Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT
Journal of Statistical Planning and Inference 123 (2004) 313–345

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Partial information reference priors: derivation and interpretations

B. Clarke[a,b,*], A. Yuan[a,b]

[a] Department of Statistics, University of British Columbia, 6356 Agricultural Road, Vancouver, Canada BC V6T 1Z2

[b] National Human Genome Center at Howard University, Washington DC, USA

Received 25 October 2001; accepted 16 March 2003

Abstract

Suppose $X_1,\ldots,X_n$ are IID $p(\cdot|\theta,\psi)$ where $(\theta,\psi)\in\mathbb{R}^d$ is distributed according to the prior density $w(\cdot)$. For estimators $S_n=S(\underline{X})$ and $T_n=T(\underline{X})$ assumed to be consistent for some function of θ and asymptotically normal, we examine the conditional Shannon mutual information (CSMI) between Θ and T_n given Ψ and S_n, $I(\Theta,T_n|\Psi,S_n)$. It is seen there are several important special cases of this CSMI. We establish asymptotic formulas for various cases and identify the resulting noninformative reference priors. As a consequence, we develop the notion of data-dependent priors and a calibration for how close an estimator is to sufficiency.

Keywords: Shannon information; Objective priors; Data dependent priors; Sufficiency; Posterior normality; Entropy asymtotics

1. Introduction

Statistical estimation can be regarded as data transmission: some unknown agent sends us multiple copies of a parameter but due to random error these copies are corrupted during transmission. The error ridden copies are the data and the goal is to uncover the message intended, namely the parameter value. This is not the usual way statisticians think of estimation but it is a valid interpretation justified by information theoretic reasoning. It is also implicitly assumed by advocates of reference priors.

Reference priors emerge by asymptotically maximizing the Shannon mutual information (SMI) between a parameter Θ and data $\underline{X}=(X_1,\ldots,X_n)$, over the distribution

* Corresponding author. Tel.: +1-604-8820570; fax: +1-604-8226960.
E-mail address: bertrand@stat.ubc.ca (B. Clarke).

0378-3758/$ - see front matter
doi:10.1016/S0378-3758(03)00157-5

for Θ. The SMI is

$$I(\Theta;\underline{X}) = \int p(\theta,\underline{x}) \log \frac{p(\theta,\underline{x})}{p(\theta)\,p(\underline{x})}\, \mathrm{d}\theta\, \mathrm{d}\underline{x}, \tag{1.1}$$

in which θ and $\underline{x} = (x_1,\ldots,x_n)$ are outcomes of Θ and $\underline{X}$ with the indicated densities. It has the form of a relative entropy between a joint distribution and the product of its marginals. In general, an SMI gives a rate of transmission, in bits per unit time. Its maximum over densities for Θ gives the capacity. This is the maximal rate of transmission across the channel defined by the conditional density $p(x_i|\theta)$. The density for Θ at which this maximum is achieved is called the reference prior by statisticians or the capacity achieving source distribution by communication engineers. In statistical terms, the capacity is the maximal amount of dependency the data can have on the parameter.

One can interpret the SMI in a data compression context as well, and argue that statistical estimation can also be regarded as data compression: The data set is "compressed" into a single optimal parameter value. Both interpretations—data compression and transmission—lead to physically plausible optimality principles which permit an important role for side information. That is, side information improves data transmission and compression procedures, so we are led to hope that it can improve estimation procedures too. Henceforth we refer only to the parallel between estimation and data transmission, neglecting data compression for brevity and ease of exposition.

To seek the improvement afforded by side information we note that in the presence of side information, the SMI generalizes to a conditional SMI, CSMI. Formally, the CSMI we examine here is defined as follows. Consider a parameter $(\Theta,\Psi) = (\Theta_1,\ldots,\Theta_{d_1},\Psi_1,\ldots,\Psi_{d_2})$ with realized values $(\theta,\psi)=(\theta_1,\ldots,\theta_{d_1},\psi_1,\ldots,\psi_{d_2}) \in \mathbb{R}^{d_1+d_2}$. The parameter is equipped with a prior density $w(\theta,\psi)$, with respect to Lebesgue measure on $\mathbb{R}^d$, where $d = d_1 + d_2$. We will interpret θ as a parameter of interest and ψ as a nuisance parameter. The entries X_i in the data vector $\underline{X}$ are conditionally independent given (θ,ψ), having a density with respect to Lebesgue measure denoted by $p(x_i|\theta,\psi)$. We extract two functions from $\underline{X}$. They are $T_n = T(\underline{X}) = (T_1(\underline{X}),\ldots,T_{d_1}(\underline{X}))$ and $S_n = S(\underline{X}) = (S_1(\underline{X}),\ldots,S_{d_2}(\underline{X}))$. We interpret T_n as the primary information we actually want to use such as an estimator, and interpret S_n as subsidiary side information such as that arises from estimating nuisance parameters or other more general modeling. Thus, T_n will be associated with Θ and S_n will be associated with Ψ. Clearly, the statistics S_n and T_n have densities derived from $p(\cdot|\theta,\psi)$.

The quantity we wish to maximize asymptotically is the CSMI between Θ and T_n given (Ψ,S_n), which we denote by $I(\Theta;T_n|\Psi,S_n)$. Thus, with some abuse of notation, the CSMI we examine is

$$\begin{aligned} I(\Theta,T_n|\Psi,S_n) &= E_{\Psi,S_n} \int p(T_n,\theta|\Psi,S_n) \log \frac{p(T_n,\Theta|\Psi,S_n)}{p(T_n|\Psi,S_n)\,p(\Theta|\Psi,S_n)}\, \mathrm{d}\theta\, \mathrm{d}T_n \\ &= E_{(T_n,S_n,\Psi,\Theta)} \log \frac{p(T_n,\Theta|\Psi,S_n)}{p(T_n|\Psi,S_n)\,p(\Theta|\Psi,S_n)}. \end{aligned} \tag{1.2}$$

It is seen that the integral is the relative entropy between the joint distribution for T_n and Θ conditional on Ψ and S_n and the product of marginals for T_n and S_n, conditional on Ψ and S_n. When we need it, we will write $I(\Theta; T_n|\Psi = \psi, S_n = s)$ to mean the integral in the middle term, without the expectation, evaluated at (ψ, s). Clearly, if Θ and T_n are independent of Ψ and S_n, the conditioning has no effect. Also, if Θ and T_n are conditionally independent of each other, the CSMI is zero.

Less trivially, partial or side information may arise from the constraint that one must use a certain estimator such as the test score in item response theory. It might arise because one is must condition on a certain statistic for other reasons, such as model selection or hyperparameter estimation. In addition, there might be an importance ranking on the parameters—a vector-valued parameter might divide into two parameters, one being much more important than the other. This is the setting studied by Berger and Bernardo (1989).

Here, we asymptotically maximize several CSMI's over choices of prior density. This parallels the optimizations commonly used to get reference priors, leading us to define partial information reference priors, PIRP's. The forms we derive will show how to incorporate side information, in the data or parameter, in prior selection. We anticipate this will lead to better inferences in statistics just as it leads to better data transmission (and decoding therefrom) in information theory. Our optimizations of CSMI's involving nuisance parameters and "nuisance" statistics gives PIRP's that are often dependent on the data through S_n. Thus, in many cases, S_n will correspond to helpful side information which should generally be used. Indeed, using a prior conditional on a function of the data is much at one with making inferences conditional on the data which is standard Bayesian practice. The interpretation, and optimality, of the data-dependent priors we find here will be argued in detail in Sections 3 and 4.

Whatever the origin of the partial information, some special cases of (1.1) are familiar quantities. For instance, if S_n is constant, Ψ does not appear, and $T_n = \underline{X}$ then we get $I(\Theta, \underline{X})$, the usual SMI between a parameter and the data. This quantity was originally proposed as the optimality criterion for defining reference priors, see Bernardo (1979). For a formal proof that Jeffreys prior (proportional to the root of the determinant of the Fisher information matrix) is the reference prior in this case see Clarke and Barron (1994).

If S_n is constant and Ψ does not appear then one has

$$I(\Theta, T_n) \leqslant I(\Theta, \underline{X}), \tag{1.3}$$

the data processing inequality, see Cover and Thomas (1991). It holds for any density $w(\cdot)$, with equality if and only if T_n is sufficient for Θ. It characterizes the loss in information as a consequence of decoding using only the statistic rather than the full data set. Since the asymptotics of the right-hand side of (1.3) are well known, an asymptotic expression for the left-hand side will permit characterization of the difference, see Section 4.3.

Observe that

$$I(\Theta; T_n) = E_{T_n} D(w(\cdot|T_n) \| w(\cdot)),$$

the expected relative entropy distance between the posterior for Θ given T_n, $w(\theta|T_n)$, and the prior for Θ, $w(\theta)$ used to form it. The outer expectation is taken with respect to the marginal for the statistic T_n. Here, the relative entropy between two densities for the same random variable is $D(p_1(\cdot)\|p_2(\cdot)) = \int p_1(x)\log(p_1(x)/p_2(x))\,\mathrm{d}\mu(x)$. Thus, maximizing over $w(\theta)$ gives the prior that is "best" in the sense that the posterior it gives will be most affected by the accumulation of data. This is the "missing data" argument used in Bernardo (1979) for $I(\Theta;\underline{X})$ but applied to T_n. Equivalently, this gives the best prior to use when one intends to obtain credibility sets from the posterior $w(\theta|T_n = t)$ rather than $w(\theta|\underline{X} = \underline{x})$. Again, it is best in the sense that it is the posterior that changes most, on average, upon receipt of the data. The justification for the PIRP's we derive here is conceptually identical to this original justification of the reference prior; the only difference is that the missing data are conditional.

If S_n is constant and $T_n = \underline{X}$ then one recognizes $I(\Theta,\underline{X}|\Psi)$ as the quantity used in Berger and Bernardo (1989) to give a reference prior in the presence of a nuisance parameter Ψ. A heuristic derivation of this reference prior is given in Mukerjee and Ghosh (1992). More recently there have been numerous investigations of exactly how the different information content of nuisance parameters and parameters of interest affect inference. In particular, reference priors are recommended in their one-at-a-time version, as explained in the ordered group reference prior work of Berger and Bernardo (1992a), with as many blocks as parameters. In this case, one has a sequence of unidimensional optimizations (in θ) of the form $I(\Theta,\underline{X}|\Psi)$ so there are as many optimizations as parameters. This is consistent with the fact that Jeffreys prior does not work well in the simplest multivariate cases (e.g., normal with both parameters unknown). The priors we derive here probably require the same stepwise treatment.

Also in this case of a nuisance parameter, observe that the Berger–Bernardo procedure leads to a conditional reference prior $w(\theta|\psi)$. This is appropriate for reference conditional posterior inferences on θ given ψ, but not for marginal inferences on θ. In this case one wants a prior of the form $w(\theta)w(\psi|\theta)$ instead of $w(\psi)w(\theta|\psi)$. The latter case is the reverse reference prior, see Mukerjee and Ghosh (1992), which does not solve the marginalization paradox unlike the direct reference prior. In these cases, when n is finite, reference priors are typically discrete, see Berger et al. (1991). However, the discrete priors often converge to Jeffreys prior, see Zhang (1994), see also Berger and Bernardo (1992b).

An important stream of inquiry has been probability matching. The work of Sun and Ye (1996) and Phillippe and Robert (1998) considered the finite sample frequentist coverage properties of reference priors. More recently, Ghosh and Kim (2001) used an asymptotic coverage probability argument in a reference prior context to propose an improved prior for the Behrens–Fisher problem and Eno and Ye (2001) developed a reference prior using probability matching in a calibration model. An overview of noninformative priors was provided by Datta and Ghosh (1995)—they investigated which proposed noninformative priors satisfied various desired conditions on priors, including probability matching. See also the extensive listing and discussion in Kass and Wasserman (1996).

Our interest here will focus on $I(\Theta, T_n|S_n)$, and $I(\Theta, T_n|S_n = s)$ when the n in S_n is fixed but T_n is permitted to depend on ever more data. This is the form of the CSMI in

(1.2) that will lead to PIRP's with new, possibly data-dependent forms. Optimization for finite n probably continues to lead to discrete, data-dependent priors for θ. However, as before, when we let n increase in T_n with $S_n = s$ held fixed we get continuous data-dependent priors. (Optimizing the integral over $S_n = s$ gives data-independent priors.) Indeed, the reasoning is given heuristically in Corollary 5 of Section 3.2, which is verified in Example 7 of Section 4.2.

Other methods for dealing with partial information in the prior context were given by Sun and Berger (1998). In their Theorem 1, for instance, they optimized $I(\Theta, \underline{X}|\Psi)$ to get a ratio of determinants of variance matrices. This is similar to the result in Theorem 2 below for a very different CSMI; the similarity arises because of the asymptotic normality we use and the symmetry of the SMI.

We have commented that we get data-dependent PIRP's. While this may be unusual to those who regard optimality in the coherency sense as essential, it would not be unusual to a communications engineer. For now, we merely note that various statistical authors have recently used data-dependent priors to achieve better results. Indeed, Wasserman (2000) produced a data-dependent prior by multiplying Jeffreys prior by the exponential of the sum of an empirical relative entropy and a maximized relative entropy. Unlike Jeffreys prior, Wasserman's prior gives a proper posterior and it satisfies a second order probability matching optimality criterion. Essentially, Wasserman (2000) shows the remarkable fact that all data-independent priors are worse than his for certain normal mixture models. (See Theorem 1 in Wasserman, 2000.)

