Higher Order Analysis of Bayesian Cross Validation in Regular Asymptotic Theory

Information Geometry and Its Applications IV

In honor of Prof. Amari's 80th birthday

June 12-17, 2016, Liblice, Czech Republic

Sumio Watanabe

Tokyo Institute of Technology

Purpose of this Research

Answer to the Bayesian question:
“Is choosing a prior by minimizing cross validation really optimal for minimizing generalization loss ?”

S. Watanabe, Bayesian Cross Validation and WAIC for Predictive Prior Design in Regular Asymptotic Theory

arXiv:1503.07970

Why Higher Order is Necessary

(1) In Bayesian statistics, it is frequently discussed how to choose (or optimize) a prior.

(2) In regular statistical models, the first order statistics does not depend on a prior.

(3) Higher order analysis is necessary to study the effect of a prior.

Optimality Measure of a Prior

In this presentation, we study the optimality of a prior on the following situations.

(1) Evaluation measure : generalization loss (= KL loss) of Bayes predictive distribution.

(2) Optimizing criteria : cross validation, information criteria, and marginal likelihood.

(3) Statistical model : regular

Contents

Notations: Model and Prior

(1) $q(x)$: an unknown true probability density on R^N.

(2) $X^n=(X_1,X_2,\ldots,X_n)$: a set of random variables which are independently subject to $q(x)$.

(3) $p(x|w)$: a probability density on R^N for a given parameter w in R^d.

Note: q(x) is not realizable by p(x|w) in general.

(4) $\varphi_0(w)$: a fixed prior on R^d (improper).

$\varphi(w)$: a candidate prior on R^d (improper).

Definition of Bayesian Estimation

(1) Posterior distribution is defined by

$$p(w|X^n) = (1/Z)\ \varphi(w) \prod_{i=1}^{n} p(X_i|w),$$

where Z is a normalizing constant.

Note: Even if a prior is improper, we assume Z is finite.

(2) $E_w[\ \ ]$ shows the expected value over $p(w|X^n)$.

$V_w[\ \ ]$ shows the variance over $p(w|X^n)$.

(3) Predictive distribution

$$p(x|X^n) = E_w[\ p(x|w)\ ].$$

Generalization and Cross Validation

(1) Random generalization loss

$$G_n(\varphi) = -\int q(x) \log p(x|X^n)\, dx.$$

(2) Average generalization loss

$$E[\, G_n(\varphi)\,].$$

(3) Cross validation loss (Leave-one-out)

$$CV_n(\varphi) = -\ (1/n) \sum_{i=1}^{n} \log p(X_i|X^n - X_i).$$

(4) Average cross validation loss

$$E[\, CV_n(\varphi)\,].$$

ISCV and WAIC

(1) Importance sampling CV (Gel'fand et. al., 1992)

$$\mathrm{ISCV}_n(\varphi) = (1/n) \sum_{i=1}^{n} \log E_w[\, 1/\, p(X_i|w)\,].$$

It is proved that $\mathrm{CV}_n(\varphi) = \mathrm{ISCV}_n(\varphi)$.

(2) Widely Applicable Information Criterion (Watanabe, 2009)

$$\mathrm{WAIC}_n(\varphi) = -\,(1/n) \sum_{i=1}^{n} \log E_w[\, p(X_i|w)\,] + (1/n) \sum_{i=1}^{n} V_w[\, \log p(X_i|w)\,].$$

In regular models, $\mathrm{CV}_n(\varphi) = \mathrm{WAIC}_n(\varphi) + O_p(1/n^3)$.

Marginal likelihood

For an improper prior $\varphi(w)$, *a priori probability* distribution is

$$\varphi(w) \,/ \int \varphi(w)\, dw.$$

The minus log marginal likelihood (I.J. Good) is

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

Note: If $\int \varphi(w)\, dw=\infty$, the marginal likelihood can not be defined, whereas CV and WAIC can be defined.

Note: If you employ the marginal likelihood as a criterion, a prior function should be proper. However, the optimal prior function that minimizes the generalization loss may be improper in general.

