Two Birational Invariants in Statistical Learning Theory

JSPS Forum, Strasbourg, August, 2009

The 5th Franco-Japanese Symposium on Singularities

Sumio Watanabe

Tokyo Institute of Technology

Contents

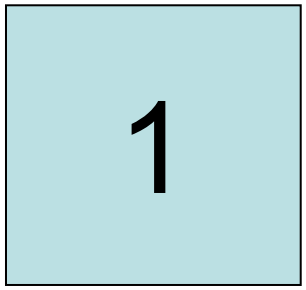
(1) Statistical Learning Theory

(2) Log Canonical Threshold

(3) Singular Fluctuation

(4) Main theorem

(5) Application to statistics



Statistical Learning Theory

Definitions

Information $x : R^N$ Parameter $w : R^d$

(1) True distribution $q(x)\,dx$

(2) Testing sample X

(3) Training samples $X_1, X_2, \ldots, X_n$

(4) Statistical model $p(x|w)$

(5) A priori distribution $\varphi(w)\,dw$

Statistical Learning

Expectation by a posteriori distribution

$$E_w[\, F(w) \,] = \frac{\int F(w) \prod_{i=1}^{n} p(X_i|w)\, \varphi(w)\, dw}{\int \prod_{i=1}^{n} p(X_i|w)\, \varphi(w)\, dw}$$

Estimated distribution

$$p^*(x) = E_w[\, p(x|w) \,]$$

Generalization and Training Errors

How accurate $p^*(x)$ for $q(x)$?

