Singularity Theory in Statistical Science

The 15th MSJ-SI

Deepening and Evolution of Applied Singularity Theory

Workpia Yokohama, 20-24, Nov. 2022

Sumio Watanabe

Tokyo Institute of Technology

Paper

Sumio Watanabe, Recent Advances in Algebraic Geometry and Bayesian Statistics.

To appear in Information Geometry.

Contents

1 Singularities in Statistical Science

In statistics, “all models are wrong”.

Models and priors are fictional candidates :

$X_i \sim p(x|w)$? : a statistical model

$w \sim \varphi(w)$? : a prior distribution

In order to find some useful model and prior,

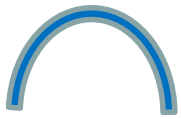
we need mathematical theory.

Distinguish model and DGP

Fictional candidates : $\{X_i\}$ is exchangeable

$X_i \sim p(x|w)$?

$w \sim \varphi(w)$?



(De Finetti Theorem) If $\{X_i\}$ is exchangeable,

then Unknown true data-generating process (DGP) exists.

$X_i \sim q(x)$: unknown distribution

$q \sim Q(q)$: unknown functional distribution

Likelihood function

Likelihood function

$$L(w) = \Pi_{i=1}^{n} \; p(X_i|w)$$

Log Likelihood function

$$\log L(w) = \Sigma_{i=1}^{n} \; \log p(X_i|w)$$

Definitions

w_0 : minimizes $- \mathbf{E}_X [\log p(x|w)]$

$$f(x,w) = \log (p(X_i|w_0) / p(X_i,|w))$$

$$K(w) = \mathbf{E}_X [f(X,w)] >= 0$$

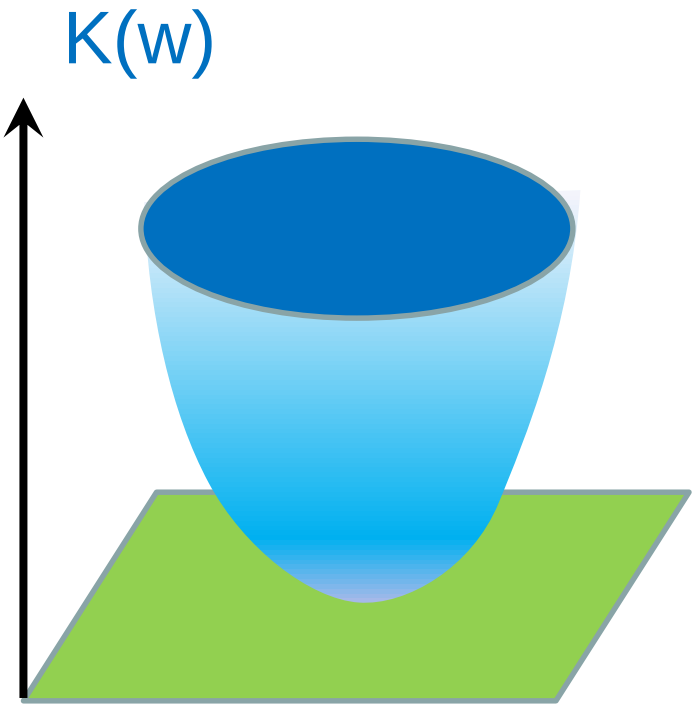
Likelihood Ratio function

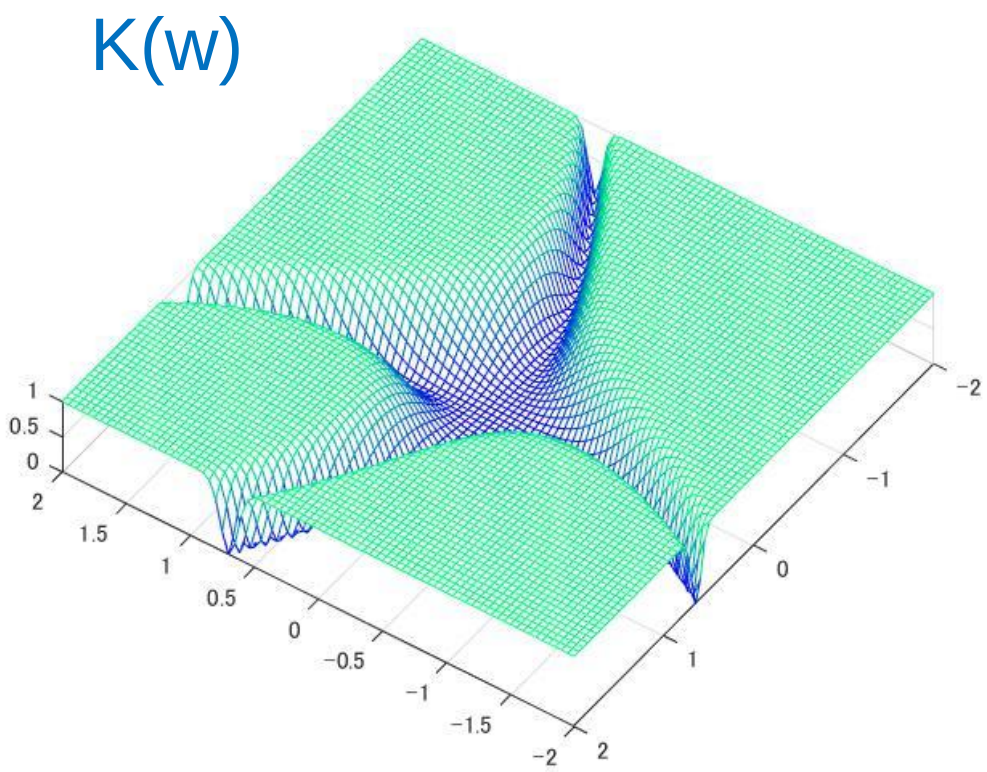
$$L(w) / L(w_0) = \exp(- H_n (w))$$

$$H_n (w) = \Sigma_{i=1}^{n} f(X_i,w)$$

Log likelihood ratio $H_n (w)$ concerns statistical problems.