In a maximum entropy context, Mazzuchi et al. (2000a, b) also constructed a data-dependent prior: To overcome model uncertainty issues they used the data to get a partition on the real line with respect to which a prior could be specified. This approach is, like the present setting, basically information theoretic. Moreover, it provides a justification for the empirical Bayes formulation in which a hyperparameter, analogous to the partition boundaries, is estimated. See also, Mazzuchi et al. (2000a, b). Other recent instances where data-dependent priors have been examined include Raftery (1996), and Richardson and Green (1997).

The information theoretic argument we develop below shows that data-dependent priors emerge naturally and their information theoretic interpretation will serve statistical purposes better than coherency-based reasoning when the statistical problem is closer to data summarization and transmission than it is to the gambling scenarios from which coherency derives. This is a different line of reasoning from that of Mazzuchi et al. (2000a, b) who argued on the basis of goodness of fit.

The structure of this paper is as follows. In Section 2 we give the foundational results needed to get asymptotic expressions for quantities such as (1.2). In Section 3, we prove our main theorem. It allows us to identify reference priors in the presence of side information and nuisance parameters when S_n and T_n are functions of a consistent and asymptotically normal statistic—where the consistency is for a reasonable function of θ. In Section 4, we defend the use of our priors by information theoretic arguments. Then, we derive some examples of our priors in special cases and show how our results give a measure of sufficiency. The proofs of the major results are relegated to Section 5, at the end.

2. Asymptotic normality of the posterior given a CAN statistic

At the heart of our treatment of all these CSMI's is the behavior of the posterior density for Θ given T_n as n increases. This posterior will typically be asymptotically normal with a variance matrix reflecting greater dispersion than one would get from use of the full data set. Indeed, the results established here are suggested by writing $I(\Theta; T_n) = H(\Theta) - H(\Theta|T_n)$ in which $H(\cdot)$ is the entropy and assuming Theorem 1 below applies to $H(\cdot|T_n)$, the expectation of the conditional entropy $H(\cdot|T_n = t)$. Satisfactory extension of Theorem 1 below would give the asymptotics of $H(\Theta|T_n)$ and thus our key results. Here, however, we present a different easier proof.

First, we state the asymptotic results we will be applying repeatedly. For simplicity, we assume the parameter θ is of interest and that the nuisance parameter ψ does not exist. Our result is for the conditional density of θ given a single statistic $T = T_n$ assumed to be consistent, a.s. P_θ, for a function of θ that we write as $\eta(\theta)$. We require T to be asymptotically normal with a rate $\sqrt{n}$, although any rate $\rho(n)$ with $\rho(n) \to \infty$ will do. That is, we assume $\{T_n\}$ satisfies

$$\sqrt{n}(T_n - \eta(\theta)) \xrightarrow{L} N(\mathbf{0}, \Omega(\theta)) \tag{2.1}$$

in which $\Omega(\theta)$ is the asymptotic variance matrix.

Let Δ be the support of $w(\cdot)$. We require Ω be continuous and have determinant bounded away from zero and infinity. That is, we require

$$0 < \inf_{\theta\in\Delta} |\Omega(\theta)| \leqslant \sup_{\theta\in\Delta} |\Omega(\theta)| < \infty \tag{2.2}$$

and that $D\eta$ exist, be continuously differentiable and satisfy

$$0 < \inf_{\theta\in\Delta} |D\eta(\theta)| \leqslant \sup_{\theta\in\Delta} |D\eta(\theta)| < \infty. \tag{2.3}$$

We write its derivative matrix as $D\eta(\theta) = (\partial\eta_i(\theta)/\partial\theta_j)$ and assume that $D\eta$ is everywhere invertible with inverse denoted $(D\eta)^{-1}(\theta)$. Note that $\dim(\theta) = \dim(T_n) = \dim(\eta)$. (In fact, this can be relaxed to $\dim(\eta) = \dim(T_n) \geqslant \dim(\theta)$, see Clarke and Ghosh, 1995.)

The proofs of our main results rely on Edgeworth expansions to control error terms. Indeed, we use an Edgeworth expansion for the density $f_{V_n}(v|\theta)$ of

$$V_n = \sqrt{n}\Omega(\theta)^{-1/2}(T_n - \eta(\theta)) \xrightarrow{L} N(\mathbf{0}, \mathbf{I_d}). \tag{2.4}$$

The main Edgeworth condition that we impose is the following.

Condition E. $\exists q(\cdot)$ which is nonnegative, bounded and integrable in $\mathbf{R}^d$, such that as $n \to \infty$

$$\sup_{\theta\in\Delta} |f_{V_n}(v|\theta) - \phi_d(v)| \leqslant \mathrm{o}(1)q(v),$$

where the o(1) does not depend on v.

Condition **E** is satisfied for smooth functions of sample means formed from IID random variables, see Yuan and Clarke (2003). See also Bhattacharya and Rao (1986,

Theorem 19.2, p. 192) and Clarke and Ghosh (1995, Proposition 2.1) for similar results. Reiss (1989, Chapter 4) permits analogous results when percentiles are used in place of sample means.

Our first result gives convergence of the marginal for T_n to a constant. It is the key result needed for the formal proofs in Section 5. We comment that the following results hold in the same mode as the convergence of T_n to $\eta(\theta_0)$, the notation P_{θ_0}, a.s. means the result is true a.s.P_{θ_0} or P_{θ_0} depending on the mode of convergence of T_n to $\eta(\theta_0)$.

Proposition 1. *Suppose that T_n is as in* (2.1) *for θ a.e. with respect to $w(\cdot)$, that Condition E is satisfied and that $\eta(\theta)$ is locally invertible on an open set containing θ_0 in the interior of Δ. Suppose w is continuous and bounded on the parameter space and positive at θ_0. Write $\Sigma(\theta) = (D\eta)^{-1}(\theta)\Omega(\theta)(D\eta)^{-1}(\theta)^T$, and $\Sigma(\theta)^{1/2} = (D\eta)^{-1}(\theta)\Omega(\theta)^{1/2}$. Let $m_{T_n}(\cdot) = \int w(\theta)\, p_{T_n}(\cdot|\theta)$ be the mixture distribution of T_n with respect to $w(\theta)$. Then, as $n \to \infty$,*

$$m_{T_n}(\cdot) \sim w(\theta_0)|(D\eta)^{-1}(\theta_0)| \quad P_{\theta_0}, \quad \text{a.s.} \tag{2.5}$$

Proof. Deferred to Section 5.

Our heuristics for the behavior of the CSMI are based on the following asymptotic normality result. We comment that a proof of it can be developed from a close study of the proof of Proposition 1.

Theorem 1. *Assume the hypotheses of Proposition* 1. *For any fixed $a, b \in \mathbb{R}^d$ we have that*

$$\int_{\eta^{-1}(T_n)+\Sigma(\eta^{-1}(T_n))^{1/2}a/\sqrt{n}}^{\eta^{-1}(T_n)+\Sigma(\eta^{-1}(T_n))^{1/2}b/\sqrt{n}} w(\theta|T_n)\,\mathrm{d}\theta \to \Phi_d(b) - \Phi_d(a), \quad P_{\theta_0}, \quad \text{a.s.} \tag{2.6}$$

as $n \to \infty$, where $\Phi_d(x)$ is the distribution function of the d-dimensional standard normal $N(\mathbf{0}, I_d)$.

Proof. For detailed proof, see Yuan and Clarke (2003).

In principle, Theorem 1 could be extended to give our main result. However, it would take a lot of work. Alternatively, we could appeal to frequentist asymptotics: recognize that $I(\Theta, S_n)$ is the integral over θ of the relative entropy between $P_\theta^m(s)$ and $M_n(s)$, $D(P_\theta^m(s)\|M_n(s))$, and then, parallel to Clarke and Barron (1990), conjecture

$$D(P_\theta^m(s)\|M_n(s)) = \frac{d}{2}\log\frac{n}{2\pi e} + \frac{1}{2}\log\det|I_S(\theta)| + \log(1/w(\theta)) + \mathrm{o}(1)$$

in which I_S is the Fisher information of the sampling distribution of the statistic S_n. If this expression was established uniformly in θ, one could integrate it to get the main results here. This approach would lead to hypotheses on the density of S_n rather than on the density of the X_i's which would be harder to verify and interpret. Rather than

using this frequentist approach or extending Theorem 1, we develop a proof based on Proposition 1.

3. Implications for reference priors

Here we generalize the established concept of reference priors to include the notion of partial information discussed in the Introduction. Thus we optimize CSMI's to find PIRP's, paralleling the way reference priors are obtained from SMI's. Recall the reference prior methodology is to maximize the constant term in an asymptotic expansion for an SMI. Earlier definitions of reference prior, see Bernardo (1979), Berger and Bernardo (1989, Eq. (5.5)) maximized first and then took limits. This is equivalent to our approach, but is much harder to establish. Others, see Berger and Bernardo (1992b) defined reference posteriors. We prefer to retain the simplicity of the methodological definition. For our purposes here, we define the PIRP to be the result of maximizing the constant term in an asymptotic approximation, accurate to order o(1), for a mutual information over a density for one of its arguments. The set over which the optimization is done may include conditioning in the density and the mutual information itself may involve conditioning. The usefulness of this definition rests on the fact that asymptotic approximations for the mutual information found to date depend on the marginal density only in the constant and lower order terms.

To go beyond the mutual informations studied up to this point, note that the general quantity (1.2) can be written

$$I(\Theta; T_n|\Psi, S_n) = I(\Theta; S_n, T_n|\Psi) - I(\Theta; S_n|\Psi). \tag{3.1}$$

From (3.1) we see that generalizing the Berger–Bernardo context to permit general statistics T_n in place of $\underline{X}$ will extend readily to the case that one has a "nuisance statistic" S_n as well as a nuisance parameter Ψ. This follows because (S_n, T_n) in the first term on the right in (3.1) is amenable to the same treatment as S_n in the second term, so that both terms are CSMI's between a parameter and a statistic.

Suppose the statistics, T_n and S_n are themselves functions of consistent and asymptotically normal statistics. Generally, this means T_n and S_n are CAN themselves, although the consistency will be for a function of the parameter and the asymptotic variance of the statistic will be a transformed version of the asymptotic variance for the consistent estimator. Write $T_n = g_1((1/n)\sum_i h_1(X_i))$ and $S_n = g_2((1/n)\sum_i h_2(X_i))$, $h = (h_1, h_2)$ and

$$\sqrt{n}\Sigma(\theta,\psi)^{-1/2}\left(\frac{1}{n}\sum_i h(X_i) - \eta(\theta,\psi)\right) \xrightarrow{L} N(\mathbf{0}, I_d) \tag{3.2}$$

for a nonsingular matrix $\Sigma(\theta,\psi)$. The functions g_1 and g_2 are assumed to have at least one continuous derivative. The relationship among T_n, S_n, and h is summarized by $\eta_1(\theta,\psi) = g_1(\theta,\psi)$, $\eta_2(\theta,\psi) = g_2(\theta,\psi)$, $\Omega_1(\theta,\psi) = (D\eta_1)(\theta,\psi)\Sigma_1(\theta,\psi)(D\eta_1)(\theta,\psi)$, and $\Omega_2(\theta,\psi) = (D\eta_2)(\theta,\psi)\Sigma_2(\theta,\psi)(D\eta_2)(\theta,\psi)$, where Σ_i means the block of Σ giving the asymptotic variance of $(1/n)\sum_j h_i(X_j)$.

Now, for $Z_n = (T_n, S_n)$. We have

$$\sqrt{n}(Z_n - \eta(\theta,\psi)) \xrightarrow{L} N(\mathbf{0}, \Omega(\theta,\psi)) \tag{3.3}$$

in which $\eta(\theta,\psi) = (\eta_1(\theta,\psi), \eta_2(\theta,\psi))$ where $\eta_1(\theta,\psi)$ is the asymptotic mean of T_n and $\eta_2(\theta,\psi)$ is the asymptotic mean of S_n. Moreover, the asymptotic variance of Z_n is

$$\Omega(\theta,\psi) = \begin{pmatrix} \Omega_1(\theta,\psi) & c(\theta,\psi) \\ c(\theta,\psi) & \Omega_2(\theta,\psi) \end{pmatrix}$$

in which $\Omega_1(\theta,\psi)$ and $\Omega_2(\theta,\psi)$ are the asymptotic variances of T_n and S_n, respectively. The function $c(\theta,\psi)$ is the unspecified asymptotic covariance between T_n and S_n. We will also use

$$D\eta(\theta,\psi) = \begin{pmatrix} D\eta_1(\theta,\psi) & \mathbf{0} \\ \mathbf{0} & D\eta_2(\theta,\psi) \end{pmatrix}.$$

For future use, we define

$$W_n = \sqrt{n}\Omega(\theta,\psi)^{-1/2}(Z_n - \eta(\theta,\psi)),$$
$$V_n = \sqrt{n}\Omega_2(\theta,\psi)^{-1/2}(S_n - \eta_2(\theta,\psi)),$$

and denote their density functions by $f_{W_n}(w|\theta,\psi)$ and $f_{V_n}(v|\theta,\psi)$, respectively. We refer to the left-hand sides as standardized sums. It is seen that W_n and V_n are asymptotically normal in distribution.

The main theorem gives the general form of PIRP's. We limit our attention to the case that T_n, S_n, and θ are continuous. It can be conjectured from the form of the proof that the result holds when T_n and S_n are discrete. Indeed, the proof of our theorem can be modified to cover those cases.

