総研大夏期大学院　2012年9月19日 20日，統数研

最尤理論、誕生から百年の統計学と

今後の方向への一考察

江口　真透

統計数理研究所

Summer-semester in Guas, 19-20 September, 2012, ISM

Maximum likelihood theory from Ronald Fisher in 1912 with discussion on future directions

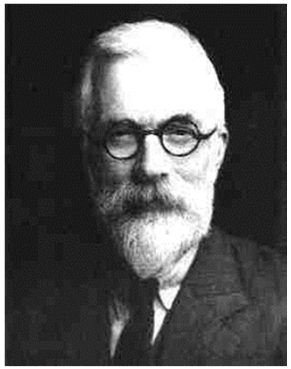
Shinto Eguchi

Institute of Statistical Mathematics

Outline

- MLE 100 years from Fisher 1912
- Information geometry
- Power entropy and divergence
- Two cultures in statistics and AdaBoost
- Poincaré conjecture and Optimal transport theory
- Likelihood for equation model

R. A. Fisher



Ronald A. Fisher
(1890-1962)

Statistics

analysis of variance

maximum likelihood estimation

Design of experiment

Randomization test

Linear discriminant analysis

Population genetics

Genetical Theory of Natural Selection

Fundamental theorem of natural selection

Wright-Fisher model

Fisher's equation

ON AN ABSOLUTE CRITERION FOR FITTING FREQUENCY CURVES.

By *R. A. Fisher*, Gonville and Caius College, Cambridge.

1. If we set ourselves the problem, in its essence one of frequent occurrence, of finding the arbitrary elements in a function of known form, which best suit a set of actual observations, we are met at the outset by an arbitrariness which appears to invalidate any results we may obtain. In the general problem of fitting a theoretical curve, either to an observed curve, or to an observed series of ordinates, it is, indeed, possible to specify a number of different standards of conformity between the observations and the theoretical curve, which definitely lead to different though mutually approximate results. This mutual approximation, though convenient in practice in that it allows a computer to make a legitimate choice of the method which is arithmetically simplest, is harmful from the theoretical standpoint as tending to obscure the practical discrepancies, and the theoretical indefiniteness which actually exist.

Messenger of Mathematics, 41: 155-160, (1912).

Maximum likelihood

Fisher (1912)

$$\int \{f(x,\theta) - y(x)\}^2 dx \qquad \sum_{i=1}^{n} \{f(x_i,\theta) - y(x_i)\}^2$$

$$\log P = \int y(x) \log f(x,\theta)\, dx$$

$$\log P' = \sum_{i=1}^{n} \log f(x_i,\theta)$$

The most probable set of values for the θ's will make P a maximum.

MLE 1912-2012, Aldrich (1997)

Efficiency, sufficiency, Fisher information,…

Boltzmann-Shannon entropy, Kullback-Leibler divergence

Max entropy distribution (exponential model)

Parametric Statistical Modeling by Minimum Integrated Square Error

David Scott Technometrics (2001)

The likelihood function plays a central role in parametric and Bayesian estimation, as well as in nonparametric function estimation via local polynomial modeling. However, integrated square error has enjoyed a long tradition as the goodness-of-fit criterion of choice in nonparametric density estimation. In this article, I investigate the use of integrated square error, or L_2 distance, as a theoretical and practical estimation tool for a variety of *parametric* statistical models. I show that the asymptotic inefficiency of the parameters estimated by minimizing the integrated square error or L_2 estimation (L_2E) criterion

$$\int \{f(x,\theta) - y(x)\}^2 \, dx \leftarrow -\overset{?}{-}- \rightarrow \int y(x) \log f(x,\theta) \, dx$$

Asymptotic efficiency

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

Maximum likelihood estimator $\hat{\theta} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log f(X_i,\theta)$

Asymptotic normal If $\{X_i\}_{i=1}^{n} \underset{\text{iid}}{\sim} f(x,\theta)$, then

$$\sqrt{n}(\hat{\theta}-\theta) \xrightarrow{D} N(0,\ I_\theta^{-1})$$

Asymptotic efficient $\text{var}_A(\hat{\theta}) \le \text{var}_A(\tilde{\theta})$ for any CAN estimator $\tilde{\theta}$

Estimative distribution $f(x,\hat{\theta})$

Confidence interval $\Pr\{C(r_\alpha) \mid f(\cdot,\theta)\} = 1-\alpha$

$$C(r_\alpha) = \{\ \theta :\ n(\theta-\hat{\theta})^{\mathrm{T}} I(\theta)(\theta-\hat{\theta}) \le r_\alpha^{\,2}\ \}$$

Most cited statisticians

1 robert tibshirani
Professor of Health Research and Policy, a
Verified email at stanford.edu
Cited by 116885

Michael I. Jordan
Professor of EECS and Professor of Statistic
Verified email at cs.berkeley.edu
Cited by 53311

2 Donald B Rubin
John L. Loeb Professor of Statistics, Harv
Verified email at stat.harvard.edu
Cited by 115695

William G Cochran 1909-1980
Professor of Statistics, Harvard University
Verified email at stat.harvard.edu
Cited by 48114

3 B Efron
Professor of statistics, Stanford University
Verified email at stat.stanford.edu
Cited by 69104

Leo Breiman 1928-2005
Professor of Statistics, UC Berkeley
Verified email at stat.berkeley.edu
Cited by 46238

Arthur P Dempster
Professor Emeritus of Statistics, Harvard Un
Verified email at stat.harvard.edu
Cited by 39427

Cf. Physics, Medicine, Mathematics

DEFINING THE CURVATURE OF A STATISTICAL PROBLEM (WITH APPLICATIONS TO SECOND ORDER EFFICIENCY)

BY BRADLEY EFRON

Stanford University

Statisticians know that one-parameter exponential families have very nice properties for estimation, testing, and other inference problems. Fundamentally this is because they can be considered to be "straight lines" through the space of all possible probability distributions on the sample space. We consider arbitrary one-parameter families $\mathcal{F}$ and try to quantify how nearly "exponential" they are. A quantity called "the statistical curvature of $\mathcal{F}$" is introduced. Statistical curvature is identically zero for exponential families, positive for nonexponential families. Our purpose is to show that families with small curvature enjoy the good properties of exponential families. Large curvature indicates a breakdown of these properties. Statistical curvature turns out to be closely related to Fisher and Rao's theory of second order efficiency.

The Annals of Statistics
1975, Vol. 3, No. 6, 1189-1242

Statistical curvature

$$g(x) = \frac{\Gamma\left(\frac{f+1}{2}\right)}{f^{\frac{1}{2}}\Gamma(\tfrac{1}{2})\Gamma\left(\frac{f}{2}\right)}\left(1+\frac{x^2}{f}\right)^{-(f+1)/2}$$

$$\gamma_0^2 = \frac{6[3f^2+18f+19]}{f(f+1)(f+5)(f+7)} :$$

f	1	2	5	10	20	$\to \infty$
γ_0^2	2.5	1.063	.306	.107	.0334	$\sim 18/f^2$

Information geometry

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

Information metric $g_{ij}(\theta) = -\mathrm{E}_{f(\cdot,\theta)} \left\{ \frac{\partial}{\partial\theta_i \partial\theta_j} \log f(x,\theta) \right\}$

Exponential connection

$$\overset{\mathrm{e}}{\Gamma}_{ij,k}(\theta) = \mathrm{E}_{f(\cdot,\theta)} \left\{ \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(x,\theta) \frac{\partial}{\partial\theta_k} \log f(x,\theta) \right\}$$

Mixture connection

$$\overset{\mathrm{m}}{\Gamma}_{ij,k}(\theta) = \overset{\mathrm{e}}{\Gamma}_{ij,k}(\theta) + \mathrm{E}_{f(\cdot,\theta)} \left\{ \frac{\partial}{\partial\theta_i} \log f(x,\theta) \frac{\partial}{\partial\theta_j} \log f(x,\theta) \frac{\partial}{\partial\theta_k} \log f(x,\theta) \right\}$$

Rao (1949) Amari (1982)

Natural Gradient Works Efficiently in Learning

Shun-ichi Amari
RIKEN Frontier Research Program, Saitama 351-01, Japan

When a parameter space has a certain underlying structure, the ordinary gradient of a function does not represent its steepest direction, but the natural gradient does. Information geometry is used for calculating the natural gradients in the parameter space of perceptrons, the space of matrices (for blind source separation), and the space of linear dynamical systems (for blind source deconvolution). The dynamical behavior of natural gradient online learning is analyzed and is proved to be Fisher efficient, implying that it has asymptotically the same performance as the optimal batch estimation of parameters. This suggests that the plateau phenomenon, which appears in the backpropagation learning algorithm of multilayer perceptrons, might disappear or might not be so serious when the natural gradient is used. An adaptive method of updating the learning rate is proposed and analyzed.

m-geodesic and e-geodesic

m-geodesic $p_t^{(\mathrm{m})}(\boldsymbol{x}) = (1-t)p(\boldsymbol{x}) + tq(\boldsymbol{x}),$

e-geodesic $p_t^{(\mathrm{e})}(\boldsymbol{x}) = c_t \exp\{(1-t)\log p(\boldsymbol{x}) + t\log q(\boldsymbol{x})\}$

A model M is **m-geodesic** if

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

A model M is **e-geodesic** if

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

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

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

Kullback-Leibler divergence

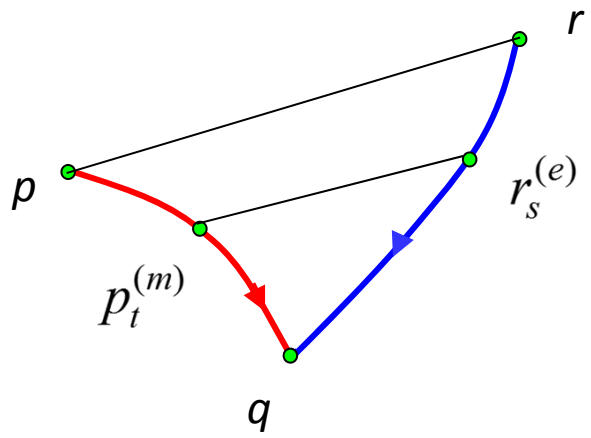
KL divergence (1951) $D(g,f)=\mathrm{E}_g \log \dfrac{g(X)}{f(X)}$

AIC $\mathrm{AIC}(M)=-2\max_{\theta\in\Theta} L_n(\theta)+2\dim(\theta)$ (Akaike, 1973)

$$D(p,r)=D(p,q)+D(q,r)$$

$$p_t^{(m)}=(1-t)p+tq \qquad r_s^{(e)}(\boldsymbol{x})=c_s\exp\{(1-s)\log r+s\log q\}$$

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



Nagaoka-Amari (1982)

MLE on Exponential family

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

Cumulant function $\kappa(\boldsymbol{\theta}) = \log \{\mathrm{E}_{p_0} \mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{t}(\boldsymbol{x})}\}$

$$\boldsymbol{\eta} := \frac{\partial}{\partial \boldsymbol{\theta}} \kappa(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}\{\boldsymbol{t}(\boldsymbol{x})\} \qquad \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} \kappa(\boldsymbol{\theta}) = \mathrm{var}_{\boldsymbol{\theta}}\{\boldsymbol{t}(\boldsymbol{x})\}$$

Log likelihood function $L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\ p(\boldsymbol{x}_i, \boldsymbol{\theta}) = n\{\boldsymbol{\theta}^T \bar{\boldsymbol{t}} - \kappa(\boldsymbol{\theta})\}$

