View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by DigitalCommons@University of Nebraska

University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

Faculty Publications, Department of Statistics

Statistics, Department of

12-2010

Reference priors for exponential families with increasing dimension

Bertrand Clarke
University of Miami, bclarke2@med.miami.edu

Subhashis Ghosal
North Carolina State University, subhashisghoshal@ncsu.edu

Follow this and additional works at: https://digitalcommons.unl.edu/statisticsfacpub

Part of the Other Statistics and Probability Commons

Clarke, Bertrand and Ghosal, Subhashis, "Reference priors for exponential families with increasing dimension" (2010). *Faculty Publications, Department of Statistics*. 93.
https://digitalcommons.unl.edu/statisticsfacpub/93

This Article is brought to you for free and open access by the Statistics, Department of at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Faculty Publications, Department of Statistics by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

Electronic Journal of Statistics
Vol. 4 (2010) 737–780
ISSN: 1935-7524
DOI: 10.1214/10-EJS569

Reference priors for exponential families with increasing dimension

Bertrand Clarke

Department of Medicine,
Department of Epidemiology and Public Health,
and the Center for Computational Sciences
e-mail: bclarke2@med.miami.edu
url: http://www.med.miami.edu/medicine/x2666.xml

and

Subhashis Ghosal*

Department of Statistics
North Carolina State University
e-mail: subhashis_ghoshal@ncsu.edu
url: http://www4.stat.ncsu.edu/∼ghoshal/

Abstract: In this article, we establish the asymptotic normality of the posterior distribution for the natural parameter in an exponential family based on independent and identically distributed data. The mode of convergence is expected Kullback-Leibler distance and the number of parameters p is increasing with the sample size n. Using this, we give an asymptotic expansion of the Shannon mutual information valid when $p = p_n$ increases at a sufficiently slow rate. The second term in the asymptotic expansion is the largest term that depends on the prior and can be optimized to give Jeffreys' prior as the reference prior in the absence of nuisance parameters. In the presence of nuisance parameters, we find an analogous result for each fixed value of the nuisance parameter. In three examples, we determine the rates at which p_n can be allowed to increase while still retaining asymptotic normality and the reference prior property.

AMS 2000 subject classifications: Primary 62F15; secondary 62C10.
Keywords and phrases: Objective prior, posterior normality, mutual information, increasing dimension, exponential family.

Received December 2009.

Contents

1 Introduction 738
2 Setting and assumptions 742
3 Statements of results 745
3.1 No nuisance parameters 745

*Research partially supported by NSF grant DMS 0349111; authorship is alphabetical. The authors are grateful to an anonymous referee and associate editor whose suggestions greatly improved this paper.

3.2 Nuisance parameters present 748
4 Examples 750
4.1 Independent normal model 750
4.2 Multinomial model 752
4.3 Dirichlet model 756
5 Discussion 763
6 Appendix A: Posterior normality 765
7 Appendix B: Detailed proof of Lemma 3.3 and Theorem 3.1 773
8 Appendix C: Useful lemmas 777
References 778

1. Introduction

The Shannon mutual information (SMI) was derived in Shannon (1948a), see also Shannon (1948b), as a rate of information transmission across an information-theoretic channel, that is, the electrical engineer's analog of a likelihood. Formally, the SMI for a random variable X distributed as P_θ with density p_θ and equipped with the prior density Π is

$$I(X \mid \Pi) = \int \Pi(\theta) D(P_\theta \| P_X) d\theta,$$

where $D(\cdot\|\cdot)$ is the relative entropy or Kullback-Leibler number

$$D(F\|G) = \int f(x) \log \frac{f(x)}{g(x)} d\nu(x)$$

and P_X is the marginal for X. Here, f and g are densities for distributions F and G with respect to a common dominating measure ν (suppressed in the notation). The interpretation is that someone draws a value of θ according to Π and transmits it over the channel defined by the likelihood so the receiver receives and outcome of X with conditional density of $p(x|\theta)$. The mutual information is then a transmission rate, in bits per symbol. So, the fastest information transmission will occur for the data source Π that maximizes the mutual information. The supremal transmission rate, over Π, is the capacity of the channel. In addition, if the relative entropy is regarded as a redundancy in noiseless source coding, i.e., it is the extra bits sent beyond what optimal coding would require, the mutual information is the Bayes redundancy and maximizing it gives the maximin redundancy.

In statistics, the widespread use of the SMI began with Lindley (1956). Since then, the SMI as a statistical quantity has been regarded as a measure of dependence between a parameter and data, a measure of distance between distributions, a mode of convergence, a measure of "information" in a data set, and as a sort of average learning rate (with respect to n). In fact, these various interpretations are much at one with Shannon's original communications theory interpretation.

