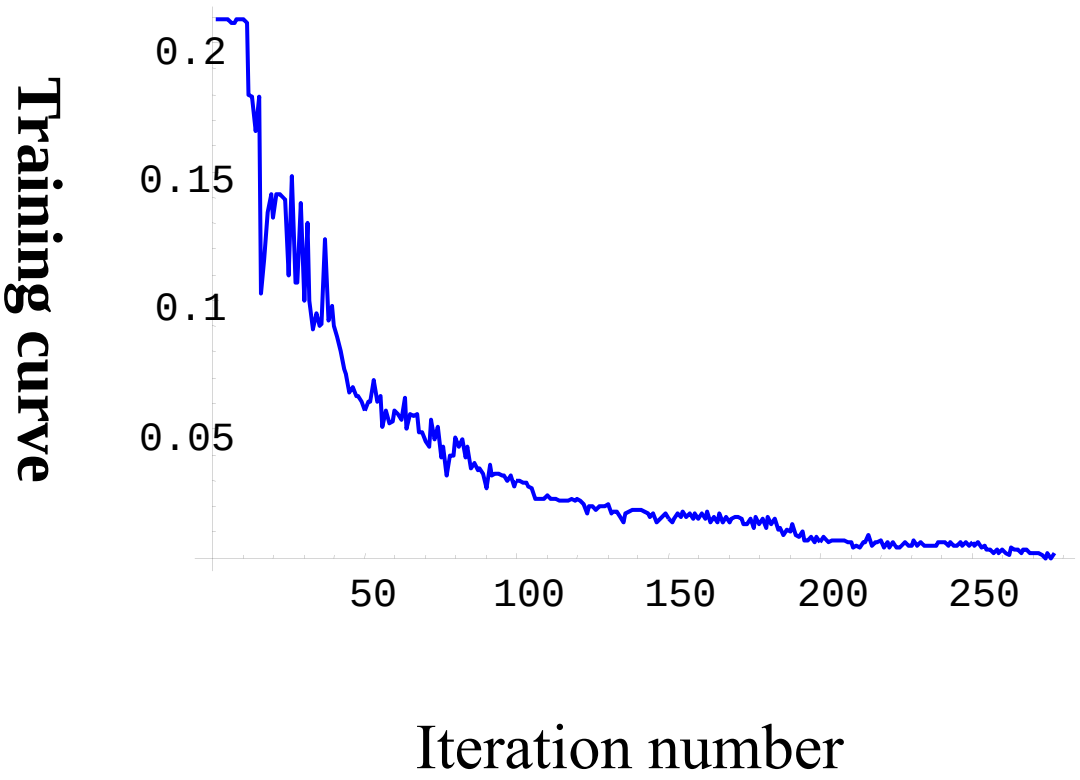


Contour of $F(\boldsymbol{x})$



Sign($F(\boldsymbol{x})$)

Learning curve



KL divergence

Label set $\boldsymbol{Y} = \{1, \cdots, G\}$ Feature space $\boldsymbol{X} \subseteq \mathbf{R}^p$

Nonnegative function $m(\boldsymbol{x}, y),\ \mu(\boldsymbol{x}, y) \quad (\boldsymbol{x} \in \boldsymbol{X}, y \in \boldsymbol{Y})$

$$D_{\mathrm{KL}}(m,\ \mu) = \int_X \sum_{y=1}^{G} \{m(\boldsymbol{x}, y) \log \frac{m(\boldsymbol{x}, y)}{\mu(\boldsymbol{x}, y)} - m(\boldsymbol{x}, y) + \mu(\boldsymbol{x}, y)\} \mathrm{d}\boldsymbol{x}$$

Note

For conditinal distribution $p(y|\boldsymbol{x})$, $q(y|\boldsymbol{x})$ given $\boldsymbol{x}$ with common marginal density $p(\boldsymbol{x})$ we can write

$$m(\boldsymbol{x}, y) = p(y \mid \boldsymbol{x}) p(\boldsymbol{x}),\ \mu(\boldsymbol{x}, y) = q(y \mid \boldsymbol{x}) p(\boldsymbol{x})$$

Then,

$$D_{\mathrm{KL}}(m,\ \mu) = \int_X \{\sum_{y=1}^{G} p(y \mid \boldsymbol{x}) \log \frac{p(y \mid \boldsymbol{x})}{q(y \mid \boldsymbol{x})}\} p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$

Twin of KL loss function

For data distribution $q(\boldsymbol{x}, y) = q(\boldsymbol{x})\, q(y|\boldsymbol{x})$ we model as

$$m_1(y \mid \boldsymbol{x}) = \sum_{g=1}^{G} \exp\{ F(\boldsymbol{x}, g) - F(\boldsymbol{x}, y)\}$$

$$m_2(y \mid \boldsymbol{x}) = \frac{\exp\{ F(\boldsymbol{x}, y)\}}{\sum_{g=1}^{G} \exp\{ F(\boldsymbol{x}, g)\}}$$

Then

$$D_{\mathrm{KL}}(q, m) = \int_X \{ \sum_{y \in \{+1, -1\}} q(y \mid \boldsymbol{x}) \log \frac{q(y \mid \boldsymbol{x})}{m(y \mid \boldsymbol{x})} - q(y \mid \boldsymbol{x}) + m(y \mid \boldsymbol{x}) \} q(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$

Log loss

Exp loss

Bound for exponential loss

Empirical exp loss $\overline{L}_{\exp}(F) = \frac{1}{n}\sum_{i=1}^{n} \exp(-y_i F(\boldsymbol{x}_i))$

Expected exp loss $L_{\exp}(F) = \mathrm{E}\{\exp(-YF(\boldsymbol{X})\}$

Theorem.

Let $\boldsymbol{F}$ be a space of all discriminant functions and

$$F_{\text{opt}} = \arg\min_{F \in \boldsymbol{F}} L_{\exp}(F).$$

Then $F_{\text{opt}}(\boldsymbol{x}) = \frac{1}{2}\log\frac{p(y=+1\,|\,\boldsymbol{x})}{p(y=-1\,|\,\boldsymbol{x})}$.

Variational calculas

$$\frac{\partial}{\partial \delta} L_{\exp}(F) = \mathrm{E}\left[\frac{\partial}{\partial F}\exp(-YF(\boldsymbol{X}))\right] \quad = \mathrm{E}\left[Y\exp(-YF(\boldsymbol{X}))\right]$$

$$= \mathrm{E}\left[\exp(-F(\boldsymbol{X}))p(y=+1\,|\,\boldsymbol{X}) - \exp(F(\boldsymbol{X}))p(y=-1\,|\,\boldsymbol{X})\right]$$

$$= \mathrm{E}\left[\exp(-F(\boldsymbol{X}))p(y=+1\,|\,\boldsymbol{X})\left\{\exp(2F(\boldsymbol{X})) - \frac{p(y=+1\,|\,\boldsymbol{X})}{p(y=-1\,|\,\boldsymbol{X})}\right\}\right]$$

Hence $$F_{\mathrm{opt}}(\boldsymbol{x}) = \frac{1}{2}\log\frac{p(y=+1\,|\,\boldsymbol{x})}{p(y=-1\,|\,\boldsymbol{x})}.$$

On AdaBoost

FDA, or Logistic regression

$$F(\boldsymbol{x}) = \boldsymbol{\alpha}_1 \boldsymbol{x} + \alpha_0 = \sum_{j=1}^{p} \alpha_{1j} x_j + \alpha_0$$

- Parametric approach to Bayes classifier

AdaBoost $F(\boldsymbol{x}) = \sum_{t=1}^{T} \alpha_{1t} f_t(\boldsymbol{x})$

- Parametric approach to Bayes classifier, but dimension and basis function are flexible
- Each $f_t(\boldsymbol{x})$ itself is a classifier, cf. real AdaBoost.
- The stopping time T can be selected according to the state of learning.

U-Boost

U-empirical loss function

$$L_U^{\text{emp}}(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n}\xi(q_{\boldsymbol{\theta}}(\boldsymbol{x}_i)) + \left\langle U(\xi(q_{\boldsymbol{\theta}}))\right\rangle$$

In a context of classification

$$L_U^{\text{emp}}(F) = -\frac{1}{n}\sum_{i=1}^{n}F(\boldsymbol{x}_i, y_i) + \frac{1}{n}\sum_{i=1}^{n}\sum_{g=1}^{G}U(F(\boldsymbol{x}_i, g))$$

Unnormalized U-loss

$$L_U^{(0)}(F) = \frac{1}{n}\sum_{i=1}^{n}\sum_{g=1}^{G}U(F(\boldsymbol{x}_i, g) - F(\boldsymbol{x}_i, y_i))$$

Normalized U-loss

$$L_U^{(1)}(F) = -\frac{1}{n}\sum_{i=1}^{n}F(\boldsymbol{x}_i, y_i) + \frac{1}{n}\sum_{i=1}^{n}\sum_{g=1}^{G}U(F(\boldsymbol{x}_i, g))$$

$$\text{subject to}\ \sum_{g=1}^{G}u(F(\boldsymbol{x}, g)) = 1$$

U-Boost (binary)

Unnormalized *U*-loss

$$L_U^{(0)}(F) = \frac{1}{n}\sum_{i=1}^{n} U(-y_i F(\boldsymbol{x}_i))$$

Note $\sum_{g=\pm 1} U(F(\boldsymbol{x}_i, g) - F(\boldsymbol{x}_i, y_i)) = U(-y_i F(\boldsymbol{x}_i))$ where $F(\boldsymbol{x}) = \frac{1}{2}\{F(\boldsymbol{x},1) - F(\boldsymbol{x},-1)\}$

Bayes risk consistency

$$\frac{\partial}{\partial F(\boldsymbol{x})} \sum_{y=\pm 1} U(-yF(\boldsymbol{x})) = \frac{\partial}{\partial F(\boldsymbol{x})} \{U(-F(\boldsymbol{x}))p(y=1|\,\boldsymbol{x}) + U(F(\boldsymbol{x}))p(y=-1|\,\boldsymbol{x})\}$$

$$\frac{u(F^*(\boldsymbol{x}))}{u(-F^*(\boldsymbol{x}))} = \frac{p(y=1|\,\boldsymbol{x})}{p(y=-1|\,\boldsymbol{x})} \quad \text{where } F^* = \arg\min_F E\{U(-yF(x))\}$$

$F^*(\boldsymbol{x})$ is Bayes risk consistent because $\frac{\partial}{\partial F}\frac{u(F)}{u(-F)} = \frac{\dot{u}(F)u(-F) + \dot{u}(-F)u(F)}{\{u(-F)\}^2} > 0$

AdaBoost with margin

Eta-loss function

$$L_\eta(F) = (1-\eta)L_{\exp}(F) + \eta L_{\text{naive}}(F),$$

with generator $U(F) = (1-\eta)\exp(F) + \eta F$

$$L_{\exp}(F) = \sum_{i=1}^{N} \exp\left(-y_i F(\boldsymbol{x}_i)\right)$$

$$L_{\text{naive}}(F) = -\sum_{i=1}^{N} y_i F(\boldsymbol{x}_i)$$

regularized

EtaBoost for Mislabels

Expected Eta-loss function

$$L_\eta(F) = (1-\eta)E\{\exp(-yF(\boldsymbol{x})) - \eta\, yF(\boldsymbol{x})\}$$

Optimal score $F^* = \arg\min_F L_\eta(F)$

The variational argument leads to

$$p(y=1\,|\,\boldsymbol{x}) = \frac{(1-\eta)e^{\frac{1}{2}F^*(\boldsymbol{x})} + \eta}{(1-\eta)(e^{\frac{1}{2}F^*(\boldsymbol{x})} + e^{-\frac{1}{2}F^*(\boldsymbol{x})}) + 2\eta}$$

$$= (1-\varepsilon(\boldsymbol{x}))\frac{e^{F^*(\boldsymbol{x})}}{1+e^{F^*(\boldsymbol{x})}} + \varepsilon(\boldsymbol{x})\frac{1}{1+e^{F^*(\boldsymbol{x})}}$$

