Algebraic Geometry and Model Selection

American Institute of Mathematics
2011/Dec/12-16

I would like to thank Prof. Russell Steele, Prof. Bernd Sturmfels, and all participants. Thank you very much.

Sumio Watanabe

Tokyo Institute of Technology

Contents

1

AIC, BIC, and DIC

Statistical Model and True Distribution

$x \in R^N$, $w \in W$ (compact) $\subset R^d$

(1) True dist. $q(x)$ i.i.d. $X_1, X_2, \ldots, X_n$

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

(3) Prior dist. $\varphi(w)$

(Remark)
$E[\quad]$ shows the expectation over $X_1,X_2,\ldots,X_n$.
$E_x[\quad]$ does that over X whose prob. dist. is $q(x)$.

Statistical Estimation

Posterior Dist.

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

Predictive Dist.

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

(Remark) In Bayesian estimation, the true distribution $q(x)$ is estimated by $p^*(x)$.

Stochastic Complexity and Generalization Loss

(1) Stochastic Complexity = - (Bayes Marginal)

$$F = -\log \int \prod_{i=1}^{n} p(X_i|w)\, \varphi(w)\, dw$$

(2) Generalization Loss

$$G = -E_x[\log p^*(x)\ ]$$

$$= S + KL(\, q(x)\, ||\, p^*(x)\,)$$

(Remark) $S = -E_x[\log q(X)]$ is the entropy of $q(x)$.

BIC(Schwarz,1978)

(1) $\mathrm{BIC} = - \Sigma \log p(X_i|w_{MLE}) + (d/2) \log n$

If the posterior ~ normal distribution.

$$F = \mathrm{BIC} + O_p(1).$$

<u>In general</u>, <u>yesterday,</u> we learned

RLCT

$$F = - \Sigma \log p(X_i|w_{MLE}) + \lambda \log n$$

$$- (m-1)\log\log n + O_p(1).$$

In our workshop,
Dr. Lin: Relation to asymptotic integral.
Dr. Drton: How to use in statistics.
Dr. Leykin: D-module and b-function.

AIC(Akaike,1974) & DIC(Spiegelhalter,et.al. 2002)

(2) $\mathrm{AIC} = - \Sigma \log p(X_i|w_{MLE}) + d$

(3) $\mathrm{DIC} = - \Sigma \log E_w[p(X_i|w)\,]$

$+ \;2\; \Sigma \{-E_w[\log p(X_i|w)]+\log p(X_i|\, E_w[w]\,)\,\}$

If the posterior ~ the normal distribution.

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

$$E[DIC]= n\, E[G]+o(1).$$

If otherwise, such relations do not hold.