Basic Question

By the definition, for an arbitrary integer n>1,

$$E[\ G_{n-1}(\varphi)\ ] = E[\ F_n(\varphi)\ ] - E[\ F_{n-1}(\varphi)\ ],$$

$$E[\ G_{n-1}(\varphi)\ ] = E[\ CV_n(\varphi)\ ].$$

However,

$G_{n-1}(\varphi)$ is not equal to $F_n(\varphi) - F_{n-1}(\varphi)$,

$G_{n-1}(\varphi)$ is not equal to $CV_n(\varphi)$.

Basic Question:

Assume $\varphi(w) = \varphi(w|\alpha)$, where α is a hyperparameter.
Does α that minimizes $CV_n(\varphi)$ or $F_n(\varphi)$ also minimizes $G_n(\varphi)$ and $E[\ G_n(\varphi)\ ]$, asymptotically ?

Contents

Notations I

(1) $\varphi_0(w)$: A fixed prior (for example, $\varphi_0(w) \equiv 1$)

(2) $L_n(w) = - (1/n) \sum_{i=1}^{n} \log p(X_i|w) - \log \varphi_0(w)$

(3) $w^* = \mathrm{argmin}\, L_n(w)$: MAP estimator for $\varphi_0(w)$

(4) $L(w) = - \int q(x) \log p(x|w)\, dx$

(5) $w_0 = \mathrm{argmin}\, L(w)$

Note: If $\varphi_0(w) = 1$ is chosen as a fixed prior, then it is improper.
$L_n(w)$ is a minus likelihood function and w^* is MLE.
w^* does not depend on a candidate prior.

Notations II

(1) For a given function f(w),

$$f_{k_1 k_2 \ldots k_m}(w) = (\partial/\partial w^{k_1})\,(\partial/\partial w^{k_2}) \cdots (\partial/\partial w^{k_m})\, f(w).$$

(2) Einstein's summation convention

$$A_{k_1 k_2} B^{k_2 k_3} = \sum_{k_2=1}^{d} A_{k_1 k_2} B^{k_2 k_3}.$$

(3) Assumption: $(L(w))_{k_1 k_2}$ is positive definite in a neighborhood of w_0

$$(g_n)^{k_1 k_2}(w) = \text{Inverse matrix of } (L_n(w))_{k_1 k_2}$$

$$(g)^{k_1 k_2}(w) = \text{Inverse matrix of } (L(w))_{k_1 k_2}$$

Note: These functions do not depend on a candidate prior.

Notations III

Correlations

$$(F_n)_{k_1, k_2}(w) = (1/n) \sum_{i=1}^{n} (\log p(X_i|w))_{k_1} (\log p(X_i|w))_{k_2}$$

$$(F_n)_{k_1 k_2, k_3}(w) = (1/n) \sum_{i=1}^{n} (\log p(X_i|w))_{k_1 k_2} (\log p(X_i|w))_{k_3}$$

Average correlations

$$(F)_{k_1, k_2}(w) = E[\ (F_n)_{k_1, k_2}(w)\ ]$$

$$(F)_{k_1 k_2, k_3}(w) = E[\ (F_n)_{k_1 k_2, k_3}(w)\ ]$$

Note: These functions do not depend on a candidate prior.

Notations IV

For higher order analysis, the following functions are necessary.

$(A_n)^{k1\,k2}(w) = (1/2)\,(g_n)^{k1k2}(w)$

$(B_n)^{k1k2}(w) = (1/2)\,\{\,(g_n)^{k1k2}(w) + (g_n)^{k1k3}(w)\,(g_n)^{k2k4}(w)\,(F_n)_{k3,\,k4}(w)\,\}$

$(C_n)^{k1}(w) = (g_n)^{k1k2}(w)\,(g_n)^{k3k4}(w)\,(F_n)_{k2k4,k3}(w)$

$-(1/2)\,(g_n)^{k1k2}(w)\,(g_n)^{k3k4}(w)\,(L_n)_{k2k3k4}(w)$

$-(1/2)\,(g_n)^{k1k2}(w)\,(g_n)^{k3k4}(w)(g_n)^{k5k6}(w)(L_n)_{k2k3k5}(w)(F_n)_{k4,k6}(w)$