Mislabel modeling

$$\text{where } \varepsilon(\boldsymbol{x}) = \frac{\eta}{(1-\eta)(e^{\frac{1}{2}F(\boldsymbol{x})} + e^{-\frac{1}{2}F(\boldsymbol{x})}) + 2\eta}$$

EtaBoost

1. Initial settings : $w_1(i) = \frac{1}{n} \quad (i = 1 \cdots n),\ F_0(\boldsymbol{x}) = 0$

2. For $m = 1, \cdots, T$ $\quad \varepsilon_m(f) \propto \sum_{i=1}^{n} \mathrm{I}(y_i \neq f(\boldsymbol{x}_i)) w_m(i),$

(a) $\quad \varepsilon_m(f_{(m)}) = \min_f \varepsilon_m(f)$

(b)

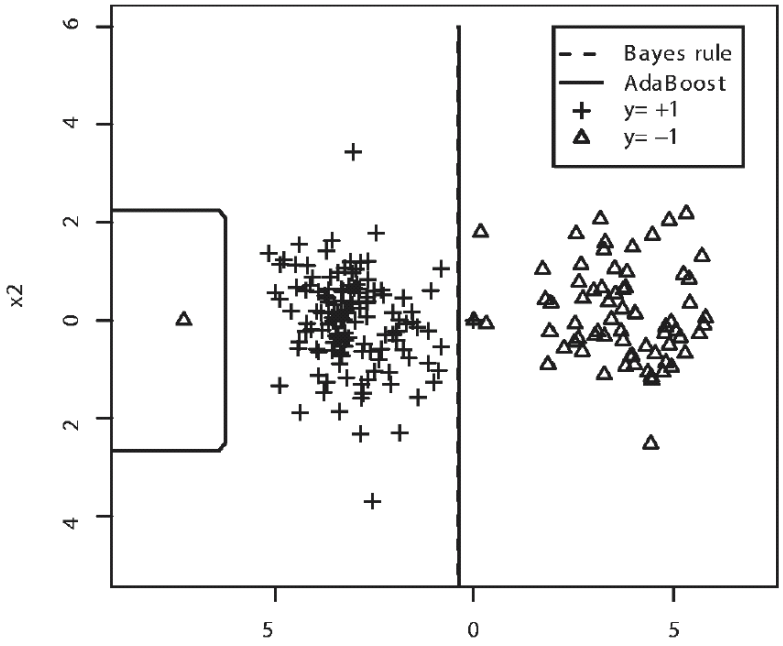
$$\alpha_m^* = \log \frac{\sqrt{1 - \varepsilon_m(f_m) + (\eta K_m)^2} + \eta K_m}{\sqrt{\varepsilon_m(f_m)}},$$

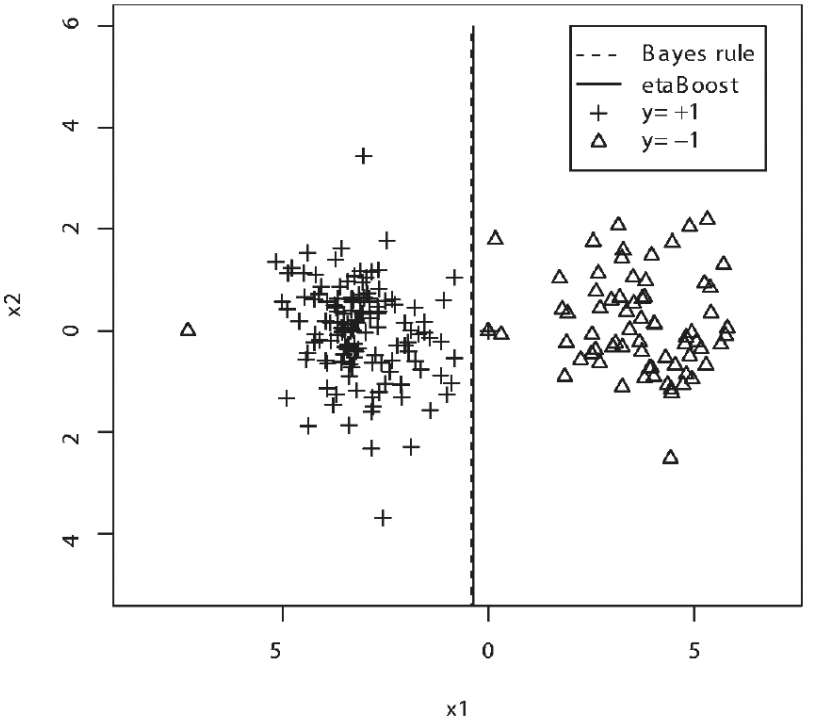
where $\varepsilon_m(f)$ is defined in AdaBoost, $K_m = \frac{(1 - 2\varepsilon_1(f_m))}{2\sqrt{\varepsilon_m(f_m)}} \left(\frac{(1-\eta) Z_m}{N} \right)^{-1}$

(c) $\quad w_{m+1}(i) \propto w_m(i) \exp(-\alpha^*_m f_{(m)}(\boldsymbol{x}_i) y_i)$

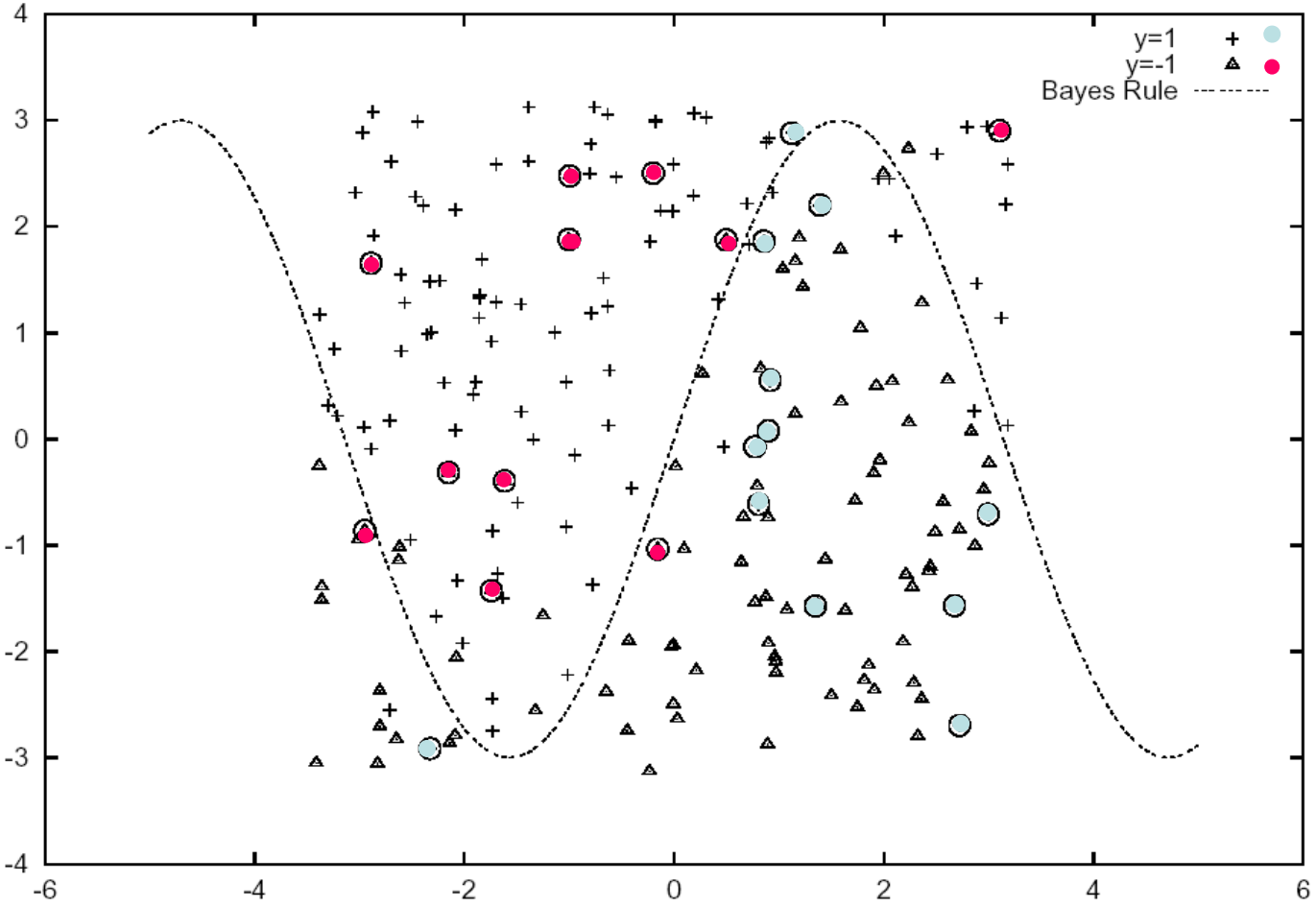
3. $\operatorname{sign}\left(F_T(\boldsymbol{x})\right)$, where $F_T(\boldsymbol{x}) = \sum_{t=1}^{T} \alpha_t f_{(t)}(\boldsymbol{x})$