Estimation of G and F in regular cases

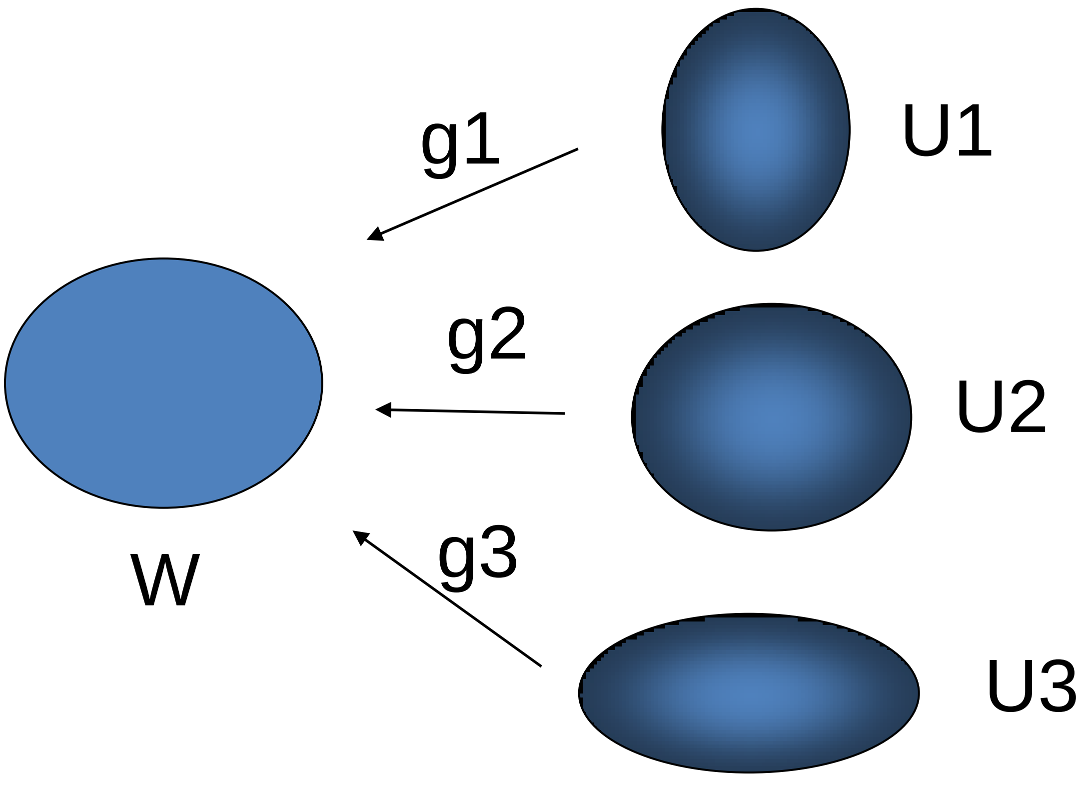
	AIC, DIC	BIC
Main Term Estimated	Random (Order 1)	Constant (log n)
Consistency in model selection	No	Yes
Unbiased Estimator of G	Yes	No

2

Birational Statistics

Birational Invariant

If a value that is defined using resolution of singularities does not depend on the choice of resolution, then it is called a birational invariant.



Nature in Statistics

It is natural that statistical theory should be made to be invariant under birational transform. Fisher's asymptotic theory does not satisfy such invariance. For example, it is not invariant under blow-up.

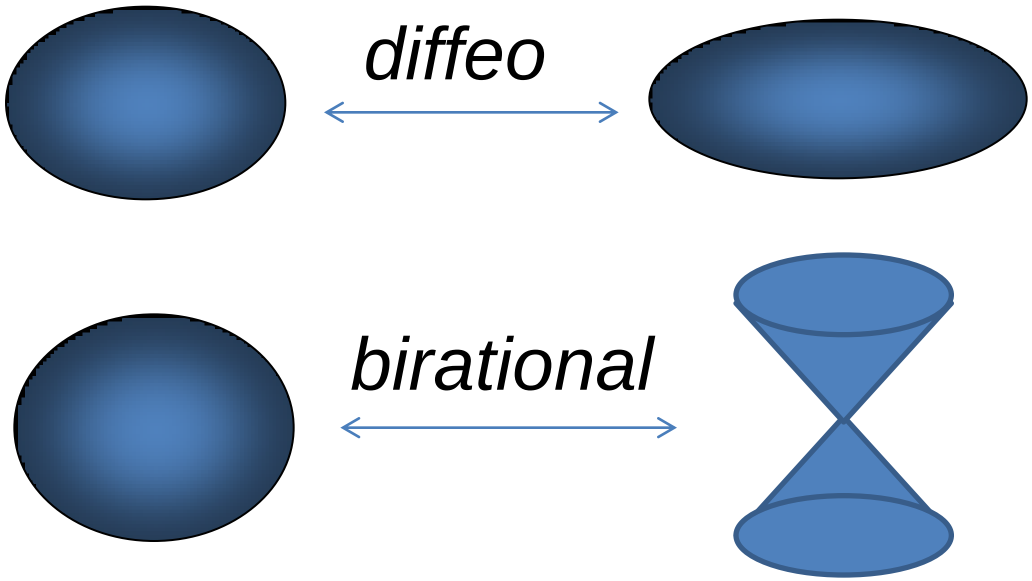
$$Y = aX + b + \text{Noise}$$

$$\begin{cases} a = c = c'd' \\ b = cd = d' \end{cases}$$

Asymptotic normality of (a,b) holds, whereas that of (c,d) in projective space not .