The seminal contribution of Bernardo (1979) was to recognize that the capacity of a channel, that is, its maximal SMI, had an important interpretation in Bayesian statistics. The capacity of a channel represents the fastest learning rate a statistician could achieve on average from a fixed likelihood and that this could be effected by finding the capacity achieving source distribution, which he termed a reference prior. In fact, prior to Bernardo (1979), Ibragimov and Hasminsky (1973) established that Jeffreys' prior is the reference prior in an asymptotic sense (in the absence of nuisance parameters) without using the term reference prior; see also Clarke and Barron (1994) for a modern formulation. The technique of their proof rested on posterior normality.

Because of its desirable properties, the concept of reference priors has received extensive development, especially in a series of papers by Berger and Bernardo and their collaborators such as Berger and Bernardo (1992b), Berger and Bernardo (1992a) and Berger and Bernardo (1991). The paper Berger and Bernardo (1989) deserves special mention because it extended the concept of reference priors to include nuisance parameter cases. The derivation of their new reference prior was not given explicitly but probably relied on a calculus of variations argument applied to a heuristic asymptotic expansion. Later, Ghosh and Mukerjee (1992) presented an argument on the basis of a formal asymptotic expansion. Moreover, in the recent paper, Berger et al. (2009) the notion of a reference prior has been formalized.

From this overall body work, it can be surmised that the usual electrical engineering treatment of information concepts which mostly, but not entirely, uses discrete random variables is less appropriate in statistics where continuous variables are common. Essentially, this meant that identifying reference priors had to be done asymptotically in the sample size n, since finite n optimizations give discrete reference priors. However, see Zhang (1994) for a convergence result that applies to the case in Berger et al. (1991).

Three important contributions that largely completed the theoretical treatment of reference priors for finite dimensions and led to the present work are the following. Let X be p dimensional and have a distribution controlled by a parameter θ. Write $X^n = (X_1, \ldots, X_n)$ to mean n independent and identical (IID) outcomes of X and let $\Pi(\theta|X^n)$ be posterior density corresponding to the prior Π. Then the SMI is given by

$$I(X^n) = I(X^n \mid \Pi) = \int \Pi(\theta \mid x^n) m_n(x^n) \log \frac{\Pi(\theta \mid x^n)}{\Pi(\theta)} d\theta d\nu_n(x^n),$$

where m_n is the mixture of the n-fold product of $f(\cdot|\theta)$s with respect to Π. Understanding the asymptotics of $I(X^n)$ as n increases proceeded from Ghosh and Mukerjee (1992) to Sun and Berger (1998) who further developed the conditional mutual information given a nuisance parameter ψ and then to Clarke and Yuan (2004) who handled the general case of $I(T_n|S_n, \psi)$, in which a conditioning statistic S_n (over which the integration is done) as well as a nuisance parameter value ψ are present. The work Berger et al. (2009) extended the class of priors over which the asymptotic optimization had been done in earlier cases.

This tour-de-force showed that the previous restrictions on the prior and likelihood were convenient but not necessary. The successful extension of reference prior concepts to distances other than the relative entropy was done for the chi-squared distance in Clarke and Sun (1997) and completed in Ghosh et al. (2010). Once these results were available, the major outstanding conceptual issues for reference priors for finite dimensional parameters with independent data were largely resolved. Admittedly, there are gaps such as dealing with nuisance parameters outside the relative entropy distance definition, but it is not clear how extensively useful this would be. Aside from the case of dependent data, which is still being studied, the frontier for reference priors has shifted to high dimensional settings.

Reference priors, or more generally objective priors, beyond the finite dimensional case, have received little attention despite the popularity of Ghosh and Ramamoorthi (2003) and the rapid development of nonparametric Bayesian methods that ensued. Apart from the generic recommendation to use a noninformative base measure in a Dirichlet process prior, the main contribution to objective prior selection in the nonparametric case seems to be Ghosal et al. (1997). There, a sequence of uniform priors on carefully selected finite subsets of a class of distributions was proposed. It was shown that when this sequence has a weak limit it can correspond to a uniform distribution and reduces to the Jeffreys prior in regular parametric settings. A variant on this construction formed by taking a convex combination of those uniform distributions leads to consistent posterior even in the nonparametric setting. The posterior usually converges at the optimal rate; see Ghosal et al. (2000).

The present paper is between the finite dimensional setting that has been well studied and the purely nonparametric approach just described. That is, we find rates of increase on the number of parameters p in terms of n so that at each stage, Jeffreys' prior is the reference prior in an asymptotic sense and the sequence of posteriors formed from these priors will be asymptotically normal.