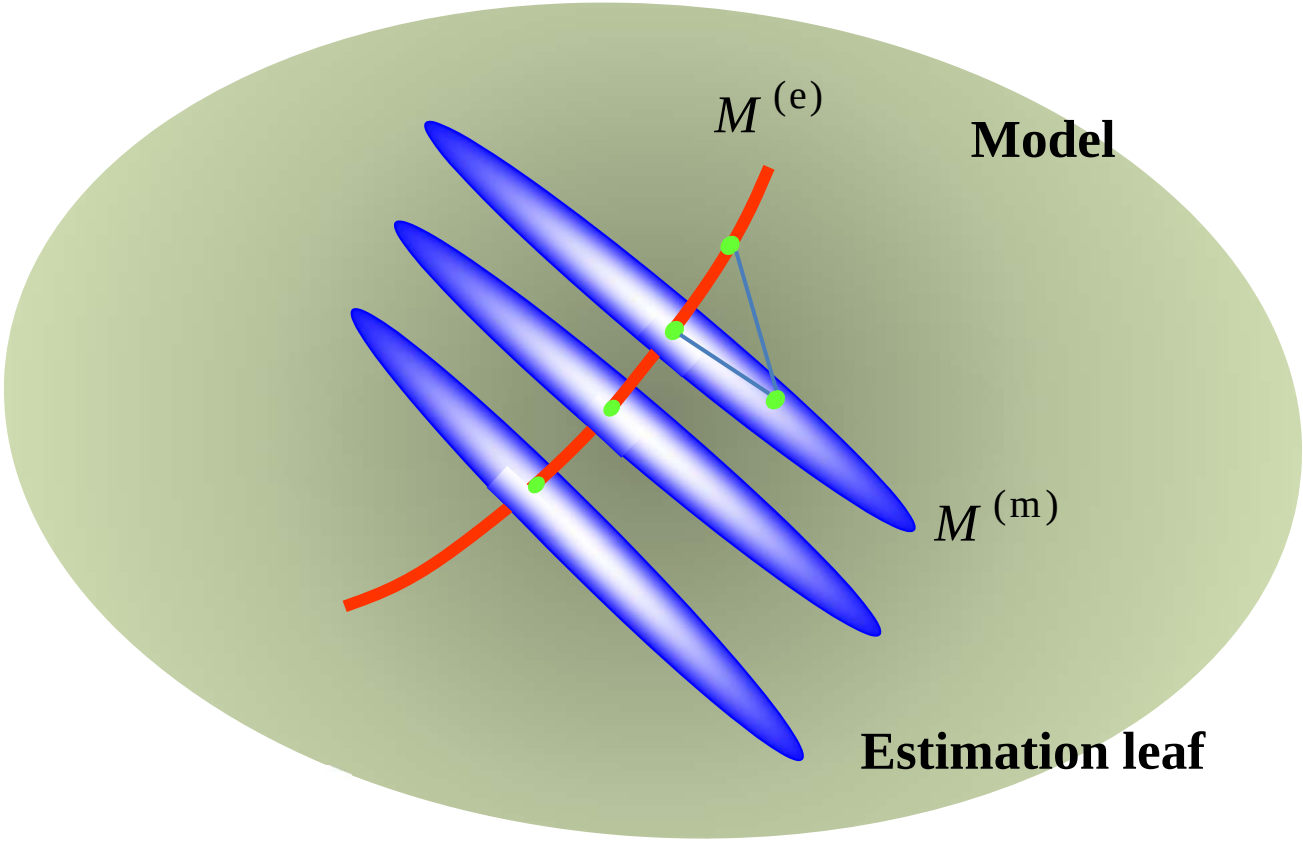
Maximum likelihood $\max_{\boldsymbol{\theta} \in \Theta} \frac{L(\boldsymbol{\theta})}{n} = \max_{\boldsymbol{\theta} \in \Theta} \{\boldsymbol{\theta}^T \bar{\boldsymbol{t}} - \kappa(\boldsymbol{\theta})\} = \kappa^*(\bar{\boldsymbol{t}})$

Likelihood equation $\frac{\partial}{\partial \boldsymbol{\theta}} \kappa(\boldsymbol{\theta}) = \bar{\boldsymbol{t}} \quad \text{or} \quad \boldsymbol{\eta}_{\mathrm{ML}} = \bar{\boldsymbol{t}}$

Mean match model $M^{(\mathrm{m})} = \{\ p(\boldsymbol{x}) : E_p\{\boldsymbol{t}(\boldsymbol{x})\} = \boldsymbol{\eta}\ \}$

Barndorff-Nielsen (1978), Brown (1986),

Pair of (model, estimation) Amari (1982)



$$\mathscr{F} = \bigcup_{\theta \in \Theta} \mathscr{L}_\theta$$

$$\mathscr{L}_\theta = \{g : \hat{\theta}(g) = \theta\}$$

Non-exponential family

t-distribution

Univariate $$f(x,\mu,\sigma^2,\nu)=\frac{\Gamma(\frac{\nu+1}{2})}{(\pi\nu)^{\frac{1}{2}}\Gamma(\frac{\nu}{2})}\left\{1+\frac{1}{\nu}\frac{(x-\mu)^2}{\sigma^2}\right\}^{-\frac{\nu+1}{2}}$$

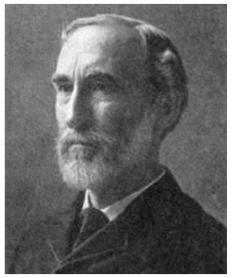
p-variate $$f_\gamma(x,\mu,\Sigma,\nu)=\frac{\Gamma(\frac{\nu+p}{2})}{(\pi\nu)^{\frac{p}{2}}\Gamma(\frac{\nu}{2})}\left\{1+\frac{1}{\nu}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right\}^{-\frac{\nu+p}{2}}$$

Generalized Pareto distribution

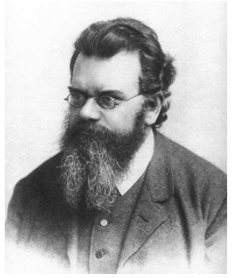
$$f(x,\theta,\xi)=\theta(1+\xi\theta\, x)^{-\frac{1}{\xi}-1}$$

Unified view to power exponential family

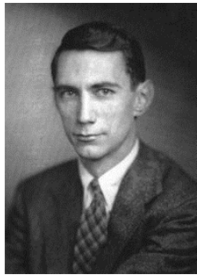
Gibbs-Boltzmann-Shannon



Josiah Willard Gibbs
(1839 - 1903)



Ludwig E. Boltzmann
(1844 - 1906)



Claude E. Shannon
(1916- 2001)

GBS-entropy $$H_{\mathrm{GBS}}(p) = -E_p \log p(X)$$

Log-likelihood $$-\frac{1}{n}\sum_{i=1}^{n} \log p(\boldsymbol{x}_i, \theta) \approx H_{\mathrm{GBS}}(p(\cdot, \theta))$$

A new estimator if we extend GBS-entropy ?

Possible Generalization of Boltzmann–Gibbs Statistics

Constantino Tsallis[1]

Journal of Statistical Physics, Vol. 52, Nos. 1/2, 1988

With the use of a quantity normally scaled in multifractals, a generalized form is postulated for entropy, namely $S_q \equiv k[1-\sum_{i=1}^{W} p_i^q]/(q-1)$, where $q \in \mathbb{R}$ characterizes the generalization and $\{p_i\}$ are the probabilities associated with W (microscopic) configurations ($W \in \mathbb{N}$). The main properties associated with this entropy are established, particularly those corresponding to the microcanonical and canonical ensembles. The Boltzmann–Gibbs statistics is recovered as the $q \to 1$ limit.

$$S_q(\boldsymbol{p}) = \frac{k}{q-1}\left(1-\sum_{j=1}^{W} {p_j}^q\right)$$

Hill's diversity index in 1973

Let p_j be a relative frequency of a speice j for $j = 1,...., S$

$$D_a(\boldsymbol{p}) = \left(\sum_{j=1}^{S} {p_j}^a \right)^{\frac{1}{1-a}}$$

$$\lim_{a \to 0} \log\{ D_a(\boldsymbol{p}) \} = \lim_{a \to 0} \frac{1}{1-a} \log \left(\sum_{j=1}^{S} {p_j}^a \right)$$

$$= -\sum_{j=1}^{S} p_j \log p_j = H(\boldsymbol{p})$$

An Analysis of Transformations

By G. E. P. Box and D. R. Cox

University of Wisconsin *Birkbeck College, University of London*

Summary

In the analysis of data it is often assumed that observations $y_1, y_2, \ldots, y_n$ are independently normally distributed with constant variance and with expectations specified by a model linear in a set of parameters $\boldsymbol{\theta}$. In this paper we make the less restrictive assumption that such a normal, homoscedastic, linear model is appropriate after some suitable transformation has been applied to the y's. Inferences about the transformation and about the parameters of the linear model are made by computing the likelihood function and the relevant posterior distribution. The contributions of normality, homoscedasticity and additivity to the transformation are separated. The relation of the present methods to earlier procedures for finding transformations is discussed. The methods are illustrated with examples.

$$y^{(\lambda)} = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & (\lambda \neq 0), \\ \log y & (\lambda = 0), \end{cases}$$

Projective Entropy

γ-cross entropy

$$C_\gamma(g,f) = -\frac{\int f(x)^\gamma g(x)dx}{\left(\int f(x)^{1+\gamma}dx\right)^{\frac{\gamma}{1+\gamma}}}$$

γ-diagonal entropy

$$H_\gamma(f) = C_\gamma(f,f) = -\left(\int f^{1+\gamma}\right)^{\frac{1}{1+\gamma}}$$

γ-divergence

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

Cf. Good (1972) Fujisawa-Eguchi (2008) Eguchi-Kato (2010) Ferrari-Yang (2010)

Remark: The diagonal entropy is equivalent to Tsallis entropy

$$S_q(f) = \frac{1}{q-1}\left(1 - \int f^q\right)$$

$$S_q(f) = \frac{1}{\gamma}\{H_\gamma(f)^q + 1\} \quad \text{where} \quad q = 1+\gamma$$

Three properties

$$C_\gamma(g,f) = -\int \left(\frac{f}{\| f \|_q}\right)^\gamma g, \qquad (q = 1+\gamma)$$

Linearity	$C_\gamma(\alpha g+\beta h, f) = \alpha C_\gamma(g,f) + \beta C_\gamma(h,f)$
scale invariance	$C_\gamma(g,\lambda f) = C_\gamma(g,f) \quad (\forall \lambda > 0)$
Lower Bound	$C_\gamma(g,\ f) \geq C_\gamma(g,g)$

These properties lead to uniqueness for C_γ ?

Theorem

Let $\Gamma(g,f)=\Phi\left(\int\rho(f)\right)\int\psi(f)g$

If $\Gamma(g,f)$ satisfies

(i) $\Gamma(g,\lambda f)=\Gamma(g,f) \quad (\forall\lambda>0)$

(ii) $\Gamma(g,f)\geq\Gamma(g,g)$

then there is a constant $\gamma>0$ suct that $\Gamma(g,f)=C_\gamma(g,f)$

Remark: If $(\Phi(Y),\rho(f),\psi(f))=(Y^{-\frac{\gamma}{1+\gamma}},f^{1+\gamma},f^\gamma)$,

then $\Gamma(g,f)=\left(\int f(x)^{1+\gamma}dx\right)^{-\frac{\gamma}{1+\gamma}}\int f^\gamma g=C_\gamma(g,f)$

Proof

(i) $\Rightarrow \quad \dfrac{\partial}{\partial\lambda}\Gamma(g,\lambda f)=0$

$$\frac{\partial}{\partial\lambda}\Gamma(g,\lambda f)=\int\left[\{\frac{\partial}{\partial\lambda}\Phi(\int\rho(\lambda f))\}\psi(\lambda f(x))+\Phi(\int\rho(\lambda f))\frac{\partial\psi(\lambda f(x))}{\partial\lambda}f(x)\right]g(x)dx$$

$$\frac{\partial\log\psi(X)}{\partial X}=-c\;X^{-1} \qquad \therefore\;\psi(X)=X^{\gamma}\;(\gamma=c),$$

(ii) $\Rightarrow \dfrac{\partial}{\partial f_i}\Gamma(g,f)\Big|_{f=g}=0$ if $\Gamma(g,f)=\Phi\Big(\sum_{j=1}^{m}\rho(f_j)\Big)\sum_{j=1}^{m}g_j\,(f_j)^{\gamma}$

(*) $\dfrac{\partial}{\partial f_i}\Gamma(g,f)\Big|_{f=g}=\Big\{\dot\Phi\Big(\sum_{j=1}^{m}\rho(g_j)\Big)\sum_{j=1}^{m}(g_j)^{\gamma+1}\Big\}\dot\rho(g_i)+\Big\{\Phi\Big(\sum_{j=1}^{m}\rho(g_j)\Big)\Big\}\gamma(g_i)^{\gamma}=0$ [$\because$ (ii)]

$$\frac{\partial\rho(g)}{\partial g}=-c\gamma\,g^{\gamma} \qquad \therefore\;\rho(g)=-c\frac{\gamma}{\gamma+1}g^{\gamma+1}$$

Taking a sum of (*) over i, $\dfrac{\dot\Phi(Y)}{\Phi(Y)}=-\dfrac{\gamma+1}{\gamma}Y,$ $\therefore\quad\Phi(Y)=Y^{-\frac{\gamma+1}{\gamma}}$

Shannon-Khinchin axiom

$$S_g[p] = \sum_{i=1}^{W} g(p_i)$$

K1: The requirement that S depends continuously on p implies that g is a continuous function.