Differential and Birational

Algebraic geometry studies mathematical properties those are invariant under the birational transform.



To construct birational statistics might be one of the purposes of algebraic statistics.

3

Singular Fluctuation and model selection

Loss for estimation

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

Generalization Loss

$$G_n = - \quad E_x[\ \log E_w[p(X|w)]\ \ ]$$

Training Loss

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

Functional Cumulant

Def. Two Cumulant Generating Functions

$$g(a) = -\ E_x[\log E_w[p(X|w)^a]\ ]$$

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

Then $g(0)=t(0)=0$,

and $G_n=g(1)$, $T_n= t(1)$.

Invariance

Two functions g(a) and t(a) are invariant under

$$w = g(u)$$

$$p(x|w) = p(x|g(u))$$

$$\varphi(w)\, dw = \varphi(g(u))\, |g'(u)|\, du$$

Cumulant generating function
= Birational invariant generating function

Example. $$\lambda = \lim_{n \to \infty} n \{ E[g(1)] - S \}$$

Notation

Def. Log density ratio function

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

Then

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

Expansion of g(a)

g(a) is rewritten as

$$g(a) = aS - E_x[\ \log E_w[\exp(-af(X,w))]\ ]$$

Therefore

$$g'(0) = S + E_x[\ E_w[f(x,w)]\ ]$$

$$g''(0) = -\ E_x[\ E_w[f(x,w)^2] - E_w[f(x,w)]^2\ ]$$

Expansion of t(a)

By the same way,

$$t(a) = aS_n - (1/n) \sum_{i=1}^{n} \log E_w[\exp(-af(X_i|w))]$$

Therefore

$$t'(0) = S_n + (1/n) \sum_{i=1}^{n} E_w[f(X_i,w)]$$

$$t''(0) = -(1/n) \sum_{i=1}^{n} \{ E_w[f(X_i,w)^2] - E_w[f(X_i,w)]^2 \}$$

Functional Variance

Def. Two random variables

$$V_1 = n E_x[E_w[f(x,w)]] - \lambda$$

$$V_2 = \sum_{i=1}^{n} \{ E_w[(\log p(X_i|w))^2] - E_w[\log p(X_i|w)]^2 \}$$

Remark. V_2 can be calculated by samples and a model without any information about true dist.. In order to calculate V_1, we need the information of the true distribution.

Singular Fluctuation

Theorem. Convergences hold,

$$V_1 \longrightarrow V_1^* \qquad\qquad E[V_1] \longrightarrow E[V_1^*]$$

$$V_2 \longrightarrow V_2^* \qquad\qquad E[V_2] \longrightarrow E[V_2^*]$$

Theorem and Def. Singular Fluctuation

$$E[V_1^*] = E[V_2^*] = 2\nu$$

<u>Remark</u>. In order to prove the above theorems, we need resolution theorem and empirical process theory. In regular cases, $\lambda=\nu=d/2$.

Outline of Proofs

Posterior distribution measure $\quad K_n(w)=(1/n)\Sigma f(X_i,w)$

$$\exp(-nK_n(w))\ \varphi(w)\ dw$$

$$= \exp(\ -\ nu^{2k} + n^{1/2}u^k\xi_n(u)\ )\ \varphi(g(u))|g'(u)|\ du$$

$$= \int_0^{\infty} dt\ \delta(t-nu^{2k})\ |u^h|\ b(u)\ \exp(\ -\ t + t^{1/2}\xi_n(u)\ )\ du$$

$$= \frac{(\log n)^{m-1}}{n^{\lambda}} \int dt\ t^{\lambda-1} \exp(\ -\ t + t^{1/2}\xi_n(u)\ )\ D(u)du$$

Outline of Proofs 2

Def. Expectation over the limit posterior distribution

$$< \quad > = \frac{\int dt \int du D(u) (\quad) \, t^{\lambda-1} \exp(-t + t^{1/2} \xi(u))}{\int dt \int du D(u) \, t^{\lambda-1} \exp(-t + t^{1/2} \xi(u))}$$