Regular and Singular



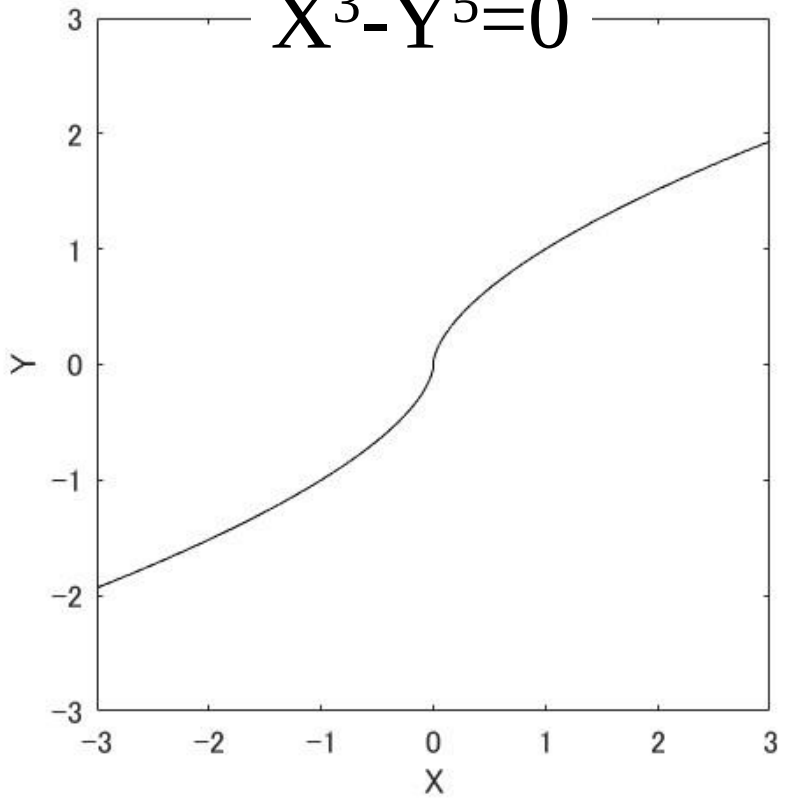


Classical Models :
linear regression,
normal distribution, …

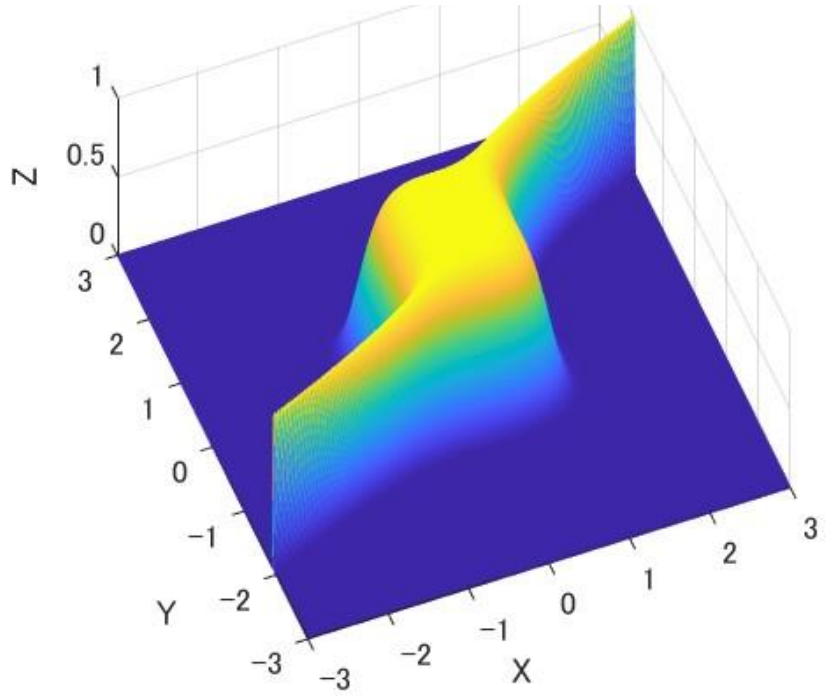
Modern models :
deep learning,
normal mixtures, …