The connection between posterior normality and reference priors has been recognized since Bernardo (1979). This was used implicitly in Berger and Bernardo (1989) and explicitly in Clarke and Barron (1990). Indeed, it is easy to see that asymptotic normality of the posterior should be equivalent to the determination of the reference prior under reasonable conditions such as the parameter having fixed finite dimension, the likelihood satisfying smoothness assumptions, and the mode of convergence being strong enough, that is, essentially equivalent to convergence in the sense of the integrated Kullback-Leibler divergence.

As in the fixed dimensional case, the root of the asymptotic expansion of the SMI lies in posterior asymptotic normality in the L_1-sense. This is possible because working with the local parameter allows explicit bounds for the error in approximating a posterior density by its limiting normal form under suitable uniform integrability conditions. The study of posterior normality in increasing p setting was pioneered by Ghosal (2000); see also Ghosal (1997), Ghosal (1999), and Boucheron and Gassiat (2009). However, the approach in Ghosal (2000) does not give an estimate of the probability of the set $W^c = W_n^c$ on which the

L_1-distance between a posterior and its limiting normal may fail to be small; Ghosal (2000) only implies that the probability of W^c converges to zero. In order to be able to use the approximate normality in the L_1-sense to derive bounds for the expected Kullback-Leibler divergence, we need an explicit bound on the probability of the set W^c. Essentially, we construct a set W with high probability on which the L_1-distance between the posterior and its approximating normal density is small.

Proofs of our results on asymptotic normality in the L_1-sense are patterned after those of Ghosal (2000), but there are important technical differences. We use a different decomposition of the integrals into central and tail regions as well as higher moments to bound probabilities. To gain the necessary control on the probability of W^c, we reduce W^c further, but then to make the L_1-distance small on the larger W, we must impose stronger conditions. Although most proofs in the section on asymptotic normality use ideas already in Ghosal (2000), for the sake of self-containedness and transparency, we shall give complete proofs of most of the results on asymptotic normality in the L_1-sense. This leads to a stronger growth restriction on the dimension p as the sample size n grows.

Once asymptotic normality is obtained, we use arguments similar to those in Clarke and Barron (1990) to make the transition from L_1-distance to the expected Kullback-Leibler divergence. The resulting analysis gives an asymptotic expansion of the SMI as the sum of a dominant term free of the prior, a term that depends on the prior but does not grow with n, and another small error term. The representation is virtually identical with that in Clarke and Barron (1990), except that p can now grow to infinity as n does. Optimizing the second term over the prior establishes Jeffreys' prior as the reference prior.

It will be seen that the growth restriction of the rate $p = p_n$ depends on the specific model under consideration. In the easiest case, all the random variables $X_1, \ldots, X_p$ are independent univariate $N(\theta_i, 1)$'s for $i = 1, \ldots, n$, and the Jeffreys prior is uniform on any compact set. Then, it is enough to choose $p = \mathcal{O}(n^{1/3-\eta})$ for any $\eta > 0$. By contrast, when the X_i's are multinomial, it will be seen that a much slower growth rate of p with n, namely $p = \mathcal{O}(n^{1/9-\eta})$, appears to be required for the reference prior to exist and give posterior normality. In our third example, a Dirichlet distribution, we find order $p = \mathcal{O}(n^{1/6-\eta})$.

We do not know if these rates are the best possible, but it appears that some restrictions like these on the growth rate of p are essential. In the next section we define our setting and notation. Then, Section 3 states our main results for identifying reference priors when nuisance parameters are not present and when they are. Section 4 presents three examples of our results, the normal, the multinomial and the Dirichlet in which explicit rates on p can be given in terms of n. Section 4 presents our three examples and Section 5 gives some concluding remarks on prior selection. Section 5 discusses extensions of the present results and their implications for prior selection in high dimensions. Appendix A states and proves an asymptotic normality theorem essential to the identification of the reference priors in Section 3 and Appendix B provides some details of proof of the results in Section 3. For convenience, Appendix C gathers together some simple lemmas we use in the various derivations.

We use the following symbols throughout this paper: "$\lesssim$" means inequality "up to a constant multiple"; I_p is the identity matrix of order p; and $x^T = (x_1, \ldots, x_p)$ (respectively, A^T) stands for the transpose of a vector x (respectively, matrix $A = ((a_{ij}))$ for $i, j = 1, \ldots, p$). The notation $\|\cdot\|$ denotes the Euclidean norm for vectors as well as the operator norm for matrices, that is, $\|A\| = \sup\{\|Ax\| : \|x\| \leq 1\}$. We use $\phi_p(\cdot|\mu, \Sigma)$ to mean the p-dimensional normal density with mean vector μ and dispersion matrix $\Sigma = ((\sigma_{ij}))$ for $i, j = 1, \ldots, p$, and $a_n = \mathcal{O}(b_n)$ (respectively, $a_n = o(b_n)$) means that a_n/b_n is bounded (respectively, $a_n/b_n \to 0$). We denote a generic constant by C, not necessarily the same from occurrence to occurrence.