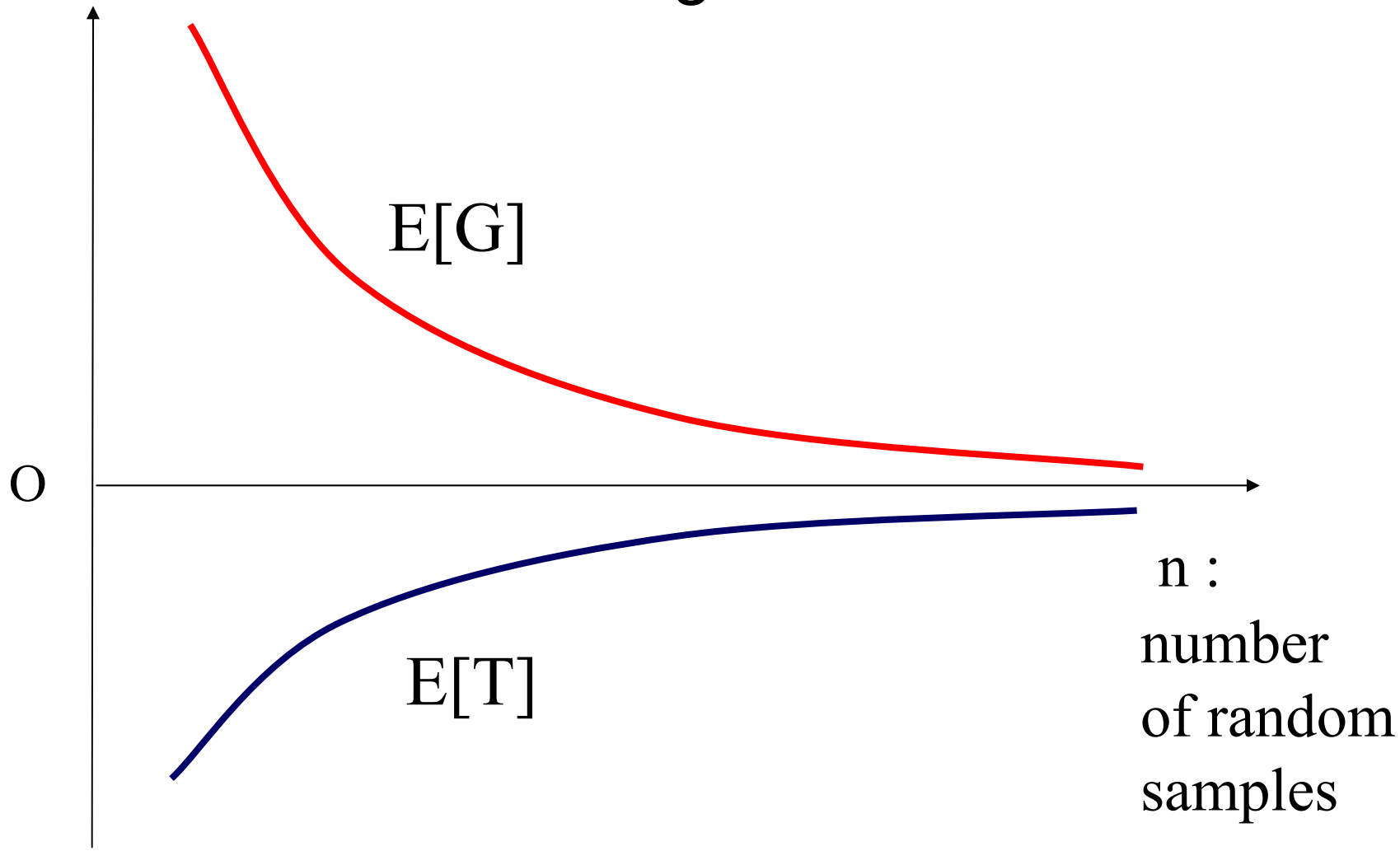
Generalization Error

$$G = \int q(x) \log \left(q(x) / p^*(x) \right) dx$$

Training Error

$$T = (1/n) \sum_{i=1}^{n} \log \left(q(X_i) / p^*(X_i) \right)$$

Learning Curves



Regular and Singular

A statistical model is called regular,

if mapping $w \mapsto p(\ |w)$ is one-to-one,

and if Fisher information matrix is positive definite.

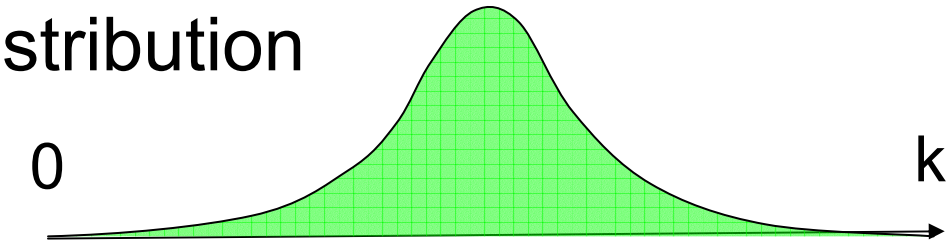
If otherwise it is called singular.

In regular models, $E[G] = d/2n$, $E[T] = - d/2n$.

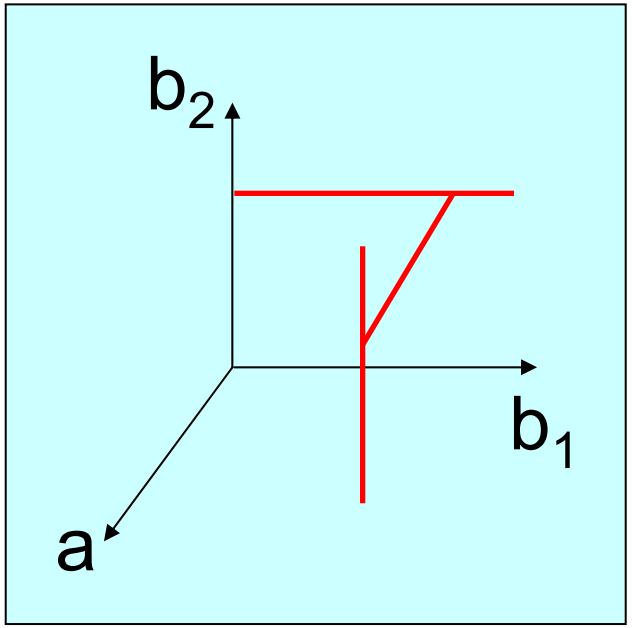
In singular models, they have been left unknown.

