INTRINSIC PRIORS VIA KULLBACK-LIEBLER GEOMETRY

by

Jacek Dmochowski
Purdue University

Technical Report #94–15

Department of Statistics
Purdue University

June 1994

INTRINSIC PRIORS VIA KULLBACK-LEIBLER GEOMETRY [1]

Jacek Dmochowski

Department of Statistics, Purdue University, West Lafayette, IN 47907

KEY WORDS: Intrinsic Bayes factor, intrinsic priors, Kullback-Leibler projection

Abstract

In Bayesian model selection or hypothesis testing problems the Bayes factor (BF) cannot be used directly if at least one of the priors is improper. Berger and Pericchi (1993) proposed the use of the so-called Intrinsic Bayes factor (IBF). The IBF is a Bayes factor (BF) multiplied by a data dependent "correction" term. The asymptotic behavior of both the BF and the "correction" term will be described in the iid case. Intrinsic priors are priors for which the resulting BF is asymptotically equivalent to the IBF. Asymptotic considerations lead to one (nested case) or two (non-nested case) functional equations. These intrinsic equations involve Kullback-Leibler projections. A geometrical interpretation and the general form of the solutions for the nested case will be presented and illustrated through examples.

1. Introduction

In the Bayesian approach to model selection or hypothesis testing with at least one of the parameter spaces being unbounded, it is typically not possible to utilize improper priors. The reason is that the Bayes factor will be defined only up to a multiplicative constant. Different solutions to this problem have been proposed. One of these is the BIC criterion of Schwarz [9]

[1] This work was supported by David Ross grant no.6901927

, based on asymptotics but without terms depending on priors. Smith and Spiegelhalter [11] and Spiegelhalter and Smith [12] proposed that one choose a multiplicative constant in the Bayes factor using an (imaginary) minimal training sample. Review of other methods can be found in Berger and Pericchi [1] . In the same paper Berger and Pericchi [1] propose a new method to determine the multiplicative constant. Consider a problem of hypothesis testing H_0 vs. H_1, where $X_1, \ldots, X_n$ is an *iid* sample from a distribution with density $f(x|\theta)$ with respect to, say, the Lebesgue measure, and H_i states that $\theta \in \Theta_i$, $i = 0, 1$. Assume that we want to use improper priors $\pi_i^N(\cdot)$. Usually there exists a minimal nonnegative integer k such that $\int_{\Theta_i} f(x_1, \ldots, x_k|\theta)\pi_i^N(\theta)d\theta < \infty$ for $i = 0, 1$. Then posteriors $\pi_i^N(\cdot|x_1, \ldots, x_k)$ can be used as proper (data dependent) priors for further analysis and we obtain

$$\begin{aligned} B_{1,0}(X_{k+1}, \ldots, X_n|X_1, \ldots, X_k) &= \frac{\int_{\Theta_1} f(X_{k+1},\ldots,X_n|\theta)\pi_1^N(\theta|X_1,\ldots,X_k)d\theta}{\int_{\Theta_0} f(X_{k+1},\ldots,X_n|\theta)\pi_0^N(\theta|X_1,\ldots,X_k)d\theta} \\ &= \frac{\int_{\Theta_1} f(X_1,\ldots,X_n|\theta)\pi_1^N(\theta)d\theta}{\int_{\Theta_0} f(X_1,\ldots,X_n|\theta)\pi_0^N(\theta)d\theta} \frac{\int_{\Theta_0} f(X_1,\ldots,X_k|\theta)\pi_0^N(\theta)d\theta}{\int_{\Theta_1} f(X_1,\ldots,X_k|\theta)\pi_1^N(\theta)d\theta} \\ &= B_{1,0}^N(X_1, \ldots, X_n)\ B_{0,1}^N(X_1, \ldots, X_k) \end{aligned}$$