2. Setting and assumptions

Let $X^n = (X_1, X_2, \ldots, X_n) \overset{\text{IID}}{\sim} f(x|\theta)$, $\theta \in \Theta \subset \mathbb{R}^p$ and suppose that the dimension of the X_i's is $p = p_n \to \infty$, where densities are with respect to a σ-finite measure ν on $\mathbb{R}^p$. Each distinct value of p is regarded as a stage in the overall structure and there is no necessary linkage from one stage to the next except that we assume that there is a true value θ_0 uniformly in the interior of the p-dimensional parameter spaces, i.e., there exists an $\epsilon_0 > 0$ (fixed) such that at the pth stage $\{\theta : \|\theta - \theta_0\| < \epsilon_0\} \subset \Theta$. This means that the dimension of the true parameter is increasing but that the extra entries thereby introduced as p increases do not move the true value outside the interior of the corresponding parameter space.

We restrict to the case of natural exponential families given by $f(x|\theta) = \exp[x^T\theta - \psi(\theta)]$. The true mean is therefore $\mu = \mathrm{E}_{\theta_0}(X) = \psi'(\theta_0) \in \mathbb{R}^p$ and the $p \times p$ Fisher information matrix is given by $F = \psi''(\theta_0)$. The maximum likelihood estimator (MLE) $\hat{\theta}$ satisfies $\psi'(\hat{\theta}) = \bar{X} = n^{-1}\sum_{i=1}^n X_i$. We use P_{θ_0} to denote the true distribution of the data, where dependence on n and p is suppressed in the notation.

Let J be any square root of F, that is, $JJ^T = F$. Then, $\|J\| = \sqrt{\|F\|}$ and $\|J^{-1}\| = \sqrt{\|F^{-1}\|}$. Note that $\|F\| = \max(\lambda_1, \ldots, \lambda_d)$, the largest eigenvalue of $\|F\|$, and $\|F^{-1}\| = \max(\lambda_1^{-1}, \ldots, \lambda_d^{-1})$.

We define the local parameter $u = \sqrt{n}J^T(\theta - \theta_0)$. Thus $\theta = \theta_0 + n^{-1/2}Hu$, where $H = (J^T)^{-1}$, and hence

$$\|u\|^2 = n(\theta - \theta_0)^T JJ^T(\theta - \theta_0) = n(\theta - \theta_0)^T F(\theta - \theta_0).$$

One consequence is that $\|\theta - \theta_0\| \leq n^{-1/2}\sqrt{\|F^{-1}\|}\|u\|$.

Define $\Delta_n = \sqrt{n}J^{-1}(\bar{X} - \mu)$, so that $\bar{X} = \mu + n^{-1/2}J\Delta_n$. Note $\mathrm{E}(\Delta_n) = 0$ and that

$$\mathrm{E}(\Delta_n\Delta_n^T) = nJ^{-1}\mathrm{E}\left[(\bar{X} - \mu)(\bar{X} - \mu)^T\right]H = nJ^{-1}\frac{F}{n}H = I_p. \tag{2.1}$$

In particular, this gives

$$\mathrm{E}\|\Delta_n\|^2 = \mathrm{E}(\mathrm{tr}(\Delta_n\Delta_n^T)) = \mathrm{tr}[\mathrm{E}(\Delta_n\Delta_n^T)] = \mathrm{tr}(I_p) = p. \tag{2.2}$$

The interplay between the parameterizations in terms of u and θ will be important for obtaining bounds.

To state the quantities on which we will impose conditions, let $X \sim f(\cdot|\theta)$ and $V = J^{-1}(X-\mu)$. Let, $V = \big(V_1, \ldots, V_p\big)'$ and for fixed $\delta > 0$ we can define

$$B_n = \sup\{\mathrm{E}_{\theta_0} \mid a^T V \mid^3 \colon \|a\| = 1\}, \tag{2.3}$$

$$B_n' = \sup\{\mathrm{E}_{\theta} \mid a^T V \mid^3 \colon \|a\| = 1, \|u\| \le p^{(1+m)/2+\delta}\}, \tag{2.4}$$

$$B_n^* = \sup\{\mathrm{E}_{\theta} \mid a^T V \mid^4 \colon \|a\| = 1, \|u\| \le p^{(1+m)/2+\delta}\}, \tag{2.5}$$

$$M_r = \max_{1\le j\le p} \mathrm{E}_{\theta_0} \mid V_j \mid^r, \tag{2.6}$$

where $m \ge 0$ is a constant related to the growth of M_r (see condition MCV4). Note that the local restriction on the parameter space appears in (2.4) and (2.5) so the expectation is indexed by the θ (which depends on u), not θ_0.