A toy example

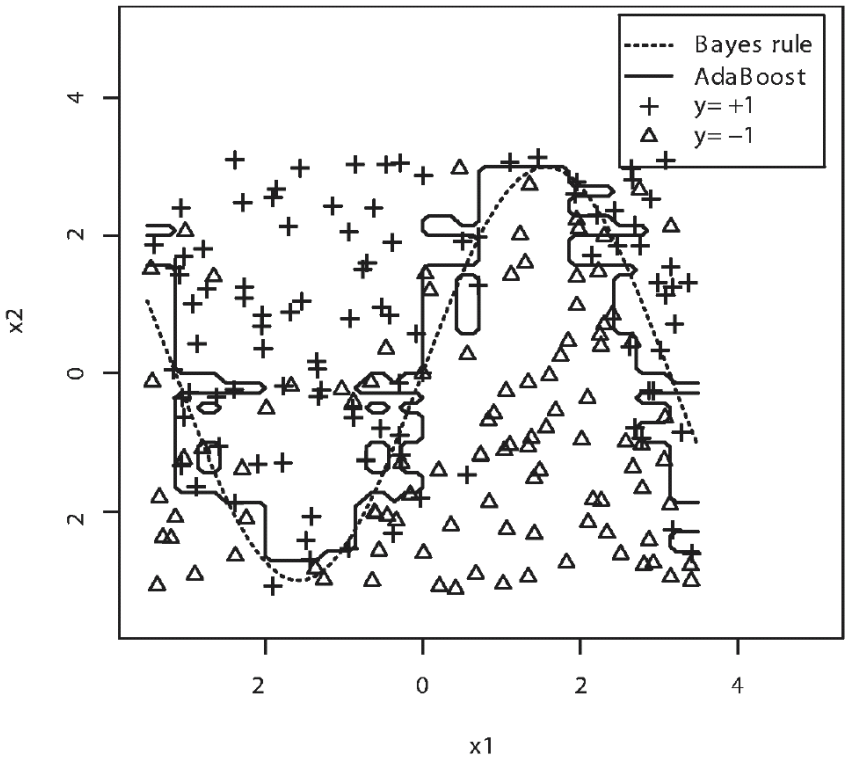




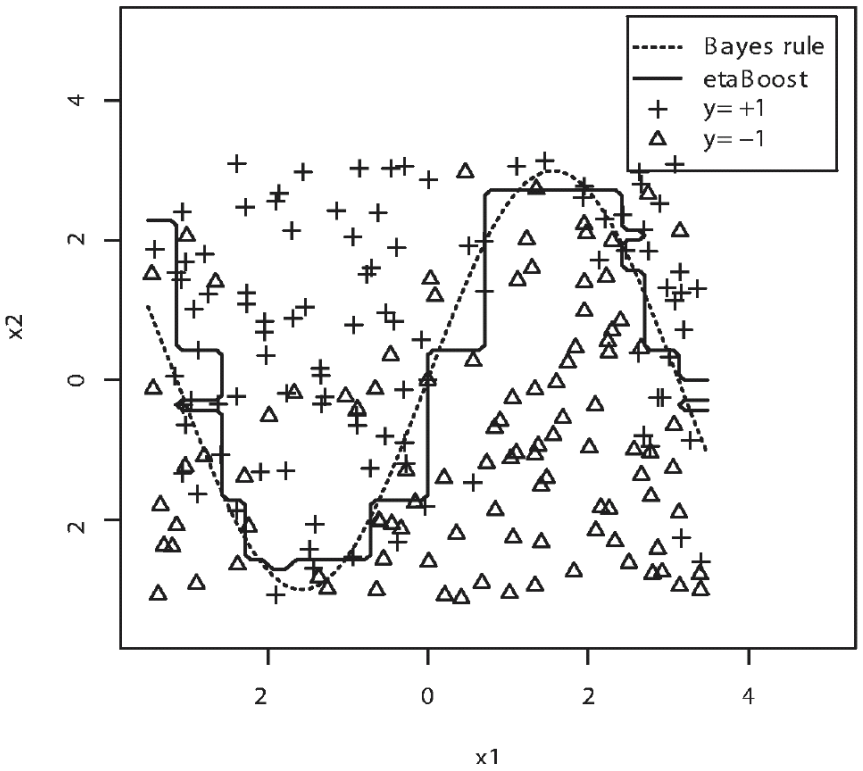
Examples partly mislabeled



AdaBoost vs. EtaBoost



AdaBoost



EtaBoost

ROC curve, AUC

$(\boldsymbol{X}, Y), \boldsymbol{X} \in \mathbb{R}^p, Y \in \{0,1\}$

判別関数 : $F(\boldsymbol{X})$, 閾値：c

$$\begin{cases} F(\boldsymbol{X}) > c \Rightarrow \text{陽性} \ (\mathrm{Y} = 1) \\ F(\boldsymbol{X}) \le c \Rightarrow \text{陰性} \ (\mathrm{Y} = 0) \end{cases}$$

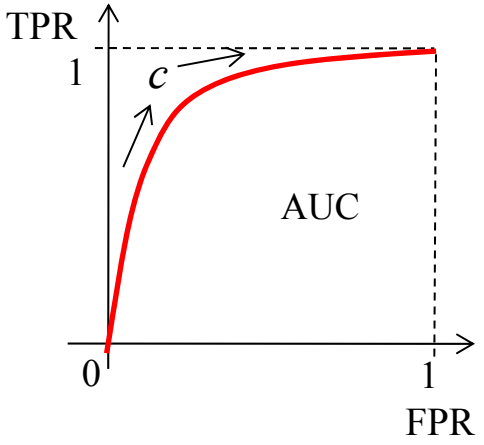
偽陽性確率 $\mathrm{FPR}(c) = P(F(\boldsymbol{X}) > c \mid \mathrm{Y} = 0)$

真の陽性確率 $\mathrm{TPR}(c) = P(F(\boldsymbol{X}) > c \mid \mathrm{Y} = 1)$

受信者動作特性（ROC）曲線 $\mathrm{ROC}(F) = \{(\mathrm{FPR}(c), \mathrm{TPR}(c)) \mid c \in \mathbb{R}\}$

Cf. Bamber (1975), Pepe et al. (2003)

下側面積（AUC） $\mathrm{AUC}(F, \alpha) = \int_{\infty}^{-\infty} \mathrm{TPR}(c')\, d\,\mathrm{FPR}(c')$



Approximate AUC

Using probability density function g_0 and g_1 for class 0 and class 1, AUC is given as

$$\mathrm{AUC}(F) = \int \int \mathrm{H}(F(\boldsymbol{x}_1) - F(\boldsymbol{x}_0)) g_0(\boldsymbol{x}_0) g_1(\boldsymbol{x}_1) d\boldsymbol{x}_0 d\boldsymbol{x}_1,$$

where H is the Heaviside function. Similarly, the approximated AUC is given as

$$\mathrm{AUC}_\sigma(F) = \int \int \mathrm{H}_\sigma(F(\boldsymbol{x}_1) - F(\boldsymbol{x}_0)) g_0(\boldsymbol{x}_0) g_1(\boldsymbol{x}_1) d\boldsymbol{x}_0 d\boldsymbol{x}_1,$$

where $\mathrm{H}_\sigma(x) = \Phi(\frac{x}{\sigma})$, with Φ being the standard normal distribution function.

Relationship between AUC approximate AUC

Theorem 1.

Let

$$\Psi(c) = \mathrm{AUC}_\sigma\big(F + c\, m(\Lambda)\big),$$

where $\Lambda(x) = g_1(x)/g_0(x)$, and m is a strictly increasing function. Then, $\Psi(c)$ is a strictly increasing function of $c \in \mathcal{R}$, and

$$\sup_F \mathrm{AUC}_\sigma(F) = \lim_{c\to\infty} \Psi(c) = \mathrm{AUC}(\Lambda)..$$

Proof.

$$\frac{\partial}{\partial c}\Psi(c) = \frac{1}{2}\int\int \big(\zeta(x_1) - \zeta(x_0)\big)\mathrm{H}'_\sigma\big(F(x_1) + c\,\zeta(x_1) - F(x_0) - c\,\zeta(x_0)\big) g_0(x_0) g_0(x_1)\big(\Lambda(x_1) - \Lambda(x_0)\big) dx_0 dx_1$$

where, $\zeta(x) = m(\Lambda(x))$. Then, we have

$$\begin{aligned}
\mathrm{AUC}_\sigma(F) &< \lim_{c\to\infty} \Psi(c) \\
&= \lim_{c\to\infty} \mathrm{AUC}_\sigma\left[c\left\{\frac{F}{c} + \zeta\right\}\right] \\
&= \lim_{c\to\infty} \mathrm{AUC}_{\frac{\sigma}{c}}\left(\frac{F}{c} + \zeta\right) \\
&= \mathrm{AUC}(\zeta) \\
&= \mathrm{AUC}(\Lambda),
\end{aligned}$$

Objective function of AUCBoost

At first we prepare a set $\mathcal{F}$, from which we choose weak classifiers to construct $F(\boldsymbol{x})$:

$$\mathcal{F} = \bigcup_{k=1}^{p} \{ f(\boldsymbol{x}) = a\mathrm{H}(x_k - b_k) + (1-a)/2 \mid a \in \{-1, 1\},\ b_k \in \mathcal{B}_k \},$$

Then, the objective function we propose is:

$$\overline{\mathrm{AUC}}_{\sigma,\lambda}(F) = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \mathrm{H}_\sigma(F(\boldsymbol{x}_{1j}) - F(\boldsymbol{x}_{0i})) - \lambda \sum_{k=1}^{p} \sum_{x_k \in \mathcal{B}_k} \left\{ F_k^{(2)}(x_k) \right\}^2,$$

where λ is a smoothing parameter and $F_k^{(2)}$ denotes the second-order difference of F_k.

Special relation between smoothing parameter (λ) and scale parameter (σ)

Without loss of generality, the objective function is equivalent to:

Objective function of AUCBoost

$$\overline{\mathrm{AUC}}_{\lambda}(F) = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \Phi(F(\boldsymbol{x}_{1j}) - F(\boldsymbol{x}_{0i})) - \lambda \sum_{k=1}^{p} \sum_{x_k \in \mathcal{B}_k} \left\{ F_k^{(2)}(x_k) \right\}^2$$

It is because the objective function is rewritten as:

$$\overline{\mathrm{AUC}}_{\sigma,\lambda}(F) = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \mathrm{H}_{\sigma}(F(\boldsymbol{x}_{1j}) - F(\boldsymbol{x}_{0i})) - \lambda\sigma^2 \sum_{k=1}^{p} \sum_{x_k \in \mathcal{B}_k} \left\{ \frac{F_k^{(2)}(x_k)}{\sigma} \right\}^2.$$

Hence we have

$$\overline{\mathrm{AUC}}_{\sigma,\lambda}(F) = \overline{\mathrm{AUC}}_{\sigma',\lambda'}\left(\frac{\sigma'}{\sigma} F\right),$$

where $\lambda\sigma^2 = \lambda'\sigma'^2$. This implies that the maximization of $\overline{\mathrm{AUC}}_{\sigma,\lambda}(F)$ is equivalent to that of $\overline{\mathrm{AUC}}_{1,\lambda\sigma^2}\left(\frac{F}{\sigma}\right)$. Therefore, we have

$$\max_{\sigma,\lambda,F} \overline{\mathrm{AUC}}_{\sigma,\lambda}(F) = \max_{\lambda,F} \overline{\mathrm{AUC}}_{1,\lambda}(F).$$

AUCBoost algorithm

1. Start with a score function $F_0(\boldsymbol{x})$.
2. For $t = 1, ..., T$
   a. Find the best weak classifier f_t and calculate the coefficient α_t as

$$f_t(\boldsymbol{x}) = \operatorname*{argmax}_{f \in \mathcal{F}} \frac{\partial}{\partial \alpha} \overline{\mathrm{AUC}}_\lambda(F_{t-1} + \alpha f)\Big|_{\alpha=0},$$

$$\alpha_t = \operatorname*{argmax}_{\alpha>0} \overline{\mathrm{AUC}}_\lambda(F_{t-1} + \alpha f_t).$$

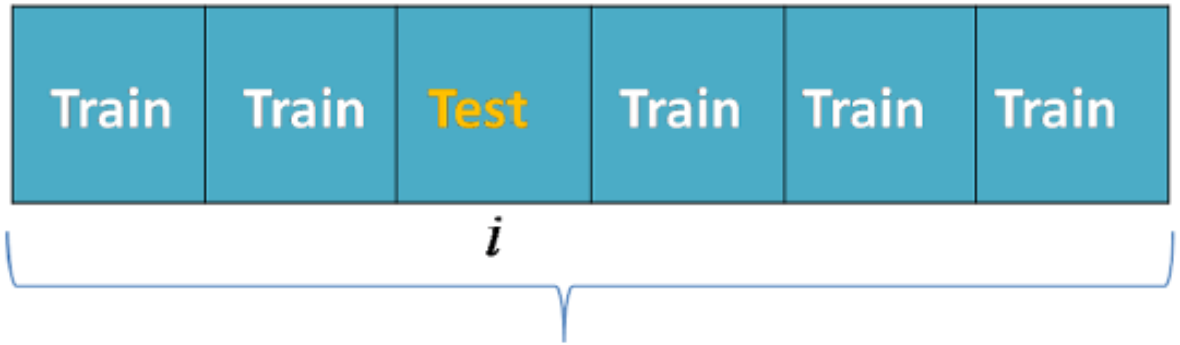
   b. Update the score function as

$$F_t(\boldsymbol{x}) = F_{t-1}(\boldsymbol{x}) + \alpha_t f_t(\boldsymbol{x}).$$

3. Finally, output the final score function

$$F(\boldsymbol{x}) = F_0(\boldsymbol{x}) + \sum_{t=1}^{T} \alpha_t f_t(\boldsymbol{x})$$