The first term looks like the Bayes factor for the full data , and the second can be treated as a data dependent multiplicative constant. Since it is impossible to decide apriori which subsample of $X_1, \ldots, X_n$ of size k should be chosen, Berger and Pericchi [1] propose to use in computations different types of averages of $B_{0,1}^N(X_{i_1}, \ldots, X_{i_k})$ over all subsamples of size k. The Bayes factor modified in this way is called *the intrinsic Bayes factor* (IBF), $B_{1,0}^I = B_{1,0}^N\ c$, where $c =$ correction term. If both priors $\pi_i^N(\cdot)$ are proper then "correction term" $\equiv 1$ and intrinsic Bayes factor is the same as the Bayes factor. The definition of the intrinsic Bayes factor will be given in Section 2. One of the justifications of the IBF method is the asymptotic correspondence of the intrinsic Bayes factor to the Bayes factor obtained from so-called intrinsic priors. The idea of intrinsic priors will be presented in Section 3. In a search for intrinsic priors, asymptotic considerations (Section 4) lead to a system of functional equations (called intrinsic equations) with intrinsic priors as their solutions. The system of intrinsic equations involves Kullback-Leibler projections. Intrinsic equations are discussed in Section 5. The main goal of the paper is to present a geometric interpretation of the IBF method.

2. Preliminaries

Let us assume the following situation. An *iid* sample $X_1, \ldots, X_n$ is available. All X_i's are distributed according to an unknown *cdf* $G(x)$ with density $g(x)$ with respect to the Lebesgue measure. Consider the model selection (hypothesis testing) problem.

$$M_0 : \ f_0(x|\theta_0), \ \theta_0 \in \Theta_0 \quad vs. \quad M_1 : \ f_1(x|\theta_1), \ \theta_1 \in \Theta_1$$

Assume vague prior knowledge about parameters. The (conditional) prior densities $\pi_i^N(\cdot)$, $i = 0, 1$ are used and at least one of them is improper. Let $m_i^N(x_1, \ldots, x_l|M_i)$ denote marginals given model M_i, $m_i^N(x_1, \ldots, x_l|M_i) = \int_{\Theta_i} f_i(x_1, \ldots, x_l|\theta_i)\pi_i^N(\theta_i)d\theta_i$. Let $\pi_i^N(\theta_i|x_1, \ldots, x_l)$ denote the posterior density of θ_i on Θ_i given the data $x_1, \ldots, x_l$, $\pi_i^N(\theta_i|x_1, \ldots, x_l, M_i) = f_i(x_1, \ldots, x_l|\theta_i)\pi_i^N(\theta_i)/m_i^N(x_1, \ldots, x_l|M_i)$. The letter k will be reserved for the minimal non-negative integer with the property that *both* posteriors $\pi_i^N(\theta_i|x_1, \ldots, x_l, M_i)$ are proper. Berk [2] observed that posteriors will be proper also for values of $l \geq k$. Let $B_{1,0}^N(x_1, \ldots, x_n)$ denote the Bayes factor for priors $\pi_1^N(\cdot)$,$\pi_0^N(\cdot)$ and data $x_1, \ldots, x_n$, $B_{1,0}^N(x_1, \ldots, x_n) = m_1^N(\cdot|M_1)/m_0^N(\cdot|M_0)$. Then the intrinsic Bayes factor (IBF) is defined as

$$B_{1,0}^I(x_1, \ldots, x_n) = B_{1,0}^N(x_1, \ldots, x_n)\text{Average}_{1 \leq l_1 < \ldots l_k \leq n} B_{0,1}^N(x_{l_1}, \ldots, x_{l_k})$$

where the average is taken over all subsamples of the data of size k. If we take the arithmetic average we get $B^{AI}_{1,0}$ - *the IBF with arithmetic correction term (AIBF)*

$$\frac{1}{\binom{n}{k}} \sum_{1 \leq l_1 < \ldots < l_k \leq n} B^N_{0,1}(x_{l_1}, \ldots, x_{l_k})$$

In the case of the geometric average we get $B^{GI}_{1,0}$ - *the IBF with geometric correction term (GIBF)*

$$\left[\prod_{1 \leq i_1 < \ldots < i_k \leq n} B^N_{0,1}(X_{i_1}, \ldots, X_{i_k}) \right]^{\frac{1}{\binom{n}{k}}}$$

Berger and Pericchi [1] discuss also computational simplifications of the formulas.