Example of a singular model

True Distribution

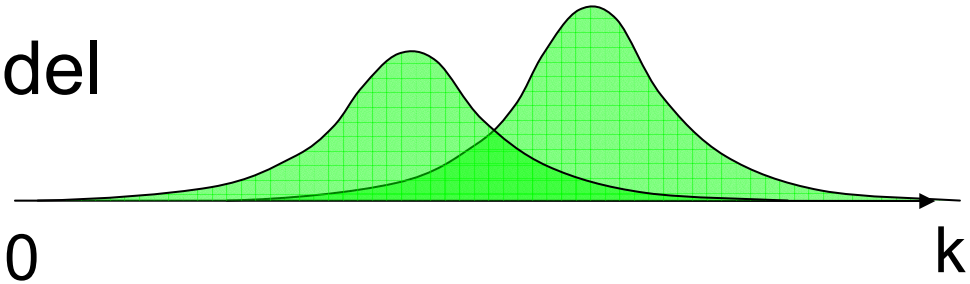


$G(x-b_0)$

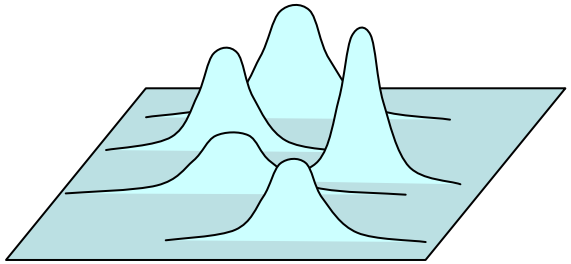


$a\,G(x-b_1) + (1-a)\,G(x-b_2)$

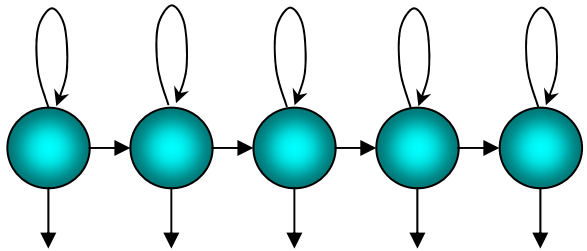
Model



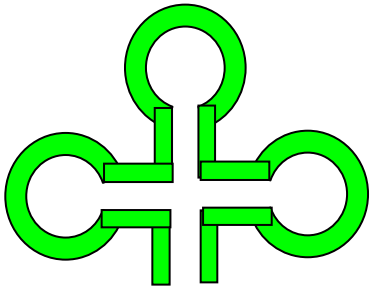
Examples of Singular Statistical Models



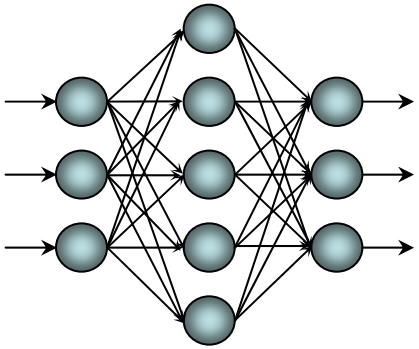
Normal Mixture



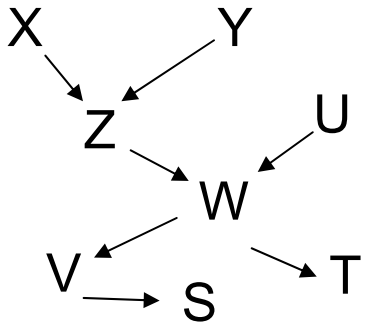
Hidden Markov Model



Stochastic CFG



Neural Networks



Bayesian Network

If a statistical model contains hierarchical modules, hidden variables, or grammatical structure, it is singular.

Kullback-Leibler distance

Log density ratio function

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

Kullback-Leibler (KL) distance

$$K(w) = E[f(X,w)]$$

Empirical KL distance,

$$K_n(w) = \frac{1}{n} \sum_{i=1}^{n} f(X_i,w)$$

Two problems of posterior distribution

The posterior distribution can be rewritten as

$$E_w[\, F(w) \,] = \frac{\int F(w) \exp(-nK_n(w)) \varphi(w)\, dw}{\int \exp(-nK_n(w)) \varphi(w)\, dw}$$

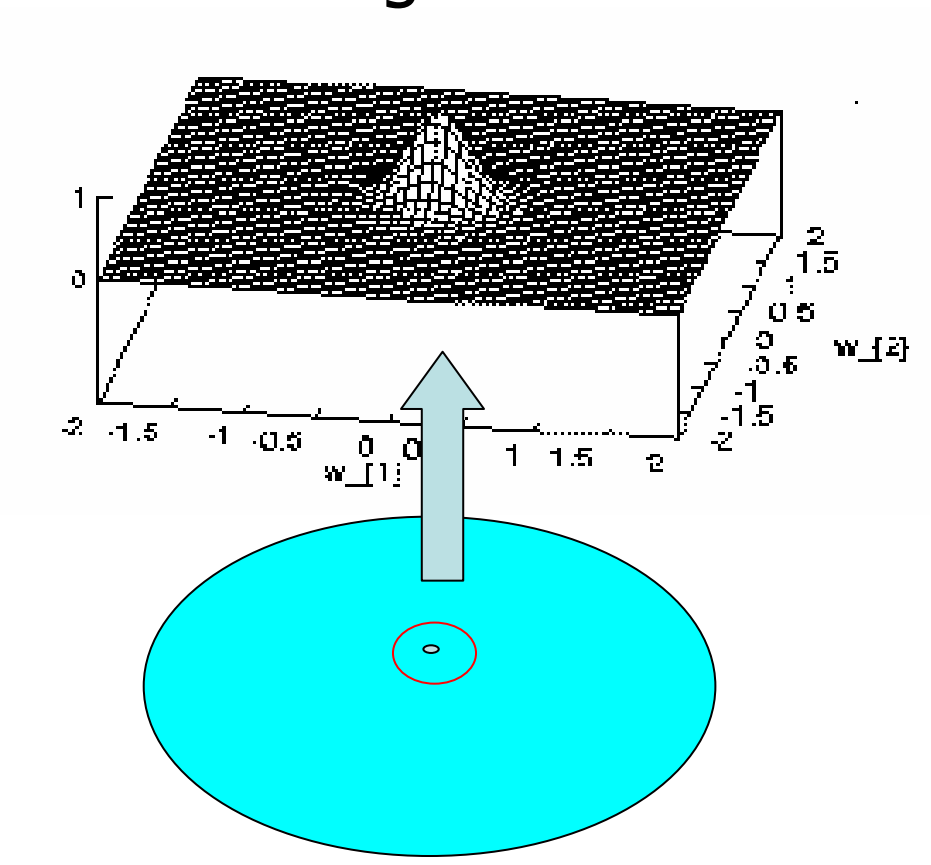
$K(w) = 0$ is an analytic set with singularities.