K2: The requirement that the entropy is maximal for the equi-distribution $p_i = 1/W$ (for all i) implies that g is a concave function.

K3: The requirement that adding a zero-probability state to a system, $W+1$ with $p_{W+1} = 0$, does not change the entropy implies $g(0) = 0$.

$$S_{c,d}[p] = \frac{e \sum_i^W \Gamma\left(1+d,\, 1-c\ln p_i\right)}{1-c+cd} - \frac{c}{1-c+cd}.$$

Cf. Hanel and Thurner (2011)

Max γ-entropy

Equal moment space $\mathscr{F}(\mu,\Sigma)=\{f(x):\mathrm{E}_f(X)=\mu,\mathrm{V}_f(X)=\Sigma\}$

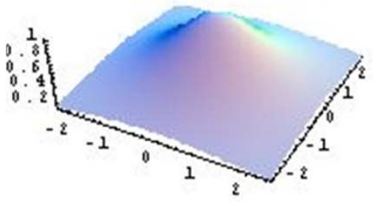
γ-entropy $H_\gamma(f)=-\|f\|_q \quad (q=1+\gamma)$

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

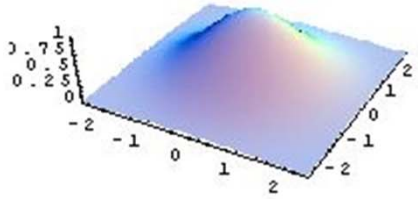
Eguchi-Komori-Kato (2011)

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

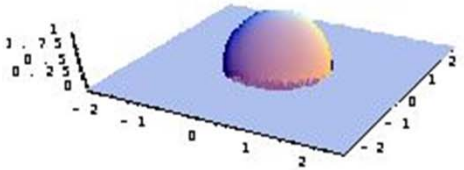
$$\mathscr{F}_\gamma(\mu,\Sigma)=\{f\in\mathscr{F}(\mu,\Sigma):\ \text{Support}\,(f)\subseteq\{x:(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\le\frac{1}{2}\kappa_\gamma\}$$



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



$\gamma=0$ (Gaussian)



$\gamma=2$ (Wigner)

γ-Estimation

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

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

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

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

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 If $\gamma > 0$, $(\hat{\mu}_\gamma, \hat{\Sigma}_\gamma)$ is robust gainst much oulying

Escort normal family

$$f_\gamma(\cdot,\mu,\Sigma)\in\mathscr{F}(\mu,\Sigma)$$

Escort distribution $e_q(f(x))=\dfrac{f(x)^{1+\gamma}}{\int f^{1+\gamma}}\qquad(q=1+\gamma)$

$$e_q(f(x,\mu,\Sigma))=c_\gamma^{\,*}\det(\Sigma)^{-\frac{1}{2}}\left\{1-\frac{\kappa_\gamma}{2}(x-\mu)\Sigma^{-1}(x-\mu)\right\}^{\frac{1+\gamma}{\gamma}}$$

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

Ξ

$$=c_\gamma^{\,*}\left\{\det(\Xi)^{\frac{2}{p\gamma+2\gamma+2}}-\frac{\gamma}{2}(x-\mu)\Xi(x-\mu)\right\}^{\frac{1+\gamma}{\gamma}}$$

$$\frac{\partial}{\partial\mu}\int e_q(f_\gamma(x,\mu,\Sigma))=0\ \Rightarrow\ \mathrm{E}_{f_\gamma(\cdot,\mu,\Sigma)}(X)=\mu$$

$$\frac{\partial}{\partial\Xi}\int e_q(f_\gamma(x,\mu,\Xi))=0\ \Rightarrow\ \mathrm{E}_{f_\gamma(\cdot,\mu,\Sigma)}\{(X-\mu)(X-\mu)^{\mathrm{T}}\}=\Sigma$$

γ-estimation on γ-normal model

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

γ-loss function

$$L_\gamma(\mu,\Sigma)=\frac{1}{n}\sum_{i=1}^{n}\det(\Sigma)^{-\frac{1}{2}\frac{\gamma}{\gamma+1}}\{1-\frac{1}{2}\kappa_\gamma(x_i-\mu)^{\mathrm{T}}\Sigma^{-1}(x_i-\mu)\}$$

$$=\det(\Xi)^{\frac{\gamma}{p\gamma+2\gamma+2}}-\frac{1}{2}\kappa_\gamma\{(\mu-\bar{x})^{\mathrm{T}}\Xi(\mu-\bar{x})+\mathrm{tr}(S\Xi)\}$$

$$L_\gamma(\bar{x},S)=\min_{(\mu,V)}L_\gamma(\mu,V)$$

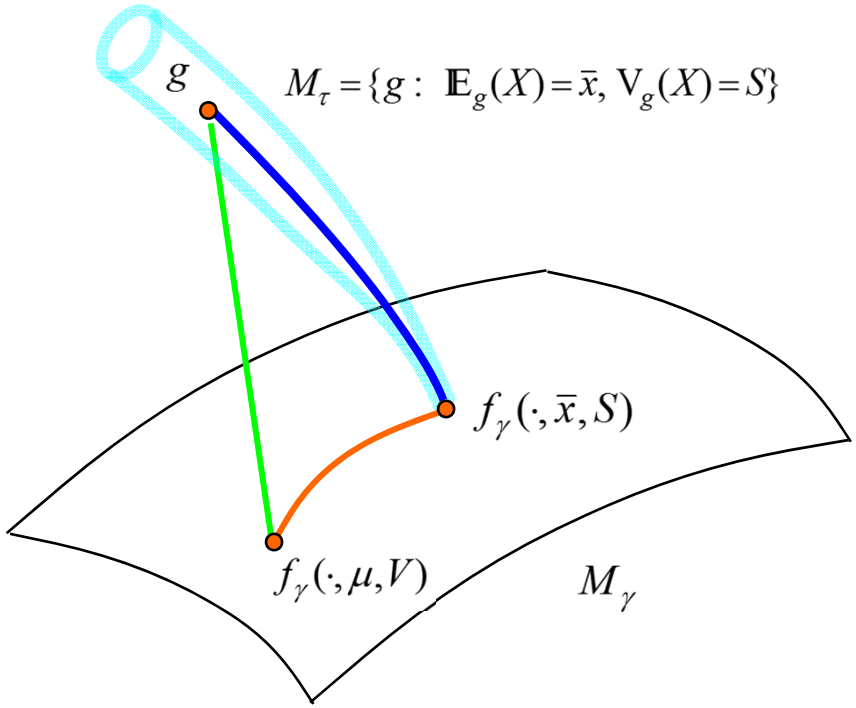
$$L_\gamma(\mu,\Sigma)-L_\gamma(\bar{x},S)=\det(\Xi)^{\frac{\gamma}{p\gamma+2\gamma+2}}-\frac{1}{2}\kappa_\gamma\{(\mu-\bar{x})^{\mathrm{T}}\Xi(\mu-\bar{x})+\mathrm{tr}(S\Xi)\}$$

$$-\det(\hat{\Xi})^{\frac{\gamma}{p\gamma+2\gamma+2}}+\frac{1}{2}\kappa_\gamma\mathrm{tr}(S\hat{\Xi})\}$$

$$=C_\gamma(f_{\bar{x},S},f_{\mu,\Sigma})-C_\gamma(f_{\bar{x},S},f_{\bar{x},S})=D_\gamma(f_{\bar{x},S},f_{\mu,\Sigma})\geq 0$$

Pythagoras triangle

Loss decomposition $L_\gamma(\mu,\Sigma)-L_\gamma(\bar{x},S)=D_\gamma(f_\gamma(\cdot,\bar{x},S),f_\gamma(\cdot,\mu,\Sigma))$



Pythagorian $D_\gamma(g,f_\gamma(\cdot,\mu,V))-D_\gamma(g,f_\gamma(\cdot,\bar{x},S))=D_\gamma(f_\gamma(\cdot,\bar{x},S),f_\gamma(\cdot,\mu,V))$

MLE for normal model

Equal moment space $\mathscr{F}(\mu,\Sigma)=\{f(x):\mathrm{E}_f(X)=\mu,\mathrm{V}_f(X)=\Sigma\}$

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

Max 0-entropy $H_0(f_{\mu,\Sigma})=\max_{f\in\mathscr{F}(\mu,\Sigma)}H_0(f)$ where $H_0(f)=-\int f\log f$

Likelihood $L_0(\mu,\Sigma)=\sum_{i=1}^{n}\log f_{\mu.\Sigma}(x_i)$

Teicher's theorem (1961)

Let $g_{\mu,\Sigma.}$ be a location-scale family.

MLE $(\hat{\mu},\hat{\Sigma})$ is the sample mean and variance $(\bar{x},S)$ if and only if $g_{\mu,\Sigma}$ is Gaussian.

where $(\bar{x},S)=\left(\frac{1}{n}\sum x_i,\ \frac{1}{n}\sum(x_i-\bar{x})(x_i-\bar{x})^{\mathrm{T}}\right)$

γ-estimator leads to γ-normal model

Let $$(\hat{\mu}_\gamma, \hat{\Sigma}_\gamma) = \underset{(\mu,V)}{\arg\min}\, L_\gamma(g(\cdot, \mu, \Sigma))$$

where $$g(x, \mu, \Sigma) = c_h \det(2\pi\Sigma)^{-\frac{1}{2}} h\left(-\tfrac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)\right)$$

Then

$$g(x, \mu, \Sigma) = f_\gamma(x, \mu, \Sigma) \iff (\hat{\mu}_\gamma, \hat{\Sigma}_\gamma) = (\bar{x}, S)$$

$\gamma = 0$ (Gaussian) Cf.Teicher (1961)

(γ-estimator , γ'-model)

loss function \ model	0-model	γ-model
0-loss (– log-likelihood)	$(\bar{x}, S)$ MLE	M-estimator
γ-loss	emergence γ-estimator	$(\bar{x}, S)$ γ-estimator

U-entropy and *U*-divergence

***U* is a convex and increasing function** Cf. Eguchi-Kano (2001)

***U*-cross entropy** $C_U(g,f) = -E_g\xi(f(X)) + \int U(\xi(f))$

***U*-entropy** $H_U(f) = -E_f\,\xi(f(X)) + \int U(\xi(f))$

where $\xi = (\dot{U})^{-1}$

***U*-divergence** $D_U(g,f) = C_U(g,f) - H_U(g)$

Example 1. **Boltzmann-Shannon entropy**.