3.1. Asymptotic form for PIRP's

Before stating the main theorem, we state a proposition that will control two of the terms arising in its proof. Let the entropy, $H(Y)$, of a random variable Y be defined by

$$H(Y) = -\int f_Y(y)\log f_Y(y)\,\mathrm{d}y.$$

Our statements follow from the main theorem in Barron (1986, p. 338) which gives the convergence of the entropy of standardized sums under hypotheses weaker than what we have assumed here. Our proposition gives the convergence of $H(S_n)$ and $H(V_n)$ when S_n is a function $g(\cdot)$ of a sum of IID random variables.

Proposition 2. *Let $g(\cdot)$ be a continuously differentiable, vector-valued function of a vector-valued argument, both of dimension d_2. Let $h(\cdot)$ be a d_2-dimensional vector-valued function of X_i, so that $h(X_1)$ has a finite second moment. Set*

$$S_n = g\left(\frac{1}{n}\sum_{i=1}^{n} h(X_i)\right).$$

Then we have the following.

(I) *If $H(S_n)$ exists and is finite for some n, then for that value of the parameter, the standardized sum satisfies*

$$H(V_n) \to H(V) = \frac{d_2}{2}\log(2\pi e). \tag{3.4}$$

(II) *In addition, for that value of the parameter, the entropy of the transformed mean satisfies*

$$H(S_n) - \frac{d_2}{2}\log\frac{n}{2\pi e} \to \frac{1}{2}\log(|\Omega(\theta,\psi)|). \tag{3.5}$$

Proof. Deferred to Section 5.

Finally, we give the main theorem. The intuitive content is that the CSMI has an asymptotic expansion of the form $(d_2/2)\log n$ plus a constant which we identify, with error o(1) as $n \to \infty$. One consequence of this is that a reference prior can be identified. It represents the source permitting the most rapid rate of transmission of outcomes of T_n, given that both the sender and receiver know the outcome of S_n. It does not depend on S_n because we have taken the limit as $n \to \infty$. To set up the statement of the theorem, suppose that Ψ does not appear and let

$$R_{(w,T_n,S_n)}(\theta) = \int\int p_{T_n,S_n}(t,s|\theta)\log\frac{p_{T_n,S_n}(t,s|\theta)m_{S_n}(s)}{m_{T_n,S_n}(t,s)p_{S_n}(s|\theta)}\,\mathrm{d}t\,\mathrm{d}s, \tag{3.6}$$

where $m_{T_n,S_n}(t,s) = \int p_{T_n,S_n}(t,s|\theta)w(\theta)\,\mathrm{d}\theta$ and $m_{S_n}(s) = \int p_{S_n}(s|\theta)w(\theta)\,\mathrm{d}\theta$. Recall that $\dim(T) = \dim(\eta_1) = \dim(h_1) = \dim(\theta) = d_1$ and $\dim(S) = \dim(\eta_2) = \dim(h_2) = d_2$. (In fact, we will use $d_2 = \dim(\psi)$ in the next section.) For convenience, we use the relation $\Omega(\theta) = D\eta(\theta)\Sigma(\theta)D\eta(\theta)$ for Ω_1, Ω_2, as well as for Ω. Now, we have the following.

Theorem 2. *Assume all the hypotheses of Propositions* 1 *and* 2, *and let Z_n and S_n be as in Proposition* 2. *In particular, assume that $|\Omega(\theta)|$ is bounded above and bounded away from zero from below, that $\eta(\theta)$ and $D\eta(\theta)$ are continuous and invertible. Suppose w has finite entropy $H(w)$ and that Condition E is satisfied for both $f_{W_n}(\cdot)$ and $f_{V_n}(\cdot)$.*

Then:

(I) *We have the asymptotic expansion*

$$\lim_{n\to\infty}\left(R_{(w,T_n,S_n)}(\theta) - \frac{d_1}{2}\log\frac{n}{2\pi e}\right) = \log\left(\frac{|\Sigma_2(\theta)|}{|\Sigma(\theta)|}\right)^{1/2} - \log w(\theta). \tag{3.7}$$

(II) *Assuming expression* (3.7) *holds uniformly in θ, we can optimize*

$$I(\Theta, T_n|S_n) = \frac{d_1}{2}\log\frac{n}{2\pi e} + \int w(\theta)\log\left(\frac{|\Sigma_2(\theta)|}{|\Sigma(\theta)|}\right)^{1/2} + H(w) + \mathrm{o}(1) \tag{3.8a}$$

to find that the PIRP $w^*(\cdot)$ *for using* T_n *conditional on* S_n *is*

$$w^*(\theta) = \frac{1}{C}\left(\frac{|\Sigma_2(\theta)|}{|\Sigma(\theta)|}\right)^{1/2} \tag{3.8b}$$

where C *is the normalizing constant.*

Proof. Deferred to Section 5.

We have limited our attention to proper priors since the information theoretic model requires that the source distribution of the θ's integrate to one. Note that the form of the optimizing prior here and in the corollaries below is a ratio of asymptotic variances. This is why Fishers information appears in cases when an efficient statistic exists as a function of $\underline{X}$. We use this with (1.3) to define relative sufficiency as a parallel to relative efficiency in Section 4.3.

Note that noninformative or objective priors such as the one identified in Theorem 2 usually are improper when taken over a whole real space. In practice, one truncates to a compact set and normalizes to get propriety. This will essentially always lead to a proper posterior. One can inquire about the limiting properties of the posterior if one permits the compact set supporting the prior to increase. In general, however, the resulting posteriors will depend on the sequence of compact sets chosen, even in normal examples, see Berger and Bernardo (1991). Although this deficiency seems unavoidable mathematically, it does make good intuitive sense: If there really is so little information that the prior is improper, at best it can provide, by the size of its density, a relative measure of how difficult it is to estimate a given parameter value. Otherwise put, there just is not enough information to permit finite amounts of data to provide useful inferences.

It seems exceptions can arise only when the tails of the likelihood are so tight that they overcome the dispersion of the prior with a finite number of data points. In these cases, the posterior will be proper even when the prior is not. Moreover, one expects the resulting inferences will be independent of any increasing sequence of compact sets because the limiting posterior will be proper.

3.2. Corollaries to the main theorem

When conditioning on two variables there are nine cases to consider: Each random variable can be absent, conditioned on as a realized value, and averaged over as a random quantity. These are: $I(\Theta, T_n)$, $I(\Theta, T_n|S_n = s)$, $I(\Theta, T_n|S_n)$; $I(\Theta, T_n|\Psi = \psi)$, $I(\Theta, T_n|\Psi = \psi, S_n = s)$, $I(\Theta, T_n|\Psi = \psi, S_n)$, $I(\Theta, T_n|\Psi)$, $I(\Theta, T_n|\Psi, S_n = s)$, and $I(\Theta, T_n|\Psi, S_n)$.

Theorem 2 handles the third information directly and is the simplest case we can use as a template to help get asymptotic expressions for the other eight. Indeed, the third information simplifies to give a form for the first, given below as Corollary 1, and the second information is seen to be a special case of the fifth (set Ψ constant), handled in Corollary 5, below. The other six informations are covered directly in the corollaries or as special cases of them. Without further comment, the proofs are deferred to Section 5.

Suppose S_n is absent, or equivalently, a constant. Then, we have the following asymptotics.

Corollary 1. *Under the assumptions of Theorem* 2, *The SMI has the asymptotic form*

$$I(\Theta, T_n) \sim \frac{d_1}{2} \log \frac{n}{2\pi e} + \int w(\theta) \log |\Sigma_1(\theta)|^{-1/2} \, \mathrm{d}\theta + H(\Theta), \tag{3.9}$$

and the reference prior is

$$w^*(\theta) = \frac{1}{C} |\Sigma_1(\theta)|^{-1/2} \tag{3.10}$$

in which C is the normalizing constant.

Remark. In Section 4.3 we will use the difference between $I(\Theta, \underline{X})$ and $I(\Theta, T_n)$ to define an index of sufficiency.

Note that this is consistent with using the chain rule for SMI, $I(\Theta; T_n) = H(\Theta) - H(\Theta|T_n)$, and approximating the second term by Theorem 1.

Corollary 2. *Let Ψ be a nuisance parameter, independent of Θ, with density $\omega(\psi)$ and suppose that* $\dim(T_n) \geqslant \dim(\theta)$ *and* $\dim(S_n) \geqslant \dim(\psi)$. *If the densities of T_n and S_n, and the densities of their normalized forms W_n and V_n, conditioning on both θ and ψ satisfy the conditions in Theorem* 2 *we have that the CSMI satisfies*

$$I(\Theta, T_n|S_n, \Psi) = \frac{d_1}{2} \log \frac{n}{2\pi e} + \int\int w(\theta)\omega(\psi) \log \left(\frac{|\Sigma_2(\theta, \psi)|}{|\Sigma(\theta, \psi)|} \right)^{1/2} \mathrm{d}\theta \, \mathrm{d}\psi + H(\Theta) + H(\Psi) + \mathrm{o}(1), \tag{3.11a}$$

and the corresponding PIRP given T_n and Ψ is

$$w^*(\theta) = \frac{1}{C} \int \left(\frac{|\Sigma_2(\theta, \psi)|}{|\Sigma(\theta, \psi)|} \right)^{1/2} \omega(\psi) \, \mathrm{d}\psi \, . \tag{3.11b}$$

Remark. This is consistent with what one would expect from using asymptotic expression on the right-hand side of (3.1).

In Corollary 2, if we let Θ and Ψ be dependent and fix $\Psi = \psi$, we get a variant on the PIRP found in Theorem 1. If we take S_n to be constant, we get the reference prior for a statistic, given the nuisance parameter, parallel to Berger and Bernardo (1989). Indeed, if we set $\Psi = \psi$, we have the following.

Corollary 3. *Asymptotically, the CSMI given a parameter is*

$$I(\Theta, T_n|\psi) \sim \frac{d_1}{2} \log \frac{n}{2\pi e} + \int w(\theta|\psi) \log |\Sigma_1(\theta, \psi)^{-1/2}| \, \mathrm{d}\theta - H(\Theta|\Psi = \psi), \tag{3.12a}$$

in which case the reference prior is

$$w^*(\theta|\psi) = \frac{1}{C} |\Sigma_1(\theta, \psi)^{-1/2}|. \tag{3.12b}$$

Corollary 3 *can be extended to allow conditioning on* S_n *as well as the fixed value of* $\Psi = \psi$. *Indeed, the CSMI between* Θ *and* T_n *given* S_n *given a fixed value* $\Psi = \psi$ *is*

$$I(\Theta, T_n|S_n, \psi) = \iiint f_{(T_n,\Theta)}(t, \theta|s, \psi) f_{S_n}(s|\psi)\omega(\psi) \times \log \frac{f_{(T_n,\Theta)}(t, \theta|s, \psi)}{f_{(T_n,\Theta)}(t|s, \psi)w(\theta|s, \psi)} \, \mathrm{d}t \, \mathrm{d}s \, \mathrm{d}\theta.$$

We have the following.

Corollary 4. *For fixed* ψ, *we have*

$$I(\Theta, T_n|S_n, \psi) = \frac{d_1}{2} \log \frac{n}{2\pi e} + \int\int w(\theta|\psi) \log \left(\left(\frac{|\Sigma_2(\theta, \psi)|}{|\Sigma(\theta, \psi)|} \right)^{1/2} \right) \mathrm{d}\theta + H(\Theta|\Psi = \psi) + \mathrm{o}(1). \tag{3.13a}$$

Thus, the reference prior is

$$w^*(\theta|\psi) = \frac{1}{C} \left(\frac{|\Sigma_2(\theta, \psi)|}{|\Sigma(\theta, \psi)|} \right)^{1/2}. \tag{3.13b}$$

As a curiosity observe that if we treat (S_m, Ψ), *for fixed* m, *as if it were a whole parameter like* Ψ, *then we have*

$$I(\Theta, T_n|S_m, \Psi) \sim \frac{d_1}{2} \log \frac{n}{2\pi e} - \iiint f_{S_m}(s|\psi)\omega(\psi)w(\theta|s, \psi) \log w(\theta|s, \psi) \, \mathrm{d}s \, \mathrm{d}\theta \, \mathrm{d}\psi + \iiint f_{S_m}(s|\psi)\omega(\psi)w(\theta|s, \psi) \log |\Omega_1(\theta, s, \psi)| \, \mathrm{d}s \, \mathrm{d}\theta \, \mathrm{d}\psi \, . \tag{3.14a}$$

Thus, it would be natural to define the reference prior given (s_m, ψ) *to be*

$$w^*(\theta|s_m, \psi) = \frac{1}{C} |\Sigma_1(\theta, s_m, \psi)|^{-1/2}. \tag{3.14b}$$

Now, consider the effect of "partial" asymptotics when $n \to \infty$, *but the value of* S_n, *say* $S_n = s$, *is fixed along with* $\Psi = \psi$. *In an information theoretic setting this corresponds to treating* T_n *as stochastic but* $S_n = s$ *as a value available to the sender and the receiver. That is, we inquire what happens when asymptotics are used on the inner integral but not on the outer integral in a CSMI. The CSMI between* T_n *and* Θ *given* $(S_m, \Psi) = (s_m, \psi)$ *is*

$$I(\Theta, T_n|s_m, \psi) = \int\int w(\theta|s_m, \psi) f_{T_n}(t|\theta, s_m, \psi) \log \frac{f_{T_n}(t|\theta, s_m, \psi)}{m_{T_n}(t|s_m, \psi)} \, \mathrm{d}t \, \mathrm{d}\psi \, ,$$

where

$$m_{T_n}(t|s_m, \psi) = \int w(\theta|\psi) f_{T_n}(t|\theta, s_m, \psi) \, \mathrm{d}\theta.$$

Let $\eta(\theta, s_m, \psi)$ *be the asymptotic mean of* T_n. *That is,* $E(T_n|\theta, s_m, \psi) \to \eta(\theta, s_m, \psi)$, $(P_{\theta,s_m,\psi})$, *a.s., with asymptotic variance given by* $\Omega(\theta, s_m, \psi)$.