Singularity determines integral

$X^3-Y^5=0$



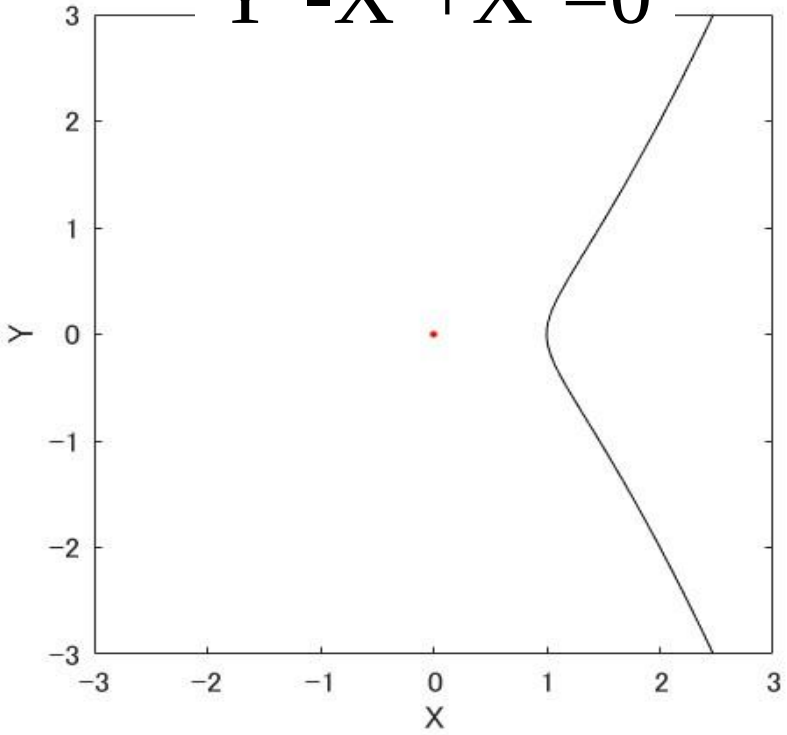
$\exp(-(X^3-Y^5)^2)$



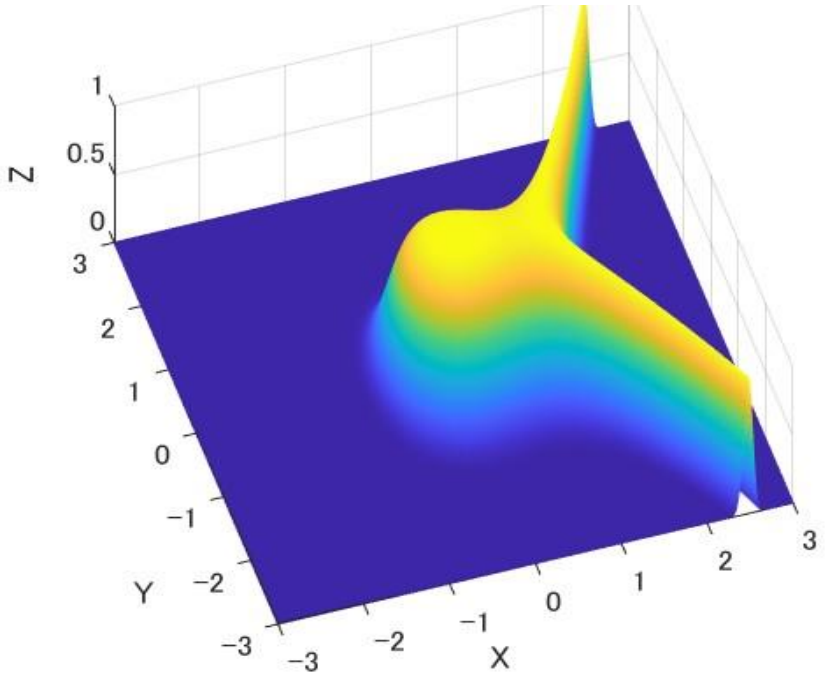
If K(w)=0 contains singularities, integration of exp(-nK(w)) is determined by singularities.

Singularity determines integral

$Y^2-X^3+X^2=0$



$\exp(-(Y^2-X^3+X^2)^2)$



We need singularity theory for evaluating these integral.

Example : Matrix factorization

$p(X|A,B) = (1/2\pi)^2 \exp(- ||X-AB||^2 /2)$,

$$A = \begin{pmatrix} a \\ b \end{pmatrix} \quad , \quad B = \begin{pmatrix} c & d \end{pmatrix}.$$

If $q(x)=p(X|0,0,0,0)$, then $\{ w ; K(w)=0 \}$ is

$\{(a,b,c,d) \text{ in } \mathbf{R}^4;\ a^2c^2+a^2d^2+b^2c^2+b^2d^2=0\}$.

Two Mathematical Problems

Log likelihood ratio

$$H_n(w) = nK(w) - (nK(w))^{1/2} \sum_{i=1}^{n} \left\{ \frac{K(w) - f(X_i, w)}{(nK(w))^{1/2}} \right\}$$

Fluctuation function

(1) $\{w ; K(w)=0\}$ contains singularities

(2) Fluctuation function is not well-defined at $K(w)=0$.

⇒ Many statistical problems were left unresolved.