V-fold cross validation

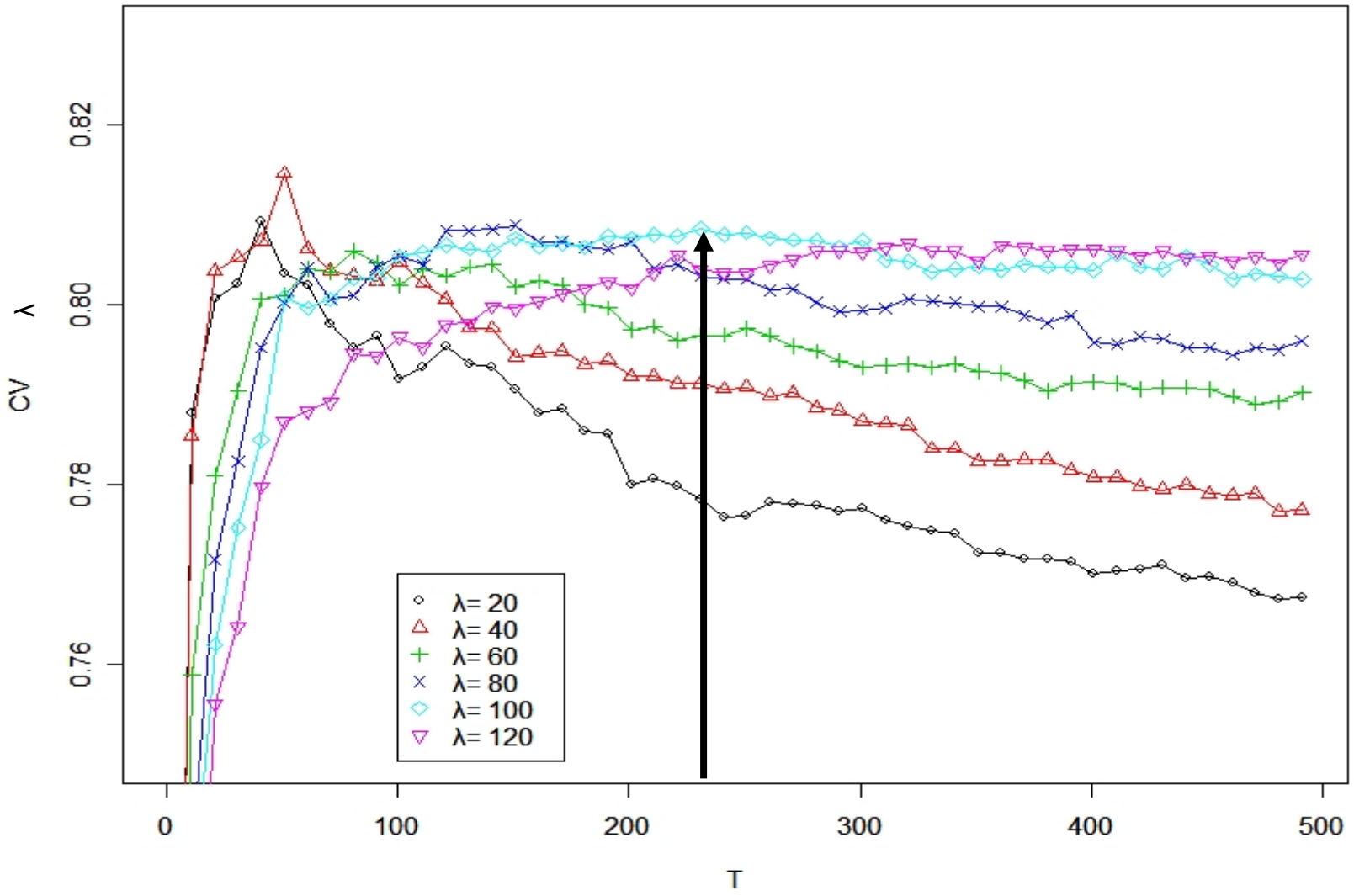


$$\mathrm{AUC}_{cv}(\lambda, T) = \frac{1}{V}\sum_{i=1}^{V} \overline{\mathrm{AUC}}_{\lambda}^{(i)}(F^{(-i)})$$

$\overline{\mathrm{AUC}}_{\lambda}^{(i)}$: $\overline{\mathrm{AUC}}_{\lambda}$ calculated based on only the i-th part of the data

$F^{(-i)}$: score function constructed with the i-th part of the data removed

5-fold CVの結果（AUCBoost）



Score plot

Score plots are useful graphical tools for representing the contribution of each marker to a total score function $F(x)$. When $F(x)$ is decomposed as

Definition

$$F(x) = \sum_{k=1}^{p} F_k(x_k), \ x\ (= (x_1, ..., x_p)) \in \mathbf{R}^p$$

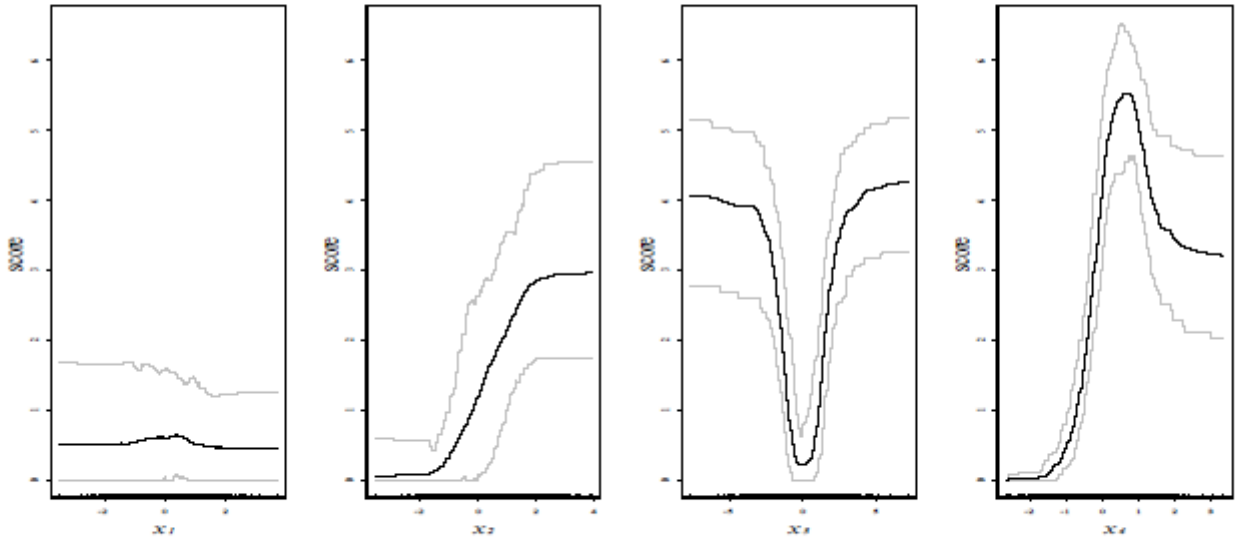
.

Each plot of $F_k(x_k)$ against x_k is called a score plot. See Friedman *and others* (2000) and Kawakita *and others* (2005).

Example of score plots under normality assumption ($X_0 \sim \mathrm{N}(\mu_0, \Sigma_0),\ X_1 \sim \mathrm{N}(\mu_1, \Sigma_1)$)

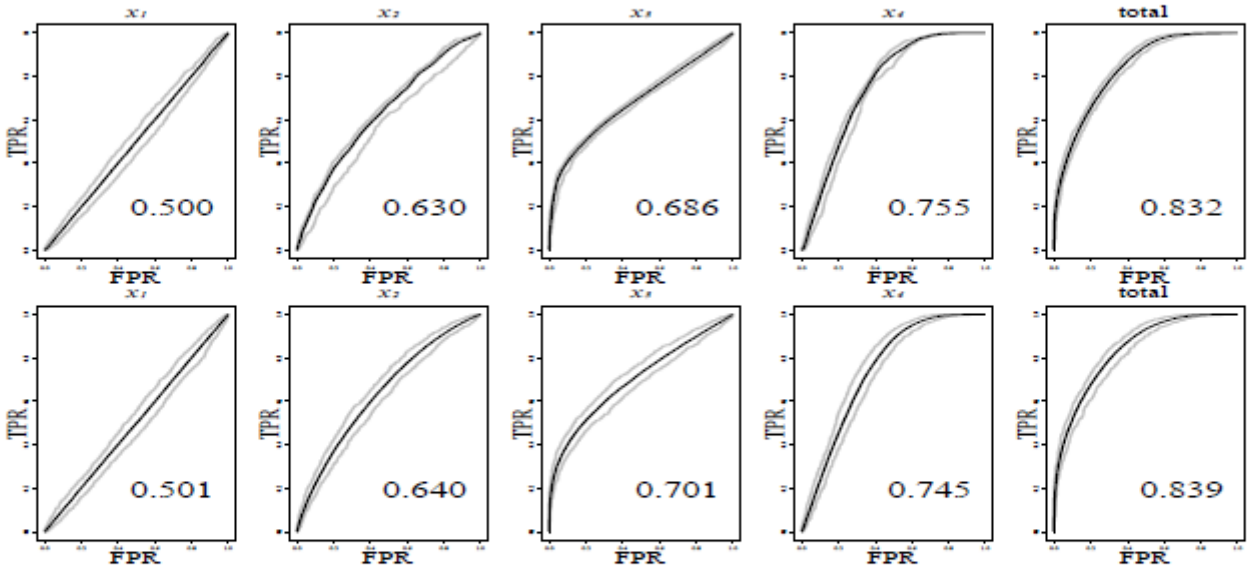
$\mu_0 = (0, 0, 0, 0)'$, $\mu_1 = (0, 0.5, 0, 0.5)'$, $\Sigma_0 = \mathrm{diag}(1, 1, 1, 1)$ and

$\Sigma_1 = \mathrm{diag}(1, 1, 4, 0.25)$



Score plots for AUCBoost. The black lines indicate mean score plots and the gray lines indicate the 95 percent confidence bands.

Corresponding score AUCs under normality assumption ($X_0 \sim \mathrm{N}(\mu_0, \Sigma_0),\ X_1 \sim \mathrm{N}(\mu_1, \Sigma_1)$)



Score ROC curves and the ROC curve constructed by AUCBoost (upper panels) and by $\hat{F}_{\mathrm{N}}$ (lower panels). The black lines indicate mean ROC curves and the gray lines indicate the 95 percent confidence bands. The corresponding mean values of AUCs are also calculated.