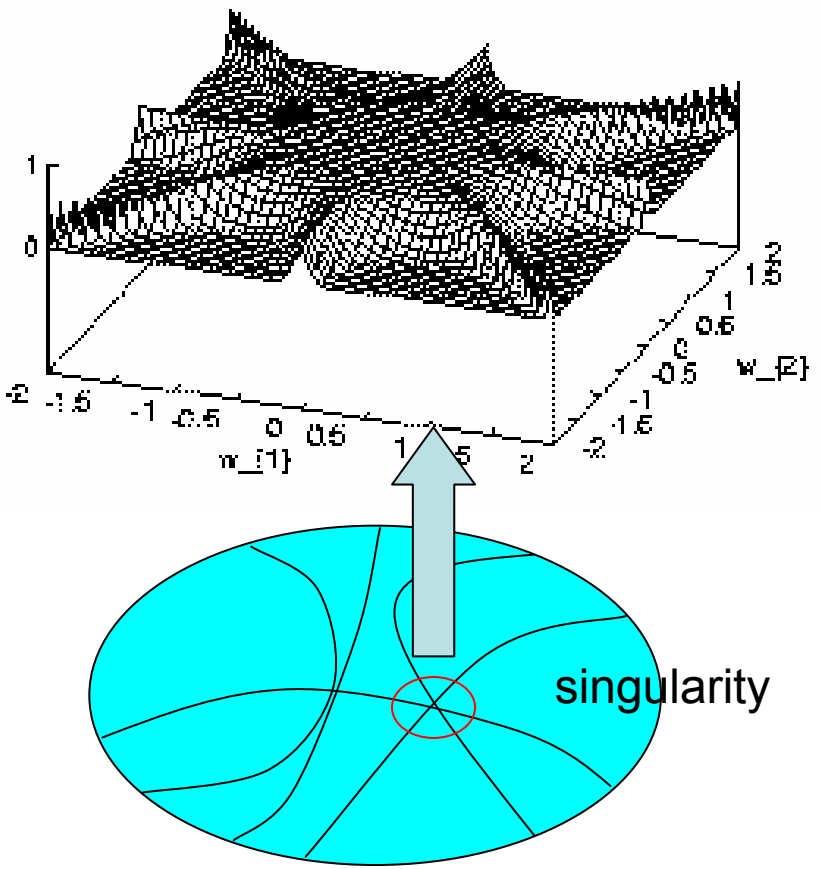
We need
- How fast $\exp(-nK(w)) \longrightarrow 0$.
- Fluctuation of $K_n(w) - K(w)$.