Now, the main hypotheses can be stated in three classes. For any fixed $M > 0$, assume the following two conditions hold uniformly for all θ_0 with $\|\theta_0\| \le M$ (i.e., the implicit constants do not depend on θ_0 as long as $\|\theta_0\| \le M$).

First we require moment controls on V to control (2.3), (2.4), (2.5), and (2.6) for $r \ge 1$.

Conditions MCV: [Moment Controls on V]

For some $\delta > 0$, $m \ge 0$,

(MCV1) $B_n p^{3(1+m)/2+3\delta}/\sqrt{n} \to 0$,
(MCV2) $B_n' p^{(1+m)/2+\delta}/\sqrt{n} \to 0$,
(MCV3) $B_n^* p^{2(1+m)+4\delta}/n \to 0$,
(MCV4) $M_{2r} = \mathcal{O}(p^{mr})$ for every integer $r \ge 1$.

It will be seen in Section 4 that $m = 0$ will suffice for the normal and Dirichlet examples whereas $m = 1$ seems to be needed for the multinomial example; the role of m partly explains the difference in the rates ranging from $n^{1/3}$ to $n^{1/9}$.

Second, we must impose conditions on the prior density Π for θ.

Conditions PDB: [Prior Density Bounds]

The prior density Π satisfies

(PDB1) $-\log \Pi(\theta_0) = \mathcal{O}(p \log p)$,
(PDB2) $\Big|\log \dfrac{\Pi(\theta)}{\Pi(\theta_0)}\Big| \le K_n \|\theta - \theta_0\|$ for all $\|\theta - \theta_0\| \le \sqrt{\|F^{-1}\|}\, p^{(1+m)/2+\delta}/\sqrt{n}$, where K_n is some constant, subject to some growth condition (see Condition (BF2) below).

Note that Conditions (PDB1) and (PDB2) ensure that $\Pi(\theta)$ remains bounded below by $e^{-cp\log p}$ for some $c > 0$, for all θ sufficiently close to θ_0. It is not hard to see that there are a large class of priors for which (PDB1) and (PDB2) are satisfied. Indeed, suppose Π is an independence prior given by a product $h_j(\theta_j)$ where the $\log h_j$'s satisfy uniform positivity condition at θ_0, i.e., $h_j(\theta_{j,0}) > \epsilon$ and

a uniform Lipschitz condition on a neighborhood of θ_0, i.e., $|h_j(\theta_{j,0}) - h_j(\theta_j)| \leq L|\theta_{j,0} - \theta_j|$, then

$$-\log \Pi(\theta) = -\sum_{j=1}^{p} \log h_j(\theta_{j,0}) \leq -\sum_{j=1}^{p} \log \epsilon = \mathcal{O}(p) \tag{2.7}$$

ensuring (PDB1) and

$$|\log \Pi(\theta) - \log \Pi(\theta_0)| \leq L \sum_{j=1}^{p} |\theta_j - \theta_{j,0}| \leq C\sqrt{p}\|\theta - \theta_0\| \tag{2.8}$$

ensuring (PDB2) for $K_n = \mathcal{O}(\sqrt{p})$ provided $\sqrt{\|F^{-1}\|}\, p^{(1+m)/2+\delta}/\sqrt{n} \to 0$.

Our third set of conditions control the growth of the norm of Fisher information or its inverse, and also involve the Lipschitz constant K_n of $\log \Pi(\theta)$ defined in (PDB2) and moment bounds B_n and B_n^*.

Conditions BF: [Bounds using F]

For some $\alpha \geq 0$, $\delta > 0$, at θ_0 we have

(BF0) $\|F\| = \mathcal{O}(p^\alpha)$ and $\|F\|/n \to 0$,
(BF1) $\log \det(F) = \mathcal{O}(p^\alpha)$,
(BF2) $K_n \sqrt{\|F^{-1}\|} p^{(1+m)/2+\delta}/\sqrt{n} \to 0$.

We further assume that $\log n = \mathcal{O}(\log p)$. If this fails, the setting is very similar to fixed dimension, and results will go through by slight variation of the arguments; see Ghosal (2000), pages 52–53, for more explanation.

We comment that (BF1) is essentially always satisfied. Indeed, if F is written as the product of its eigenvalues, λ_j for $j = 1, \ldots, p$, the geometric mean-arithmetic mean inequality gives

$$(\det F)^{1/p} = \left(\prod_{j=1}^{p} \lambda_j\right)^{1/p} \leq \frac{\sum_{j=1}^{p} \lambda_j}{p} = \frac{\mathrm{tr}(F)}{p}.$$

So, taking logarithms and rearranging terms gives

$$\log \det F \leq p \log \frac{\mathrm{tr}(F)}{p} = \mathcal{O}(p \log p), \tag{2.9}$$

provided the diagonal entries of F are uniformly bounded by a polynomial in p. The same condition clearly implies $\|F\| \leq \mathrm{tr}(F) = \mathcal{O}(p^\alpha)$. This occurs in the normal, multinomial, and Dirichlet examples in Section 4. In addition, for priors satisfying uniform positivity and Lipschitz conditions as above (so that $K_n = \mathcal{O}(\sqrt{p})$), (BF2) is always satisfied for some rate.