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

$$C_0(g,f) = -\int g \log f + 1 \qquad H_0(f) = -\int f \log f + 1$$

$$D_0(g,f) = \int g(\log g - \log f)$$

Example 2. **Power entropy**. Tsallis (1988) Basu et al (1998) Cf. Box-Cox (1964)

$$(U(t),\ \xi(s))=(\frac{(1+\beta t)^{\frac{1+\beta}{\beta}}}{1+\beta},\frac{s^{\beta}-1}{\beta}) \quad (0\le\beta\le 1)$$

$$C_\beta(g,f)=\int -g\frac{f^{\beta}-1}{\beta}+\frac{g^{\beta+1}}{\beta+1} \qquad H_\beta(f)=\int (-\frac{f^{\beta+1}}{\beta(\beta+1)}+\frac{f}{\beta})$$

$$D_\beta(g,f)=\int g\frac{g^{\beta}-f^{\beta}}{\beta}-\frac{g^{\beta+1}-f^{\beta+1}}{\beta+1}$$

Tsallis entropy ($q=\beta+1$)

Example 3. **Sigmoid entropy**. Takenouchi-Eguchi (2004)

$$(U(t),\xi(s))=(e^{t}-\eta t,\ \log(s+\eta)) \quad (0<\eta<\frac{1}{2})$$

$$C_\eta(g,f)=-\int (g+\eta)\log(f+\eta) \quad H_\eta(f)=-\int (f+\eta)\log(f+\eta)$$

$$D_\eta(g,f)=\int (g+\eta)\{\log(g+\eta)-\log(f+\eta)\}$$

max *U*-entropy model

Let $p_0(\boldsymbol{x})$ be a pdf and let $\boldsymbol{t}(\boldsymbol{x})$ be a statistic of dimension *d.*

Consider a problem $\max_{p \in \mathscr{L}(p_0)} H_U(p)$

where $\mathscr{L}(p_0) = \{ p : E_p \boldsymbol{t}(X) = E_{p_0} \boldsymbol{t}(X) \}$

Mean-equal space

The Lagrangian $L(p, \boldsymbol{\theta}, \lambda) = \int \{ -p\xi(p) + U(\xi(p)) + \boldsymbol{\theta}^{\mathrm{T}} (\boldsymbol{t} - \boldsymbol{\tau}_0) p \} d\boldsymbol{x}$

leads to $\xi(p(\boldsymbol{x})) = \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{t}(\boldsymbol{x}) + \text{const.}$

So $p(\boldsymbol{x}) = \dot{U}(\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{t}(\boldsymbol{x}) - \kappa_U(\boldsymbol{\theta}))$, we call ***U*-model**

U-estimation

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be from $p(\boldsymbol{x})$ and let $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ be a model function.

_U_-loss function

$$L_U(p_{\boldsymbol{\theta}}) = -\frac{1}{n}\sum_{i=1}^{n} \xi(p_{\boldsymbol{\theta}}(\boldsymbol{x}_i)) + \int U(\xi(p_{\boldsymbol{\theta}}))$$

_U_-estimator

$$\hat{\boldsymbol{\theta}}_U = \arg\min_{\boldsymbol{\theta}\in\Theta} L_U(p_{\boldsymbol{\theta}})$$

Note: $\mathrm{E}_p L_U(p_{\boldsymbol{\theta}}) = C_U(p, p_{\boldsymbol{\theta}})$; $L_U(p_{\boldsymbol{\theta}}) \underset{\text{a.s.}}{\to} C_U(p, p_{\boldsymbol{\theta}})$ as $n \to \infty$

Consistency

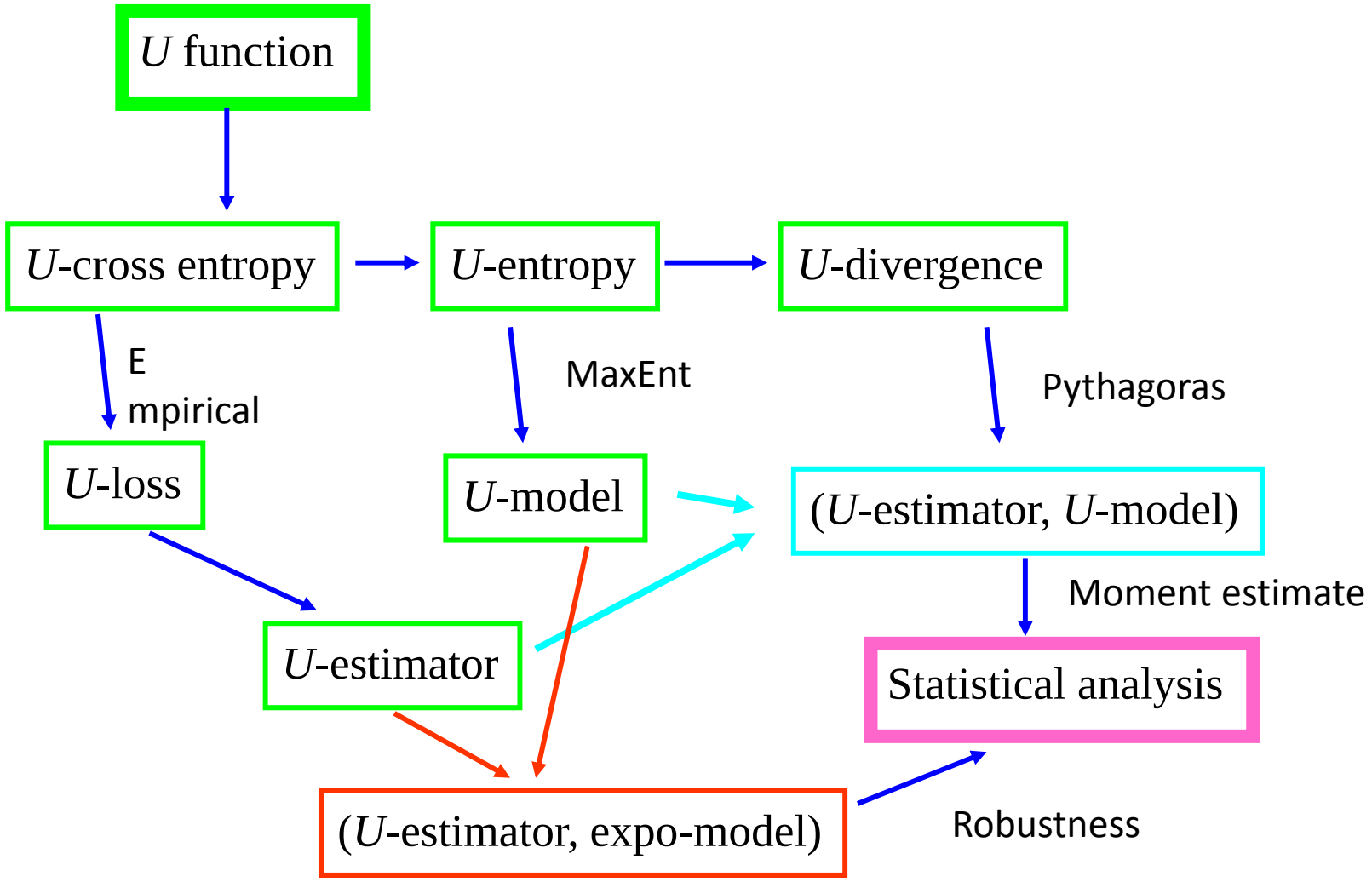
$$\hat{\boldsymbol{\theta}}_U \underset{\text{a.s.}}{\to} \boldsymbol{\theta}(p) \quad \text{as } n \to \infty$$

$$\text{where } \boldsymbol{\theta}(p) = \arg\min_{\theta\in\Theta} C_U(p, p_{\boldsymbol{\theta}})$$

Asymptotic normality

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_U - \boldsymbol{\theta}(p)) \underset{D}{\to} N(0, V_U(\boldsymbol{\theta}, p)) \quad \text{as } n \to \infty$$

U-method


U function
U-cross entropy
U-entropy
U-divergence
E
mpirical
MaxEnt
Pythagoras
U-loss
U-model
(*U*-estimator, *U*-model)
Moment estimate
U-estimator
Statistical analysis
(*U*-estimator, expo-model)
Robustness

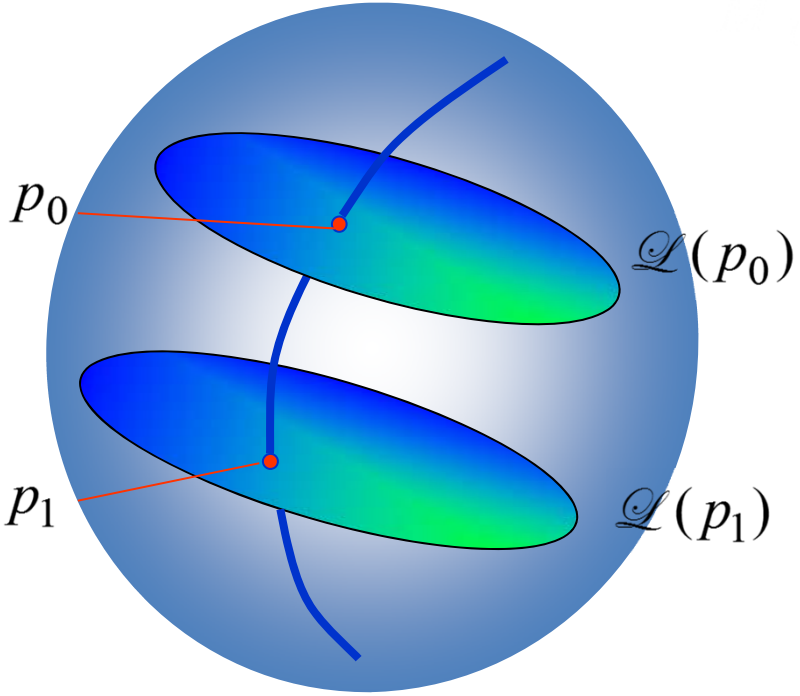
U_1-estimator under exp-model

model / loss function	Exponential	U_1-model
U_{exp} = likelihood	**MLE**	**Robust M-estimate**
U_1-loss	**Robust U_1-estimotor**	**Moment estimator**

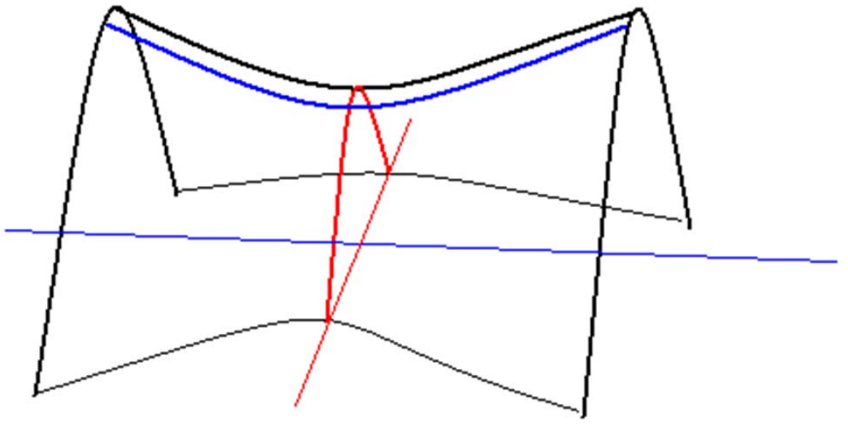
Minimax game

$$\max_{p \in \mathscr{L}(p_0)} \min_{q \in \mathscr{P}} C_U(p,q) = C_U(p^*,p^*) = \min_{q \in \mathscr{P}} \max_{p \in \mathscr{L}(p_0)} C_U(p,q)$$

$$\text{where } p^* = \arg\max_{p \in \mathscr{L}(p_0)} H_U(p)$$



Cf. Grunwald and Dawid (2004)



Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

Statistical Science
2001, Vol. 16, No. 3, 199–231

Pattern recognition

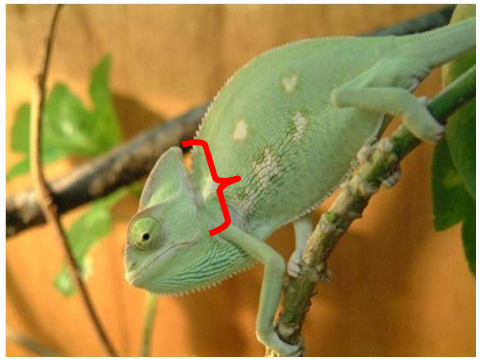
Feature vector $\boldsymbol{x} = (x_1, \ldots, x_p)$

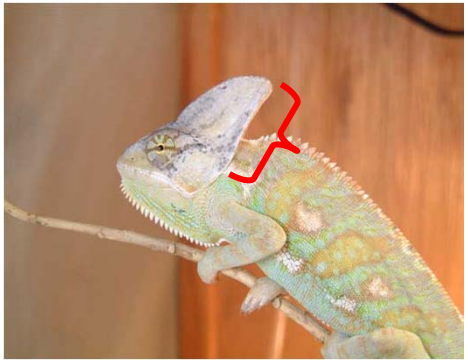
Class label $y \in \{-1, +1\}$

Classifier $\boldsymbol{x} \mapsto y = f(\boldsymbol{x})$

Training data $D = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$

Statistical classifier $f(\boldsymbol{x}) = \operatorname{sgin}(F(\boldsymbol{x}))$





Set of weak learners

Decision stamp

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

Linear classifiers

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

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

No stronger, but exhaustive!