Very different posterior distributions

regular

singular





2 Log Canonical Threshold

Zeta function

State density function

$$s(t) = \int \delta(t-K(w))\, \varphi(w)\, dw$$

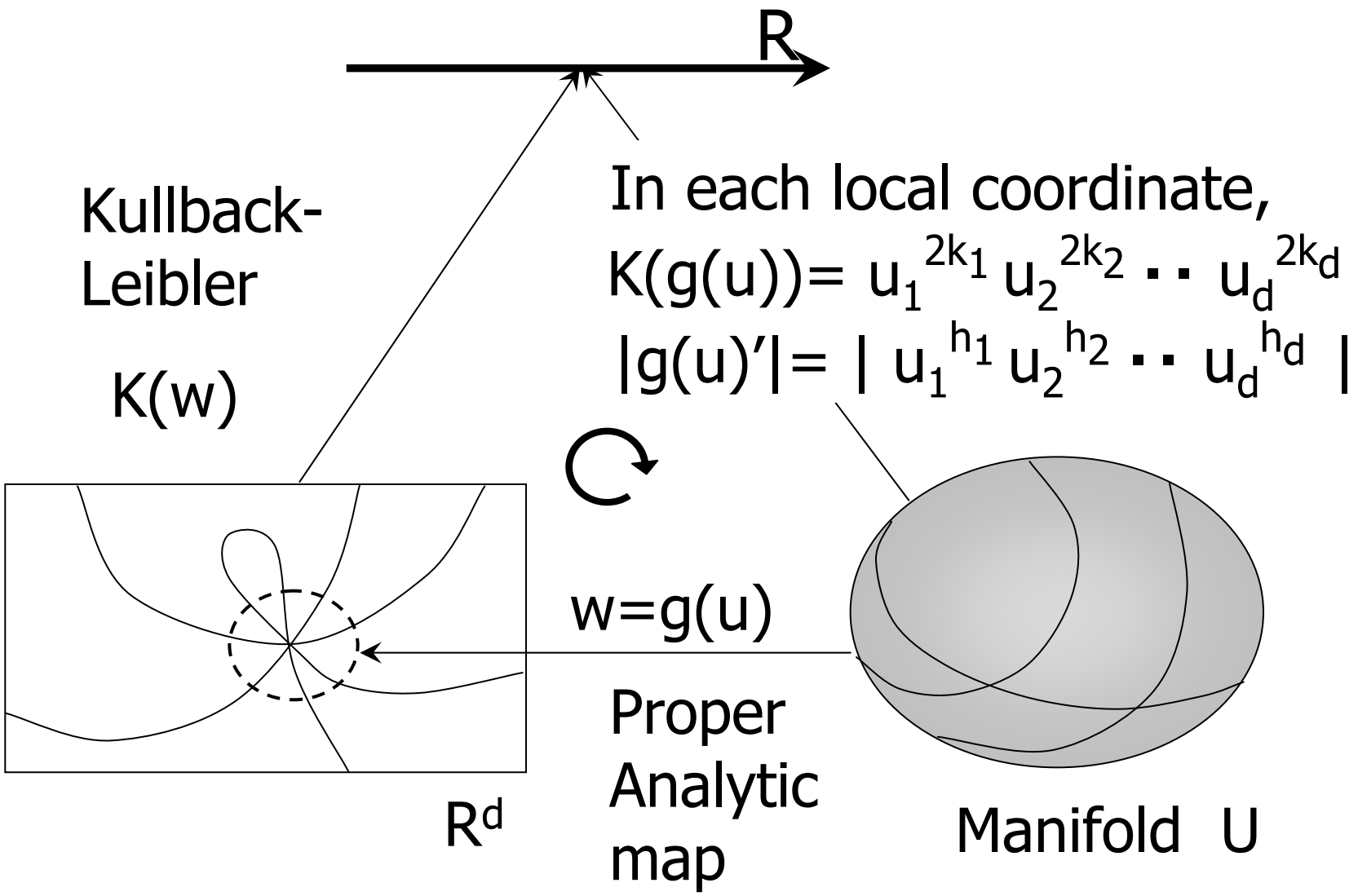
Posterior is Laplace transformation of s(t),