Given Π, the posterior density of θ in an exponential family assumes the convenient form

$$\Pi_n(\theta) \propto \Pi(\theta) \prod_{i=1}^{n} f(X_i; \theta) = \Pi(\theta) \exp[n(\bar{X}^T\theta - \psi(\theta))].$$

The mixture of densities over the whole parameter space, i.e., the marginal density of X^n, will be denoted $m_n(\cdot)$. That is,

$$\begin{aligned} m_n(X^n) &= \int p(X^n|\theta)\Pi(\theta)d\theta \\ &= n^{-p/2}(\det(F))^{-1/2}\int p(X^n|\theta_0 + n^{-1/2}Hu)\Pi(\theta_0 + n^{-1/2}Hu)du, \end{aligned}$$

where the second expression follows from a change of variables.

By contrast, for examining local behavior, we define the local likelihood ratio process, that is, the likelihood ratio in terms of the local parameter u, by

$$\begin{aligned} Z_n(u) &= \prod_{i=1}^{n} \frac{f(X_i|\theta_0 + n^{-1/2}Hu)}{f(x_i;\theta_0)} \\ &= \frac{\exp[n(\bar{X}^T\theta_0 + \bar{X}^T n^{-1/2}Hu) - n\psi(\theta_0 + n^{-1/2}Hu)]}{\exp[n\bar{X}^T\theta_0 - n\psi(\theta_0)]} \\ &= \exp[\sqrt{n}\bar{X}^T Hu - n\{\psi(\theta_0 + n^{-1/2}Hu) - \psi(\theta_0)\}] \\ &= \exp[\sqrt{n}u^T J^{-1}\bar{X} - n\{\psi(\theta_0 + n^{-1/2}Hu) - \psi(\theta_0)\}]. \end{aligned} \tag{2.10}$$

Consequently, the posterior density whose asymptotics we want to find, is given in terms of u by

$$\Pi_n^*(u) = \frac{\Pi(\theta_0 + n^{-1/2}Hu)Z_n(u)}{\int \Pi(\theta_0 + n^{-1/2}Hw)Z_n(w)\,dw}.$$

When no confusion will result, we drop the subscript n, writing only Π^* for the posterior and we use Π to denote both the prior probability and its density.

3. Statements of results

In this section, we state our three main results for the increasing p setting. First, we give an asymptotic expression for the relative entropy between the n-fold product of densities and their mixture distribution. From this we derive a reference prior in the absence of nuisance parameters. Then, equipped with these results we identify reference priors in the presence of nuisance parameters.

3.1. No nuisance parameters

In the absence of nuisance parameters, we can derive reference priors from an asymptotic expression for the relative entropy. Our result is the following.

Theorem 3.1. *Under Conditions (PDB1), (PDB2), (MCV1)–(MCV4), and (BF1)–(BF2), the relative entropy between p_{θ_0} and the mixture of density m_n*

is given by

$$D(p^n_{\theta_0}\|m_n) = \mathrm{E}_{\theta_0}\left[\log\frac{p(X^n|\theta_0)}{m_n(X^n)}\right]$$
$$= \frac{p}{2}\log\frac{n}{2\pi} - \log\frac{\Pi(\theta_0)}{(\det(F))^{1/2}} - \frac{p}{2} + o(1),$$

as $n, p \to \infty$*, uniformly for all* $\theta_0 \in \Theta$ *such that* $\|\theta_0\| \le M$*.*

A sketch of the proof is given below; some details are relegated to Section 7.

To use Theorem 3.1, let Π be a prior density satisfying Conditions (PDB1) and (PDB2) uniformly in θ_0 and concentrated on $\{\|\theta_0\| \le M\} \cap \Theta$. The SMI is given by

$$I(X^n|\Pi) = \int D(p_{\theta_0}\|m_n)\Pi(\theta_0)d\theta_0 = \int D(\Pi_n\|\Pi)m(X^n)d\nu_n,$$

the expected Kullback-Leibler divergence between the posterior and the prior. By the uniformity in Theorem 3.1, we obtain the following result.