2 Resolution of Singularities

Hironaka Resolution Theorem 1964

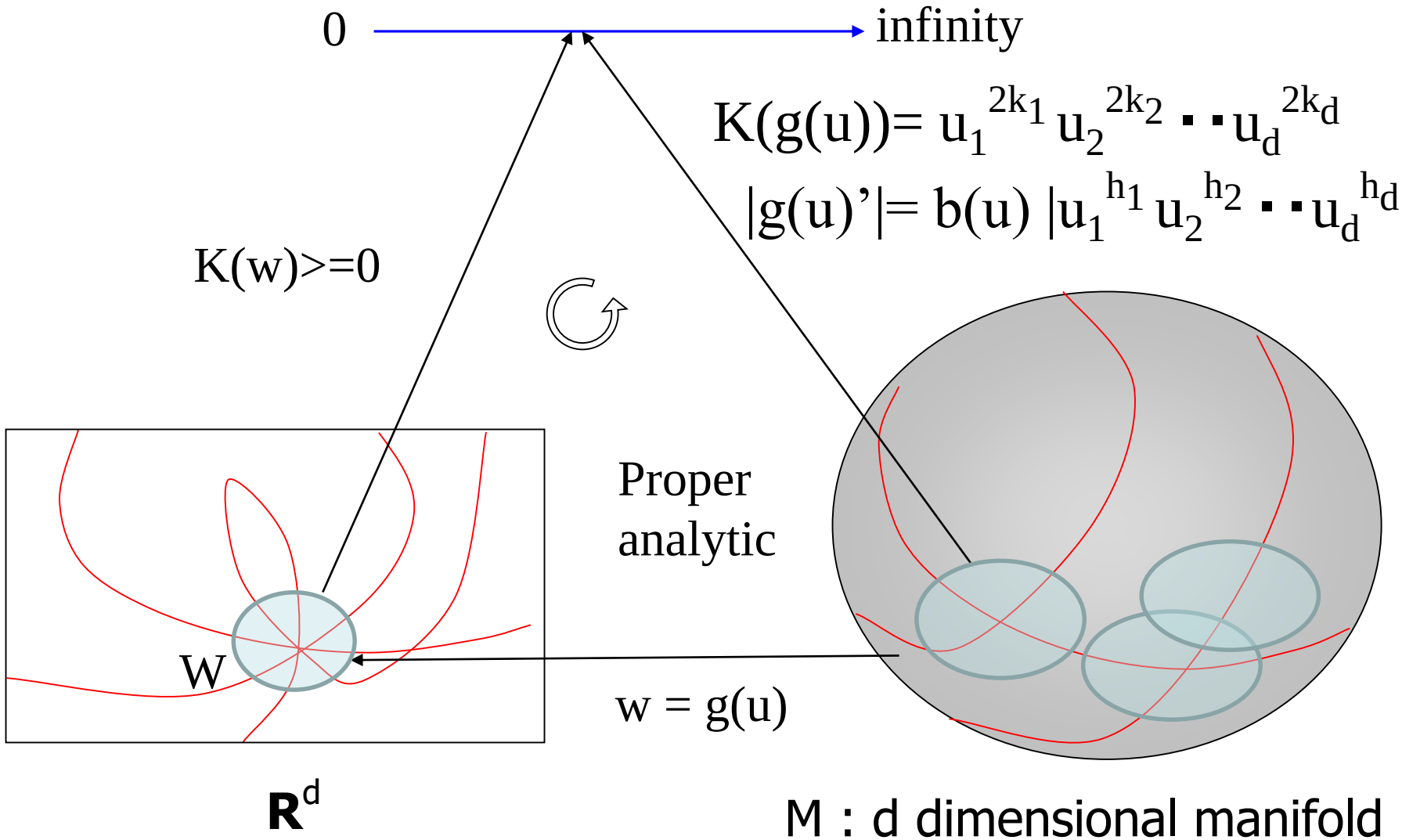
Theorem. Let K(w) be a nonnegative analytic function from an open set W in $\mathbf{R}^d$ to **R**, which satisfies $K(w_0)=0$ for some w_0 in W. Then there exist a d-dimensional analytic manifold M and a proper analytic map w=g(u) from M to W, such that, in each local coordinate of M,

$$K(g(u)) = u_1^{2k_1} u_2^{2k_2} \ldots u_d^{2k_d} ,$$

$$|g'(u)| = b(u) \, | \, u_1^{h_1} u_2^{h_2} \ldots u_d^{h_d} \, |,$$

where b(u)>0 and $(k_1,k_2,\ldots,k_d)$ and $(h_1,h_2,\ldots,h_d)$ are sets of nonnegative integers which depend on a local coordinate (at least one $k_i>0$) .

Hironaka Theorem (Resolution of Singularities)



Definition : Real Log Canonical Threshold

Definition. Assume that an analytic function K(w) has a resolution of singularities in each local coordinate,

$$K(g(u)) = u_1^{2k_1} u_2^{2k_2} \cdots u_d^{2k_d}$$

$$|g(u)'| = b(u)\,|u_1^{h_1} u_2^{h_2} \cdots u_d^{h_d}|, \quad (b(u)>0)$$

The real log canonical threshold and its multiplicity are defined by (Lo.co. means local coordinate)

RLCT $$\lambda = \min_{\text{Lo.co.}} \min_{j=1,2,\ldots,d} (h_j+1)/(2k_j)$$

multiplicity $$m = \max_{\text{Lo.co.}} \#\{\ j\ ;\ \lambda = (h_j+1)/(2k_j)\ \}.$$

Note : for $k_j=0$, $(h_j+1)/(2k_j)$ is defined as infinity.

Example : a statistical model

Let w=(a,b,c) be a parameter. A statistical model

$$p(y|x,w) = (1/2\pi)^{1/2}\exp(- \{ y- a\, s (bx) - cx \}^2/2),$$

is studied, where $s(x) = x+ x^2$.