$$Z(n) = \int \exp(-nK(w))\, \varphi(w)\, dw$$

Zeta is Mellin transformation of s(t),

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

Application of Hironaka's Theorem (1964) to Statistics



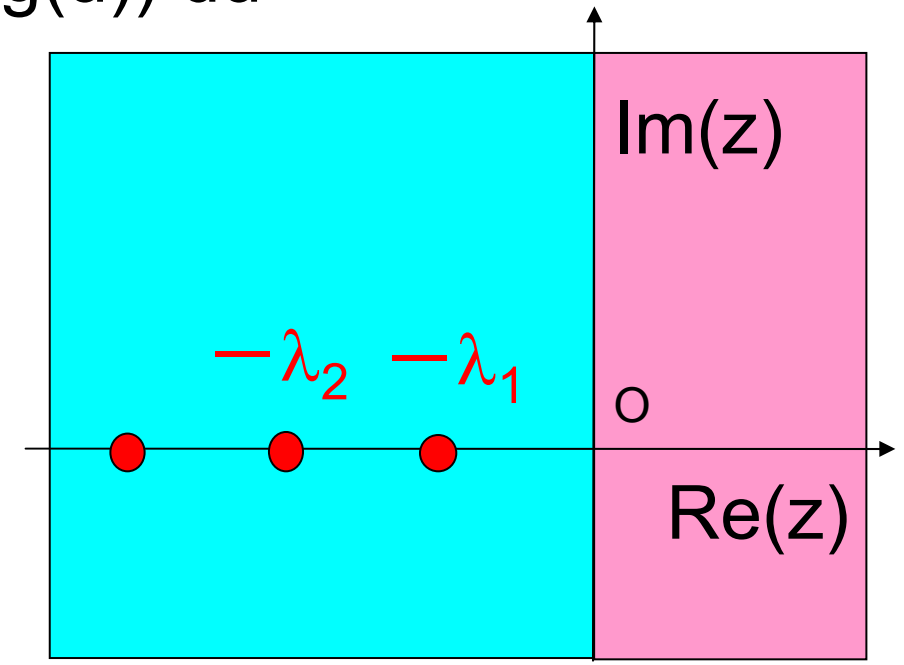
Analytic continuation of Zeta function

$$\zeta(z) = \Sigma \int K(g(u))^z \varphi(g(u)) |g'(u)| du$$

$$= \Sigma \int \Pi u_j^{2k_j z + h_j} \varphi(g(u)) du$$

$$= \frac{C_1}{(z+\lambda_1)^{m_1}} + \frac{C_2}{(z+\lambda_2)^{m_2}} + \cdots$$

$\zeta(z)$, a holomorphic function in Re(z)>0, can be analytically continued to entire complex plane as a meromorphic function.



Definition. $\lambda = \lambda_1$: Log Canonical Threshold.

Asymptotic expansion of Posterior

Lemma.1

There exists a measure D(du) on U such that

$$\exp(-nK(w)) \varphi(w)\, dw$$

$$= \sum \frac{(\log n)^{m-1}}{n^{\lambda}} \int dt\ e^{-t}\, t^{\lambda-1}\, D(du) + (\text{small}),$$

3 Singular Fluctuation

Decomposition of Empirical KL

From $K(g(u)) = u^{2k}$, there exists $a(x,u)$ such that

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

Definition. Empirical Process,

$$\xi_n(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{ a(X_i,u) - E_x[a(X,u)] \}$$

Empirical Process

Empirical process shows fluctuation of $K_n(g(u))$,

$$nK_n(g(u)) = nK(g(u)) - \sqrt{nK(g(u))}\,\xi_n(u)$$

Lemma.2 $\xi_n \longrightarrow \xi$: gaussian process

The convergence in law is proved as a random variable in Banach space of uniformly bounded functions on U.

Renormalized distribution

Definition. Renormalized posterior distribution

$$E_{t,u}[\, F(t,u)\,] = \frac{\Sigma \int D(du) \int F(t,u)\, t^{\lambda-1} e^{-t-\xi(u)t^{1/2}}\, dt}{\Sigma \int D(du) \int t^{\lambda-1} e^{-t-\xi(u)t^{1/2}}\, dt}$$

Lemma. 3 For arbitrary $s > 0$,

$$n^{s/2} E_w [\; f(x,w)^s \;] \to E_{t,u}[\; t^{s/2} a(x,u)^s \;]$$

Singular Fluctuation

Definition. Singular Fluctuation is defined by

$$\nu = \frac{1}{2} E_\xi E_x \{ E_{t,u}[a(X,u)^2 t] - E_{t,u}[a(X,u) t^{1/2}]^2 \}$$

There are infinitely many resolutions of singularities.
However, neither λ nor ν depends on the choice.
Therefore, λ and ν are birational invariants.

If p(x|w) is statistically regular, $\lambda = \nu = d/2$,

If p(x|w) is singular, they are different.

The functional variance

Definition. The functional variance is defined by

$$V = \sum_{i=1}^{n} \{ E_w[(\log p(X_i|w))^2] - E_w[\log p(X_i|w)]^2 \}$$

Lemma. 4 $E[V] = 2\nu + o(1)$

Singular fluctuation can be estimated from random samples.

4 Main Theorem

Main Theorem (2009, Watanabe)

Generalization and training errors are given by real log canonical threshold λ and singular fluctuation ν.

$$E[\,G\,] = \lambda / n + o(1/n),$$

$$E[\,T\,] = (\,\lambda - 2\nu\,) / n + o(1/n),$$

$$E[\,V\,] = 2\nu + o(1).$$

Explanation

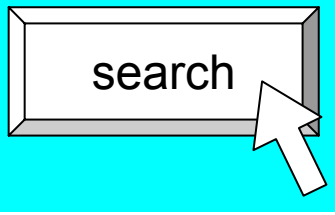
(1) The log canonical threshold λ shows how fast posterior distribution shrinks to analytic set $K(w)=0$.