Theorem 3.2. *Assume that the conditions of Theorem 3.1 hold uniformly for all* $\theta_0 \in \Theta$ *with* $\|\theta_0\| \le M$ *and that the support of* Π *is in* $\{\theta_0|\ \|\theta_0\| \le M\}$*. Then we have*

$$I(X^n|\Pi) = \frac{p}{2}\log\frac{n}{2\pi e} + \int \Pi(\theta_0)\log\frac{\sqrt{\det(F(\theta_0))}}{\Pi(\theta_0)}d\theta_0 + o(1). \tag{3.1}$$

Consequently, on $\{\theta_0|\ \|\theta_0\| \le M\}$*, Jeffreys' prior,*

$$\Pi(\theta_0) \propto \sqrt{\det(F(\theta_0))},$$

asymptotically maximizes the SMI.

The proof of Theorem 3.1 rests on the posterior normality established in Appendix A as well as bounding the density ratio $\frac{m_n(X^n)}{p(X^n|\theta_0)}$ formally shown in Appendix B. We begin our sketch of the proof of Theorem 3.1 by stating the bounds for the density ratio. Taken together, these permit general upper and lower bounds on the relative entropy between P^n_θ and m_n. Note that two auxiliary bounds λ_n and λ^*_n appear in this result. They are formally defined in Appendix A, Lemma 6.7 and in Corollary 6.2, respectively. Here, it is enough to observe $p\lambda_n \to 0$ and $\lambda^*_n \to 0$ as $n \to \infty$ in P_{θ_0}-probability, see Lemma 6.8.

Lemma 3.3. *Upper bound on* $\frac{m_n(X^n)}{p(X^n|\theta_0)}$*: Assume Conditions (MCV1), (MCV3), (BF2) and (PDB2). Then on* W*, we have the bound*

$$\frac{m_n(X^n)}{p(X^n|\theta_0)} \le (2\pi)^{p/2}(1-\lambda^*_n)^{-1}\exp[K_n\sqrt{\|F^{-1}\|}\frac{p^{(1+m)/2+\delta}}{\sqrt{n}}]\Pi(\theta_0)n^{-p/2}$$
$$\times \exp[\frac{\|\Delta_n\|^2}{2(1-2\lambda_n)}](\det(F))^{-1/2}(1-2\lambda_n)^{-p/2}.$$

Lower bound on $\frac{m_n(X^n)}{p(X^n|\theta_0)}$*: Assume Conditions (MCV1) and (MCV3). Then on* W*, we have the bound*

$$\frac{m_n(X^n)}{p(X^n|\theta_0)} \geq \exp[-K_n\sqrt{\|F^{-1}\|}\frac{p^{(1+m)/2+\delta}}{\sqrt{n}}]\Pi(\theta_0)n^{-p/2}(\det(F))^{-1/2}(2\pi)^{p/2}$$
$$\times \exp\,[\frac{\|\Delta_n\|^2}{2(1+2\lambda_n)}](1+2\lambda_n)^{-p/2}(1-e^{-c_2p^{1+m+2\delta}}).$$

Now, we provide a sketch of the proof of Theorem 3.1.

Sketch of proof of Theorem 3.1:

The strategy of the proof is to define an error expression

$$R_n = \log\frac{p(X^n|\theta_0)}{m_n(X^n)} - \frac{p}{2}\log\frac{n}{2\pi} + \log\frac{\Pi(\theta_0)}{(\det(F))^{1/2}} + \frac{\|\Delta_n\|^2}{2}$$

and show it goes to zero in L_1. Then, the result will follow from the fact that $\mathrm{E}(\|\Delta_n\|^2) = p$.

To bound $|R_n|$ on W, note that Lemma 3.3 may be written as

$$\begin{aligned}
\log\frac{p(X^n|\theta_0)}{m_n(X^n)} &\geq -K_n\sqrt{\|F^{-1}\|}\frac{p^{(1+m)/2+\delta}}{\sqrt{n}} - \log(\Pi(\theta_0)) + \frac{p}{2}\log\frac{n}{2\pi} \\
&\quad +\frac{p}{2}\log(1-2\lambda_n) + \frac{1}{2}\log\det(F) - \frac{\|\Delta_n\|^2}{2(1-2\lambda_n)} + \log(1-\lambda_n^*); \\
\log\frac{p(X^n|\theta_0)}{m_n(X^n)} &\leq K_n\sqrt{\|F^{-1}\|}\frac{p^{(1+m)/2+\delta}}{\sqrt{n}} - \log\Pi(\theta_0) + \frac{p}{2}\log\frac{n}{2\pi} + \frac{1}{2}\log\det(F) \\
&\quad +\frac{p}{2}\log(1-2\lambda_n) - \log(1-e^{-cp^{1+m+2\delta}}) - \frac{\|\Delta_n\|^2}{2(1+2\lambda_n)}.
\end{aligned}$$