2標本問題とAUCブーストの関係

AUCブーストの学習アルゴリズムは逐次的に各 t-ステップで
得られた判別関数 $F_t(x)$ のウィルコクソン検定統計量

$$\hat{C}(F_t) = \frac{1}{n_0\, n_1} \sum_{k=1}^{n_1} \sum_{i=1}^{n_0} I(F_t(\boldsymbol{x}_{1k}) \ > F_t(\boldsymbol{x}_{0i}) \ )$$

を$(t+1)$-ステップ

$$F_{t+1}(\boldsymbol{x}) = F_t(\boldsymbol{x}) + \alpha_{t+1} f_{t+1}(\boldsymbol{x})$$

で最大方向に増大するよう設計されている.
このように

$$\cdots \ < \hat{C}(F_t) < \hat{C}(F_{t+1}) < \ \cdots$$

となる中で全ての単点解析が統合されている.
この意味で全ての単点解析はブースティングに
組むことは可能である. Eguchi-Komori (2009)

The partial area under the ROC curve

Setting

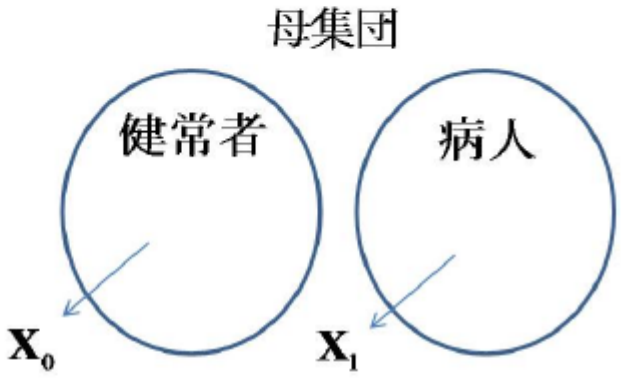
$\boldsymbol{x} \in \mathbf{R}^p$: marker vector
$y \in \{0, 1\}$: class label
$F(\boldsymbol{x})$: score function, c: threshold

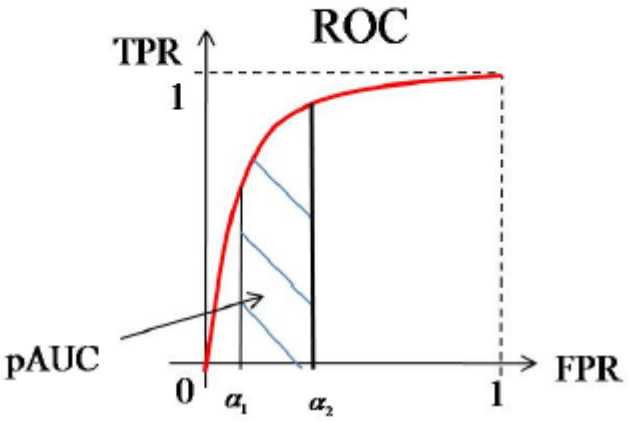
$$\begin{cases} F(\boldsymbol{x}) > c \Rightarrow positive \\ F(\boldsymbol{x}) \leq c \Rightarrow negative \end{cases}$$

$$\mathrm{FPR}(c) = P(F(\boldsymbol{X}) > c \,|y = 0)$$
$$\mathrm{TPR}(c) = P(F(\boldsymbol{X}) > c \,|y = 1)$$

$$\Downarrow$$

$$\mathrm{ROC} = \{(\mathrm{FPR}(c), \mathrm{TPR}(c)) \,|c \in \mathbf{R}\}$$





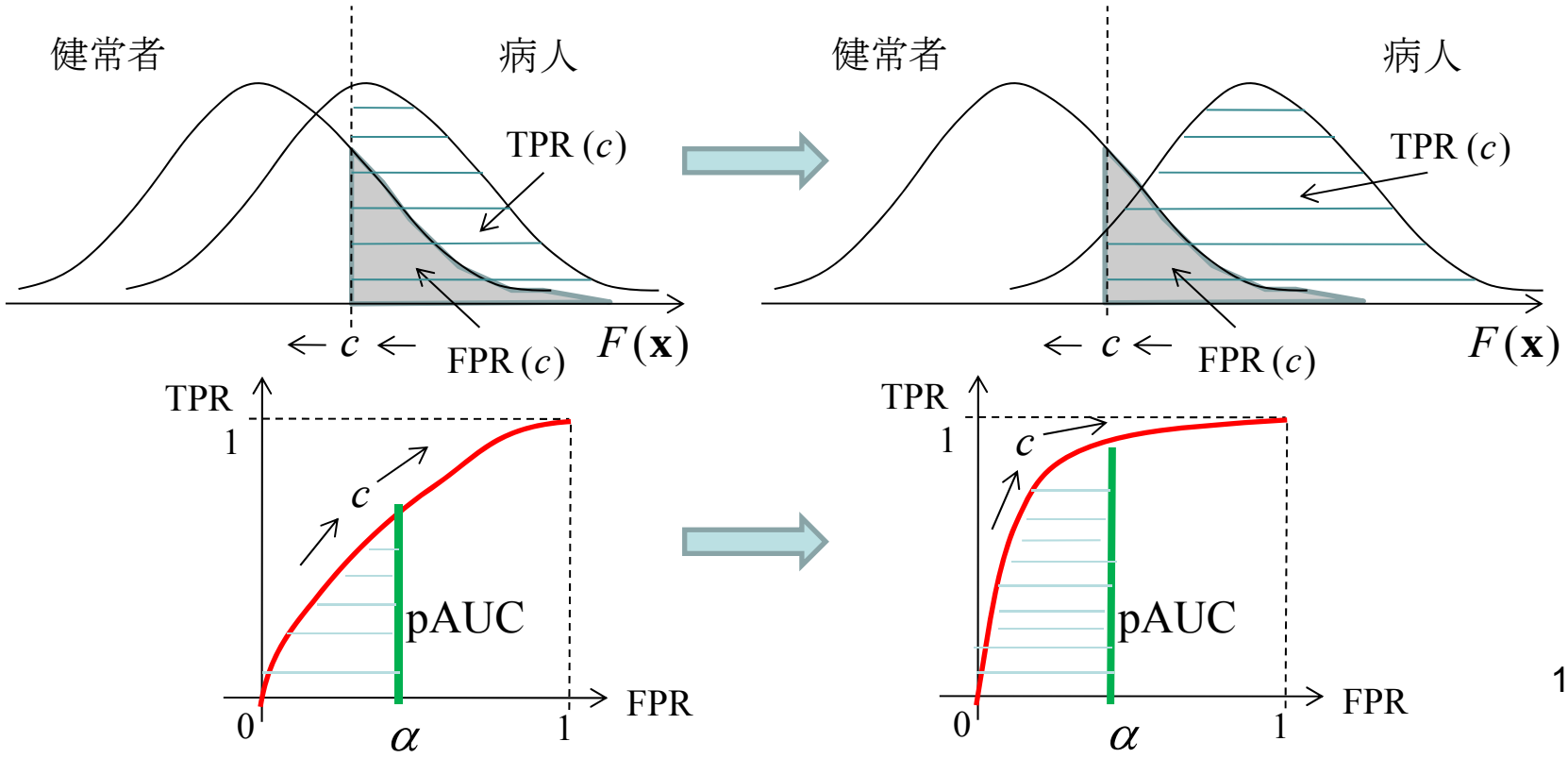
From AUC to pAUC

- The AUC has been severely criticized for inconsistencies arising between the statistical significance and the corresponding clinical significance (Cook, 2007).
- Recently, Pencina *and others* (2008) proposed a criterion termed integrated discriminant improvement, and showed its advantage over the AUC in the assessment of a new marker.

Partial AUC

$$\mathrm{pAUC} = \int_{c}^{-\infty} \mathrm{TPR}(c)\, d\,\mathrm{FPR}(c)$$

$$= \mathrm{P}(F(\mathrm{X}_0) < F(\mathrm{X}_1), c \le F(X_0)\ )$$



pAUCBoost algorithm

1. Set the initial value of score function as $F_0(\boldsymbol{x}) = 0$, and the absolute value of the initial coefficient as $|\beta_0(f)| = 1$.

2. For $t = 1, ..., T$,
   a. For all f's $\in \mathcal{F}$, calculate the values of thresholds $\overline{c}_1$ and $\overline{c}_2$ of $F_{t-1} + \beta_{t-1}(f)f$
   b. Update the coefficient with one-step Newton-Raphson iteration:

$$\beta_{t-1}(f) \rightarrow \beta_t(f)$$

   c. Find the best weak classifier f_t.

$$f_t = \underset{f \in \mathcal{F}}{\operatorname{argmax}} \, \overline{\mathrm{pAUC}}_{\lambda}(F_{t-1} + \beta_t(f)f, \alpha_1, \alpha_2)$$

   d. Update the score function as.

$$F_t(\boldsymbol{x}) = F_{t-1}(\boldsymbol{x}) + \beta_t(f_t)f_t(\boldsymbol{x}).$$

3. Finally, output a final score function as

Robust for noninformative genes

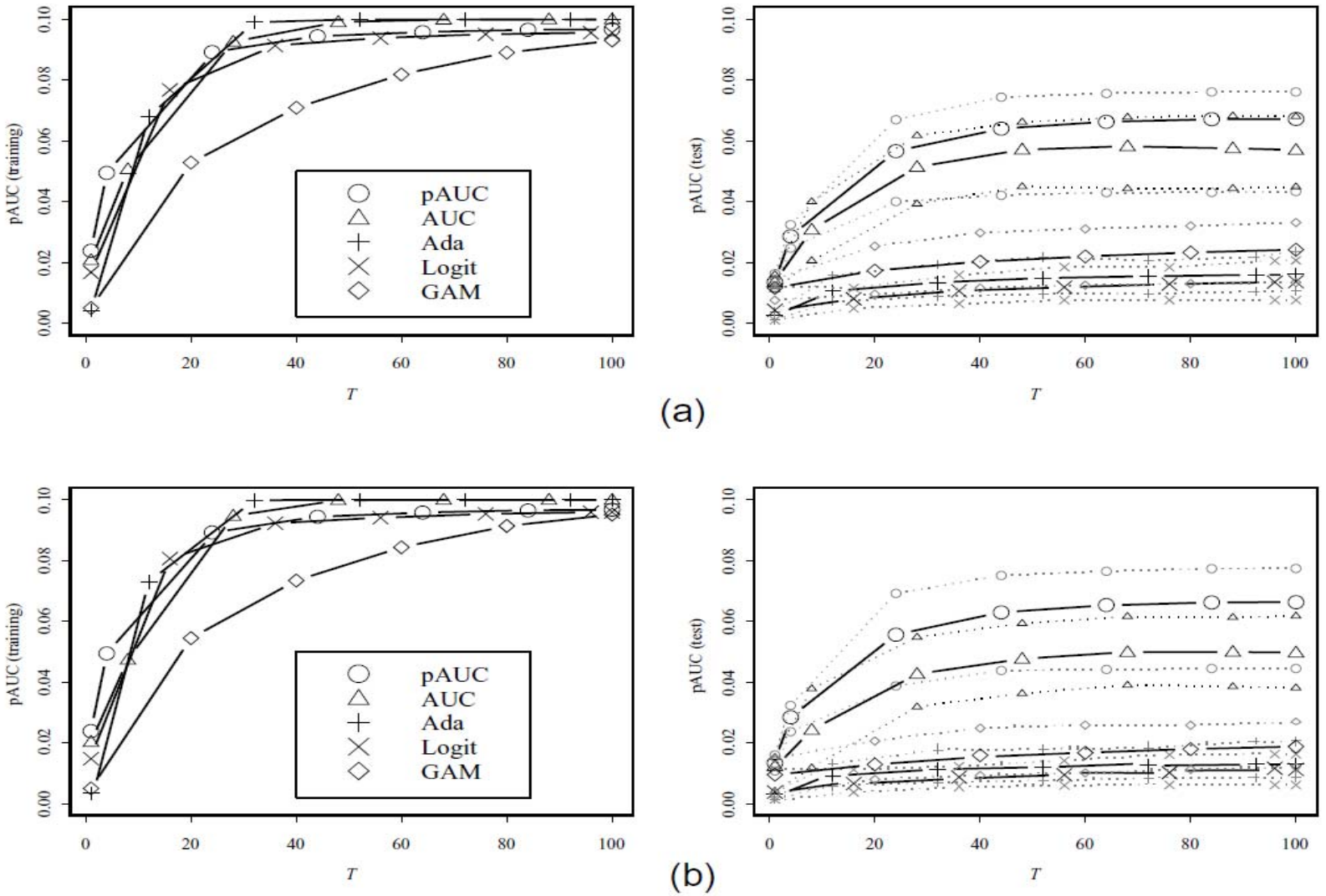


Figure 3 Results of simulation study based on the values of the pAUC. (a) The results of the pAUC with FPR between 0 and 0.1 for training data (left panel) and test data (right panel) with only informative genes. The gray dashed lines indicate the 95% confidence bands. (b) the results of the pAUC with noninformative genes added.