Assume that

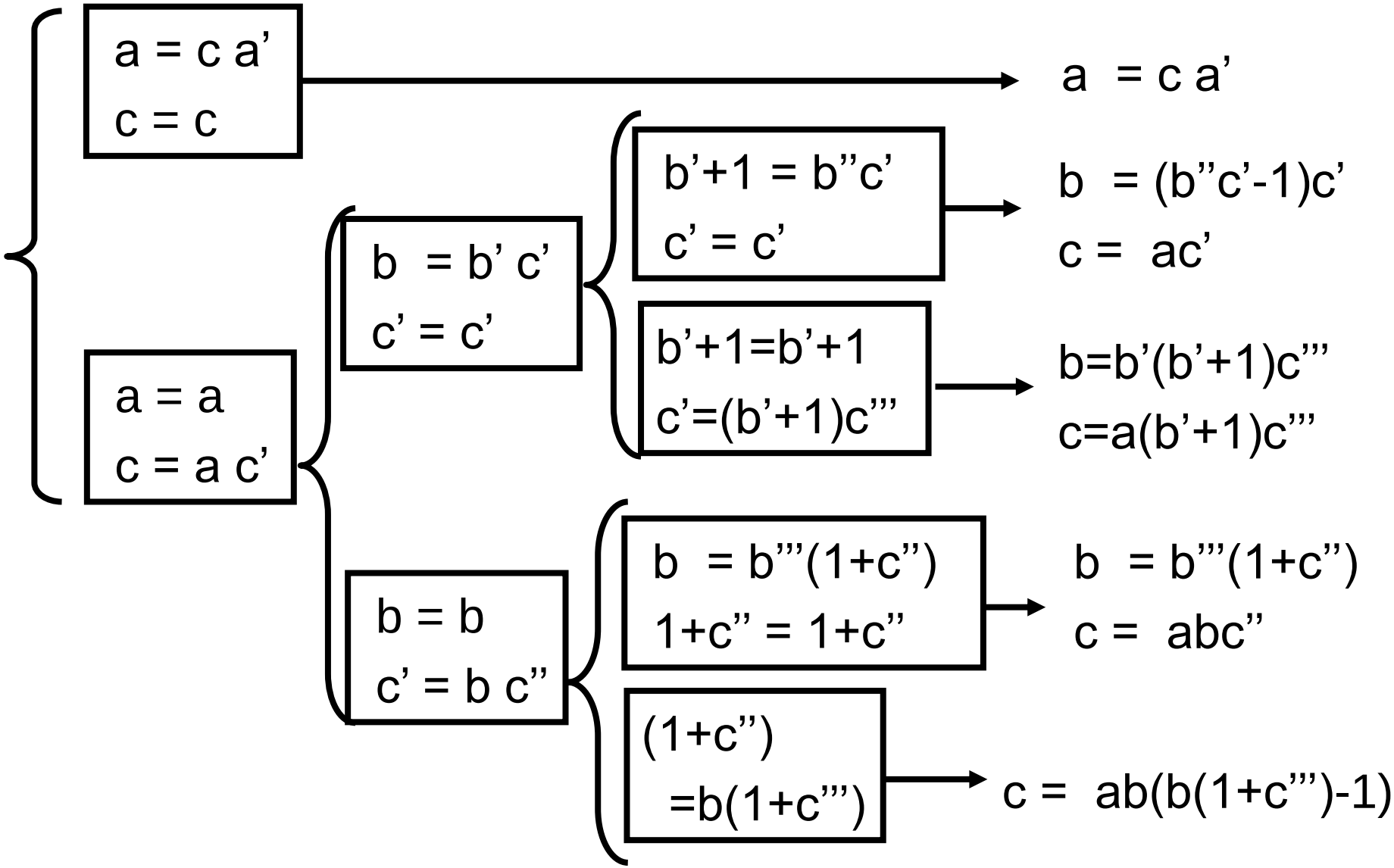
True : $N(x|0,1)\ p(y|x,(0,0,0))$,

Model : $p(y|x,w)$.

Then KL distance is

$$K(a,b,c) = \{(ab+c)^2+3a^2b^4\}/2.$$

Example : recursive blowups



Example : RCLT and multiplicity

RLCT multiplicity		K(g(u)) : normal crossing	$\|g(u)'\|$
local 1, 1	RLCT =3/4, Multi = 1	$2K = c^2 \{(a'b+1)^2+3a'^2b^4\}$	$\|c\|$
local 3/4, 1		$2K = a^2 c'^4 \{(b'')^2+3(b''c'-1)'^4\}$	$\|ac'^2\|$
local 3/4, 1		$2K = a^2 (b'+1)^4 c'''^2 \{1+3b'^4\}$	$\|ac'''(b'+1)^2\|$
local 1, 3		$2K = a^2 b'''^2 (1+c'')^2\{1+3b'''^2\}$	$\|ab'''(1+c'')\|$
local 3/4, 1		$2K = a^2 b^4 \{(1+c'')^2+3)$	$\| ab^2\|$

3 Two mathematical solutions

Two Mathematical Solutions (1)

(1) Asymptotic expansion of

$$\varphi(w) \exp(- n K(w))$$

(2) Weak Convergence of Fluctuation function

$$\Sigma_{i=1}^{n} \left\{ \frac{K(w) - f(X_i, |w)}{(nK(w))^{1/2}} \right\}$$

Zeta function ois Statistics

Definition. Gel'fand zeta function of a statistical model 1954 is defined by

$$\zeta(z) = \int K(w)^z \, \varphi(w) \, dw,$$

where z in **C**.

In Re(z)>0, $\zeta(z)$ is a holomorphic function, which can be analytically continued to a unique meromorphic function onto the entire complex plane (M.F.Atiyah, 1970).

Bridge between algebraic geometry and statistics

Resolution Theorem

$$K(g(u))= \Pi\ (u_j)^{2k_j}$$

Analytic cont. of Zeta func.

$$\int K(w)^{z}\ \varphi(w)\ dw \quad = \quad c_1\ /(z+\lambda)^{m}$$

Inverse Mellin transform

$$\int \delta(\ t - K(w)\ )\ \varphi(w)\ dw\ = c_2\ t^{\ \lambda\ -1}(-\log\ t)^{m-1}$$

Laplace transform

$$\int \exp(\ -\ n\ K(w)\ )\ \varphi(w)\ dw = c_3(\log\ n)^{\ m-1}\ /n^{\ \lambda}$$

Example : Singular Schwartz distribution

$$\frac{n^{\lambda}}{(\log n)^{m-1}}\ \delta(t- n x^4 y^6 z^8)\ x^1 y^2 z^6 \rightarrow \frac{1}{24}\ t^{\lambda-1}\ \delta(x)\delta(y) z^2$$

holds on $[0,1]^3$, where

$\lambda = \min\{ (1+1)/4, (2+1)/6, (6+1)/8 \} = 1/2$

$m = 2$

Two Mathematical Solutions (2)

(1) Asymptotic expansion of

$$\varphi(w) \exp(- n K(w))$$

(2) Weak Convergence of Fluctuation function

$$\Sigma_{i=1}^{n} \left\{ \frac{K(w) - f(X_i, |w)}{(nK(w))^{1/2}} \right\}$$

Relatively finite variance and Factor Theorem

Assume that f(x,w) is an $L^r(q)$-valued analytic has a relatively finite variance,

$$K(w)=\mathbf{E}_X[\, f(X,w)\,] >= c_0\, \mathbf{E}_X[\, f(X,w)^2\,].$$

By using normal crossing property of K(g(u)),

$$K(g(u)) = u^{2k} = u_1^{2k_1} u_2^{2k_2} \cdots u_d^{2k_d}.$$

It follows that $u^{2k} >= c_0\, E_X[\, f(X,g(u))^2\,]$. By using factor theorem, there exists an $L^r(q)$-valued analytic function a(x,u) such that

$$f(x,g(u)) = a(x,u)\, u^k.$$

Empirical Process

Definition. An empirical process is defined by

$$\xi_n(u) = (1/n^{1/2}) \sum_{i=1}^{n} \{ \mathbf{E}[a(X,u)] - a(X_i,u) \},$$

which is a C(M)- valued random variable, where C(M) is a set of all continuous function on a compact subset M.