As a final corollary to Theorem 2, *let both* ψ *and* $S = s$ *be fixed. We verify that* (3.28a, b) *are valid more formally.*

Corollary 5. *As* $n \to \infty$,

$$I(\Theta, T_n|\psi, s_m) \sim \frac{d_1}{2} \log \frac{n}{2\pi e} + \int w(\theta|s_m, \psi) \log |\Sigma(\theta, s_m, \psi)|^{-1/2} \, \mathrm{d}\theta$$

$$- \int w(\theta|s_m, \psi) \log w(\theta|s_m, \psi) \, \mathrm{d}\theta. \tag{3.15}$$

So, the PIRP given $(S_m, \psi) = (s_m, \psi)$ *is the same as in* (3.14b).

4. Examples and implications

In the first subsection we give more details on the information theoretic interpretation for PIRP's. We argue that our priors are good in both a data compression sense of higher compression rates and in a data transmission sense of higher channel capacity. This justifies our presentation of a few standard examples in the next subsection. In the final subsection, we use our notion of PIRP's to measure how close to sufficiency a statistic is.

4.1. Information theoretic motivation

It is mathematically apparent that optimization of $I(\Theta; T_n|S_n = s)$ will lead to data-dependent priors. If the n in the T_n is allowed to increase, these priors will be continuous. The data summarization interpretation of these priors would be that one which is designing a code for the repeated compression of T_n's conditional on the random value of Θ and on the fixed value of $S_n = s$. This is a perfectly sensible coding strategy in some settings. In addition, there is a transmission interpretation of such priors. It is that many messages θ drawn from the data-dependent prior with $S_n = s$ are to be sent and the receiver will decode them from the corresponding T_n's received, again conditional on $S_n = s$. This latter case is recognized by electrical engineers as part of a multiple access channel and so is perfectly reasonable in an information theoretic sense. Here, we prefer the data transmission interpretation because it is easier to describe although the data compression interpretation may be more germane statistically. Conventional Bayesians would reject these interpretations even though reference priors that are data-independent have an analogous interpretation in terms of repeated transmission, apart from the conditioning.

Indeed, more general mutual informations permitted by (1.2) arise in network information theory contexts, such as multiple access channels, broadcast channels, relay channels, source coding with side information, and rate distortion with side information. The general point is that S_n represents side information which on average improves

data compression and transmission. The effect of conditioning on fixed values depends delicately on the data and the likelihood, as one would expect.

To a communications engineer, data dependence in source distributions, or priors, occurs quite naturally as standard modeling in many network information theory settings. Thus, to an engineer, data-dependent priors would not be surprising even though orthodox Bayesians prefer data-independent priors. The seeming conflict between the engineer and the orthodox Bayesian is resolved by realizing that the Bayesian invokes an optimality principle derived from gambling scenarios, whereas the engineer invokes an optimality principle from data transmission and compression. Under their respective principles both are right. The question remains to determine which approach—gambling or information—is appropriate for a given modeling situation. Indeed, if one chooses the form of the dependence of the prior on the data pre-experimentally then it is not clear that the coherency arguments are germane. In particular, proofs of the optimality of data independence do not explicitly construct a betting strategy that achieves the infinite gain proved to be possible in the presence of incoherency. One can argue that a sort of weak coherency such as regarding the prior density itself as a stochastic function of the data, may be a satisfactory practical resolution to the Bayesian's proper concern that the experimenter not use the data to choose a prior to get whatever result he wants. In particular, one can anticipate a sort of asymptotic coherency if the dependence of the prior on a stochastic function drops out. For instance, this may occur if the prior depends on a statistic that converges to a fixed value. It is this approach that we implicitly assume here.

Information theoretic thinking distinguishes between side information in the prior, or source, and side information in the likelihood, or receiver. Usually, information theory only supposes side information in the source because there are standard extensions of the source coding theorem and the rate distortion theorem to give optimality in this case, see Cover and Thomas (1991, Chapter 14.8,9), Blahut (1991, Chapter 9). By the symmetry of the SMI this is equivalent to considering side information only for the receiver. The presence of side information can only decrease the entropy or decrease the rate distortion function lower bound. This means that side information decreases code lengths or increases compression rates—both of which are desirable.

By contrast, side information is a comparatively unformulated concept in statistics, despite being well established. Its most obvious form may be the concept of ancillarity. An ancillary statistic has no direct bearing on estimating a parameter, but may be helpful because it identifies which part of the sample space is relevant to the problem at hand. Orthodox Bayesian thinking usually eschews side information, believing one should condition on all of the data.

The SMI arises in data transmission because its maximum is the capacity of a channel. The SMI also arises in data compression because optimizing it defines the rate distortion function, see Cover and Thomas (1991, Chapters 8 and 13). The extension of the SMI to the CSMI to include conditioning occurs in the analysis of communication networks and in rate distortion with side information, see Cover and Thomas (1991, Chapter 14). These extensions are more realistic than examining single user channels or rate distortion properties in isolation.

Inference can be regarded as data compression because information from several random variables is to be combined into one statement in a way that ensures every message that might be sent, i.e., a data set, can be well represented by a single message that might be received, i.e., a parameter value. Inference can also be regarded as data transmission in the sense that one can regard data points as the messages received with random error by a receiver when the sender really sent the true value of the parameter many times. In this case, the receiver wants to infer the parameter sent from the data received. Consequently, the receiver wants the agent sending the data to draw it from a distribution permitting the highest rate of transmission across the channel defined by the conditional density for $(X|\theta)$. This will mean that the receiver receives the bits of data as fast as possible so that inference—uncovering the parameter value—will be as rapid, in terms of sample size n, as possible too. This is another statistical meaning for the capacity, equivalent to the notion of dependency used at the beginning of Section 1.

The multiple access channel supposes several broadcasters each sending data to a common receiver. The broadcasters are the random variables the statistician gets and the receiver is the statistician's client who wants an estimate of a parameter. Thus, the multiple access channel can be regarded as a sort of crude model for what the statistician actually does.

The multiple access channel we consider here is defined as follows. Two senders X_1 and X_2 want to send information to a common receiver Y. The X_i's can be from independent trials or they can be statistics calculated from the same data. The senders must deal with background noise and interference from each other. The conditional density defining the channel is $p(y|x_1,x_2)$ and it is assumed that the candidate marginals for the pair (X_1,X_2) all satisfy $p(x_1,x_2) = p(x_1)p(x_2)$. The capacity of a channel is the maximal rate it can transmit information. The capacity of the multiple access channel is the closure of the convex hull of all (R_1,R_2) in the positive quadrant of $\mathbb{R}^2$ satisfying

$$R_1 < I(X_1;Y|X_2), \quad R_2 < I(X_2;Y|X_1), \quad \text{and} \quad R_1 + R_2 < I(X_1;X_2|Y).$$

The value $I(X_1;Y|X_2)$ is the maximum rate achievable from sender 1 to the receiver when sender 2 is not sending any information. The maximal value for this rate is $I(X_1;Y|X_2 = x_2)$ where x_2 is the value achieving the maximal conditional information between X_1 and Y. Regarding the multiple access channel as an indexed set of single user channels (with x_2 as the index) $I(X_1;Y|X_2)$ is the transmission rate achieved. Indeed, the boundary of the region of (R_1,R_2)'s has a point which represents the maximum rate sender 2 can send when sender 1 sends at his maximum rate. A more complete treatment of this is found in Cover and Thomas (1991, p. 396).

In the present context, we argue for the following procedure. Consider the results in Corollary 4 or Corollary 5, but assume ψ no longer appears. That is, we have asymptotics, and a reference prior from $I(\Theta,T_n|s_n)$. As a practical matter, we would suggest choosing n large enough that it is reasonable to assume the convergence to an analog of Jeffreys prior of the density for Θ has occurred but that the dependence of this continuous density for Θ depends on the statistic s_n. As we have seen, the typical form of reference priors is ratios of standard deviations. So, we suggest that conditional

on the data dependence, the priors resulting from this procedure will be variants of Jeffreys prior, thereby inheriting invariance properties and matching frequentist coverage properties much like Jeffreys does, but in a conditional sense. Moreover, the other optimality properties of Jeffreys priors, such as good performance in practice, should carry over as well. Indeed, the only difference is the conditioning. We remark that this amounts to an empirical Bayes approach, which has been well studied in the literature.

Why not go fully Bayesian? The key reason is that in many cases you can get a higher rate of data transmission if you condition on some of the data. Moreover, this makes good sense if the extra information S is not as important as the information in T. More formally, the expected value of $I(\Theta; T_n|S_n = s)$ over S is $I(\Theta; T_n|S_n)$. When $I(\Theta; T_n|S_n = s)$ is greater than $I(\Theta; T_n|S_n)$, optimizing $I(\Theta; T_n|S_n = s)$ will give a higher capacity and it is this capacity which is achieved by a reference prior. In such cases, the reference prior analysis incorporating dependence on the data will outperform the reference prior analysis that does not use the extra data. Of course, when $I(\Theta; T_n|S_n = s)$ is less than $I(\Theta; T_n|S_n)$, the extra information in $S = s$ is not helpful. Doubtless there are cases where $P(I(\Theta; T_n|S = s) > I(\Theta; T_n|S)) > 0.5$ indicating even odds that data dependence in the prior is likely to be beneficial. This amounts to a decision rule for when to go fully Bayes and when to remain empirical Bayes.

Historically, Bayesians have argued against data dependence in their priors. The arguments generally take the form of an analysis of betting strategies. A sophisticated treatment can be found in Purves and Freedman (1969), see also Bernardo and Smith (1994). There, betting arguments show that as a long run strategy one must post odds on the basis of a data-independent probability—essentially the prior—or be certain to go broke. Other arguments can be found in Bernardo and Smith (1994).

Here, we focus on settings in which the task of the statistician is closer to data compression and data transmission than to gambling. In such contexts, coherency arguments remain mathematically true but fail to encapsulate accurately what the actual modeling criterion really is. In particular, our procedure merely uses a different optimality criterion which is information theoretic. It remains decision theoretic in that information theory uses an entropy loss rather than a monetary loss. Moreover, information theory has a natural interpretation for conditioning on a statistic, as we have done here, rather than on the full data set. Note there is nothing sacred about entropy loss either: one can imagine settings in which criteria different from entropy loss and monetary loss are physically plausible. Such criteria would lead to different classes of priors, some of which might be data dependent.

4.2. Normal and exponential family examples

In this section we verify that the proposed priors often give recognizable quantities in typical cases. For completeness, we give one example for Theorem 2, and one for each of the five corollaries, in order. The first two are routine in that we obtain priors that reduce to the usual reference priors. The later ones are new—especially the optimal data-dependent prior in the last example. In these examples, we assume the X_i's are IID according to a parametric family of exponential form, with density function of the

form

$$f(x|\theta) = C(\theta)\exp\left(\sum_{i=1}^{d}\theta_i h_i(x)\right), \tag{4.1}$$

with respect to a dominating measure $\mu(x)$ assumed to have absorbed a data-dependent factor $h(x)$ already. It is seen that

$$E(h_i(X)) = -\frac{\partial \log C(\theta)}{\partial\theta_i}, \quad \mathrm{Cov}(h_i(X), h_j(X)) = -\frac{\partial^2 \log C(\theta)}{\partial\theta_i\partial\theta_j}. \tag{4.2}$$

Note that we derive only the form of the density of the priors without normalizing. It is only in specific applications that one would truncate to get proper posteriors, or condition on some of the data to ensure propriety of the posterior.

Example 1. Consider the normal family $N(\mu,\sigma^2)$. Let $\theta=(\mu,\sigma^2)$. $T_n=\sum_{i=1}^n X_i/n$, $S_n=(\sum_{i=1}^n X_i^2/n, \sum_{i=1}^n X_i^3/n)$. Let $\Sigma(\theta)$ and $\Sigma_2(\theta)$ be the asymptotic variance matrices for $\sqrt{n}(T_n,S_n)$ and $\sqrt{n}S_n$, respectively. Using the formula $E(X_1-\mu)^k=(k!/2^n n!)\sigma^{2n}$ if $k=2n$, and 0 if $k=2n-1$, we have

$$\Sigma_2(\theta)=\begin{pmatrix} 2\sigma^2(\sigma^2+2\mu^2) & 12\sigma^4\mu+6\sigma^2\mu^3 \\ 12\sigma^4\mu+6\sigma^2\mu^3 & 15\sigma^6+36\sigma^4\mu^2+9\sigma^2\mu^4 \end{pmatrix}$$

and

$$\Sigma(\theta)=\begin{pmatrix} \sigma^2 & 2\sigma^2\mu & 3\sigma^2(\sigma^2+\mu^2) \\ 2\sigma^2\mu & 2\sigma^2(\sigma^2+2\mu^2) & 12\sigma^4\mu+6\sigma^2\mu^3 \\ 3\sigma^2(\sigma^2+\mu^2) & 12\sigma^4\mu+6\sigma^2\mu^3 & 15\sigma^6+36\sigma^4\mu^2+9\sigma^2\mu^4 \end{pmatrix}.$$

By Theorem 2, the reference prior for Θ given S_n is

$$w^*(\theta) \propto \left(\frac{|\Sigma_2(\theta)|}{|\Sigma(\theta)|}\right)^{1/2} \propto \left|\frac{15\sigma^4-6\sigma^2\mu^2+9\mu^4}{6\sigma^2(\sigma^4-2\mu^4)}\right|^{1/2},$$

which is proportional to $1/\sigma$ when $\mu=0$. This shows that the reference prior in this case reduces to the usual reference prior one expects. More generally, when μ cannot be set equal to zero, the criterion for the prior to be well defined amounts to saying that $\mu/\sigma<1$, which is much like the signal to noise ratios regularly occurring in multiple user channels.