Robustness for noninformative genes

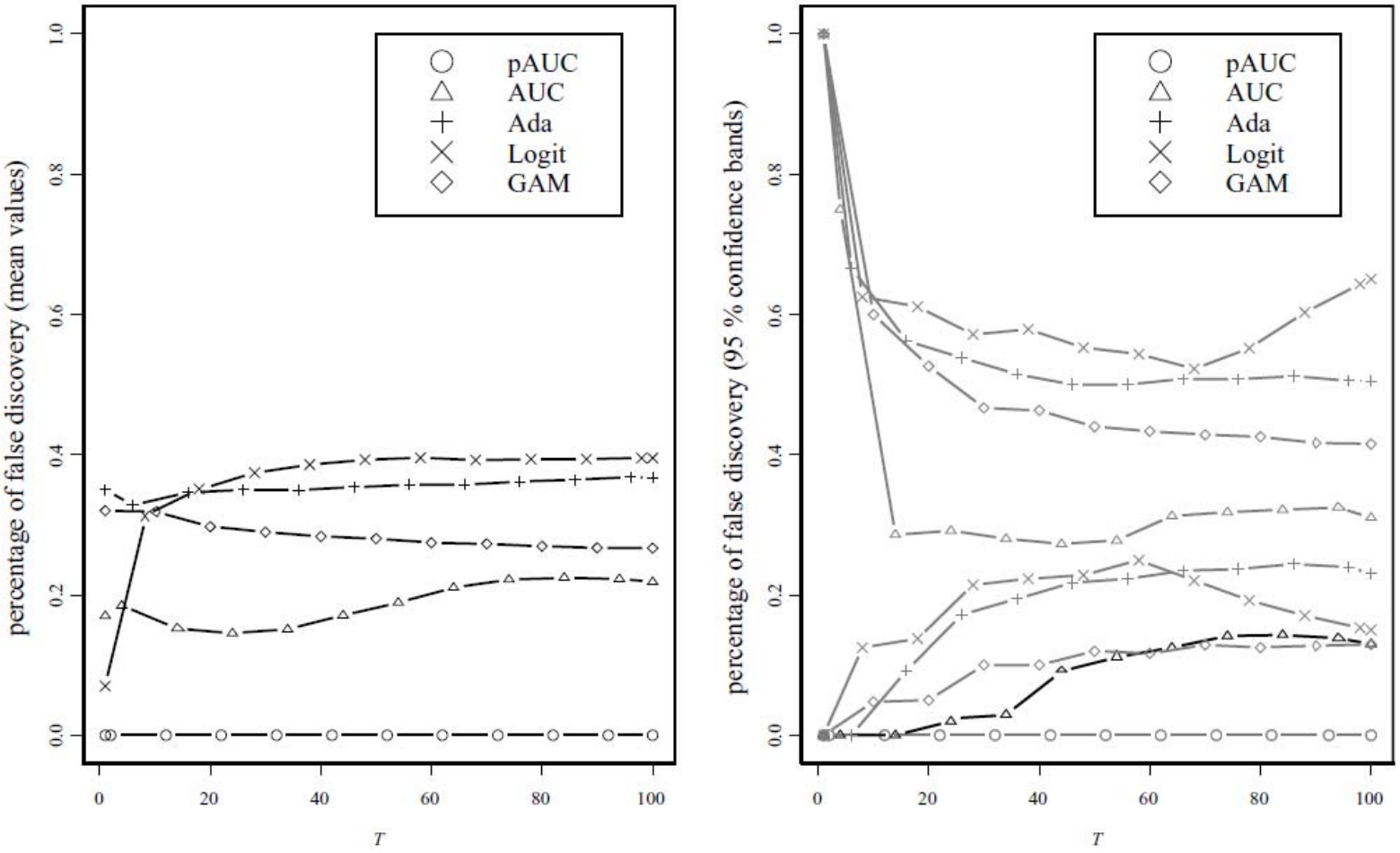


Figure 4 Results of simulation study based on the marker selection. The mean values of percentage of false discovery (left panel) and 95% confidence bands (right panel) for each boosting method. The horizontal axis denotes the iteration number of *T*. The lower sides of the 95% confidence bands of AUCBoost are shown by the heavy black line to emphasize the difference from those of pAUCBoost.

LETTER

An Extension of the Receiver Operating Characteristic Curve and AUC-Optimal Classification

Takashi Takenouchi　Osamu Komori　Shinto Eguchi

Neural Computation **24**, 2789–2824 (2012)

生成関数　$U(0) = 0, \quad U(z) \leq 1, \quad \lim_{z \to \infty} U(z) = 1.$

真のU陽性確率　$\mathrm{TPU}_U(F, c) = \int_x U(F(x) - c) p_{+1}(x)\, \mathrm{d}x,$

$$\mathrm{A}_U(F) = \int_{c:\infty}^{-\infty} \mathrm{TPU}_U(F, c) \mathrm{dFPR}(F, c)$$
$$= \int_x \int_{x'} U(F(x) - F(x')) p_{+1}(x) p_{-1}(x')\, dx dx'$$

$$\tilde{\mathrm{A}}_U(F) = \frac{1}{n_{+1} n_{-1}} \sum_{i:y_i=+1} \sum_{j:y_j=-1} U(F(x_i) - F(x_j))$$

Boost learning algorithm

1. Initialize a weight for a pair of examples as $w_1(i, j) = \frac{1}{n_{+1} n_{-1}}$
2. For $t = 1, \ldots, T$,
   a. Find a base classifier as

$$f_t(x) = \operatorname*{argmin}_{f \in \mathcal{F}} \left(\mathrm{err}_t(f) - \mathrm{acc}_t(f) \right),$$

$$\mathrm{acc}_t(f) = \sum_{i:y_i=+1} \sum_{j:y_j=-1} w_t(i, j) \mathrm{I}(f(x_i) = +1) \mathrm{I}(f(x_j) = -1),$$

$$\mathrm{err}_t(f) = \sum_{i:y_i=+1} \sum_{j:y_j=-1} w_t(i, j) \mathrm{I}(f(x_i) = -1) \mathrm{I}(f(x_j) = +1).$$

   b. $$\alpha_t = \operatorname*{argmax}_{\alpha'} \tilde{\mathrm{A}}_U(F_{t-1} + \alpha' f_t).$$

   c. Update the weight as

$$w_{t+1}(i, j) = \frac{U'(F_t(x_i) - F_t(x_j))}{Z_t},$$

3. Output the discriminant function $F_T(x)$.

Bayes risk consistency

Let $F^* = \underset{F}{\operatorname{argmax}} A_U(F)$.

Then the likelihood ratio is

$$\frac{p_{+1}(x)}{p_{-1}(x)} = exp(2\Psi_{F^*}(F^*(x))) \quad \text{where} \quad \Psi_F(z) = \frac{1}{2} log \frac{\int U'(F(x') - z) p_{+1}(x')\, dx'}{\int U'(z - F(x')) p_{-1}(x')\, dx'}.$$

Remark. The conditinal probabilit y of $Y = y$ given $\boldsymbol{x}$ is in logistic model

$$p(y|x) = \frac{1}{1 + \exp\left(-2y\left(\Psi_{F^*}(F^*(x)) + \log\sqrt{\frac{\pi_{+1}}{\pi_{-1}}}\right)\right)}$$

ベンチマークデータ

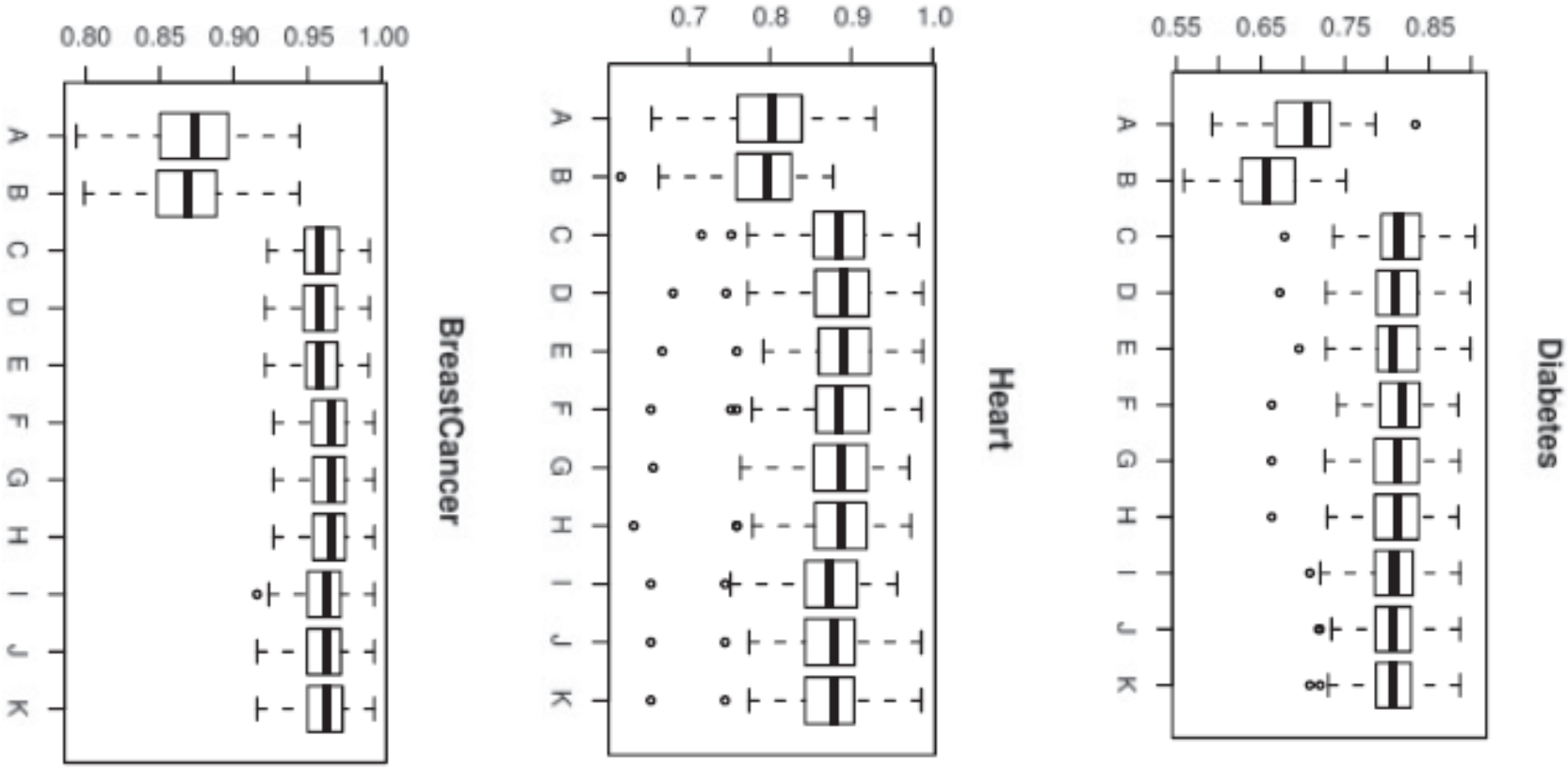
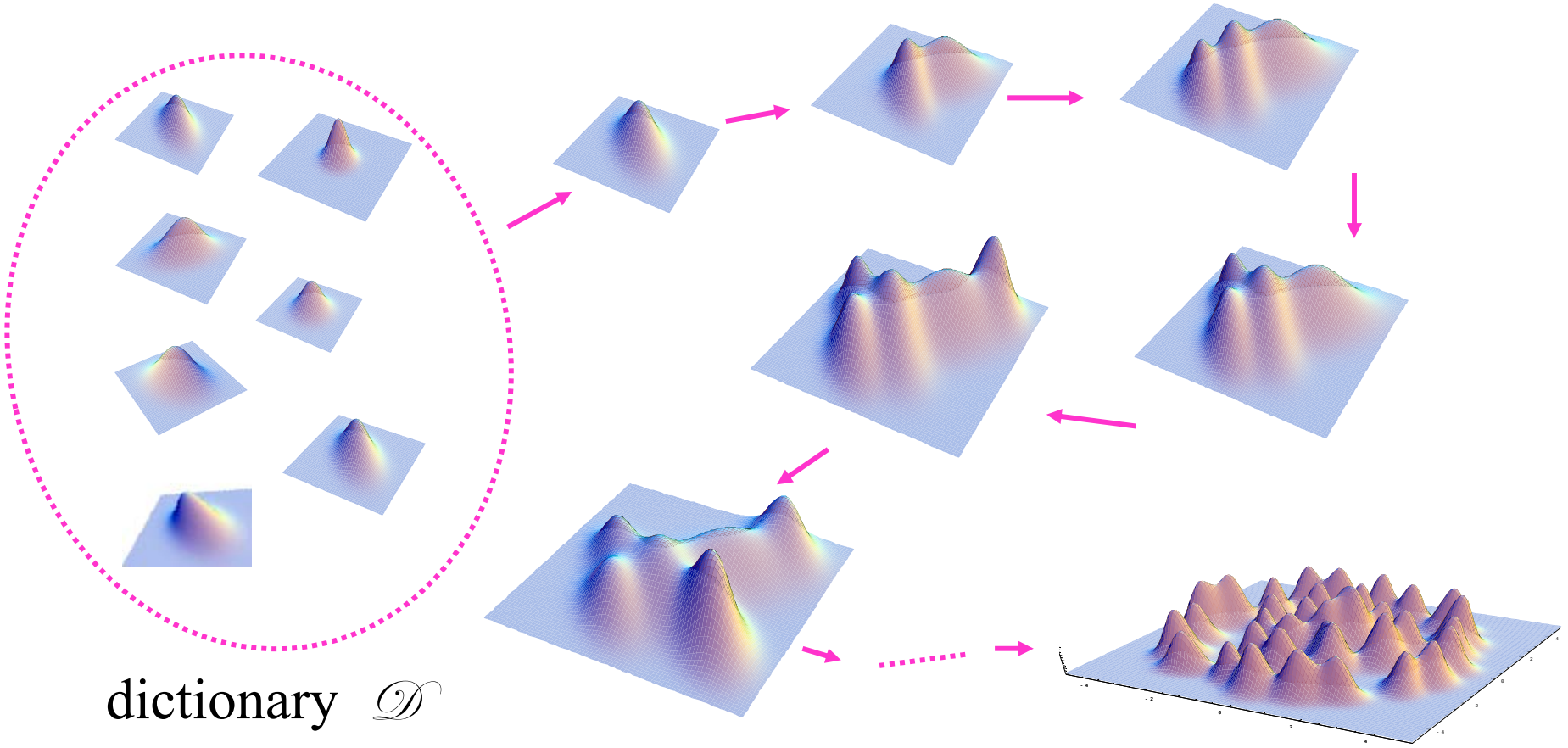