Then we can show the convergence in distribution

$$\xi_n (w) \quad \text{-- >} \quad \xi (w)$$

in C(M).

4 Three Applications to Statistics

Three Applications to Statistics (1)

(1) Renormalized Posterior Distribution

(2) Free energy

(3) Generalization loss and Estimators

Average by Posterior Distribution

The average of an arbitrary function $\Psi(w)$ by the posterior distribution is defined by

$$E_w[\,\Psi(w)\,] = \frac{\int \Psi(w)\,\varphi(w)\,\Pi_i\, p(X_i|w)\,dw}{\int \varphi(w)\,\Pi_i\, p(X_i|w)\,dw}\,.$$

$$= \frac{\int \Psi(w)\,\varphi(w)\,\exp(-nK_n(w))\,dw}{\int \varphi(w)\,\exp(-nK_n(w))\,dw}\,.$$

Posterior Distribution (2)

The posterior distribution is written by using w=g(u),

$$\mathbf{E}_w[\Psi(w)] =$$

$$\frac{\Sigma \int_{[0,1]^d} \Psi(g(u)) \exp(- n u^{2k} + n^{1/2} u^k \xi_n(u)) b(u) u^h du}{\Sigma \int_{[0,1]^d} \exp(- n u^{2k} + n^{1/2} u^k \xi_n(u)) b(u) u^h du}$$

The above expectation is denoted by $\mathbf{E}_u[\Psi(g(u))]$.

Posterior Average of log density ratio function

Recall that $f(x,w) = u^k a(x,u)$.

$$\mathbf{E}_w[\, f(x,w)^\alpha \,] = \mathbf{E}_u[\, f(x,g(u)) \,] =$$

$$\frac{\Sigma \int_{[0,1]^d} u^{\alpha k} a(x,u)^\alpha \exp(- n u^{2k} + n^{1/2} u^k \xi_n(u)) b(u) u^h \, du}{\Sigma \int_{[0,1]^d} \exp(- n u^{2k} + n^{1/2} u^k \xi_n(u)) b(u) u^h \, du}$$

Asymptotics of Schwartz Distriution

By using asymptotic property of singular Schwartz distribution,

$$\mathbf{E}_w[f(x,w)^{\alpha}] =$$

$$\frac{\sum \dfrac{(\log n)^{m-1}}{n^{\lambda+\alpha/2}} \iint_{[0,1]^d} a(x,u)^{\alpha}\, t^{\alpha/2}\, t^{\lambda-1} D(u) \exp(-t+t^{1/2}\xi_n(u))\,du\,dt}{\sum \dfrac{(\log n)^{m-1}}{n^{\lambda}} \iint_{[0,1]^d} t^{\lambda-1}\, D(u)\, \exp(-t+t^{1/2}\xi_n(u))\,du\,dt}$$

$$+ \text{smaller order term}$$

Asymptotic behavior of Posterior Average

At last, we derived

$$\mathbf{E}_w[\, f(x,w)^{\alpha} \,] = \frac{1}{n^{\alpha/2}} \left(\frac{\Sigma \iint_{[0,1]^d} a(x,u)^{\alpha}\, t^{\alpha/2}\, t^{\lambda-1} D(u) \exp(-t+t^{1/2}\xi_n(u))\, du\, dt}{\Sigma \iint_{[0,1]^d} t^{\lambda-1}\, D(u)\, \exp(-t+t^{1/2}\xi_n(u))\, du\, dt} \right) + \text{smaller order term}$$

Note. Only local coordinates in which local RCLT is smallest and local multiplicity is largest remain in the main term.

Renormalized Posterior Average

Definition. Let $\Psi(u,t)$ be an arbitrary function on $g^{-1}(W)$ times $\mathbf{R}$. Then the average of $\Psi(u,t)$ by the renormalized posterior distribution is defined by

$$< \Psi(u,t) | \xi > = \left(\frac{\Sigma \int\int_{[0,1]^d} \Psi(u,t)\, t^{\lambda-1} D(u) \exp(- t + t^{1/2} \xi(u)) du dt}{\Sigma \int\int_{[0,1]^d} t^{\lambda-1} D(u) \exp(- t + t^{1/2} \xi(u)) du dt} \right)$$

Asymptotic Posterior Distribution

Theorem. Let $\alpha>0$. The following convergence in distribution holds as n tends to infinity.

$$n^{\alpha/2}\ \mathbf{E}_w[\ f(x,w)^{\alpha}\ ]\ \text{converges to}\ \ <\ a(x,u)^{\ \alpha}\ t^{\alpha/2}\ |\ \xi\ >.$$

Definition. The singular fluctuation is defined by using the renormalized posterior distribution,

$$\mathrm{Fluc}(\xi) = \mathbf{E}_x[\ <t\ a(X,u)^2|\ \xi\ > - <t^{1/2}a(X,u)\ |\ \xi\ >^2\ ]\ ,$$

$$\nu\ = (1/2)\ \mathbf{E}_{\xi}[\ \mathrm{Fluc}(\xi)\ ]\ .$$

Three Applications to Statistics (2)

(1) Renormalized Posterior Distribution

(2) Free energy

(3) Generalization loss and Estimators

Free Energy

Free energy (=minus log marginal likelihood) is defined by

$$F_n = -\log \int \varphi(w) \prod p(X_i|w)\, dw.$$

Remark: In statistics, the free energy is often employed for model and prior evaluation. However, to calculate Fn Numerically needs heavy computational cost.