Example 2. Consider (4.1) with $d=1$, $Y_i=h(X_i)$, $T_n=\overline{Y}$, and set $k(\theta)=\log C(\theta)$ so that $\Sigma_1(\theta)=\mathrm{Var}_\theta(Y_1)=k''(\theta)$. Now, $\sqrt{n}\Sigma_1^{-1/2}(\theta)(T_n-\eta_1(\theta))\xrightarrow{L} N(0,1)$, so by Corollary 1 the reference prior in this case is

$$w^*(\theta) \propto |\Sigma_1^{-1/2}(\theta)| = k''(\theta)^{-1/2}.$$

For a normal family with known variance σ_0^2 and unknown mean θ, we have $h_1(X)=X$, $C(\theta)=\exp(-\theta^2/(2\sigma_0^2))$, and so $w^*(\theta)$ is proportional to a constant, as expected.

Example 3. To see that this procedure does give new forms for priors, let $\theta=(\mu,\sigma^2)$, $\mu=E(X)$ and $\sigma^2=\mathrm{Var}(X)$ and write $T_n=(\bar{X},S_n^2)$, where S_n^2 is the sample variance. Define $\eta(\theta)=\mu/\sigma^2$ and set $Y_i=(X_i,X_i^2)$. Now, Y_i has mean $\zeta(\theta)=(\mu,E_\theta(X_1^2))$, and variance $\mathrm{Var}(Y_1)=\Sigma(\theta)$. Standard asymptotics gives $\overline{Y}\overset{\text{a.s.}}{\to}\zeta$ and $\sqrt{n}(\overline{Y}-\zeta(\theta))\overset{L}{\to}N(0,\Sigma(\theta))$. For $g(y)=(y_1,y_2-y_1^2)$, we have $\eta(\theta)=g(\zeta)$ and $T_n=g(\overline{Y})$. We get $T_n\overset{\text{a.s.}}{\to}\eta(\theta)$, and $\sqrt{n}(T_n-\eta(\theta))\overset{L}{\to}N(0,\Omega(\theta))$, where $\Omega(\theta)=Dg(\theta)\Sigma(\theta)Dg(\theta)'$. Now, Corollary 1 gives

$$w^*(\theta)\propto|\Sigma^{-1/2}(\theta)|.$$

For a normal family with unknown mean μ and unknown variance σ^2, we have $h_1(X)=X$, $h_2(X)=X^2$, and $C(\theta)=\exp(-\mu^2/(2\sigma^2))/\sqrt{2\pi\sigma^2}=\sqrt{-\theta_2/\pi}\exp(\theta_1^2/(4\theta_2))$, where $\theta_1=\mu/\sigma^2$, $\theta_2=-1/(2\sigma^2)$. Let $J=1/(2(\sigma^2)^3)$ be the Jacobian of the transformation from (μ,σ^2) to (θ_1,θ_2). By (4.2), we have

$$\Sigma(\theta)=\begin{pmatrix}-1/(2\theta_2) & \theta_1/(2\theta_2^2)\\ \theta_1/(2\theta_2^2) & (1-\theta_1^2/\theta 2)/(2\theta_2^2)\end{pmatrix}.$$

Thus, $w^*(\theta)\propto J|(-\theta_2^3)^{1/2}|\propto(\sigma^2)^{3/2}$.

Example 4. Now, consider Example 1 but let σ^2 be a nuisance parameter with prior density $\pi(\sigma^2)=\exp(-\sigma^2)$ and let $w^*(\theta)=w^*(\mu,\sigma^2)$ be the PIRP from Example 1. By Corollary 2, the PIRP for μ is

$$w^*(\mu)\propto\int w^*(\mu,\sigma^2)\omega(\sigma^2)\,\mathrm{d}\sigma^2\propto\int\left|\frac{15\sigma^4-6\sigma^2\mu^2+9\mu^4}{6\sigma^2(\sigma^4-2\mu^4)}\right|^{1/2}\exp(-\sigma^2)\,\mathrm{d}\sigma^2.$$

Example 5. Consider the normal family in Example 1. Let $T_n=(\sum_{i=1}^n X_i/n,\sum_{i=1}^n X_i^2/n)$, we have

$$\Sigma_1(\theta)=\begin{pmatrix}\sigma^2 & 2\sigma^2\mu\\ 2\sigma^2\mu & 2\sigma^2(\sigma^2+2\mu^2)\end{pmatrix}.$$

Given σ^2, by Corollary 3, the PIRP for μ is

$$w^*(\mu|\sigma^2)\propto|\Sigma_1(\theta)|^{-1/2}\propto(\sigma^2)^{-3/2}.$$

Example 6. Consider the normal family and T_n and S_n as in Example 1. By Corollary 4, the PIRP for μ given σ^2 is

$$w^*(\mu|\sigma^2)\propto\left(\frac{|\Sigma_2(\theta)|}{|\Sigma(\theta)|}\right)^{1/2}\propto\left|\frac{15\sigma^4-6\sigma^2\mu^2+9\mu^4}{6\sigma^2(\sigma^4-2\mu^4)}\right|^{1/2}.$$

Note that we get the same prior as in Example 1, but the interpretation is different. The comparison suggests a flat prior for σ.

Example 7. To illustrate Corollary 5, let $X_1,\ldots,X_n$ be IID with respect to a mixture of normal densities given by

$$\alpha\phi(x|\theta_1,\sigma^2)+(1-\alpha)\phi(x|\theta_2,\sigma^2),$$

where $\phi(x|\theta,\sigma^2)$ is the density of a $N(\theta,\sigma^2)$. Here we have a three-dimensional parameter, so we need a three-dimensional statistic; a one-dimensional nuisance parameter and a one-dimensional statistic S_m for it. Set $T_n=(\sum_{i=1}^n X_i/n, \sum_{i=1}^n X_i^2/n, \sum_{i=1}^n X_i^3/n)$ and $\theta=(\theta_1,\theta_2)$. Setting $\mu'=E(X_1)=\alpha\theta_1+(1-\alpha)\theta_2$ and $(\sigma')^2=\mathrm{Var}(X_1)=\sigma^2+\alpha(1-\alpha)(\theta_1-\theta_2)^2$, Example 1 implies that the asymptotic variance matrix of $\sqrt{n}T_n$ is $\Sigma(\theta,\alpha,\sigma^2)$ equal to

$$\begin{pmatrix} (\sigma')^2 & 2(\sigma')^2\mu' & 3(\sigma')^2((\sigma')^2+(\mu')^2) \\ 2(\sigma')^2\mu' & 2(\sigma')^2((\sigma')^2+2(\mu')^2) & 12(\sigma')^4\mu'+6(\sigma')^2(\mu')^3 \\ 3(\sigma')^2((\sigma')^2+(\mu')^2) & 12(\sigma')^4\mu'+6(\sigma')^2(\mu')^3 & 15(\sigma')^6+36(\sigma')^4(\mu')^2+9(\sigma')^2(\mu')^4 \end{pmatrix}.$$

Now,

$$\begin{aligned} |\Sigma(\theta,\alpha,\sigma^2)| &= 12(\sigma')^8((\sigma')^4-2(\mu')^4) \\ &= 12(\sigma^2+\alpha(1-\alpha)(\theta_1-\theta_2)^2)^4[(\sigma^2+\alpha(1-\alpha)(\theta_1-\theta_2)^2)^2 \\ &\quad -2(\alpha\theta_1+(1-\alpha)\theta_2)^4)]. \end{aligned}$$

Let S_m be an estimate of σ^2 based on $X_1,\ldots,X_m$, and regard α be a nuisance parameter. By Corollary 5, the PIRP for θ given $(S_m,\alpha)=(s_m,\alpha)$ is

$$\begin{aligned} w^*(\theta|s_m,\alpha) &\propto |\Sigma(\theta,\alpha,s_m)|^{-1/2} \\ &\propto (s_m+\alpha(1-\alpha)(\theta_1-\theta_2)^2)^{-2}[(s_m+\alpha(1-\alpha)(\theta_1-\theta_2)^2)^2 \\ &\quad -2(\alpha\theta_1+(1-\alpha)\theta_2)^4)]^{-1/2}. \end{aligned}$$

Setting $\alpha=0,1$ gives special cases. In view of the work by Wasserman (2000), the appearance of a data-dependent prior in a mixture context is no surprise. On the other hand, the prior here emerged from an optimality criterion different from that used by Wasserman (2000). Indeed, Wasserman's prior is optimal but not necessarily unique even under his criterion. Although any small perturbation of Wasserman's prior will also satisfy second order probability matching, it does not appear that our prior can be regarded as a perturbation of his.

4.3. An index of sufficiency

In this section we explicitly indicate the dependence of $I(\Theta,T_n)$ on $w(\cdot)$ by writing $I_w(\Theta,T_n)$. It is well known, see Kullback and Leibler (1951) that for any $w(\cdot)$

$$I_w(\Theta,T_n)\leqslant I_w(\Theta,\underline{X}),$$

with equality iff T_n is sufficient for Θ. Intuitively, the larger $I_w(\Theta,T_n)$ is, the more information T_n contains about Θ in the sense that there is a code for the channel defined by (Θ,T_n) with a higher rate of transmission. That is, the more sufficient T_n is for Θ, the faster we can transmit over the channel. We use the SMI as a measure of how close to sufficiency a statistic is by defining

$$\alpha_n=\alpha_{w,n}=\exp\{2(I_w(\Theta,T_n)-I_w(\Theta,\underline{X}))\} \tag{4.3}$$

as an index of sufficiency for T_n given w. Thus, a sufficient statistic has $\alpha_n = 1$, for all w and any other statistic is partially sufficient. We see that $0 \leqslant \alpha_n \leqslant 1$, $\forall w(\cdot)$, with "=1" iff T_n is sufficient. Taking a supremum over priors w results in a prior $w^*(\cdot)$ which is the prior with respect to which a given statistic T_n is most sufficient.

Let $I(\theta)$ be the Fisher information matrix of X, then Clarke and Barron (1994) give conditions so that

$$I_w(\Theta, \underline{X}) = \frac{d}{2}\log\frac{n}{2\pi e} + \int w(\theta)\log\sqrt{|I(\theta)|}\,\mathrm{d}\theta - \int w(\theta)\log w(\theta)\,\mathrm{d}\theta + \mathrm{o}(1),$$

asymptotically. By using this and Corollary 1, we get

$$\alpha_n = \exp\left\{\int w(\theta)\log\frac{|\Omega_1^{-1}(\theta)\|D\eta_1(\theta)|^2}{|I(\theta)|}\,\mathrm{d}\theta\right\} + \mathrm{o}(1), \tag{4.4}$$

a quantity reminiscent of the term optimized in Berger and Bernardo (1989). We denote the leading term in (4.4) by α and call it the asymptotic index of sufficiency.

This definition is reasonable because when T_n and V_n are functions of CAN statistics and satisfy $I(w;T_n) = I(w;V_n)$ they have the same α. Indeed, suppose T_n can be expressed as a function of CAN estimator of θ: $\sqrt{n}\Sigma(\theta)^{-1/2}(g(\underline{X}) - \theta) \xrightarrow{L} N(\mathbf{0}, I_d)$, and $T_n = h_1(g(\underline{X}))$, $\eta_1(\theta) = h_1(\theta)$. Then $\Omega_1(\theta) = D\eta_1(\theta)\Sigma(\theta)D\eta_1(\theta)'$. This leads to $w^*(\theta) = |\Omega_1^{-1}(\theta)|^{1/2}|D\eta_1(\theta)|^2/c_2$, and $\alpha_n \sim \exp\{\int w(\theta)\log(|\Omega^{-1}(\theta)\|D\eta_1(\theta)|^2/I(\theta)\,\mathrm{d}\theta\}$.

If $\eta(\theta)$ is invertible, then $|\Omega_1^{-1}(\theta)|^2 D\eta_1(\theta) = \Sigma^{-1}(\theta)$. So we get $\alpha(T) = \alpha(V) = \exp\{\int w(\theta)\log|\Sigma^{-1}(\theta)|/I(\theta))\,\mathrm{d}\theta\}$, i.e., when $I_w(\Theta, T_n) = I_w(\Theta, V_n)$, T_n and V_n have the same asymptotic coefficient of sufficiency.