Definitions of $(A)^{k1\,k2}(w)$, $(B)^{k1\,k2}(w)$, and $(C)^{k1}(w)$

$(A)^{k1\,k2}(w)$, $(B)^{k1\,k2}(w)$, and $(C)^{k1}(w)$ are defined by the same equations as $(A_n)^{k1\,k2}(w)$, $(B_n)^{k1\,k2}(w)$, and $(C_n)^{k1}(w)$ by using $(g)^{k1k2}(w)$, $(F)_{k_1,\,k_2}(w)$, and $(F)_{k_1k_2,\,k_3}(w)$ in stead of $(g_n)^{k1k2}(w)$, $(F_n)_{k_1,\,k_2}(w)$, and $(F_n)_{k_1k_2,\,k_3}(w)$.

Note: These functions do not depend on a candidate prior.

Notations V

For higher order analysis, the followings are necessary.

Mathematical relations between priors $\varphi(w)$ and $\varphi_0(w)$.

$\Phi(w) = \varphi(w)/\varphi_0(w)$. Ratio of candidate and fixed priors.

$$M_n(\varphi,w) = (A_n)^{k1k2}(w)\ (\log \Phi)_{k1}(\log \Phi)_{k2}$$
$$+ (B_n)^{k1k2}(w)\ (\log \Phi)_{k1k2} + (C_n)^{k1}(w)\ (\log \Phi)_{k1}$$

$$M(\varphi,w) = (A)^{k1k2}(w)\ (\log \Phi)_{k1}(\log \Phi)_{k2}$$
$$+ (B)^{k1k2}(w)\ (\log \Phi)_{k1k2} + (C)^{k1}(w)\ (\log \Phi)_{k1}$$

Note: Neither $(A)^{k1,k2}(w)$, $(B)^{k1,k2}(w)$, $(C)^{k1}(w)$, $(A_n)^{k1,k2}(w)$,$(B_n)^{k1,k2}(w)$, nor $(C_n)^{k1}(w)$ depends on the candidate prior $\varphi(w)$.
A candidate prior affects only $(\log \Phi)$.

Theorem

w^* = MAP estimator for $\varphi_0(w)$

(1) Mathematical relations asymptotically satisfy

$$M_n(\varphi, w^*) = M(\varphi, w_0) + O_p(1/n^{1/2}),$$

$$E[\, M_n(\varphi, w^*)\,] = M(\varphi, w_0) + O(1/n).$$

Note: Minimizing $M_n(\varphi, w^*)$ is asymptotically equivalent to minimizing $E[\, M_n(\varphi, w^*)\,]$ and $M(\varphi, w_0)$.

Theorem

(2) Cross validation asymptotically satisfies

$$CV(\varphi) = CV(\varphi_0) + (1/n^2)\, M_n(\varphi,w^*) + O_p(1/n^3)$$

$$E[CV(\varphi)] = E[CV(\varphi_0)] + (1/n^2)\, M(\varphi,w_0) + O(1/n^3)$$

Note: Minimizing $CV(\varphi)$ is asymptotically equivalent to minimizing $M_n(\varphi,w^*)$.

Note: Minimizing $CV(\varphi)$ is asymptotically equivalent to minimizing $E[CV(\varphi)]$.

Theorem

(3) Generalization loss asymptotically satisfies

$$G_n(\varphi) = G_n(\varphi_0) + O_p(1/n^{3/2})$$

$$E[\, G_n(\varphi)\,] = E[\, G_n(\varphi_0)\,] + (1/n^2)\, M(\varphi, w_0) + O(1/n^3)$$

Note: Minimizing $CV_n(\varphi)$ is not asymptotically equivalent to minimizing $G_n(\varphi)$.

Note: Minimizing $CV_n(\varphi)$ is asymptotically equivalent to minimizing $E[\, G_n(\varphi)\,]$.