Asymptotic Behaviors of Random Variables

Theorem. Let λ and m be RLCT and its multiplicity.

It follows that

$$F_n = n L_n + \lambda \log n - (m-1) \log\log n + O_p(1),$$

In regular case, $\lambda = d/2$, $m=1$, which is called BIC (1978).

Widely Applicable BIC

By $\beta* = 1/\log n$, WBIC is defined by

$$\mathrm{WBIC} = \frac{\int \{ - \Sigma \log p(X_i|w) \} \varphi(w) \Pi\, p(X_i|w)^{\beta*} \varphi(w)\, dw}{\int \varphi(w) \Pi\, p(X_i|w)^{\beta*} \varphi(w)\, dw}$$

It follows that

$$F_n = \mathrm{WBIC} + O_p((\log n)^{1/2}).$$

This property is used for calculating the free energy.

sBIC （singular BIC)

Drton-Plummer (2017) proposed a method by which RLCT is estimated, and the singular BIC is defined by

$$\mathrm{sBIC} = n L_n(w^*) + \lambda \log n.$$

If a true distribution is realizable by a statistical model, and if MLE satisfies $nL_n(w^*) = o_p(\log n)$,

$$F_n = \mathrm{sBIC} + o_p(\log n).$$

Three Applications to Statistics (3)

(1) Renormalized Posterior Distribution

(2) Free energy

(3) Generalization loss and Estimators

Definition of Generalization Loss

Definition. The generalization loss is defined by

$$G_n = - \mathbf{E}_x \log p(X \mid X^n)$$

Remark : In statistics, the smaller G_n means the more precise prediction by a statistical model and a prior. However, G_n needs the expectation over the unknown data-generating distribution.

Estimators of Generalization Loss

Definition. The training loss, leave-one-out cross validation, and widely applicable information criterion are defined by

$$T_n = - (1/n) \sum_{i=1}^{n} \log p(X_i \mid X^n)$$

$$C_n = - (1/n) \sum_{i=1}^{n} \log p(X_i \mid X^n - X_i)$$

$$W_n = T_n + (1/n) \sum_{i=1}^{n} \mathbf{V}_w[\log p(X_i|w)],$$

Asymptotic Losses

Theorem. Asymptotic behaviors are given by

$$G_n = L_0 + (1/2n)\{ 2\lambda + \langle t^{1/2}\xi_n(u)\rangle - \mathrm{Fluc}(\xi_n)\} + o_p(1/n),$$
$$T_n = L_n + (1/2n)\{ 2\lambda - \langle t^{1/2}\xi_n(u)\rangle - \mathrm{Fluc}(\xi_n)\} + o_p(1/n),$$
$$C_n = L_n + (1/2n)\{ 2\lambda - \langle t^{1/2}\xi_n(u)\rangle + \mathrm{Fluc}(\xi_n)\} + o_p(1/n),$$
$$W_n = L_n + (1/2n)\{ 2\lambda - \langle t^{1/2}\xi_n(u)\rangle + \mathrm{Fluc}(\xi_n)\} + o_p(1/n).$$

Their averages are given by

$$\mathbf{E}[\,G_n\,] = L_0 + \lambda / n + o(1/n),$$
$$\mathbf{E}[\,T_n\,] = L_0 + (\lambda - 2\nu) / n + o(1/n),$$
$$\mathbf{E}[\,C_n\,] = L_0 + \lambda / n + o(1/n),$$
$$\mathbf{E}[\,W_n\,] = L_0 + \lambda / n + o(1/n).$$

Example.1 Reduced Rank Regression

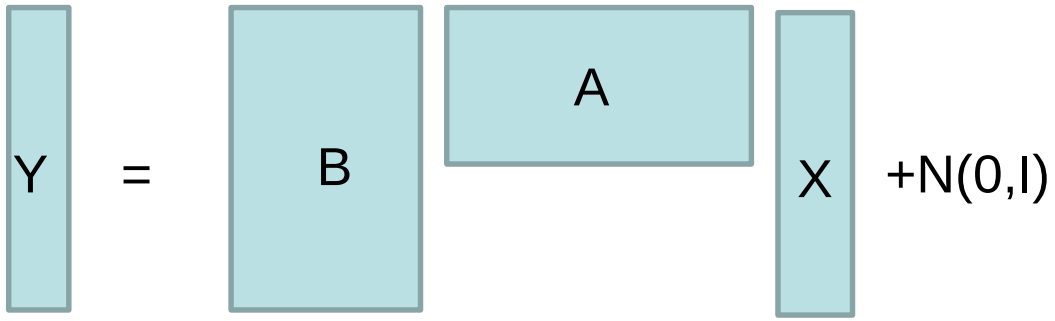
X in $\mathbf{R}^M$, Y in $\mathbf{R}^N$, B in $\mathbf{R}^{NH}$, A in $\mathbf{R}^{HM}$

A statistical model of reduced rank regression is defined by

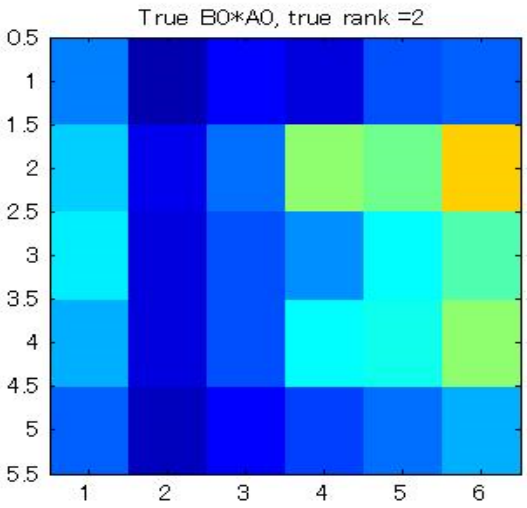
model $p(y|x,A,B) = (1/2\pi)^{N/2} \exp(-(1/2)|| y-BAx||^2)$

prior $\varphi(A,B)$: positive on sufficiently large set.