(2) The singular fluctuation ν shows the variance of $K_n(w)$ in the neighborhood of singularities.

(3) Statistical learning process is determined by λ and ν.

Proof: Mathematically rigorous proof can be found by

algebraic geometry and statistical learning theory



5 Application to Statistics

Equation of State

Theorem. By eliminating λ and ν,

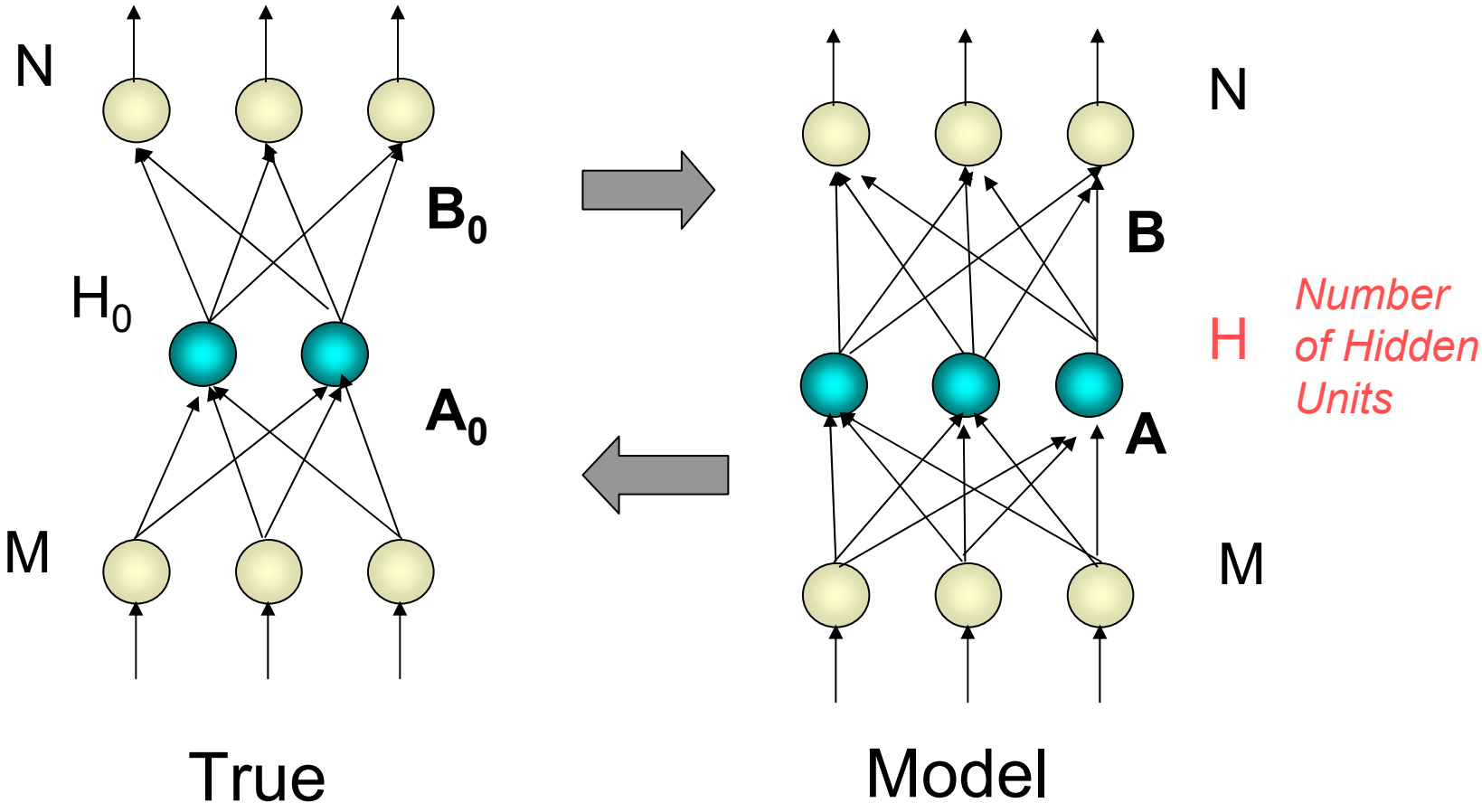
$$E[G] = E[\, T + V/n \,] + o(1/n)$$

This relation holds for any statistical model, any true distribution, any a priori distribution, and any singularities.