Note: Minimizing $E[\, G_n(\varphi)\,]$ can be performed by minimizing $CV_n(\varphi)$.

Note: Minimizing $G_n(\varphi)$ seems to be impossible if we do not know the true distribution.

Contents

Contents

1. Basic Bayesian procedures

2. Main Theorem

3. Proof

4. Example

An Example

Model $p(x|s,m) = (s/2\pi)^{1/2} \exp(- (s/2)(x-m)^2)$

True $q(x) = p(x|1,1)$

Prior $\varphi(s,m|\mu, \lambda) = s^{\mu} \exp(- \lambda\, s(m^2 +1))$

$\varphi(\mu, \lambda)$ is a set of hyperparameters

Proper $\Leftrightarrow$ $\mu > -1/2,\ \lambda>0$

Fixed Prior $\varphi_0(s,m) =1$

(w^*,s^*) : MAP = MLE

An Example

$$(A_n)^{k1,\, k2}(w^*) = \begin{pmatrix} 1/(2s^*) & 0 \\ 0 & s^{*2} \end{pmatrix}$$

$$(B_n)^{k1,\, k2}(w^*) = \begin{pmatrix} 1/(s^*) & -s^{*2}M_3/2 \\ -s^{*2}M_3/2 & (s^{*2} + s^{*4}M_4)/2 \end{pmatrix}$$

$$(C_n)^{k1}(w^*) = (0,\ s^* + s^{*3}M_3)$$

An Example

Mathematical relation between priors $\varphi(s,m)$ and $\varphi0(s,m)$

results in

$$M_n(\varphi,m^*,s^*) = (1/2)\,\lambda^2 s^* m^{*2} + (-\lambda s^* m^{*2}/2 + \mu - \lambda s^*/2)^2$$