ANNs SVMs *k*-NNs

Adaboost

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

2. For $t = 1, \cdots, T$ $\qquad \varepsilon_t(f) = \sum \mathrm{I}(y_i \neq f(\boldsymbol{x}_i))\, w_t(i)$,

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

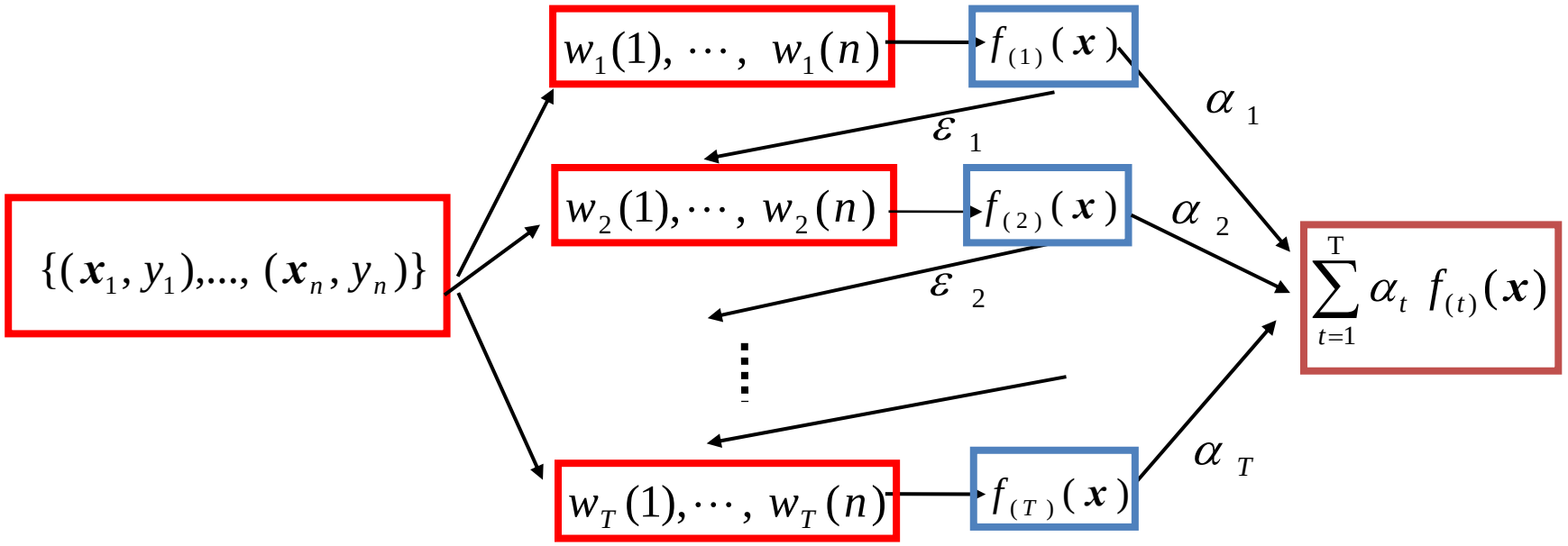
(b) $\qquad \alpha_t = \frac{1}{2} \log \frac{1 - \varepsilon_t(f_{(t)})}{\varepsilon_t(f_{(t)})}$

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

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

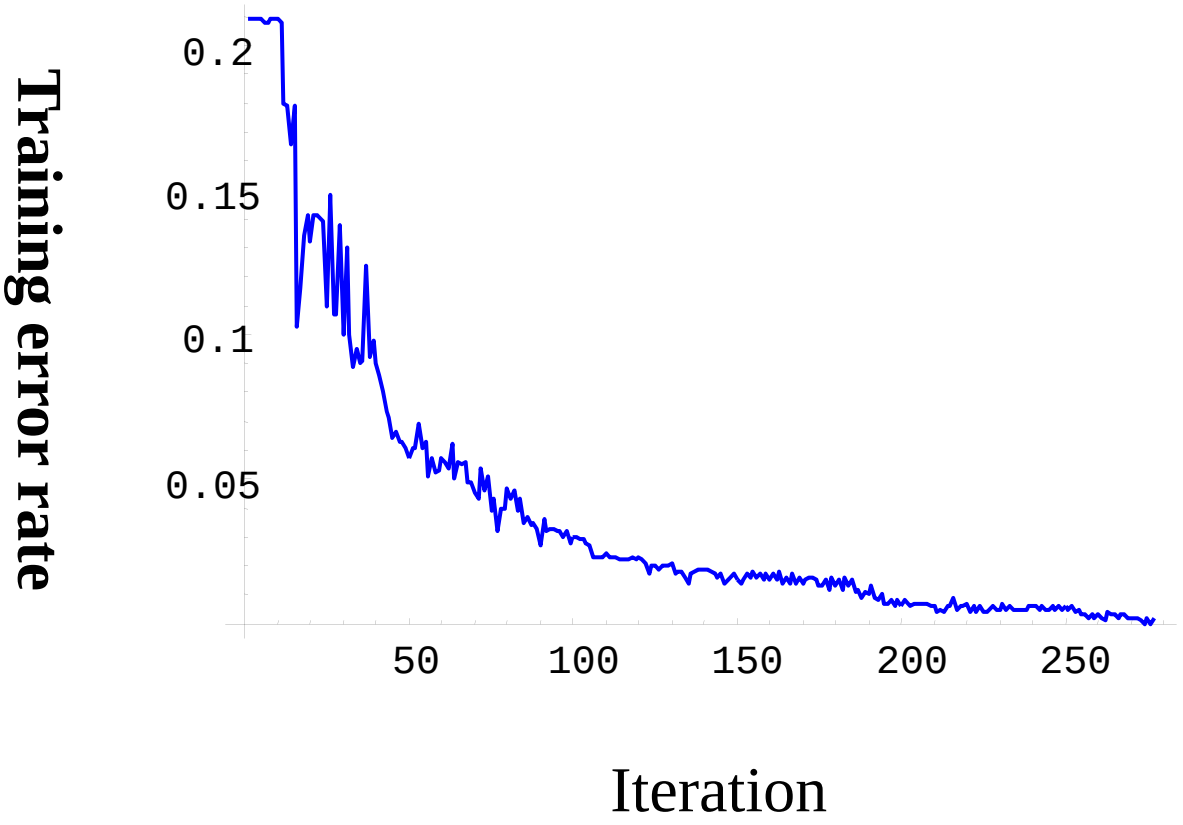
Freund-Schapire (1997).

Learning process



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

Learning curve



Stopping rule

Training data $D_{\text{train}} = \{ (\boldsymbol{x}_i, y_i) : \ i = 1, \ldots, n \}$

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

How to find ***T*** ?

X $T_{\text{opt}} = \arg\min_{T>0} \text{Err}^{\text{train}}(h_{F_T})$

X $T_{\text{opt}} = \arg\min_{T>0} \text{Err}^{\text{test}}(h_{F_T})$

○ $T_{\text{opt}} = \arg\min_{T>0} \text{CVErr}^{\text{train}}(h_{F_T})$

10-fold CV Error rate

$$D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}, \; F_T(\boldsymbol{x}) = \sum_{t=1}^{T} \alpha_t f_{(t)}(\boldsymbol{x})$$

$$\varepsilon(1)\ \varepsilon(2)\ \varepsilon(3)$$

$$\Rightarrow \quad \text{Averaing by} \quad \frac{1}{10}\sum_{k=1}^{10}\varepsilon(k)$$

Update of exp-loss

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

Consider one update as $F(\boldsymbol{x}) \rightarrow F(\boldsymbol{x}) + \alpha\, f(\boldsymbol{x})$

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

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

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

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 minimization

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

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

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

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

$$\text{Equality iff} \quad \alpha_{\text{opt}} = \frac{1}{2}\log\frac{1-\varepsilon(f)}{\varepsilon(f)}$$

Adaboost = sequential min expo-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) $$w_{t+1}(i) \propto w_t(i) \exp\{\alpha_t y_i f_t\ (x_i)\}$$

Bayes risk consistency

Expected loss functional

$$\mathbb{L}_{\exp}(F) = \sum_{y=\pm 1} \int \exp\{-y\,F(\boldsymbol{x})\} dP(y,\boldsymbol{x})$$

$$= \int \{e^{-F(\boldsymbol{x})} p(+1|\,\boldsymbol{x}) + e^{F(\boldsymbol{x})} p(-1|\,\boldsymbol{x})\}\, f(\boldsymbol{x})\, d\boldsymbol{x}$$

Optimal discriminant function

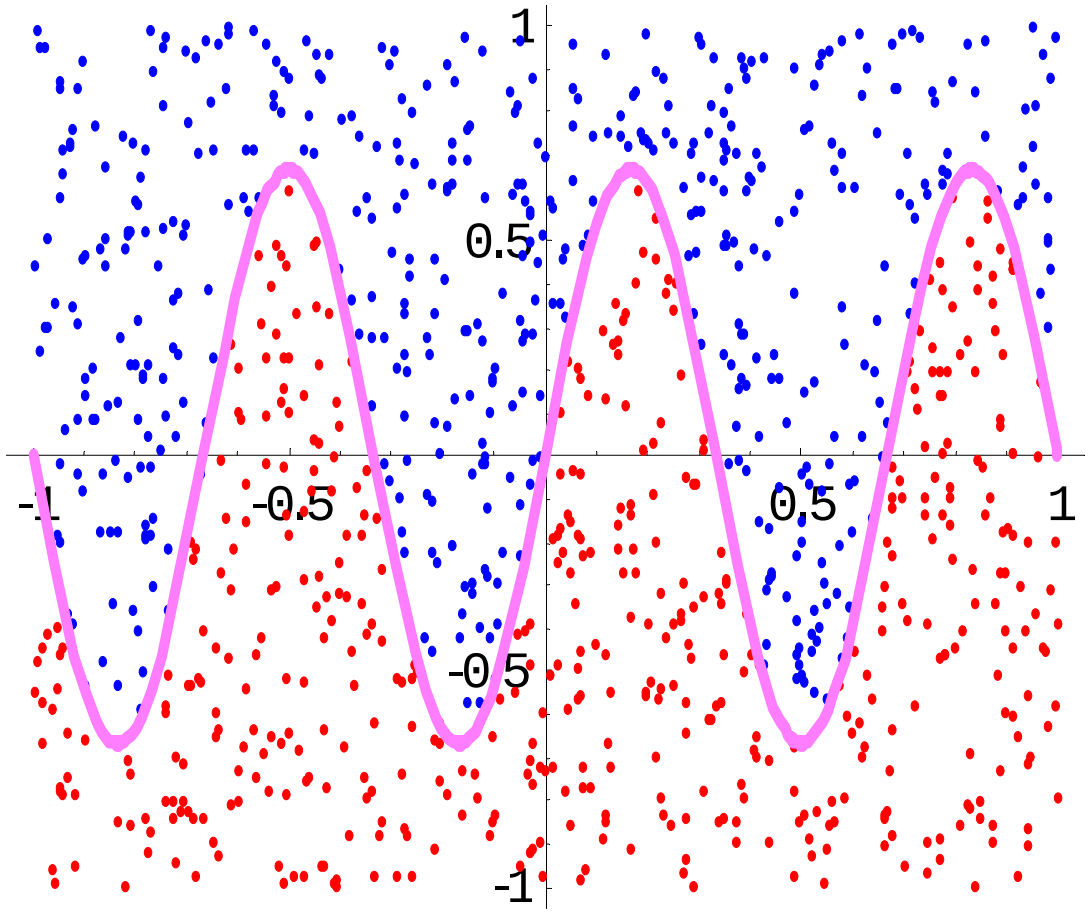
$$F_{\text{opt}}(\boldsymbol{x}) = \frac{1}{2} \log \frac{p(+1|\boldsymbol{x})}{p(-1|\boldsymbol{x})}$$

$$\mathbb{L}_{\exp}(F) - \mathbb{L}_{\exp}(F_{\text{opt}})$$

$$= \int \{e^{-F(\boldsymbol{x})} p(+1|\,\boldsymbol{x}) + e^{F(\boldsymbol{x})} p(-1|\,\boldsymbol{x}) - 2\sqrt{p(+1|\,\boldsymbol{x})\, p(-1|\,\boldsymbol{x})}\}\, f(\boldsymbol{x})\, d\boldsymbol{x}$$

$$= \int \{\sqrt{e^{-F(\boldsymbol{x})} p(+1|\,\boldsymbol{x})} - \sqrt{e^{F(\boldsymbol{x})} p(-1|\,\boldsymbol{x})}\}^2\, f(\boldsymbol{x}) d\boldsymbol{x} \geq 0$$