Now we consider the relationship between the absolute efficiency of a CAN statistic and the sufficiency of a CAN statistic. Note that efficiency is a local property, but that sufficiency is a global property. We show that

$$\alpha = \exp\left\{\int w(\theta)\log e^d_{T,T'}(\theta)\,\mathrm{d}\theta\right\}, \tag{4.5}$$

in which

$$e_{T,T'} = e_{T,T'}(\theta) = \left(\frac{|I(\theta)|}{|\Omega_1(\theta)|}\right)^{1/d},$$

is the ratio of asymptotic variances called the absolute efficiency. Here, T_n' is the best estimator of $\eta(\theta)$, with the inverse Fisher information matrix $I^{-1}(\theta)$ as its asymptotic variance and both T_n and T_n' are CAN estimators of $\eta(\theta)$. The asymptotic variance matrix of $T(\theta)$ is $\Omega_1(\theta)$. If the asymptotic variance of T_n' is $\Omega_1'(\theta)$, then we have, more generally, that

$$e_{T,T'} = \frac{|\Omega_1'(\theta)|}{|\Omega_1(\theta)|}.$$

Indeed, if δ_n and $\delta_{n'}'$ are two CAN estimators of $\eta(\theta)$,

$$\sqrt{n}(\delta_n - \eta(\theta)) \xrightarrow{L} N(0, \Sigma(\theta)),$$

$$\sqrt{n'}(\delta_{n'}' - \eta(\theta)) \xrightarrow{L} N(0, \Sigma(\theta)),$$

where $n' = n'(n)$, then the *asymptotic relative efficiency*, see Serfling (1980), of $\{\delta_n\}$ with respect to $\{\delta_{n'}\}$ is defined to be

$$e_{\delta,\delta'} = \lim_{n\to\infty} \frac{n'(n)}{n}.$$

Now, if

$$\sqrt{n}(\delta'_n - \eta(\theta)) \xrightarrow{L} N(0, \Sigma'(\theta)),$$

we get

$$e_{\delta,\delta'} = \left(\frac{|\Sigma'(\theta)|}{|\Sigma(\theta)|}\right)^{1/d},$$

and $\Sigma(\theta) - I^{-1}(\theta) \geqslant 0$, for all asymptotic variance matrices $\Sigma(\theta)$ in the sense of nonnegative definite.

Note that from (4.5), when $w(\cdot)$ is degenerate, or $e_{T,T'}(\theta)$ is independent of θ we get $\alpha = e^d_{T,T'}$. In this sense, we can think of a statistic T_n with coefficient of sufficiency α_n roughly as equivalent to a sub-sample from $\underline{X}$ with size $\alpha_n^{1/d} n$.

5. Proofs of major results

Here we have gathered the proofs of Propositions 1 and 2, Theorem 2, and Corollaries 1–5. We present them without comment.

Proof of Proposition 1. Let

$$L(n,a) = \eta^{-1}(T_n) + \Sigma(\eta^{-1}(T_n))^{1/2} a/\sqrt{n}.$$

We first prove

$$R(n) := \int_{L(n,a)}^{L(n,b)} f_{T_n}(T_n|\theta) w(\theta)\,\mathrm{d}\theta \sim w(\theta_0)|(D\eta)^{-1}(\theta_0)|(\Phi_d(b) - \Phi_d(a)),$$
$$P_{\theta_0} \quad \text{a.s.} \tag{5.1}$$

In fact, for fixed θ and n, consider the transformation on the sample space defined by

$$v' = \sqrt{n}\Omega(\theta)^{-1/2}(t - \eta(\theta)). \tag{5.2}$$

The volume elements transform as $\mathrm{d}v' = \sqrt{n}\Omega(\theta)^{-1/2}\,\mathrm{d}t$. Now,

$$\begin{aligned} f_{V_n}(v'|\theta)\,\mathrm{d}v' &= \sqrt{n}^d |\Omega(\theta)^{-1/2}| f_{V_n}(\sqrt{n}\Omega(\theta)^{-1/2}(t - \eta(\theta))|\theta)\,\mathrm{d}t \\ &= f_{T_n}(t|\theta)\,\mathrm{d}t, \end{aligned} \tag{5.3}$$

so we have a useful expression for f_{T_n}. By use of (5.3) we get that $R(n)$ is

$$R(n) = \int_{L(n,a)}^{L(n,b)} \sqrt{n}^d |\Omega(\theta)^{-1/2}| f_{V_n}(\sqrt{n}\Omega(\theta)^{-1/2}(T_n - \eta(\theta))|\theta) w(\theta)\,\mathrm{d}\theta. \tag{5.4}$$

Now, having transformed on the sample space to get an equivalent form for $R(n)$ we recognize that the integral is over the parameter space. We will transform the integral over θ to an integral over β where for fixed T_n

$$\beta = \sqrt{n}(T_n - \eta(\theta)) \tag{5.5}$$

so that $\mathrm{d}\beta = \sqrt{n}|(D\eta)(\theta)|\,\mathrm{d}\theta$ and for each fixed β, we have the inverse transform

$$\theta = \theta_n(\beta) = \eta^{-1}(T_n - \sqrt{n}^{-1}\beta) \to \theta_0, \quad P_{\theta_0} \quad \text{a.s.}$$

In (5.5), unlike (5.2), we have not used the variance matrix Ω.

To apply (5.5) in (5.4) we transform the domain of integration. Let $\overline{\beta_n}$ be the upper limit of the transformed domain of integration. We solve for $\overline{\beta_n}$ in terms of b from

$$\eta^{-1}(T_n - \sqrt{n}^{-1}\overline{\beta_n}) = \eta^{-1}(T_n) + \sqrt{n}^{-1}\Sigma(\eta^{-1}(T_n))^{1/2}b. \tag{5.6}$$

Applying $\eta(\cdot)$ to both sides of (5.6), rearranging terms and using $T_n = \eta(\eta^{-1}(T_n))$ gives

$$\overline{\beta_n} = \sqrt{n}[\eta(\eta^{-1}(T_n)) - \eta(\eta^{-1}(T_n) + \sqrt{n}^{-1}\Sigma(\eta^{-1}(T_n)^{1/2}b))]. \tag{5.7}$$

Taylor expanding η at $\eta^{-1}(T_n)$ in the right-hand side of (5.7) gives

$$\overline{\beta_n} = (D\eta)(\overline{\zeta})\Sigma(\eta^{-1}(T_n))^{1/2}b, \tag{5.8}$$

since the $\sqrt{n}$'s cancel, for some $\overline{\zeta}$ on the straightline joining $\eta^{-1}(T_n)$ and $\eta^{-1}(T_n) + \sqrt{n}^{-1}\Sigma(\eta^{-1}(T_n))^{1/2}b$.

Similarly, the lower limit of integration is

$$\underline{\beta_n} = (D\eta)(\underline{\zeta})\Sigma(\eta^{-1}(T_n))^{1/2}a, \tag{5.9}$$

where $\underline{\zeta}$ lies on the straight line joining $\eta^{-1}(T_n)$ and $\eta^{-1}(T_n) + \sqrt{n}^{-1}\Sigma(\eta^{-1}(T_n))^{1/2}a$.

Now, (5.5), (5.8), and (5.9) give that (5.4) is

$$\int_{\underline{\beta_n}}^{\overline{\beta_n}} |(D\eta)^{-1}(\theta_n(\beta))\|\Omega(\theta_n(\beta))^{-1/2}|f_{V_n}(\Omega(\theta_n(\beta))^{-1/2}\beta|\theta_n(\beta))w(\theta_n(\beta))\,\mathrm{d}\beta. \tag{5.10}$$

To verify that (5.10) converges to the right-hand side of (5.1), choose ε_1 and ε_2 positive and let N be so large that

$$U_n = \{\|T_n - \eta(\theta_0)\| < \varepsilon_1\}$$

has P_{θ_0}-probability close to one and that $\max(|a|, |b|)/\sqrt{n} < \varepsilon_2$. Henceforth, we assume $x^n \in U_n$.

For $x^n \in U_n$, for fixed a, b, and for η, $D\eta$, η^{-1}, and $(D\eta)^{-1}$ with the continuity and derivative assumptions as above we can choose $\varepsilon_3 > 0$ as a function of n to be decreasing to zero so that

$$\|\theta_n(\beta) - \theta_0\| < \varepsilon_3. \tag{5.11}$$

(Using the definition of $\theta_n(\beta)$, adding and subtracting $\eta^{-1}(T_n)$, Taylor expanding η^{-1}, and using the bounds on β from the domain of integration.)

Moreover, we also have that

$$\underline{\beta_n} \to \underline{\beta} \equiv (D\eta)(\theta_0)\Sigma(\theta_0)^{1/2}a, \quad \overline{\beta_n} \to \bar{\beta} \equiv (D\eta)(\theta_0)\Sigma(\theta_0)^{1/2}b, \quad P_{\theta_0} \quad \text{a.s.} \tag{5.12}$$

This follows from using the convergence of T_n to $\eta(\theta_0)$, the local invertibility of η, the convergence of $\eta^{-1}(T_n)$ to θ_0, and therefore the convergence of $\Sigma(\eta^{-1}(T_n))^{1/2}b$ to zero. (The latter forces $\overline{\zeta}$ and $\underline{\zeta}$ to converge to θ_0.)

Next we control the appearance of β in the conditioning argument of f_n in (5.10). We note that β is an element of $[\underline{\beta_n}, \overline{\beta_n}]$, a compact set of uniformly bounded size for $x^n \in U_n$ for fixed a and b. In fact, since β only appears as the argument of θ_n which is always within ε_3 of θ_0, to get an upper bound for (5.10) we take suprema over the values assumed by $\theta_n(\beta)$ in w and $(D\eta)$. This gives

$$\sup_{\|\theta-\theta_0\|<\varepsilon_3} w(\theta) \sup_{\|\theta-\theta_0\|<\varepsilon_3} (D\eta)^{-1}(\theta)$$

$$\times \int_{D\eta(\theta_0)\Sigma(\theta_0)^{1/2}a}^{D\eta(\theta_0)\Sigma(\theta_0)^{1/2}b} (|\Omega(\theta_n(\beta))^{-1/2}| f_{V_n}(\Omega(\theta_n(\beta))^{-1/2}\beta|\theta_n(\beta))) \,\mathrm{d}\beta. \tag{5.13}$$

We have also replaced the upper and lower limits of integration from (5.8) and (5.9) in (5.10), obtained from Taylor expansions, with their limits from (5.12). This follows by adding and subtracting (5.10) in the integral in (5.13): The remainder is negligible because the integrand is bounded by condition **E** and the domain of integration has Lebesgue measure or counting measures going to zero.

Now, we use condition **E** again. Add and subtract $|\Omega(\theta_n(\beta))^{-1/2}|\phi_d(\Omega(\theta_n(\beta))^{-1/2}\beta)$ in the integrand and then take a supremum over $\|\theta-\theta_0\| < \varepsilon_3$, where $\varepsilon_3 > 0$ is fixed. By condition **E**, the term with the difference between f_{V_n} and ϕ_d goes to zero as n increases. Letting n go to infinity, and then letting ε_3 go to zero gives the expression

$$w(\theta_0)(D\eta)^{-1}(\theta_0) \int_{D\eta(\theta_0)\Sigma(\theta_0)^{1/2}a}^{D\eta(\theta_0)\Sigma(\theta_0)^{1/2}b} \Omega(\theta_0)^{-1/2}\phi_d(\Omega(\theta_0)^{-1/2}\beta) \,\mathrm{d}\beta$$

$$= w(\theta_0)(D\eta)^{-1}(\theta_0)(\Phi(b) - \Phi(a)),$$

so (5.1) is verified.

Now we prove the conclusion of Proposition 1. First, choose an $\varepsilon > 0$ and find an M so that $\int_{-M}^{M} \phi(v)\,\mathrm{d}v \geqslant 1-\varepsilon$. Now write

$$m(T_n) = R_2(n) + R_3(n),$$

where

$$R_2(n) = \int_{[M_{1,n}, M_{2,n}]} \sqrt{n}^d |\Omega(\theta)^{-1/2}| f_{V_n}(\sqrt{n}\Omega(\theta)^{-1/2}(T_n - \eta(\theta))|\theta) w(\theta)\,\mathrm{d}\theta, \tag{5.14}$$

in which

$$M_1 = M_{1,n} = \eta^{-1}(T_n) - \sqrt{n}^{-1}\Sigma(\eta^{-1}(T_n))^{1/2}M,$$

$$M_2 = M_{2,n} = \eta^{-1}(T_n) + \sqrt{n}^{-1}\Sigma(\eta^{-1}(T_n))^{1/2}M$$

and

$$R_3(n) = \int_{[M_1,M_2]^c} \sqrt{n}^d |\Omega(\theta)^{-1/2}| f_{V_n}(\sqrt{n}\Omega(\theta)^{-1/2}(T_n - \eta(\theta))|\theta) w(\theta)\,\mathrm{d}\theta. \tag{5.15}$$

The same proof as in (5.1), with a and b replaced by $-M$ and M, gives that

$$R_2(n) \sim w(\theta_0)|(D\eta)^{-1}(\theta_0)|(\Phi(M) - \Phi(-M)), \quad P_{\theta_0} \quad \text{a.s.}$$

Consequently, as $M \to \infty$, we have

$$R_2(n) = w(\theta_0)|(D\eta)^{-1}(\theta_0)|(1 + \mathrm{o}(1)), \quad P_{\theta_0} \quad \text{a.s.} \tag{5.16}$$

Now, it remains to show that $R_3(n)$ goes to zero because its domain of integration excludes the true value θ_0. Multiplying and dividing by $D\eta$, it is straightforward to see that $R_3(n)$ is bounded from below by zero and from above by

$$\sup_\theta |(D\eta)^{-1}(\theta)| \int_{[M_1,M_2]^c} \sqrt{n}^d |\Omega(\theta)^{1/2}||D\eta(\theta)| \times f_{V_n}(\sqrt{n}\Omega(\theta)^{-1/2}(T_n - \eta(\theta))|\theta) w(\theta)\,\mathrm{d}\theta. \tag{5.17}$$

As before, we use transformation (5.5). This time however, the lower limit of integration is defined from $M_1 = \theta = \eta^{-1}(T_n - \beta/\sqrt{n})$ so that $\beta = \sqrt{n}(\eta(\eta^{-1}(T_n)) - \eta(M_1))$. From the form of M_1 we see we can use a Taylor expansion of η at $\eta^{-1}(T_n)$ as we did in (5.7). Doing the same for the upper limit of integration, we see that the error terms in that Taylor expansion can be controlled by letting $0 < \underline{\xi} < \overline{\xi} < \infty$ be the infimum and supremum of the absolute value of $(D\eta)(\theta)\Sigma(\theta)^{1/2}$. Now (5.17) is upper bounded by

$$\begin{aligned}
&\sup_\theta |(D\eta)^{-1}(\theta)| \int \chi_{[-\underline{\xi}M,\overline{\xi}M]^c}(\beta)|\Omega((\theta_n(\beta))^{1/2}| \\
&\quad \times f_{V_n}(\Omega(\theta_n(\beta))^{-1/2}\beta|\theta_n(\beta)) w(\theta_n(\beta))\,\mathrm{d}\beta \\
&\qquad \leqslant \sup_\theta |(D\eta)^{-1}(\theta)| \sup_\theta w(\theta) \int \chi_{[-\underline{\xi}M,\overline{\xi}M]^c}(\beta)|\Omega((\theta_n(\beta))^{1/2}| \\
&\quad \times f_{V_n}(\Omega(\theta_n(\beta))^{-1/2}\beta|\theta_n(\beta))\,\mathrm{d}\beta,
\end{aligned} \tag{5.18}$$

in which we have converted to M from M_1 and M_2.