Figure 2: Box plots of AUC values for each data set. (A) SVM with RBF kernel. (B) SVM with linear kernel. (C) Boosting with U_{rank}. (D) Boosting with U_{logit}. (E) Boosting with U_{mada}. (F) U-AUCBoost with U_{rank}. (G) U-AUCBoost with U_{logit}. (H) U-AUCBoost with U_{mada}. (I) pU-AUCBoost with U_{rank}. (J) pU-AUCBoost with U_{logit}. (K) pU-AUCBoost with U_{mada}.

Mach Learn (2013) 90:29–57
DOI 10.1007/s10994-012-5298-3

Density estimation with minimization of U-divergence

Kanta Naito · Shinto Eguchi



U-Boost learning

U-loss function $L_U(f) = -\frac{1}{n}\sum_{i=1}^{n} \xi(f(\boldsymbol{x}_i)) + \int_{\mathbb{R}^p} U(\xi(f(\boldsymbol{x})))\mathrm{d}\boldsymbol{x}$

Dictionary of density functions

$$\mathscr{D} = \{\, g_\lambda(\boldsymbol{x}) : g_\lambda(\boldsymbol{x}) \geq 0,\ \int g_\lambda(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1,\ \ \lambda \in \Lambda\}$$

Learning space = U-model

$$\mathscr{D}_U{}^* = \xi^{-1}\big(\mathrm{co}(\xi(\mathscr{D}))\big) = \{\, \xi^{-1}\big(\sum_{\lambda \in \Lambda} \pi_\lambda\, \xi(g_\lambda(\boldsymbol{x}))\big)\}$$

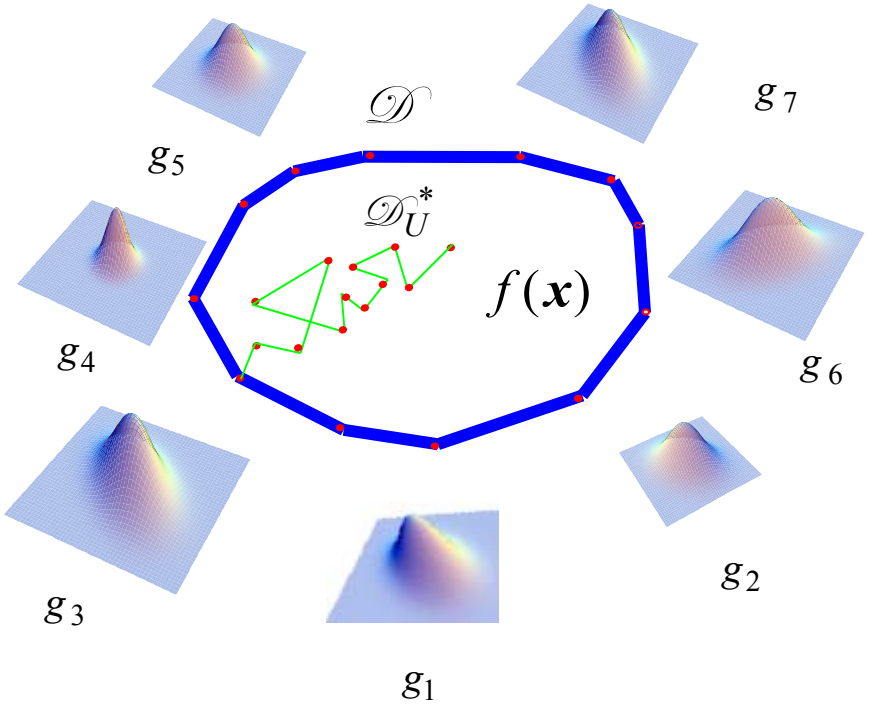
- Let $f(\boldsymbol{x}, \boldsymbol{\pi}) = \xi^{-1}\big(\sum_{\lambda \in \Lambda} \pi_\lambda\, \xi(g_\lambda(\boldsymbol{x}))\big)$. Then $f(\boldsymbol{x}, \underset{(\lambda)}{(0,\cdots,1,\cdots 0)}) = g_\lambda(\boldsymbol{x})$
- $\mathscr{D}_U{}^* \supseteq \mathscr{D}$

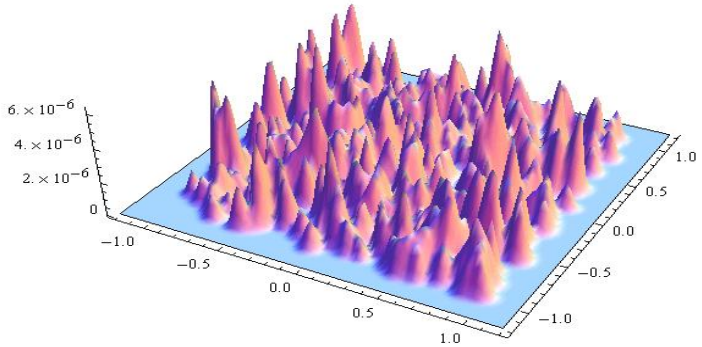
Goal : find $f^* = \underset{f \in \mathscr{D}_U{}^*}{\mathrm{argmin}}\, L_U(f)$

Inner step in the convex hull

$$\mathscr{D} = \{\phi(x,\lambda) : \lambda \in \Lambda\} \longrightarrow \mathscr{D}_U^{\,*} = \xi^{-1}(\mathrm{co}\,\xi(\mathscr{D}))$$

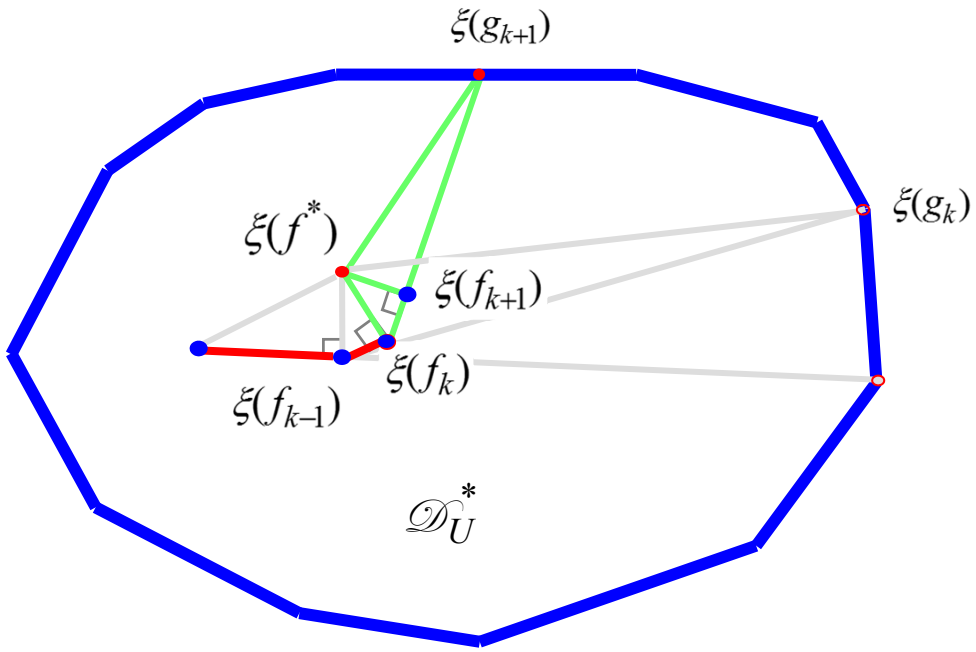
Goal : $f^* = \underset{f \in \mathscr{D}_U^{\,*}}{\mathrm{argmin}}\, L_U(f)$ $\quad f^*(\boldsymbol{x}) = \xi^{-1}(\hat{\pi}_1 \xi(\hat{f}_1(\boldsymbol{x})) + \cdots + \hat{\pi}_{\hat{k}} \xi(\hat{f}_{\hat{k}}(\boldsymbol{x})))$