3. Intrinsic Priors

One of the motivations for use of the IBF was stated in the Introduction. Namely, the correction of the Bayes factor is needed if improper priors are used. Suppose that we can find priors $\pi_i(\cdot)$, $i = 0, 1$ such that we have the following expansion

$$B_{1,0} = B^N_{1,0} \, \frac{\pi_1(\hat{\theta}_1)\pi^N_0(\hat{\theta}_0)}{\pi^N_1(\hat{\theta}_1)\pi_0(\hat{\theta}_0)} \, (1 + o(1))$$

where, as in Berger and Pericchi [1], $\hat{\theta}_i$ is a maximum likelihood estimator (MLE) of the appropriate θ_i and $o(1) \to 0$ in probability under M_1 and M_0, as $n \to \infty$. Recall that $B^I_{1,0} = B^N_{1,0} \; c$, where $c =$ correction term, so if we want to find intrinsic priors, we should have asymptotically $B^I_{1,0} \approx B_{1,0}$. This would lead to

$$\frac{\pi_1(\hat{\theta}_1)\pi^N_0(\hat{\theta}_0)}{\pi^N_1(\hat{\theta}_1)\pi_0(\hat{\theta}_0)}(1 + o(1)) = \text{correction term}$$

The natural tendency is to take the limit of both sides when sample size $n \to \infty$, but such limits would depend on a true value of the parameter. So practically we have to look at two limits, one when M_1 is the true model, and one when M_0 is true. In the next section we will identify the limits for the MLEs and the correction terms. Here we would like to mention why the above asymptotic expansion is a reasonable assumption.

$$\begin{aligned} \frac{\int_{\Theta_1} f_1(x_1,\ldots,x_n|\theta_1)\pi_1(\theta_1)d\theta_1}{\int_{\Theta_0} f_0(x_1,\ldots,x_n|\theta_0)\pi_0(\theta_0)d\theta_0} &= \frac{\int_{\Theta_1} f_1(x_1,\ldots,x_n|\theta_1)\pi_1^N(\theta_1)\frac{\pi_1(\theta_1)}{\pi_1^N(\theta_1)}d\theta_1}{\int_{\Theta_0} f_0(x_1,\ldots,x_n|\theta_0)\pi_0^N(\theta_0)\frac{\pi_0(\theta_0)}{\pi_0^N(\theta_0)}d\theta_0} \\ &\asymp \frac{\pi_1(\hat\theta_1)\pi_0^N(\hat\theta_0)}{\pi_1^N(\hat\theta_1)\pi_0(\hat\theta_0)}B_{1,0}^N \end{aligned}$$

provided that π_i/π_i^N, $i = 0, 1$ can be treated as constants in places where likelihoods have peaks as $n \to \infty$. For further details see examples in Berger and Pericchi [1]; also Haughton [5] , Haughton and Dudley [6] and Erkanli [4] have relevant material.

4. Asymptotics

4.1 MLE

In order to get a more precise definition of intrinsic priors we have to look at the asymptotic behavior of the MLEs and of the correction term as n becomes large. For the MLE we rediscovered an observation of Huber [7] that the Wald [13] proof of consistency of MLE's can be applied to our situation with minor changes. We have the following

Theorem 1 *Under classical Wald [13] assumptions with the only differences that the "true" distribution, say $G(x)$, is outside the family $\{f(x|\theta),\ \theta \in \Theta\}$ and adding the following condition (instead of proving Lemma 1 of Wald) $(\exists!\ \theta^* \in \Theta)$ $E_G \log f(X|\theta^*) = \sup_{\theta\in\Theta} E_G \log f(X|\theta)$ we have $\hat\theta_n \to \theta^*$, G- a.s. as $n \to \infty$.*

Proof. Exactly along the lines of Wald [13], changing every occurrence of Wald's true $f(x|\theta_0)$ to $G(x)$. □

The problem of existence of such θ^* will be discussed elsewhere. For now it seems that the minimal requirement is convexity of Θ, if we restrict ourselves to exponential families with canonical parametrization or to normal families with mean value parametrizations.

Let us define Kullback-Leibler projections.

Definition 1 *Let $f(x|\theta)$, $\theta \in \Theta$ be a family of densities. Let $g(x)$ be another density. Then θ^* is a Kullback-Leibler projection of $g(\cdot)$ onto Θ iff θ^* is an element of Θ such that*

$$E_g \log f(X|\theta^*) = \sup_{\theta \in \Theta} E_g \log f(X|\theta)$$

provided that such θ^* is unique. In such a case we will write $P_\Theta(g) = \theta^*$.