Denote the integral in (5.18) by J_n. The integrand of J_n is nonnegative and for each fixed β in the domain of integration we have that $\theta_n(\beta) \to \theta_0$ a.s. P_{θ_0} as $n \to \infty$. In particular, this means that for n large enough $\theta_n(\beta)$ eventually ends up in $N(\theta_0, \delta)$. So, Condition **E** can be applied pointwise in θ.

Indeed, since $\Omega(\cdot)$ is bounded and continuous, Condition **E** gives that

$$|\Omega(\theta_n(\beta))^{-1/2}| f_{V_n}(\Omega(\theta_n(\beta))^{-1/2}\beta|\theta_n(\beta)) \to |\Omega(\theta_0)^{-1/2}|\phi_d(\Omega(\theta_0)^{-1/2}\beta).$$

By Condition E and the Dominated Convergence Theorem, we have

$$\lim_n J_n \leqslant \int \lim_n (\chi_{[-\underline{\xi}M,\bar{\xi}M]^c}(\beta)|\Omega(\theta_{V_n}(\Omega(\theta_n(\beta))^{-1/2}\beta|\theta_n(\beta)))\,\mathrm{d}\beta$$

$$= \int_{[-\underline{\xi}M,\bar{\xi}M]^c} \Omega(\theta_0)^{-1/2}\phi_d(\Omega(\theta_0)^{1/2}\beta)\,\mathrm{d}\beta \tag{5.19}$$

which is arbitrarily small for large M. Thus (5.16) and (5.19) together gives Proposition 1. □

Proof of Proposition 2. (I) First consider the sum of IID random variables $(1/n)\sum_{i=1}^n h(X_i)$, and denote its standardized sum by U_n. Letting $\zeta(\theta,\psi)=E_{\theta,\psi}h(X_1)$, and $\Sigma(\theta,\psi)=\mathrm{Var}_{\theta,\psi}\,h(X_1)$, the standardized sum is

$$U_n = \sqrt{n}\Sigma(\theta,\psi)^{-1/2}\left(\frac{1}{n}\sum_{i=1}^n h(X_i) - \zeta(\theta,\psi)\right).$$

By Barron (1986, Theorem, p. 338), we have that

$$H(U_n) \to \frac{d_2}{2}\log(2\pi e). \tag{5.20}$$

It is seen that S_n is g applied to $(1/n)\sum_{i=1}^n h(X_i)$ and V_n is the standardized sum associated with S_n. Thus, we can use a Taylor expansion to express V_n in terms of U_n, i.e., $V_n = Q^{-1}(\zeta_n,\theta,\psi)U_n$, where ζ_n lies between $(1/n)\sum_{i=1}^n h(X_i)$ and $\zeta(\theta,\psi)$. Since $Q^{-1}(\zeta_n,\theta,\psi)$ is a random variable, it is not easy to get the density of V_n in terms of this Taylor expansion. Now we upper and lower bound $H(U_n)$, and prove that both the bounds tend to $-(d/2)\log(2\pi e)$. Here, we only construct the upper bound, the lower bound is similar. Let $\{\zeta_n'\}$ be a sequence of random variables each independent of ζ_n, defined on the same measurable space as ζ_n and also converge to $\zeta(\theta)$ in distribution uniformly in θ. We assume $\{\zeta_n'\}$ is chosen so that the entropy of $V_n' = Q^{-1}(\zeta_n',\theta,\psi)S_n$ converges to the normal entropy no faster than that of V_n does. More formally, we suppose that ζ_n' satisfies

$$\overline{\lim_n} H(V_n) \leqslant \overline{\lim_n} H(V_n').$$

Now, let $F_{\zeta_n'}$ be the distribution function of ζ_n'. We have

$$f_{V_n'}(x|\theta,\psi) = \int |Q(t,\theta,\psi)| f_{U_n}(Q(t,\theta,\psi)x|\theta,\psi)\,\mathrm{d}F_{\zeta_n'}(t)$$

$$= \int |Q(t,\theta,\psi)|(f_{U_n}(Q(t,\theta,\psi)x|\theta,\psi) - \phi_{d_2}(Q(t,\theta,\psi)x))\,\mathrm{d}F_{\zeta_n'}(t)$$

$$+ \int |Q(t,\theta,\psi)|\phi_{d_2}(Q(t,\theta,\psi)x))\,\mathrm{d}F_{\zeta_n'}(t)$$

$$= \mathrm{o}(1) + \int |Q(t,\theta,\psi)|\phi_{d_2}(Q(t,\theta,\psi)x)\,\mathrm{d}F_{\zeta_n'}(t)$$

$$\to \phi_{d_2}(Q(\theta,\psi;\theta,\psi)x) = \phi_{d_2}(x),$$

where the o(1) is by Condition E, and the resulting limiting density is by Helly's theorem (Serfling, 1980, Theorem A (iii), p. 16). Thus, we have

$$\overline{\lim_n} H(V_n') \leqslant -\int \overline{\lim_n} f_{V_n'}(x|\theta,\psi) \log\left(\overline{\lim_n} f_{V_n'}(x|\theta,\psi)\right) \mathrm{d}x$$

$$= -\int \phi_{d_2}(x) \log \phi_{d_2}(x)\, \mathrm{d}x = \frac{d_2}{2} \log(2\pi e).$$

The same lower bound is established by choosing a sequence $\{\zeta_n''\}$ which converges more quickly in entropy than $\{\zeta_n\}$ does.

(II) Since the density of S_n can be represented as

$$p_{S_n}(s|\theta,\psi) = n^{d_2/2}|\Omega(\theta,\psi)^{-1/2}| f_{V_n}(\sqrt{n}\Omega(\theta,\psi)^{-1/2}(s-\eta_2(\theta,\psi))|\theta,\psi)|,$$

we have that

$$H(S_n) = -\int n^{d_2/2}|\Omega(\theta,\psi)^{-1/2}| f_{V_n}(\sqrt{n}\Omega(\theta,\psi)^{-1/2}(s-\eta(\theta,\psi))|\theta,\psi)$$

$$\times \log(n^{d_2/2}|\Omega(\theta,\psi)^{-1/2}| f_{V_n}(\sqrt{n}\Omega(\theta,\psi)^{-1/2}(s-\eta(\theta,\psi))|\theta,\psi))\, \mathrm{d}s.$$

Transforming the integral and taking the nonstochastic factors out of the argument of the logarithm gives that the last expression is

$$-\frac{d_2}{2}\log n + \log|\Omega(\theta,\psi)|^{1/2} - \int f_{V_n}(v|\theta,\psi) \log f_{V_n}(v|\theta,\psi)\, \mathrm{d}v.$$

By (I), the last term is $(d_2/2)\log 2\pi e$ asymptotically so part (II) follows. □

Proof of Theorem 2. Observe that

$$I(\Theta, T_n|S_n) = \int w(\theta) R_{(w,T_n,S_n)}(\theta)\, \mathrm{d}\theta. \tag{5.21}$$

Now, to use Proposition 1, let $\theta' = (\theta,\theta)$ so that θ' and Z_n have the same dimension. Let $w'(\theta') = w(\theta)w(\theta)$, $p_{T_n,S_n}(t,s|\theta') = p_{T_n,S_n}(t,s|\theta)$ and $D\eta(\theta') = D\eta(\theta)$. Now,

$$m_{T_n,S_n}(t,s) = \int p_{T_n,S_n}(t,s|\theta') w'(\theta')\, \mathrm{d}\theta'. \tag{5.22}$$

Part I of the Theorem will follow if we establish an asymptotic form for $R_{(w,T_n,S_n)}(\theta)$. For any fixed θ, we have

$$R_{(w,T_n,S_n)}(\theta) = \int\int p_{T_n,S_n}(t,s|\theta) \log \frac{m_{S_n}(s)}{m_{T_n,S_n}(t,s)}\, \mathrm{d}t\, \mathrm{d}s$$

$$+ \int\int p_{T_n,S_n}(t,s|\theta) \log p_{T_n,S_n}(t,s|\theta)\, \mathrm{d}t\, \mathrm{d}s$$

$$- \int p_{S_n}(s|\theta) \log p_{S_n}(s|\theta)\, \mathrm{d}s, \tag{5.23}$$

so we can apply part (II) of Proposition 2. It gives

$$\int\int p_{T_n,S_n}(t,s|\theta)\log p_{T_n,S_n}(t,s|\theta)\,\mathrm{d}t\,\mathrm{d}s \sim \frac{(d_1+d_2)}{2}\log\frac{n}{2\pi e} - \log|\Omega(\theta)|^{1/2} \tag{5.24}$$

and

$$\int p_{S_n}(s|\theta)\log p_{S_n}(s|\theta)\,\mathrm{d}s \sim \frac{d_2}{2}\log\frac{n}{2\pi e} - \log|\Omega_2(\theta)|^{1/2}. \tag{5.25}$$

When (5.24) and (5.25) are used in (5.23), we get

$$R_{(w,T_n,S_n)}(\theta) = \int\int p_{T_n,S_n}(t,s|\theta)\log\frac{m_{S_n}(s)}{m_{T_n,S_n}(t,s)}\,\mathrm{d}t\,\mathrm{d}s + \frac{d_1}{2}\log\frac{n}{2\pi e} + \log\left(\frac{|\Omega_2(\theta)|}{|\Omega(\theta)|}\right)^{1/2} + \mathrm{o}(1). \tag{5.26}$$

Now, to get the asymptotics for $R_{(w,T_n,S_n)}(\theta)$ we need the asymptotic behavior of the first term in (5.26). When θ is the "true" value, Proposition 1 gives

$$m_{T_n,S_n}(t,s) \sim w'(\theta') = w^2(\theta)|D^{-1}\eta(\theta)| \tag{5.27}$$

and

$$m_{S_n}(s) \sim w(\theta)|D_2^{-1}\eta(\theta)|. \tag{5.28}$$

Now, for the first term in (5.26), suppose $\log|D\eta_2(\theta)|/w(\theta)|D\eta(\theta)| > 0$. (The proof is similar when it is less than zero.) Now, for large n, $\log m_{S_n}(s)/m_{T_n,S_n}(t,s)$ is continuous (a.e.) and by Proposition 1 it is bounded. So, since $(T_n,S_n)|\theta$ converges, P_θ, a.s., to some limit $(T,S)|\theta$, it also converges to the same limit in distribution. Thus

$$\int \log\frac{m_{S_n}(s)}{m_{T_n,S_n}(t,s)}\,\mathrm{d}F_{T_n,S_n}(t,s|\theta) \to \int \log\frac{|D\eta_2(\theta)|}{w(\theta)|D\eta(\theta)|}\,\mathrm{d}F_{T,S}(t,s|\theta) = \log\frac{|D\eta_2(\theta)|}{w(\theta)|D\eta(\theta)|}. \tag{5.29}$$

Using (5.29), (5.26) and the relationship $\Omega(\theta) = D\eta(\theta)\Sigma(\theta)D\eta(\theta)$, we get (3.7) for any θ in the interior of the support of w.