$$+ (-\lambda s^* m^{*2}/2 + \mu/2 - \lambda s^*/2)(1 + s^{*2} M_4)$$

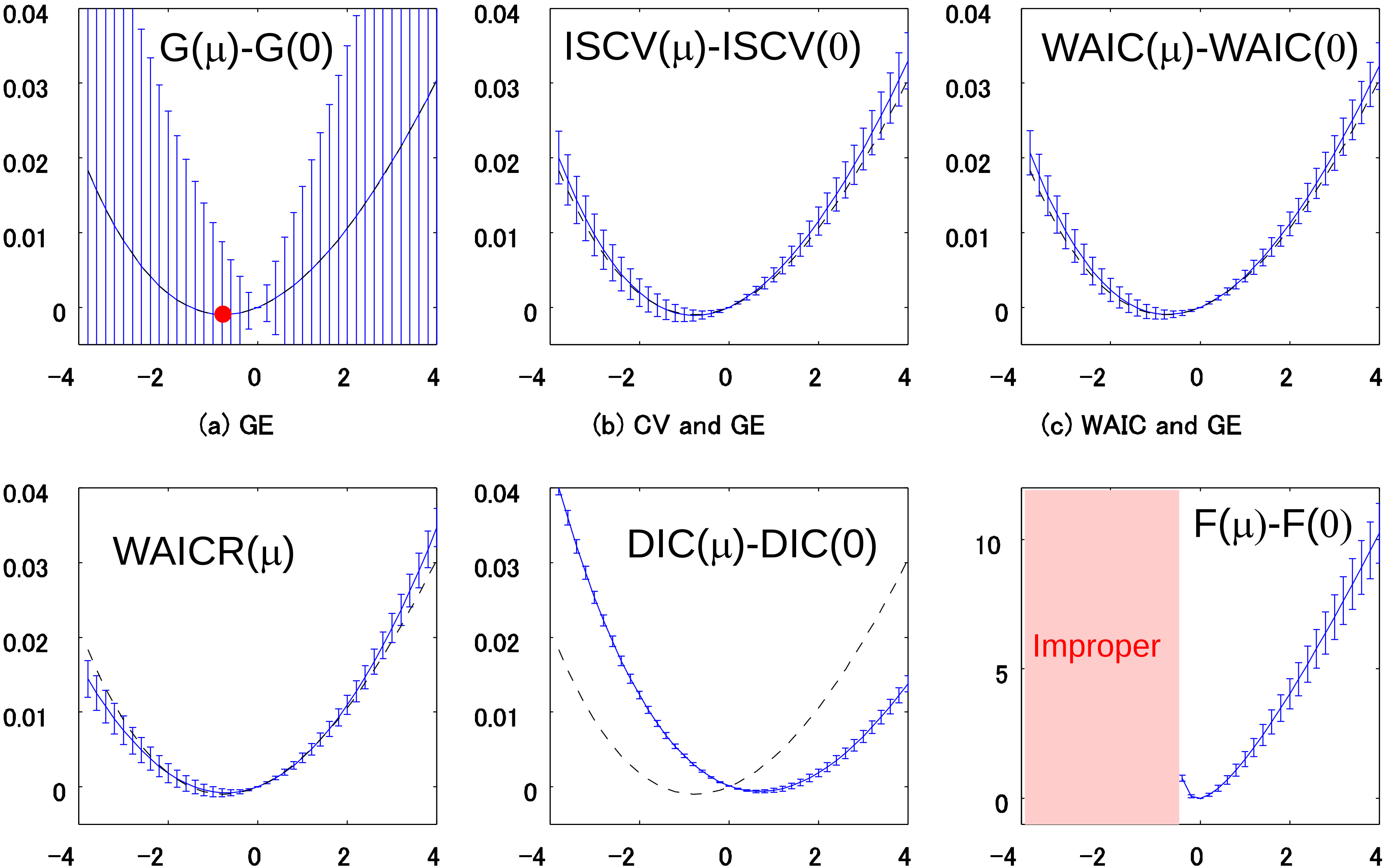
$$- \lambda + \lambda m^* s^{*2} M_3$$

Simulation : λ is fixed and μ is optimized.

Information Criteria

Generalization loss	$G_n(\mu) = - E_x[\log p(x\|X^n)]$
Importance sampling cross validation	$ISCV_n(\mu) = (1/n) \Sigma \log E_w[1/ p(X_i\|w)]$
Widely Applicable Information Criterion	$WAIC_n(\mu) = - (1/n) \Sigma \log E_w[p(X_i\|w)] +(1/n) \Sigma V_w[\log p(X_i\|w)]$
Deviance Information Criterion (Spiegelhalter et.al.)	$DIC_n(\mu) = (1/n) \Sigma \log p(X_i\| E_w[w]) - (2/n) \Sigma \log E_w[p(X_i\|w)]$
Minus log marginal Likelihood	$F_n(\mu) = - \log \int \varphi(w) \Pi p(X_i\|w)dw + \log \int \varphi(w) dw$
Higher order CV	$WAICR_n(\mu) = (1/n^2) M_n(\varphi,w^*)$

Simulation Results



(a) GE

(b) CV and GE

(c) WAIC and GE

(d) WAICR and GE

(d) DIC and GE

(e) Free Energy

Experimental Discussion

Model $p(x|s,m) = (s/2\pi)^{1/2} \exp(- (s/2)(x-m)^2)$

True $q(x) = p(x|1,1)$

Prior $\varphi(s,m|\mu, \lambda) = s^{\mu} \exp(- \lambda s(m^2 +1))$

From the view point of hyperparameter optimization,

(1) The variance of the random generalization loss is far larger than cross validation and information criteria.

(2) $\varphi(s,m|\mu, \lambda)$ is improper at the optimal μ that minimizes the average generalization loss. It can not be found by maximizing the marginal likelihood.

Conclusion

1. Higher order asymptotic theory of Bayesian cross validation is established.

2. Average generalization loss is minimized by minimizing the cross validation or WAIC.

3. Average generalization loss is not minimized by using the marginal likelihood or DIC.

4. Random generalization loss is not minimized by any criteria. It seems to be impossible.

Future Study

1. Understanding the results from the viewpoint of information geometry.

2. In singular models, choosing a prior often affects the first order statistics.