A simulation



Feature space

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

Decision boundary

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

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

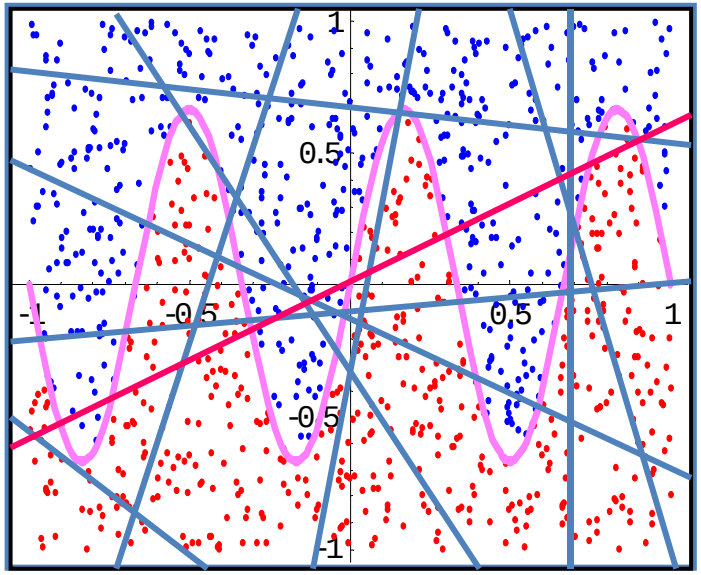
$\boldsymbol{x}_i \in [-1, 1] \times [-1, 1]$

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

Set of linear classifiers

Linear classifier

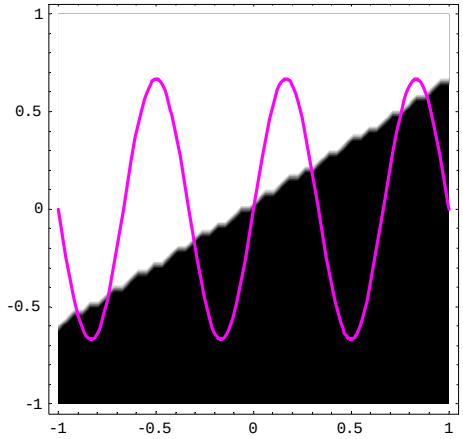
$$f(x_1, x_2) = \operatorname{sgn}(r_1 x_1 + r_2 x_2 + r_3) = \begin{cases} +1 & \text{if } r_1 x_1 + r_2 x_2 + r_3 \geq 0 \\ -1 & \text{if } r_1 x_1 + r_2 x_2 + r_3 < 0 \end{cases}$$



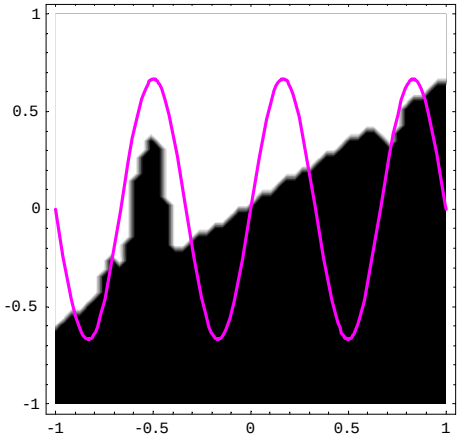
Random generation

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

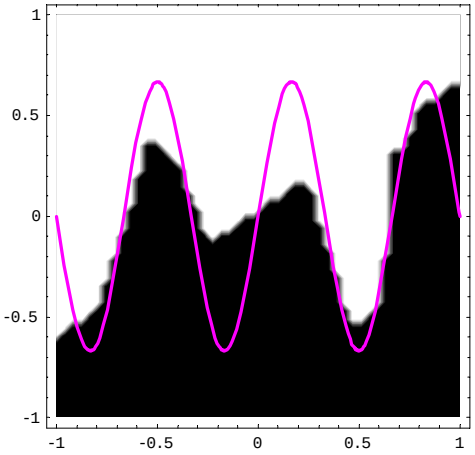
Learning process



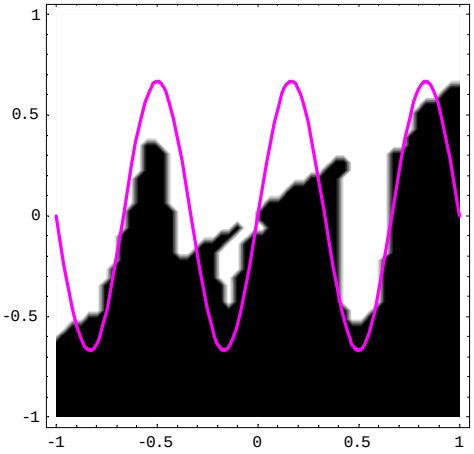
Iter = 1, train err = 0.21



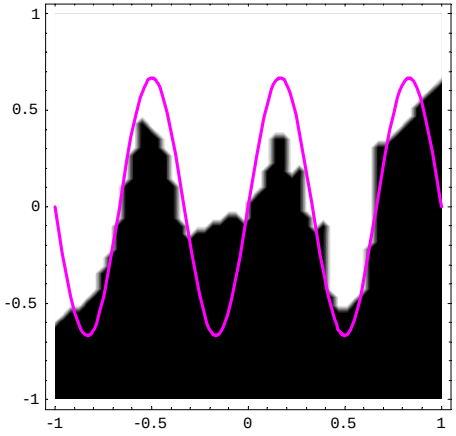
Iter = 13, train err = 0.18



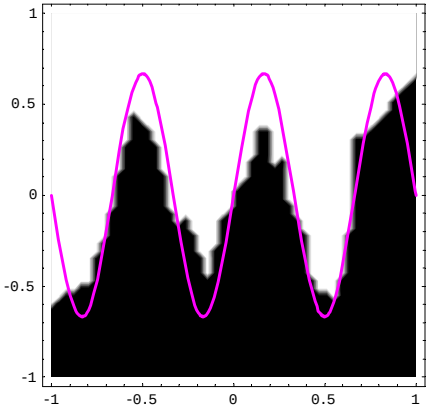
Iter = 17, train err = 0.10



Iter = 23, train err = 0.10

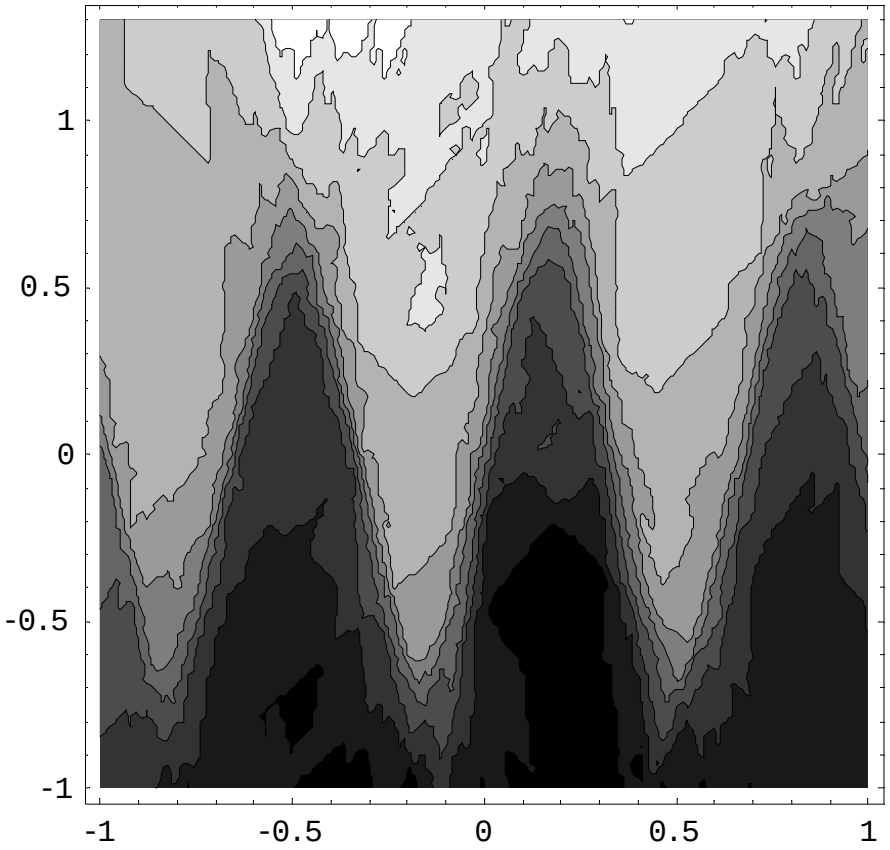


Iter = 31, train err = 0.095

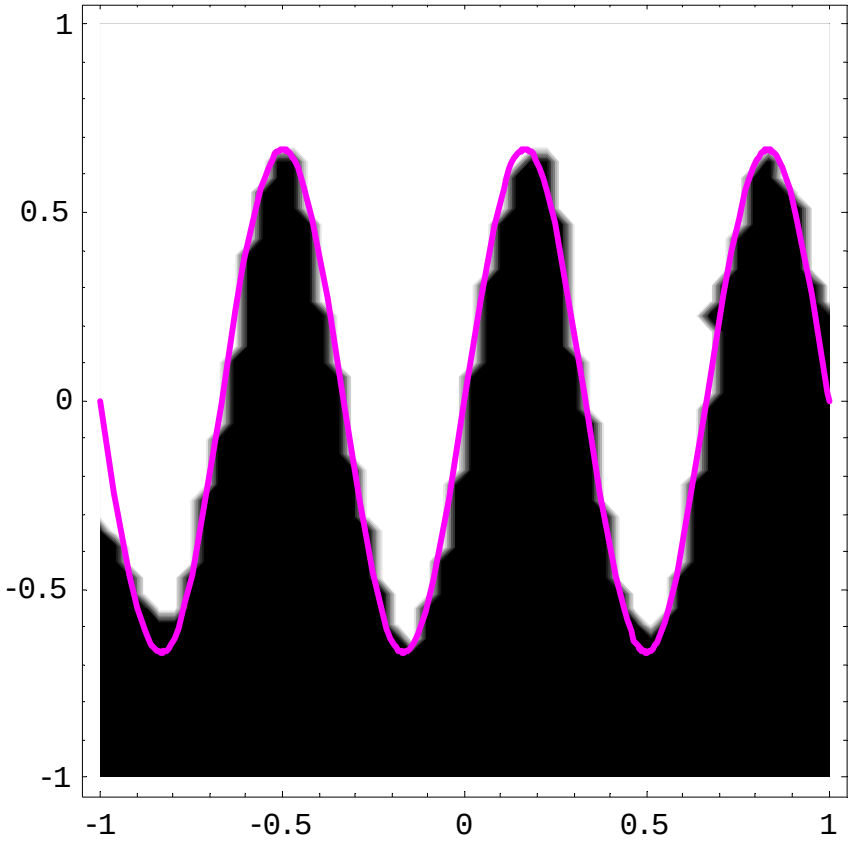


Iter = 47, train err = 0.08

Final decision



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



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

U-boost for classification

Let $(\boldsymbol{x}, y)$ be a variable with feature vector $\boldsymbol{x}$ and label y