The definition implies that $P_\Theta(g)$ is the limiting value of the MLE restricted to Θ, if samples are obtained from $g(x)$. Of course consistency of MLE's gives $P_\Theta(\ f(\cdot|\theta_0)\ ) = \theta_0$ if $g(\cdot) \equiv f(\cdot|\theta_0)$. If $g(\cdot)$ is outside $\{f(\cdot|\theta),\ \theta \in \Theta\})$ the MLE will approach the maximizer of $E_g \log f(X|\theta)$. Providing that $E_g \log g(X)$ exists, we get also that $P_\Theta(g)$ is a minimizer of $E_g \log[g(X)/f(X|\theta)]$ which is just Kullback-Leibler divergence.

4.2 Correction Terms

For the correction term in the *iid* case we have the following result.

Theorem 2 *Let $X_1, \ldots, X_n$ be an iid sample from the distribution with cdf $G(x)$. Let $h_A(X_1, \ldots, X_l) = B^N_{0,1}(X_1, \ldots, X_l)$ and $h_G(X_1, \ldots, X_l) = \log B^N_{0,1}(X_1, \ldots, X_l)$ where k is the minimal training sample size and $n \geq l \geq k$ and k. Then*

$$\frac{1}{\binom{n}{l}} \sum_{1 \leq i_1 < \ldots < i_l \leq n} h_A(X_{i_1}, \ldots, X_{i_l}) \rightarrow E_G h_A(X_1, \ldots, X_l) \text{ and}$$

$$\exp\left[\frac{1}{\binom{n}{l}}\sum_{1\leq i_1<\ldots<i_l\leq n}\{h_G(X_{i_1},\ldots,X_{i_l})\}\right]\rightarrow \exp E_G[h_G(X_1,\ldots,X_l)]$$

G - a.s. as $n \longrightarrow \infty$, provided that the limits exist.

Proof. It is simple application of U-statistics theory. Both functions h_{A} and h_{G} are symmetric with respect to the permutations of the arguments. Applying standard a.s. convergence results for U-statistics proves the theorem. □

Theorem 2 describes the a.s. behavior of the correction terms. Using U-statistics theory, as in Serfling [10] or Lee [8] , it is easy also to obtain a Central Limit Theorem for both forms of the correction term.

5. Intrinsic Equations

Let us come back to the asymptotic expansion for the IBF and intrinsic priors. If model (hypothesis) M_1 with the value of the parameter θ_1 is true, then using Theorem 1 and Theorem 2 and MLE consistency we get $\hat{\theta}_1 \longrightarrow \theta_1$, $\hat{\theta}_0 \longrightarrow P_{\Theta_0}(\theta_1)$ and (correction term) $\longrightarrow H_1(\theta_1)$ a.s. $f_1(X|\theta_1)$, where $H_1(\cdot)$ is one of the limits in Theorem 2 (arithmetic or geometric) with cdf $G(x)$ corresponding to the density $f_1(x|\theta_1)$. Instead of the somewhat cumbersome $P_{\Theta_0}(\theta_1)$ we will write $P_0(\theta_1)$. Assuming continuity of the priors we get

$$\frac{\pi_1(\theta_1)\ \pi_0^N(P_0[\theta_1])}{\pi_1^N(\theta_1)\ \pi_0(P_0[\theta_1])} = H_1(\theta_1).$$

Similarly if model M_0 with the parameter value θ_0 is true we obtain

$$\frac{\pi_1(P_1[\theta_0])\ \pi_0^N(\theta_0)}{\pi_1^N(P_1[\theta_0])\ \pi_0(\theta_0)} = H_0(\theta_0).$$

To simplify notation we will denote by $\pi_i^R \equiv \pi_i/\pi_i^N$ for $i = 0, 1$, the *relative* (to non-informative) intrinsic priors. So intrinsic priors are the solutions of the system of functional equations

$$\begin{cases} \pi_1^R(\theta_1) & = & \pi_0^R(P_0[\theta_1])H_1(\theta_1) \\ \pi_0^R(\theta_0) & = & \pi_1^R(P_1[\theta_0])/H_0(\theta_0) \end{cases}$$