The generalization error G can be estimated from the training error T and the functional variance V.

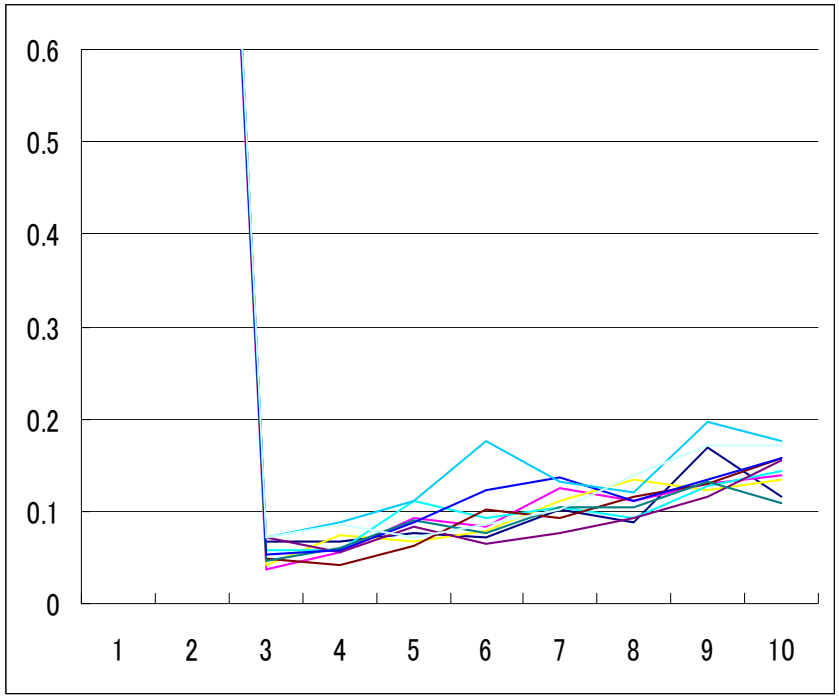
Application to Model Evaluation

Reduced rank regression : How to choose model



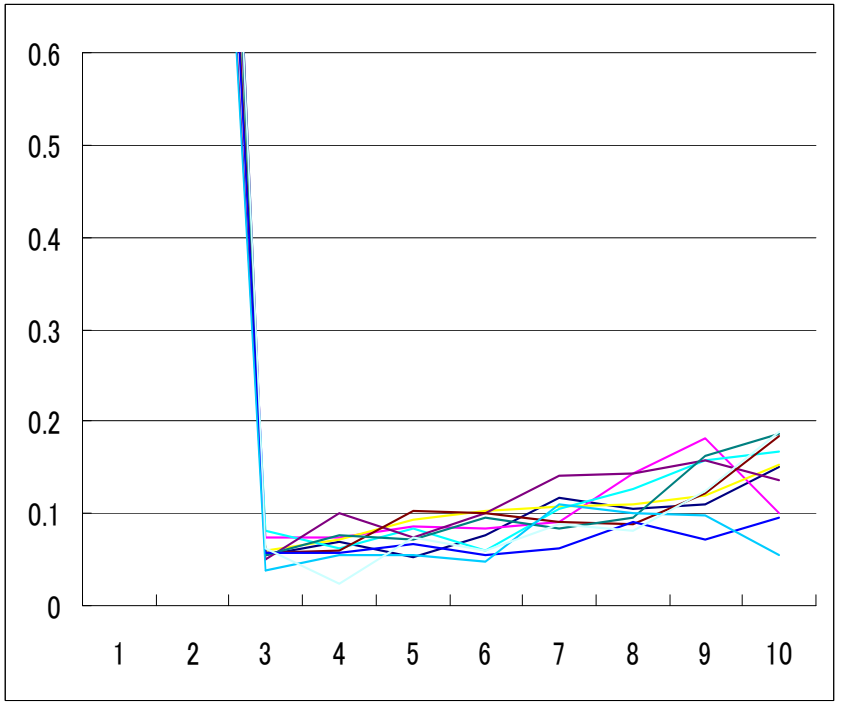
Experimental Results

G



Number of Hidden Units

T + V/n



Number of Hidden Units

M=N=10, True =3, n=500, Para=2000, MCMC=400000

Conclusion

1. Statistical models in information science and biostatistics are not regular but singular.

2. Two birational invariants are introduced.

3. Statistical learning process are determined by two birational invariants.

4. A new methodology in statistics is provided by singularity theory.

Thank you

JSPS Forum, Strasbourg, August, 2009

The 5th Franco-Japanese Symposium on Singularities