Lemma. For $s \geqq 0$,

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

Cumulants --- Random Variables

$$g'(0) = S + (\lambda+V_1/2)/n + o_p(1/n)$$

$$g''(0)= - V_2/n + o_p(1/n)$$

$$t'(0) = S_n + (\lambda-V_1/2)/n + o_p(1/n)$$

$$t''(0) = - V_2 /n + o_p(1/n)$$

Cumulants --- Birational invariants.

$$E[\, g'(0)\,] = S + (\lambda+\nu)/n + o_p(1/n)$$

$$E[\, g''(0)\,] = -\, 2\nu/n + o_p(1/n)$$

$$E[\, t'(0)\,] = S + (\lambda-\nu)/n + o_p(1/n)$$

$$E[\, t''(0)\,] = -\, 2\nu\,/n + o_p(1/n)$$

Theorem

Generalization Loss

$$G_n = S + [\, \lambda + V_1/2 - V_2/2 \,] / n + o_p(1/n)$$

Training Loss

$$T_n = S_n + [\, \lambda - V_1/2 - V_2/2 \,] / n + o_p(1/n)$$

Theorem

When n tends to infinity,

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

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

$$E[\,V_2\,] = 2\nu + o(1).$$

WAIC

Def. WAIC is defined by

$$W_n = T_n + V_2/n.$$

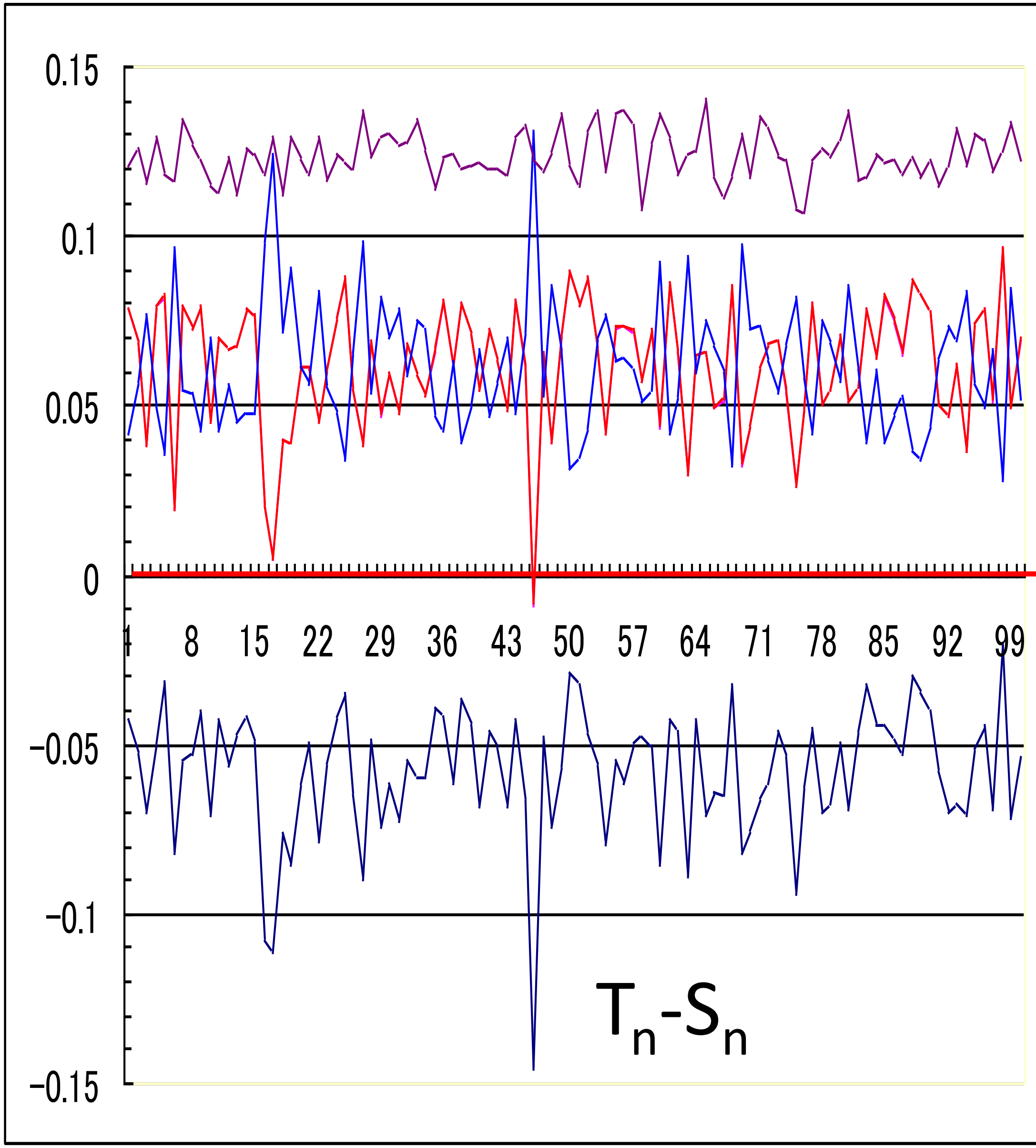
Theorem

For arbitrary set $(q(x), p(x|w), \varphi(w))$,

$$E[\,G_n\,] = E[\,W_n\,] + o(1/n^2).$$

$$(G_n - S) + (W_n - S_n) = 2\lambda/n + o_p(1/n).$$

Remark. WAIC is asymptotically equivalent to Bayes cross validation (Watanabe, JMLR, Vol.11, 3571-3594, 2010)