Decision rule $$h(\boldsymbol{x}) = \underset{y \in \{1,\ldots,g\}}{\operatorname{argmax}}\ \hat{q}(y \mid \boldsymbol{x})$$

U-boost: $$q_t(y \mid \boldsymbol{x}) = u(\alpha_t f_t(\boldsymbol{x}, y) + \xi(q_{t-1}(y \mid \boldsymbol{x})))$$

Step 1. $$f_t = \underset{f}{\operatorname{argmin}} \varepsilon_t(f \mid q_{t-1})$$

Step 2. $$\alpha_t = \underset{\alpha}{\operatorname{argmin}}\ L_U^{\text{emp}}(\alpha f_t + \xi(q_{t-1}))$$

$$L_U^{\text{emp}}(\xi(q_t)) - L_U^{\text{emp}}(\xi(q_{t+1})) = D_U(q_{t+1}, q_t)$$

Murata *et al* (2004)

U-Boost for density estimation

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

U-Boost algorithm

(A) Find $f_1 = \arg\min_{g \in \mathscr{D}} L_U(g)$

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

$$(\alpha_{k+1}, g_{k+1}) = \arg\min_{(\alpha, g) \in (0,1) \times \mathscr{D}} L_U(\xi^{-1}((1-\alpha)\xi(f_k) + \alpha\xi(g)))$$

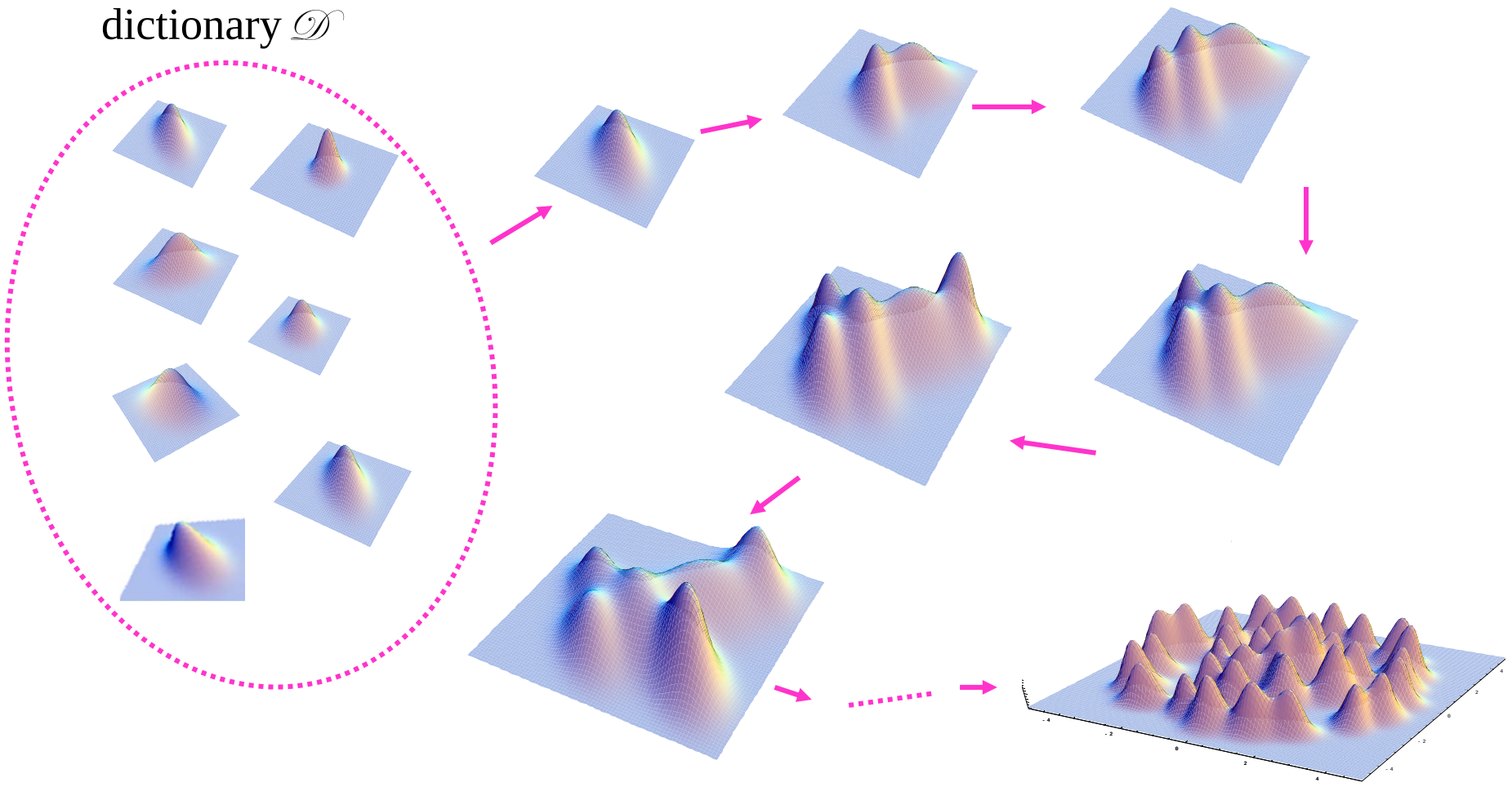
(C) Select K, and $\hat{f} = \xi^{-1}((1-\alpha_K)\xi(f_{K-1}) + \alpha_K \xi(g_K))$

Example 2. Power entropy $f^*(x) = \left(\sum_k \pi_k \, g_k(x)^\beta\right)^{\frac{1}{\beta}}$

If $\beta = 0$, $f^*(\boldsymbol{x}) = \exp\left(\sum_k \pi_k \, \log g_k(\boldsymbol{x})\right) = \prod_k g_k(\boldsymbol{x})^{\pi_k}$ Friedman et al (1984)

If $\beta = 1$, $f^*(\boldsymbol{x}) = \sum_k \pi_k \, g_k(\boldsymbol{x})$ Klemela (2007)

Learning?

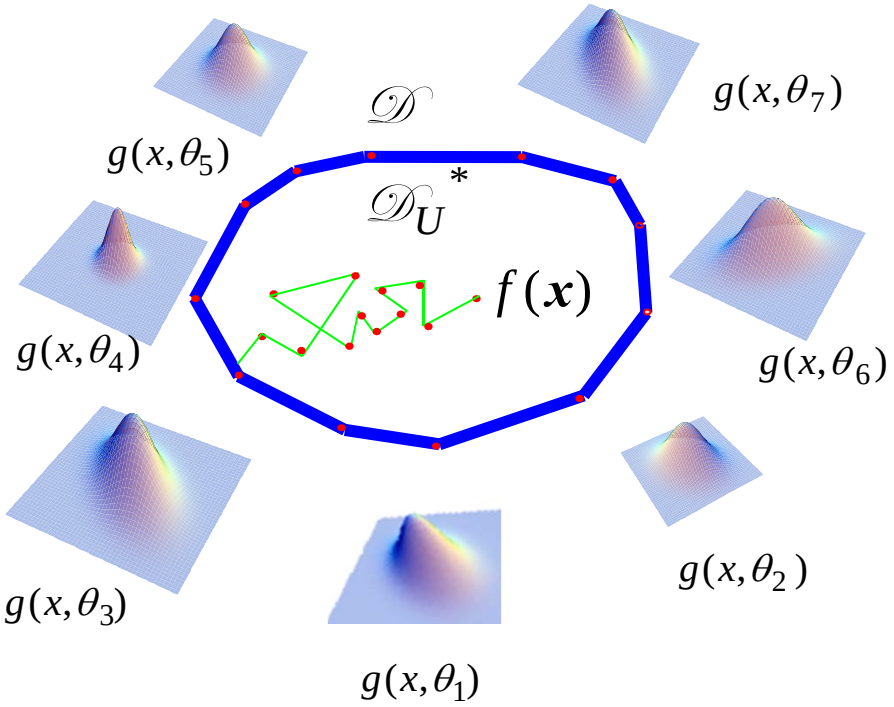


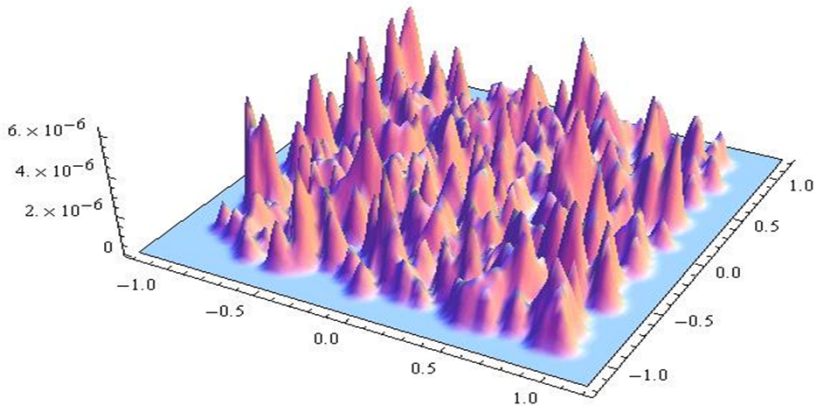
Boosting is a forward stagewise algorithm

Inner step in the convex hull

$$\mathscr{D} = \{g(x,\theta) : \theta \in \Theta\} \longrightarrow \mathscr{D}^* = \text{convex}(\mathscr{D})$$

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





$f(\boldsymbol{x})$

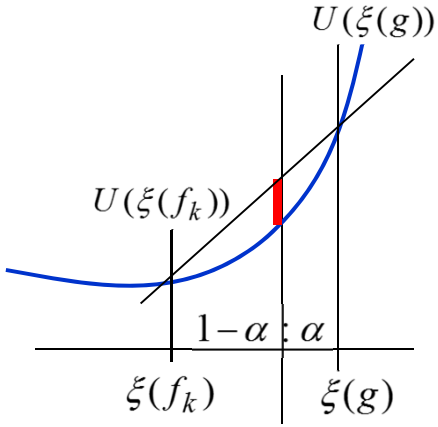
Loss decomposition

$$L_U(\xi^{-1}((1-\alpha)\xi(f_k)+\alpha\xi(g)))=(1-\alpha)L_U(f_k)+\alpha L_U(g)-\Delta_U(f_k,g,\alpha)$$

where $\Delta_U(f_k,g,\alpha)=\int\{(1-\alpha)U(\xi(f_k))+\alpha U(\xi(g))-U((1-\alpha)\xi(f_k)+\alpha\xi(g))\}$

Note 1. (1) $\Delta_U(f_k,g,\alpha)\geq 0,$

(2) $\Delta_U(f_k,g,\alpha)=0 \Leftrightarrow f_k=g$ (a.e.)



Min problem: $\min_{g\in\mathscr{D}}\{\alpha L_U(g)-\Delta_U(f_k,g,\alpha)\}$

Make $L_U(g)$ smaller, and $\Delta_U(f_k,g,\alpha)$ larger in g simultaneously.