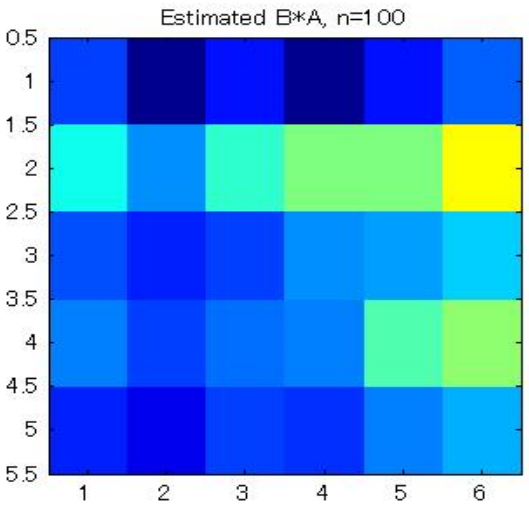
true $q(y|x)=p(y|x,A_0,B_0)$, $q(x) =N(x,0,I)$.



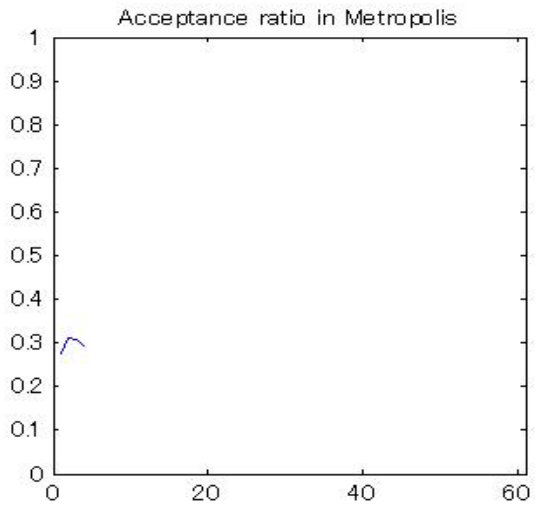
True : B_0A_0 5 times 6
rank(B_0A_0)=2



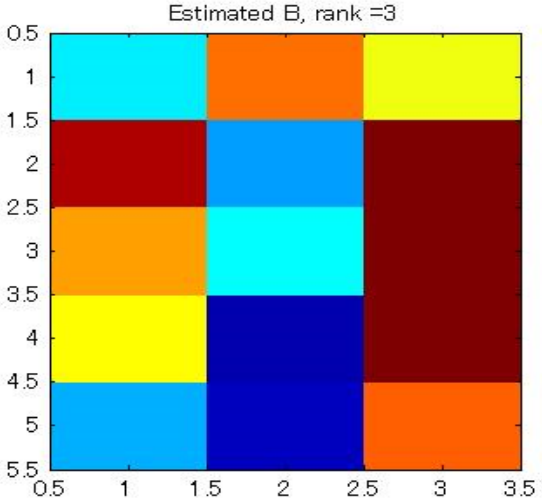
Posterior
Average (BA)



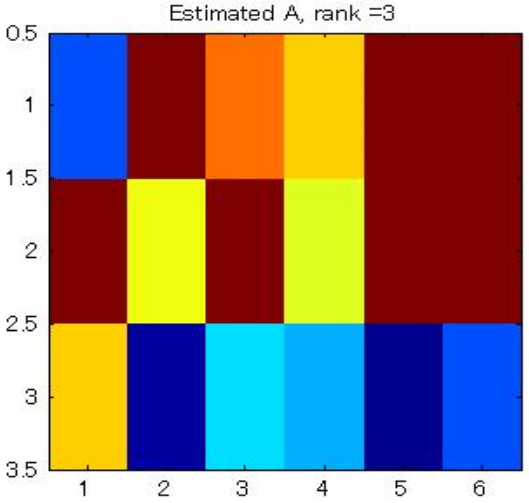
Acceptance
Prob in MCMC



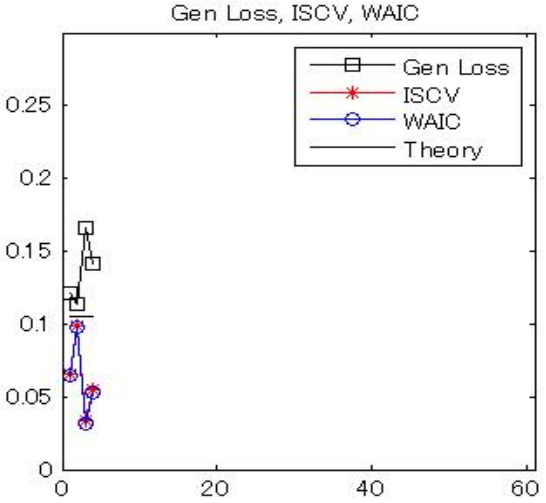
B(5 times 3)



A(3 times 6)



G, C, and W



http://watanabe-www.math.dis.titech.ac.jp/users/swatanab/red_rank_reg.mp4

Conclusion

Singularity theory in statistical science was reported.

1. It has been difficult to study statistical problems if statistical models contain singularities.

2. Resolution theorem gives two mathematical solutions and three statistical applications.

These theoretical results are now being used in statistical science and machine learning.