Let us see what happens in the case of nested models, say $\Theta_0 \subset \Theta_1$. Then $P_1(\theta_0) = \theta_0$. Also it is not hard to see that in this case $H_0(\theta_0) = H_1(\theta_0)$. So we obtain just one functional equation

$$\pi_1^R(\theta_1) = \pi_0^R(P_0[\theta_1])H_1(\theta_1).$$

The second one is just restriction of the first equation to Θ_0. The general form of the solution is

$$\begin{cases} \pi_0^R(\theta_0) & = & u(\theta_0) \\ \pi_1^R(\theta_1) & = & u(P_0[\theta_1])H_1(\theta_1) \end{cases}$$

where $u(\cdot)$ is an arbitrary nonnegative continuous function. So on the smaller space the intrinsic priors should satisfy $\pi_1^R(\theta_0)/\pi_0^R(\theta_0) = H_1(\theta_0) \equiv H_0(\theta_0)$ The second equation says that on the level sets $P_0(\theta_1) =$ constant we have $\pi_1^R(\theta_1) \propto H_1(\theta_1)$. In the non-nested case situation is much more complicated and will be discussed elsewhere.

6. Examples

We will present three simple examples illustrating different features of the IBF method. The first two examples are taken from Berger and Pericchi [1] but they are used to emphasize different points.

Example 1. (lack of uniqueness). Let $X_1, \ldots, X_n$ be an *iid* sample from $N(\theta, 1)$.

$$M_0 : \theta \leq 0 \text{ vs. } M_1 : \theta \in \mathbf{R}$$

We will use (conditional) priors $\pi_0^N \equiv 1$ and $\pi_1^N \equiv 1$. Simple calculation for the Arithmetic IBF leads to $H_1(\theta) = \Phi(-\theta/\sqrt{2})$, where Φ-cdf of $N(0,1)$. Kullback-Leibler projections are given by

$$P_0(\theta) = \begin{cases} \theta & \text{if } \theta \le 0 \\ 0 & \text{otherwise} \end{cases}$$

The general solution of the intrinsic equations is

$$\begin{cases} \pi_0(\theta) & = & u(\theta) \text{ for } \theta \le 0 \\ \pi_1(\theta) & = & \begin{cases} u(\theta)\Phi(-\theta/\sqrt{2}) & \text{if } \theta \le 0 \\ u(0)\Phi(-\theta/\sqrt{2}) & \text{otherwise} \end{cases} \end{cases}$$

Berger and Pericchi [1] propose the solution

$$\begin{cases} \pi_0(\theta) & = & 1 = \pi_0^N(\theta) \text{ for } \theta \le 0 \\ \pi_1(\theta) & = & \Phi(-\theta/\sqrt{2}) \end{cases}$$

Another reasonable solution would be $\pi_0(\theta) = 1/\Phi(-\theta/\sqrt{2})$,which corresponds to $\pi_1 \equiv \pi_1^N$ on Θ_0 and results in

$$\begin{cases} \pi_0(\theta) & = & 1/\Phi(-\theta/\sqrt{2}) \text{ for } \theta \le 0 \\ \pi_1(\theta) & = & \begin{cases} 1 & \text{if } \theta \le 0 \\ 2\Phi(-\theta/\sqrt{2}) & \text{otherwise} \end{cases} \end{cases}$$

Example 2. (Kullback-Leibler geometry structure). Let $X_1, \ldots, X_n$ be an *iid* sample from $N(\mu, \sigma^2)$.

$$M_0 : \mu = 0,\ \sigma > 0 \ \text{ vs. } \ M_1 : \mu \in \mathbf{R},\ \sigma > 0$$

We will use (conditional) priors $\pi_0^N(\sigma) = 1/\sigma$ and $\pi_1^N(\sigma) = 1/\sigma$. Also $H_1(\mu, \sigma) = E\{(X_1 - X_2)^2/[\sqrt{\pi}(X_1^2 + X_2^2)]\}$ for the Arithmetic IBF. The Kullback-Leibler projection is $P_0(\mu, \sigma) =$

$\sqrt{\mu^2+\sigma^2}$. The general solution is

$$\begin{cases} \pi_0(\sigma) & = & u(\sigma)/\sigma \\ \pi_1(\mu,\sigma) & = & u(\sqrt{\mu^2+\sigma^2})H_1(\mu,\sigma)/\sigma \end{cases}$$

Berger and Pericchi [1] select $u \equiv 1$. The Kullback-Leibler geometry of the problem indicates that $\pi_1(\mu,\sigma) \propto H_1(\mu,\sigma)/\sigma$ along the level sets of projection which are in this case circles $\sqrt{\mu^2+\sigma^2} = \text{const}$.