W_n-S_n + G-S

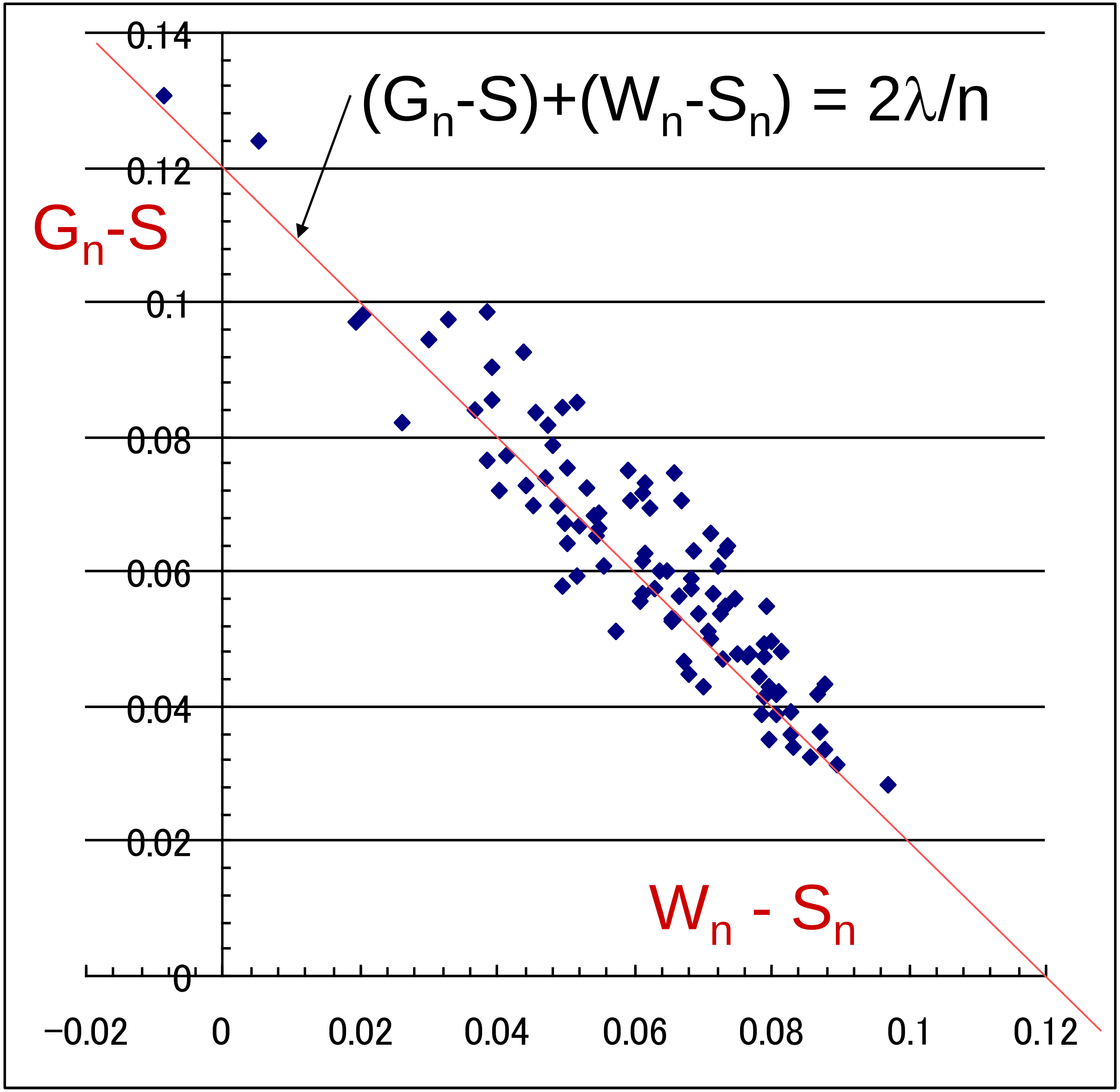
Red $=W_n$-S_n

Theory=λ/n

Blue $=G_n$- S

100
Indep.
Trials

Reduced Rank
Regression 5-5-5
True 5-3-5
By Theory, λ=12
n=200
Metropolis
100000-200000



$(G_n\text{-}S)+(W_n\text{-}S_n) = 2\lambda/n$
$G_n\text{-}S$
$W_n - S_n$
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
−0.02
0
0.02
0.04
0.06
0.08
0.1
0.12

WAIC

(1) $E[\ W_n\ ]= E[\ G_n\ ]+o(1/n^2)$ holds
even if q(x) is not realizable by p(x|w).

(2) The essential main term is fluctuated.
Inconsistency in model selection.

(3) If the posterior can be approximated by
a normal distribution, WAIC is equivalent to
AIC and DIC as a random variable.

(4) If otherwise, WAIC is unbiased estimator of
generalization error, whereas either AIC
or DIC not.

4 Open Problems

Two Biratinal Invariants

If the regularity condition is satisfied,

$$\lambda = \nu = d/2.$$

In general, they are different.

λ = Dimension that shows how fast the posterior shrinks.

ν = Dimension that shows how strong the posterior fluctuates.

Q. Mathematically, what is ν ?

Generalized RLCT

$$g(a) = -E_x[\log E_w[p(x|w)^a]]$$

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

Q. $E[(d/da)^k g(0)]$ and $E[(d/da)^k t(0)]$ $(k=1,2,\ldots,)$ are all birational invariants. They are statistically generalized ones from real log canonical threshold. What are they ?

Variance of Information Criterion

$$E[\,G_n\,] = E[\,W_n\,] + o(1/n^2).$$

$$(G_n - S) + (\,W_n - S_n) = 2\lambda/n + o_p(1/n).$$

Q. These equations show that WAIC seems to have the smallest variance among the information criteria whose averages are equal to G_n. Can we prove this ?

Bayes hypothesis testing

For a given prior $\varphi(w)$, we define

$$F(\varphi) = - \log \int \prod_{i=1}^{n} p(X_i|w)\, \varphi(w)\, dw$$

If the null hypothesis is $\varphi_0(w)$, and if the alternative is $\varphi_1(w)$, then the most powerful test is given by $DF = F(\varphi_1)-F(\varphi_0)$.

Q. In order to make the most powerful hypothesis test, probability distribution of DF is necessary.