Non-asymptotic bound

Naito and Eguchi (2012)

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

Then we have

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

where

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

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

$$\mathrm{IE}(K) = \frac{b_U}{K + c_U} \quad (b_U, c_U \text{ are constants}) \qquad \text{(Iteration effect)}$$

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

Statistical applications

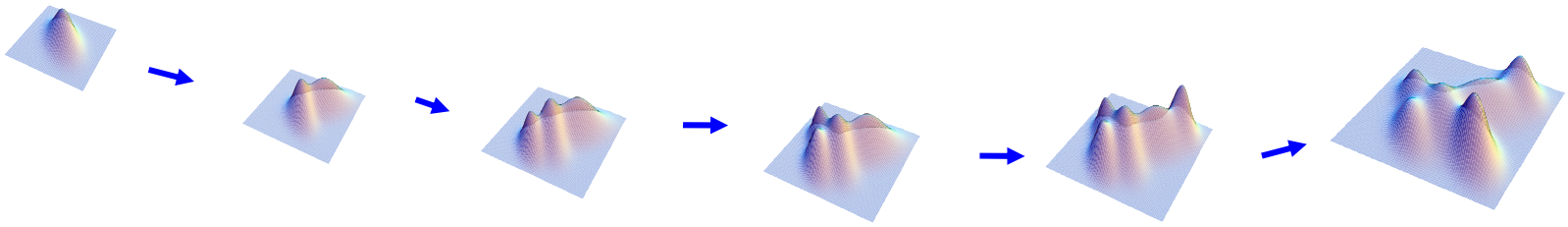
(kernel) *U*-PCA *U*-ICA (mixture)

U-Boost U-SVM U-AUC U-cluster

Robustness (outliers) (Exp-model, U-loss)

Redescendency (hidden structure) (U-model, U-estmate)

U-Boosting density Evolve (U-model, U-loss)



Drawback of Boosting

$(\hat{\delta}, \hat{f}_*)$ is suffered from over - learning by employing the same dataset D

because $$(\hat{\delta}, \hat{f}_*) = \underset{(\delta, f_*) \in F}{\arg\min} L_1((1-\delta)\hat{f}_t + \delta f_*, D)$$

- Bagging: $(\hat{\delta}, \hat{f}_*)$ is learned from bootstraped samples D^*'s and avaraging

  Breiman (1996)

- Predetermi nd mixing ratio as $\delta = (\delta_1, ..., \delta_t, ...)$

  Early stopping (Zhang-Yu,2005), Clemela (2008)

- Boosted Lasso (Zhao - Yu, 2004), Bühlmann (2006)

  Meinshausen and Buhlmann (2011), Stadle-Buhlmann-van de Geer (2010)

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

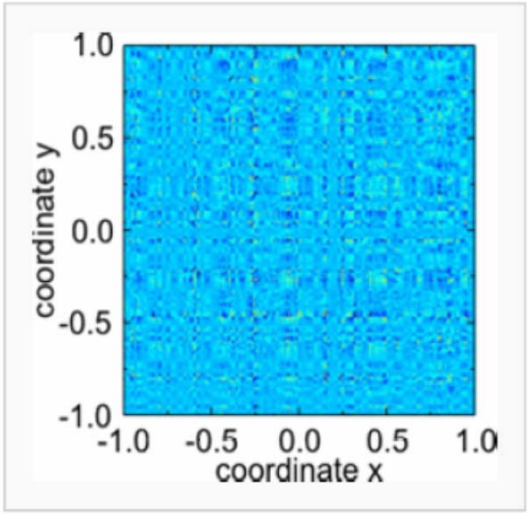
Geometers consider not a space, but a distribution family on the space.

Fisher's equation in 1937

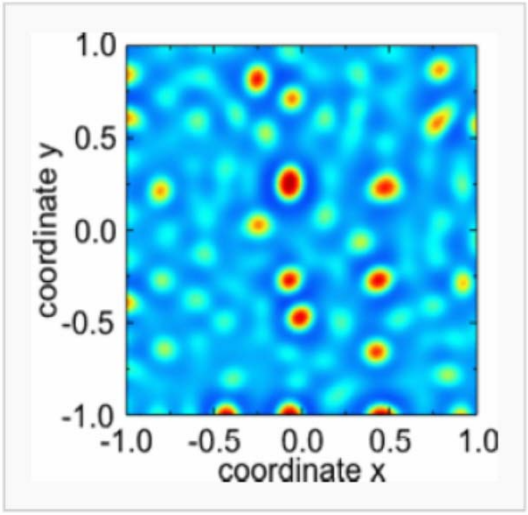
Fisher-Kolmogorov equation $\frac{\partial u}{\partial t} = \Delta u + u(1-u)$

Reaction–diffusion systems $\frac{\partial u}{\partial t} = \Delta u + f(u)$

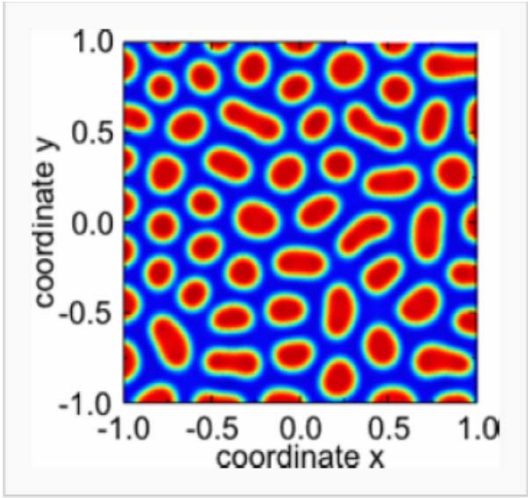
Ecology, physiology, combustion, crystallization, plasma physics, phase transition problems cf. Wiki (reaction diffusion process)



Noisy initial conditions at $t = 0$.



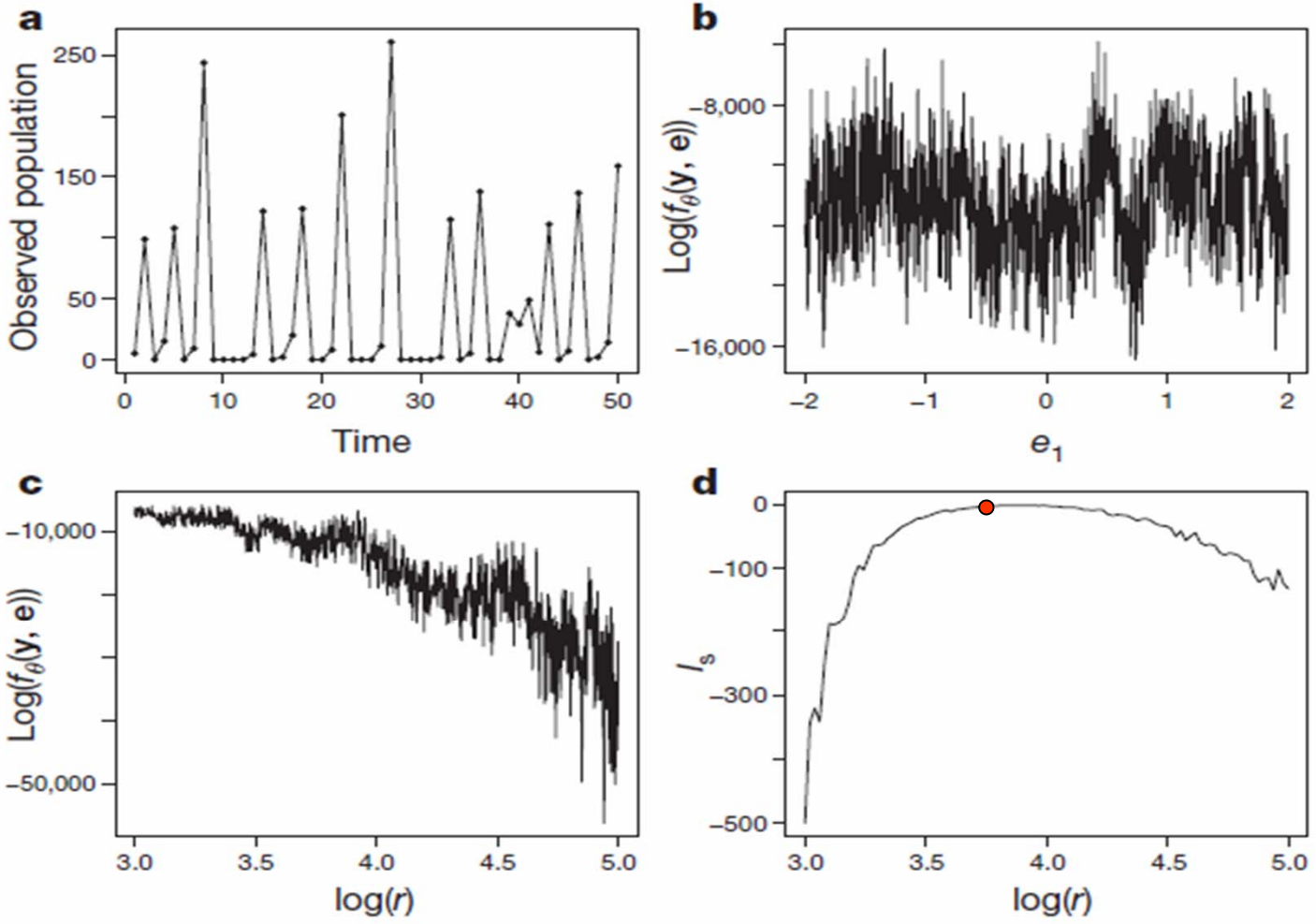
State of the system at $t = 10$.



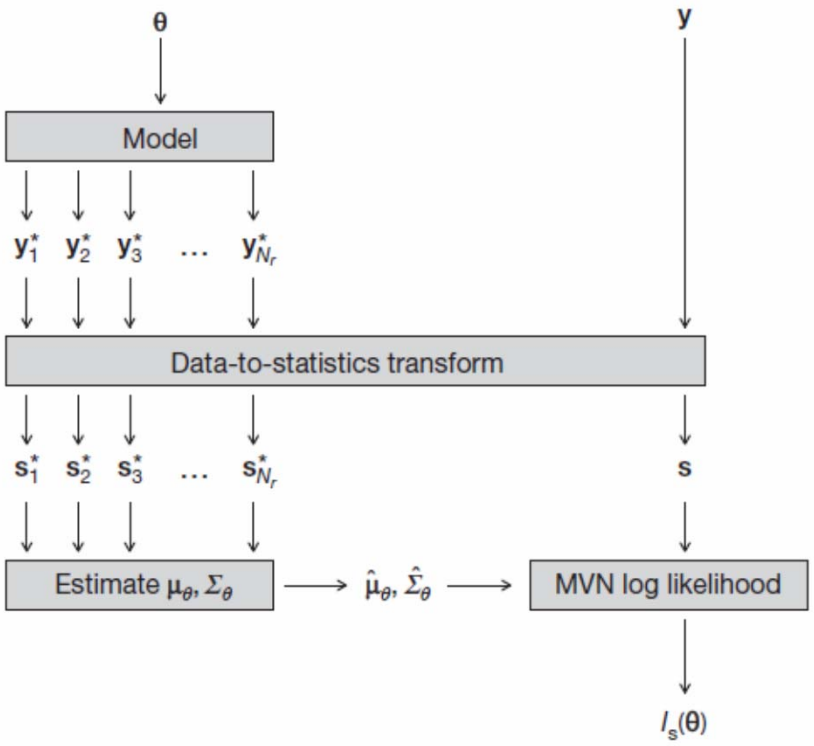
Almost converged state at $t = 100$.

Statistical inference for noisy nonlinear ecological dynamic systems

Simon N. Wood



Statistical algorithm



Sufficient statistic

$$\boldsymbol{s} = (s_1, s_2, s_3)^{\mathrm{T}}$$

$$y_t^{\alpha} = s_1 y_{t-1}^{\alpha} + s_2 y_t^{2\alpha} + s_3 y_t^{3\alpha} + \varepsilon_t$$

Simulated curved normal model

Ricker map $N_{t+1} = rN_t \exp(-N_t + e_t)$

where $\{e_t\}$ are independent $N(0, \sigma_e^2)$

Sampling $y_t \sim \text{Poisson}(\phi N_t)$

Statistical parameter $\boldsymbol{\theta} = (r, \sigma_e^2, \phi)$ defines $f(\boldsymbol{y}, \boldsymbol{e}, \boldsymbol{\theta})$

$$\boldsymbol{\theta} \underset{\text{simulate}}{\longrightarrow} (\boldsymbol{y}_1^*, \ldots, \boldsymbol{y}_M^*) \longrightarrow (\boldsymbol{s}_1^*, \ldots, \boldsymbol{s}_M^*)$$

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} = \sum_{i=1}^{M} \boldsymbol{s}_i^*, \quad \boldsymbol{V}_{\boldsymbol{\theta}} = \sum_{i=1}^{M} (\boldsymbol{s}_i^* - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}})(\boldsymbol{s}_i^* - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}})^{\mathrm{T}}$$

$$L(\boldsymbol{\theta}) = -\frac{1}{2}(\boldsymbol{s} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}})^{\mathrm{T}} \boldsymbol{V}_{\boldsymbol{\theta}}^{-1} (\boldsymbol{s} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}) - \frac{1}{2} \log \det(2\pi \boldsymbol{V}_{\boldsymbol{\theta}})$$

Conclusions and Future directions

Release Fisher's mind control