Example 3. (non-existence of intrinsic priors in simple non-nested case). Let $X_1,\ldots,X_n$ be an *iid* sample from $N(\theta,1)$.

$$H_0 : \theta \leq 0 \text{ vs. } H_1 : \theta \geq 1$$

We will use (conditional) priors $\pi_0^N \equiv 1$ and $\pi_1^N \equiv 1$. The limiting version of the AIBF does not exist. The expected value of $\Phi(-X)/\Phi(X-1)$ does not exist under one of the hypothesis. Under another hypothesis the reciprocal of this expression has infinite expected value. But we have $H_i(\theta) = \exp\{E_{N(\theta,1)}\log[\Phi(-X)/\Phi(X-1)]\}$ for the Geometric IBF. Kullback-Leibler projections are $P_0(\theta_1) = 0$ for $\theta_1 \geq 0$ and $P_1(\theta_0) = 1$ for $\theta_0 \leq 0$. In order for the intrinsic equations to have solutions the following consistency conditions should be satisfied

$$\begin{cases} \pi_1(1) & = & \pi_0(0)H_1(1) \\ \pi_0(0) & = & \pi_1(1)/H_0(0) \end{cases}$$

which implies $H_1(1) = H_0(0)$, but this is not satisfied so intrinsic equations do not have a solution.

7. Conclusions

The IBF method is an attractive new approach to model selection and hypothesis testing problems. One can look at the IBF as a way of scaling the Bayes factor when improper (conditional) priors are used. Another justification of the method is the existence of the intrinsic priors for which the resulting Bayes factor is asymptotically equivalent to the IBF. In the nested case such priors can be found. The intrinsic priors along the level sets of the Kullback-Leibler projections are proportional to $H_1(\theta_1)$, the limiting value of the correction term. The situation for non-nested case requires further studies and will be dealt with elsewhere.

Acknowledgements. I am grateful to Jim Berger, Burgess Davis and especially to my advisor Tom Sellke for many fruitful discussions.

References

[1] Berger, J.O. and Pericchi L.R. (1993) The intrinsic Bayes factor for model selection and prediction., Purdue University,*Technical Report 93-43C.*

[2] Berk, R.H. (1966). Limiting behavior of posterior distributions when the model is incorrect.,*Ann. Math. Statist.* **37**,51-58.

[3] Berk, R.H. (1970), Consistency a posteriori, *Ann. Math. Statist.*, **41**,894-906.

[4] Erkanli, A. (1994). Laplace approximations for posterior expectations when the mode occurs at the boundary of the parameter space. *JASA*, **89**, 250-258.

[5] Haughton, D.M.A. (1988). On the choice of a model to fit data from an exponential family.,*Ann. Statist.*, **16**,342-355.

[6] Haughton, D. and Dudley, R. (1992). Information criteria for multiple data sets and restricted parameters. Technical Report, Dept. of Mathematical Sciences, Bentley College, Waltham.

[7] Huber, P. (1967). The behavior of maximum likelihood estimators under nonstandard conditions., *Proc. Fifth Brkeley Symp. Math. Statist. Probab.*, **1**, 221-233. Univ. California Press.

[8] Lee, A.J. (1990), *U-Statistics. Theory and Practice*, Dekker, New York.

[9] Schwarz,G. (1978). Estimating the dimension of a model. ,*Ann. Statist.*, **6** , 461-464.

[10] Serfling,R.J. (1980),*Approximation Theorems of Mathematical Statistics*, New York: John Wiley.

[11] Smith, A. F. M. and Spiegelhalter, D. J. (1980). Bayes factors and choice criteria for linear models. *J. Royal Statist. Soc. B*, **42**, 213-220.

[12] Spiegelhalter, D. J. and Smith, A. F. M. (1982). Bayes factors for linear and log-linear models with vague prior information. *J. Royal Statist. Soc. B*, **44**, 377-387.

[13] Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.*, **20** 595-601.