(II) This is a standard application of the calculus of variations. Let

$$A(\theta) = \left(\frac{|\Sigma_2(\theta)|}{|\Sigma(\theta)|}\right)^{1/2} \tag{5.30}$$

and let

$$b(w) = \int w(\theta)\log A(\theta)\,\mathrm{d}\theta - \int w(\theta)\log w(\theta)\,\mathrm{d}\theta. \tag{5.31}$$

Any prior $w(\cdot)$ can be written in the form $w(\theta) = w^*(\theta) + \lambda\zeta(\theta)$, where λ is a constant and $\zeta(\theta)$ satisfies $\int\zeta(\theta)\,\mathrm{d}\theta = 0$. Since $w^*(\cdot)$ achieves the extrema in $g(w)$ with the constraint $\int w^*(\theta)\,\mathrm{d}\theta = 1$, consider

$$B(w) = b(w) + C\int w(\theta)\,\mathrm{d}\theta.$$

Thus, we seek a constrained extrema with a Lagrange multiplier term. The calculus of variations gives that a solution must satisfy

$$0 = \left.\frac{\partial B(w^* + \lambda\zeta)}{\partial\lambda}\right|_{\lambda=0} = \int (\log A(\theta) - \log w^*(\theta) - C)\zeta(\theta)\,\mathrm{d}\theta. \tag{5.32}$$

Since $\zeta(\cdot)$ is arbitrary, we must have, for all θ, that

$$\log A(\theta) - \log w^*(\theta) - C = 0,$$

which is equivalent to

$$w^*(\theta) = A(\theta)/C \quad \forall\theta. \tag{5.33}$$

Note that this C is different from the previous one, and the probability density constraint on $w^*(\cdot)$ gives $C = \int A(\theta)\,\mathrm{d}\theta$. □

Outline of Proof of Corollary 1. We have that

$$I(\Theta, T_n) = \int w(\theta) R_{w,T_n}(\theta)\,\mathrm{d}\theta,$$

in which

$$R_{w,T_n}(\theta) = \int f_{T_n}(t|\theta)\log\frac{f_{T_n}(t|\theta)}{m_{T_n}(t)}\,\mathrm{d}t = H(T_n|\Theta=\theta) - \int f_{T_n}(t|\theta)\log m_{T_n}(t)\,\mathrm{d}t.$$

By Proposition 2, the first term above is asymptotically equivalent to

$$\frac{d_1}{2}\log\left(\frac{n}{2\pi e}\right) - \frac{1}{2}\log|\Omega_1(\theta)|,$$

and by (5.28), the second term above is asymptotically equivalent to

$$\log w(\theta) - \log|D\eta_1(\theta)|. \qquad \square$$

Outline of Proof of Corollary 2. We rewrite $I(\Theta, T_n|S_n, \Psi)$ as

$$I(\Theta, T_n|S_n, \Psi) = \iint w(\theta)\omega(\psi) R_{(w,T_n,S_n)}(\theta,\psi)\,\mathrm{d}\theta\,\mathrm{d}\psi\ ,$$

where

$$\begin{aligned} R_{(w,T_n,S_n)}(\theta,\psi) &= \iint f_{(T_n,S_n)}(t,s|\theta,\psi)\log\frac{f_{(T_n,S_n)}(t,s|\theta,\psi) f_{S_n}(s|\psi)}{f_{S_n}(s|\theta,\psi) f_{(T_n,S_n)}(t,s|\psi)}\,\mathrm{d}t\,\mathrm{d}s \\ &= \iint f_{(T_n,S_n)}(t,s|\theta,\psi)\log f_{(T_n,S_n)}(t,s|\theta,\psi)\,\mathrm{d}t\,\mathrm{d}s \\ &\quad - \int f_{S_n}(s|\theta,\psi)\log f_{S_n}(s|\theta,\psi)\,\mathrm{d}s \\ &\quad + \iint f_{(T_n,S_n)}(t,s|\theta,\psi)\log\frac{f_{S_n}(s|\psi)}{f_{(T_n,S_n)}(t,s|\psi)}\,\mathrm{d}t\,\mathrm{d}s. \end{aligned}$$

Now, by reasoning similar to that in (5.21)–(5.29), the result is true. □

Outline of Proof of Corollary 3. Observe that

$$I(\Theta, T_n|\psi) = \int w(\theta|\psi) \left(\int f_{T_n}(t|\theta,\psi) \log f_{T_n}(t|\theta,\psi)\, \mathrm{d}t \right) \mathrm{d}\theta - \int w(\theta|\psi) \left(\int f_{T_n}(t|\theta,\psi) \log f_{T_n}(t|\psi)\, \mathrm{d}t \right) \mathrm{d}\theta.$$

Asymptotically, the innermost integral in the first term is

$$\frac{d_1}{2} \log \frac{n}{2\pi e} - \log |\Omega_1(\theta,\psi)|^{1/2}.$$

As in the proof of Theorem 2, we use

$$f_{T_n}(t|\psi) \sim w(\theta|\psi)|(D_1\eta_1(\theta,\psi)|)^{-1},$$

in the innermost integral of the second term to get

$$w(\theta|\psi)|(D_1\eta_1(\theta,\psi))^{-1}|,$$

asymptotically, thereby giving (3.12a). Maximizing (3.12a) over $w(\cdot|\psi)$ for each fixed ψ, we get (3.12b). □

Outline of Proof of Corollary 4. We rewrite $I(\Theta, T_n|S_n,\psi)$ as

$$I(\Theta, T_n|S_n,\psi) = \iiint f_{(T_n,S_n)}(t,s|\theta,\psi) w(\theta|\psi) \times \log \frac{f_{(T_n,S_n)}(t,s|\theta,\psi) f_{S_n}(s|\psi)}{f_{(T_n,S_n)}(t,s|\psi) f_{S_n}(s|\theta,\psi)}\, \mathrm{d}t\, \mathrm{d}s\, \mathrm{d}\theta = \int w(\theta|\psi) R'_{(w,T_n,S_n)}(\theta|\psi)\, \mathrm{d}\theta,$$

where

$$R'_{(w,T_n,S_n)}(\theta|\psi) = \iint f_{(T_n,S_n)}(t,s|\theta,\psi) \log \left(\frac{f_{(T_n,S_n)}(t,s|\theta,\psi)}{S_n f(s|\theta,\psi)} \frac{f_{S_n}(s|\psi)}{f_{(T_n,S_n)}(t,s|\psi)} \right) \mathrm{d}t\, \mathrm{d}s$$
$$= \iint f_{(T_n,S_n)}(t,s|\theta,\psi) \log \frac{f_{S_n}(s|\psi)}{f_{(T_n,S_n)}(t,s|\psi)}\, \mathrm{d}t\, \mathrm{d}s + \iint f_{(T_n,S_n)}(t,s|\theta,\psi) \log f_{(T_n,S_n)}(t,s|\theta,\psi)\, \mathrm{d}t\, \mathrm{d}s - \iint f_{S_n}(s|\theta,\psi) \log f_{S_n}(s|\theta,\psi)\, \mathrm{d}s.$$

As in the proof of Theorem 2,

$$\iint f_{(T_n,S_n)}(t,s|\theta,\psi)\log f_{(T_n,S_n)}(t,s|\theta,\psi)\,\mathrm{d}t\,\mathrm{d}s$$
$$\sim \frac{(d_1+d_2)}{2}\log\frac{n}{2\pi e} - \log|\Omega(\theta,\psi)^{1/2}|,$$
$$\int\int f_{S_n}(s|\theta,\psi)\log f_{S_n}(s|\theta,\psi)\,\mathrm{d}s \sim \frac{d_2}{2}\log\frac{n}{2\pi e} - \log|\Omega_2(\theta,\psi)^{1/2}|,$$

and

$$f_{(T_n,S_n)}(t,s|\psi) \sim w^2(\theta|\psi)|(D_1\eta_2(\theta,\psi))^{-1}|,$$
$$f_{S_n}(s|\psi) \sim w(\theta|\psi)|(D_1\eta(\theta,\psi))^{-1}|.$$

Then we assemble the above to get the result. □

Outline of Proof of Corollary 5. Write

$$I(\Theta,T_n|\psi,s_m) = \int w(\theta|s_m,\psi)\left(\int f_{T_n}(t|\theta,s_m,\psi)\log\frac{f_{T_n}(t|\theta,s_m,\psi)}{m_{T_n}(t|s_m,\psi)}\,\mathrm{d}t\right)\mathrm{d}\theta$$
$$\times\int w(\theta|s_m,\psi)J_n(\theta,s_m,\psi).$$

By Proposition 1

$$m_{T_n}(t|s_m,\psi) \sim w(\theta|s_m,\psi)|D_1^{-1}\eta(\theta,s_m,\psi)|, \quad P_{(\theta,s_m,\psi)}, \quad \text{a.s.}$$

so we have

$$J_n(\theta,s_m,\psi) = \int f_{T_n}(t|\theta,s_m,\psi)\log f_{T_n}(t|\theta,s_m,\psi)\,\mathrm{d}t$$
$$-\int f_{T_n}(t|\theta,s_m,\psi)\log m_{T_n}(t|s_m,\psi)\,\mathrm{d}t$$
$$\sim \frac{d_1}{2}\log\frac{n}{2\pi e} - \log|\Omega_1(\theta,s_m,\psi)|^{1/2}$$
$$-\log(w(\theta|s_m,\psi)|D_1^{-1}\eta_1(\theta,s_m,\psi)|)$$
$$= \frac{d_1}{2}\log\frac{n}{2\pi e} + \log\frac{|D_1\eta_1(\theta,s_m,\psi)|}{|\Omega_1(\theta,s_m,\psi)|^{1/2}} - w(\theta|s_m,\psi).$$

By (II) of Proposition 2, the corollary holds by using the calculus of variations. □

References

Barron, A.R., 1986. Entropy and the central limit theorem. Ann. Probab. 14 (1), 336–342.
Berger, J., Bernardo, J., 1989. Estimating a product of means: Bayesian analysis with reference priors. J. Amer. Statist. Assoc. 84, 200–207.
Berger, J., Bernardo, J., 1991. Reference priors in a variance components problem. In: Goel, P., Iyengar, N.S. (Eds.), Bayesian Inference in Statistics and Econometrics. Springer, New York.
Berger, J., Bernardo, J., 1992a. Ordered group reference priors with application to the multinomial. Biometrika 25, 25–37.
Berger, J., Bernardo, J., 1992b. On the development of reference priors. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting. Clarendon Press, Oxford, UK.
Berger, J., Bernardo, J., Mendoza, M., 1991. On priors that maximize expected information. In: Klein, J.P., Lee, J.C. (Eds.), Recent Developments in Statistics and Their Applications. Freedom Academy Pub. Co.: Seoul. distributions for Bayesian inference. J. R. Statist. Soc. B 41 (2), 113–147.
Bernardo, J.M., 1979. Reference posterior distributions for Bayesian inference, with discussion, J. Roy. Statist. Soc. Ser. B 41, 113–147.
Bernardo, J.M., Smith, A.F.M., 1994. Bayes Theory. Wiley, Chichester.
Bhattacharya, R.N., Rao, R.R., 1986. Normal Approximation and Asymptotic Expansions. Robert E. Kreiger Publishing Company, Malabar.
Blahut, R.E., 1991. Principles and Practice of Information Theory. Addison-Wesley, Reading, MA.
Clarke, B., Barron, A.R., 1990. Information theoretic asymptotics of Bayes methods. IEEE Trans. Inform. Theory 36 (3), 453–471.
Clarke, B., Barron, A.R., 1994. Jeffrey's prior is asymptotically least favorable under entropy risk. J. Statist. Inference Planning 41, 37–60.
Clarke, B., Ghosh, J.K., 1995. Posterior normality given the mean with applications. Ann. Statist. 23, 2116–2144.
Cover, T.M., Thomas, J.A., 1991. Elements of Information Theory. John Wiley, Chichester.
Datta, G.S., Ghosh, M., 1995. Some remarks on noninformative priors. J. Amer. Statist. Assoc. 90, 1357–1363.
Eno, D.R., Ye, K., 2001. Probability matching priors for an extended statistical calibration model. Can. J. Statist. 29, 19–36.
Ghosh, M., Kim, Y.-H., 2001. The Behrens–Fisher problem revisited: a Bayes-frequentist analysis. Can. J. Statist. 29, 5–18.
Kass, R., Wasserman, L.A., 1996. The selection of prior distributions by formal rules. J. Amer. Statist. Soc. 91, 1343–1370.
Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Statist. 22, 79–86.
Mazzuchi, T.A., Soofi, E.S., Soyer, R., 2000a. Computation of maximum entropy Dirichlet for modeling lifetime data. Comput. Statist. Data Anal. 32, 361–378.
Mazzuchi, T.A., Soofi, E.S., Soyer, R., Retzer, J.J., 2000b. Maximum entropy modeling of consumer choice. ASA Proc. Section Bayesian Statist. Sci., to appear.
Mukerjee, R., Ghosh, J.K., 1992. Noninformative prior. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting. Clarendon Press, Oxford, UK.
Phillippe, A., Robert, C.P., 1998. A note on the confidence properties of reference priors for the calibration model. Test 7, 147–160.
Purves, R., Freedman, D., 1969. Bayes method for Bookies. Ann. Inst. Math. Statist. 40 (4), 1177–1186.
Raftery, A., 1996. Hypothesis testing and model selection via posterior simulation. In: Gilks, W.R., Richardson, S., SpiegelHalter, D.J. (Eds.), Markov Chain Monte Carlo in Practice. Chapman & Hall, London, pp. 163–188.
Reiss, R.D., 1989. Approximate Distributions of Prder Statistics. Springer, New York.
Richardson, S., Green, P.J., 1997. On Bayesian analysis of mixtures with an unknown number of components, with discussion. J. Roy. Statist. Soc. Ser. B 59, 731–792.
Serfling, R.J., 1980. Approximation Theorems of Mathematical Statistics. Wiley, Chichester.

Sun, D., Berger, J.O., 1998. Reference priors with partial information. Biometrika 85, 55–71.
Sun, D., Ye, K., 1996. Reference prior Bayesian analysis for normal mean products. J. Amer. Statist. Assoc. 90, 589–597.
Wasserman, L.A., 2000. Asymptotic inference for mixture models using data dependent priors. J. Roy. Statist. Ser. B 62, 159–180.
Yuan, A., Clarke, B., 2003. Asymptotic normality of the posterior given a statistic. Can. J. Statist., to appear.
Zhang, Z., 1994. Discrete Non-Informative Priors. Ph.D. Thesis, Department of Statistics, Yale University.