$f(\boldsymbol{x})$

Convergence of Algorithm



$$L_U(f_k) = L_U(f_{k+1}) + D_U(f_k, f_{k+1})$$

$$\xi(f_1) \to \cdots \to \xi(f_{k-1}) \to \xi(f_k) \to \xi(f_{k+1}) \to \cdots \to \xi(f^*)$$

$$L_U(f_1) = L_U(f_{k+1}) + \sum_{j=1}^{k} D_U(f_j, f_{j+1})$$

ξ-mixture

$$\xi(f_K) = (1-\alpha_{K-1})\xi(f_{K-1}) + \alpha_{K-1}\xi(g_K)$$

$$= \pi_1\xi(g_1) + \cdots\cdots + \pi_K\xi(g_K) \;\to\; \xi(f^*)$$

Simplified boosting

$$\hat{f}_K = \xi^{-1}\Big(\sum_{k=1}^{K} \pi_k \xi(g_k)\Big)$$

For $k = 0, \cdots, K-1$,

$$f_k \longrightarrow f_{k+1} = \xi^{-1}((1-\alpha_{k+1})\xi(f_k) + \alpha_{k+1}\xi(g_{k+1}))$$

such that

$$g_{k+1} = \arg\min_{g \in \mathscr{D}} L_U(\xi^{-1}((1-\alpha)\xi(f_k) + \alpha\xi(g))),$$

$$\alpha_{k+1} = \frac{c}{k+c}$$

with the initial $f_1 = \arg\min_{g \in \mathscr{D}} L_U(g)$

Non-asymptotic bound

Theorem. Assume that a data distribution has a density $g(\boldsymbol{x})$ and that

$$\text{(A)} \qquad \sup_{(\psi,\phi,\varphi)\in \mathrm{co}(\xi(\mathscr{D}))\times\mathscr{D}\times\mathscr{D}} \int \ddot{U}(\psi)\{\xi(\phi)-\xi(\varphi)\}^2 \le b_U$$

Then we have

$$\mathbb{E}_g D_U(g,\hat{f}_K) \le \mathrm{FA}(g,\mathscr{D}_U{}^*) + \mathrm{EE}(g,\mathscr{D}) + \mathrm{IE}(K,\mathscr{D}),$$

where

$$\mathrm{FA}(g,\mathscr{D}_U{}^*) = \inf_{f\in\mathscr{D}_U{}^*} D_U(g,f) \qquad \text{(Functional approximation)}$$

$$\mathrm{EE}(g,\mathscr{D}) = 2\,\mathbb{E}_g\Big\{\sup_{f\in\mathscr{D}} \Big|\frac{1}{n}\sum_{i=1}^{n}\xi(f(\boldsymbol{x}_i)) - \mathbb{E}_p\xi(f)\Big|\Big\} \qquad \text{(Estimation error)}$$

$$\mathrm{IE}(K,\mathscr{D}) = \frac{c^2 b_U{}^2}{K+c-1} \quad (c:\text{step-lengthconstant}) \qquad \text{(Iteration effect)}$$

Remark. Trade between $\mathrm{FA}(p,\mathscr{D}_U{}^*)$ and $\mathrm{EE}(p,\mathscr{D})$

Lemmas

Lemma 1 Let $\tilde{f}$ be a density estimator and $f^* = \xi^{-1}(\sum_{j=1}^{N} p_j \xi(\phi_j(\boldsymbol{x})))$. Then

$$\sum_{j=1}^{N} p_j L_U(\xi^{-1}((1-\pi)\xi(\tilde{f}(\boldsymbol{x})) + \pi\xi(\phi_j(\boldsymbol{x})))) - L_U(f^*) \leq (1-\pi)\{L_U(\tilde{f}) - L_U(f^*)\} + \Delta_U(\tilde{f}, f^*, \pi).$$

Lemma 2 Let $\tilde{f}$ be in $\mathrm{co}(\xi(\mathscr{F}))$. Then

$$\Delta_U(f, f^*, \pi) \leq \pi^2 {b_U}^2 \quad (\forall \pi \in [0,1])$$

Lemma 3 Under assumption (A)

$$L_U(\hat{f}_K) \leq \inf_{\lambda} L_U(f_\lambda) + \frac{c^2 {b_U}^2}{K + c - 1} \quad \text{where } f_\lambda = \xi^{-1}(\sum \lambda_j \xi(\phi_j))$$

Lemma 4 If $L_U(\hat{f}_K) \leq \inf_{\lambda} L_U(f_\lambda) + \tau$, then

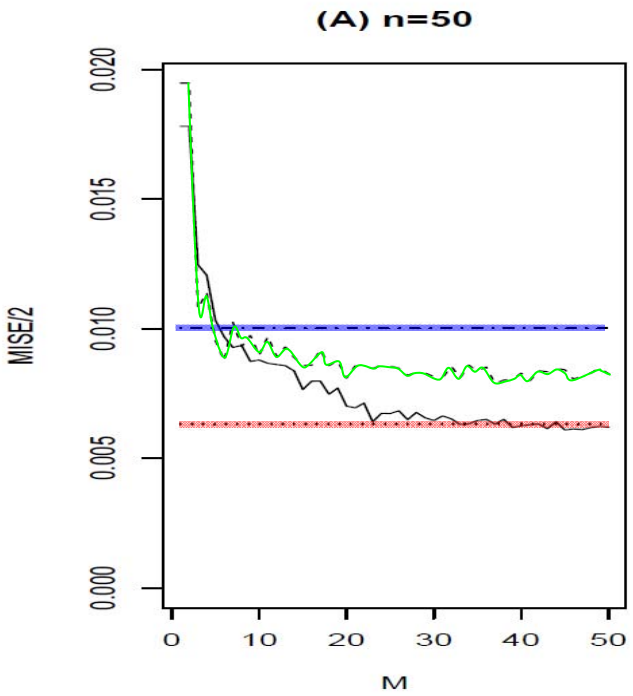
$$\mathbb{E}_g D_U(g, \hat{f}_K) \leq \inf D_U(g, f_\lambda) + \tau + 2\,\mathbb{E}_g \sup_{f \in \mathscr{F}} \| \frac{1}{n} \sum \xi(f(\boldsymbol{x}_i)) - \mathbb{E}_g \xi(f) \|$$

β-Boost (β=1)

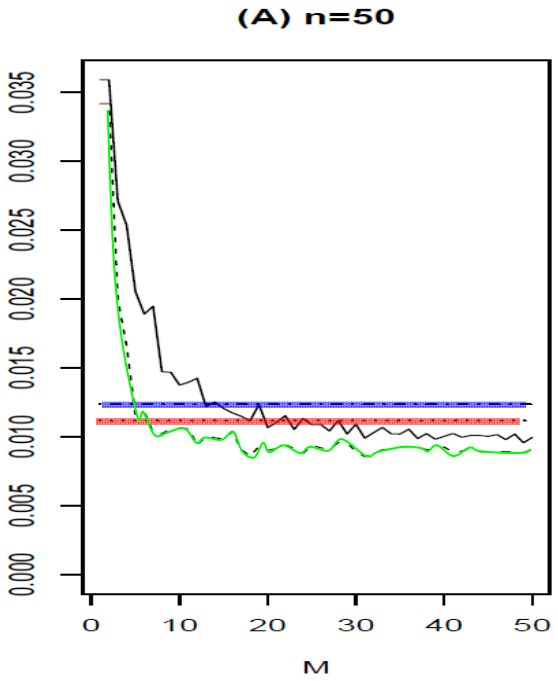
β-Boost (β=1/2)

KDE

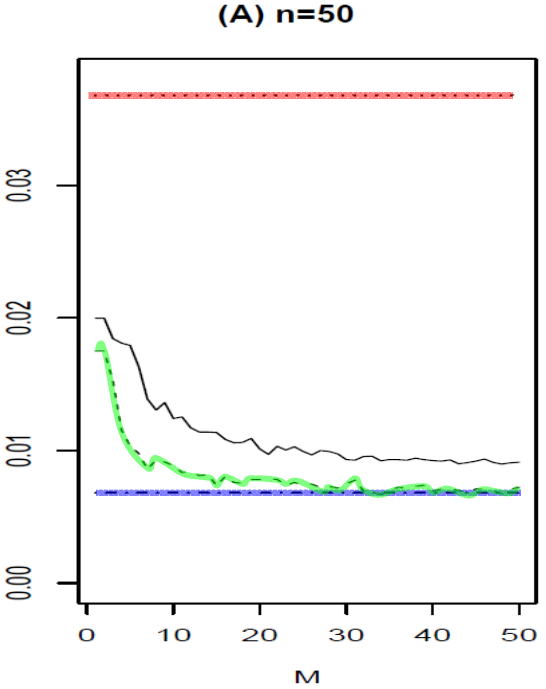
RSDE



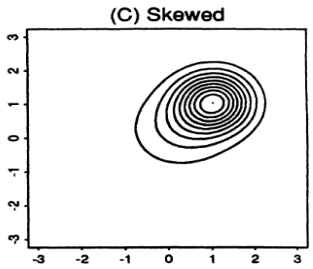
C (skewed-unimodal)

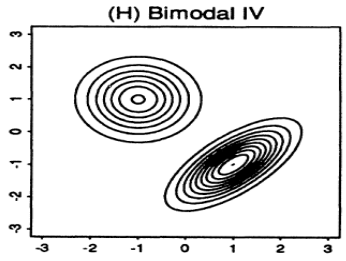


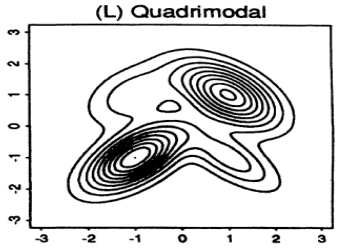
H (bimodal)



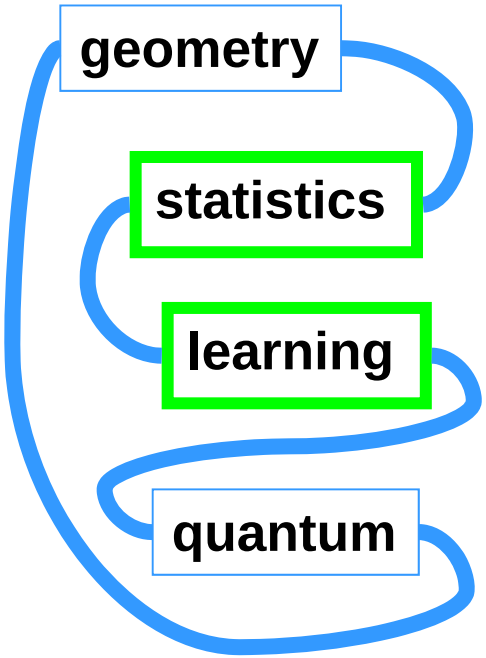
L (quadrimodal)







Future of information geometry



Deepening

Poincare conjecture

Optimal transform

Free probability

Compressed sensing

Semiparametrics

Dynamical systems

Mathematical evolutionary theory

Dempster-Shafer theory

Poincaré conjecture 1904

Every simply connected, closed 3-manifold is homeomorphic to the 3-sphere

Ricci flow by Richard Hamilton in 1981

Perelmann 2002, 2003

Optimal control theory due to Pontryagin and Bellman

Wasserstein space $d_{\mathrm{W}}(f,g) = \min\{\sqrt{\mathrm{E}\|X-Y\|^2} : X \sim f, Y \sim g\}$

Optimal transport $\phi_{f,g} = \arg\min\{\sqrt{\mathrm{E}\,\|X-\phi(X)\|^2} : X \sim f, \phi(X) \sim g\}$

Optimal transport

Theorem (Brenier, 1991) There exists a convex function Φ such that $\nabla\Phi = \phi_{f,g}$

Assume that $\exists K > 0$ st $-\boldsymbol{u}^{\mathrm{T}}\left\{\frac{\partial}{\partial \boldsymbol{x}\partial \boldsymbol{x}^{\mathrm{T}}}\log g(\boldsymbol{x})\right\}\boldsymbol{u} \geq K \parallel \boldsymbol{u} \parallel^2$

Talagrand inequality $D(f,g) \geq \frac{K}{2} d_{\mathrm{W}}(f,g)^2$

Log Sovolev inequality $D(f,\Phi^2 g) \leq \frac{2}{K} \mathrm{E}_g \parallel \nabla\Phi \parallel^2$

Optimal transport theory is extending on a general manifold

C. Villani (2010) Optimal transport theory as Fields medal

Geometers consider not a space, but a distribution family on the space.

Thank you