Now, under the Conditions (MCV1)–(MCV4) and (BF1) and (BF2), restricting to the set W gives the bound

$$\begin{aligned}
|R_n| &\lesssim K_n\sqrt{\|F^{-1}\|}\frac{p^{(1+m)/2+\delta}}{\sqrt{n}} + |\log(1-\lambda_n^*)| + \frac{p}{2}|\log(1-2\lambda_n)| \\
&\quad +\frac{p^{1+m+2\delta}}{16}|(1-2\lambda_n)^{-1}-1| + e^{-cp^{1+m+2\delta}} \\
&\lesssim K_n\sqrt{\|F^{-1}\|}\frac{p^{(1+m)/2+\delta}}{\sqrt{n}} + p\lambda_n + \lambda_n^* + p^{1+m+2\delta}\lambda_n + e^{-cp^{1+m+2\delta}} \\
&\lesssim K_n\sqrt{\|F^{-1}\|}\frac{p^{(1+m)/2+\delta}}{\sqrt{n}} + \frac{p^{3(1+m)/2+3\delta}}{\sqrt{n}}B_n + \frac{p^{2+2m+4\delta}}{n}B_n^*, \quad (3.2)
\end{aligned}$$

which goes to 0 as $n \to \infty$. Thus it remains to show that

$$\mathrm{E}_{\theta_0}|R_n| \to 0, \quad (3.3)$$

and that the convergence is uniform over $\theta_0 \in \Theta$ satisfying $\|\theta_0\| \leq M$.

Given (3.2), to show (3.3), it suffices to show

$$\mathrm{E}_{\theta_0}\left\{\left(\log_+ \frac{p(X^n|\theta_0)}{m_n(X^n)}\right)\mathbb{1}_{W^c}\right\} \to 0; \tag{3.4}$$

$$\mathrm{E}_{\theta_0}\left\{\left(\log_- \frac{p(X^n|\theta_0)}{m_n(X^n)}\right)\mathbb{1}_{W^c}\right\} \to 0; \tag{3.5}$$

$$\left\{\frac{p}{2}\log\frac{n}{2\pi} + \log\frac{\Pi(\theta_0)}{\sqrt{\det(F)}}\right\} P_{\theta_0}(W^c) \to 0; \tag{3.6}$$

$$\mathrm{E}_{\theta_0}\left(\|\Delta_n\|^2 \cdot \mathbb{1}\{\|\Delta_n\| > p^{(1+m)/2+\delta}/4\}\right) \to 0, \tag{3.7}$$

where $\mathbb{1}_A$ is the indicator function for the set A. These four convergences are verified in Appendix B. □

3.2. Nuisance parameters present

While it is often reasonable to use a reference prior for reference purposes, or even directly as an objective prior, it is also common for nuisance parameters to appear. This is particularly common in high dimensional parameters. Thus, reference priors have been extensively studied in nuisance parameter contexts. It will be seen below that our results extend readily to the setting of Berger and Bernardo (1989), Ghosh and Mukerjee (1992), and Clarke and Yuan (2004). For instance, in the case of a nuisance parameter ψ, the conditional mutual information given a nuisance parameter ξ is

$$I_\xi(X^n|\Pi) = \int \Pi(\theta|\xi) D(P^n_{\theta,\xi}\|P^n_\xi)d\theta \tag{3.8}$$

and integrating over ξ gives the conditional SMI, where $P^n_\xi = \int P^n_{\theta,\xi}\Pi(\theta|\xi)d\theta$. If ψ is fixed dimensional and varies over a compact set, it is enough to verify that the expansion for the information inside the integral is uniform in ξ. If this is done, then we obtain an analog to the prior proposed in Berger and Bernardo (1989) and Ghosh and Mukerjee (1992).

The typical situation is that the limiting form in Theorem 3.2 is an improper density. When Jeffreys' prior is not proper, it is routinely truncated. In such cases conditions like "maximizing mutual information" and "permissibility", or their increasing-dimensional analogs, must be imposed, see Berger et al. (2009). A separate problem is that many inference settings have nuisance parameters. That is, prior selection for the parameter of interest must be done conditionally on the value of some other parameter, say ψ. However, the inferential goal is not to estimate ψ, only θ. We are unconcerned about the value of ψ except that it may affect our inferences on θ. The classic example of this is estimating a normal mean without being concerned about the variance. When the variance is unknown, the intervals from a t_{n-1}-distribution for estimating θ are wider than those from a $N(\theta, \sigma_0^2)$ with σ_0 known.

To state the nuisance parameter setting formally, augment ξ to θ, where $p = p_n = \dim(\theta)$ and $q = q_n = \dim(\xi)$. That is, in principle, there may be countably many nuisance parameters in the limit of large n. Assume that the data remain IID and there are true values θ_0 and ξ_0 uniformly in the interior of $p+q$ dimensional parameter spaces. That is, there is a fixed ϵ_0 so that for any stage $p+q$, $\{(\theta,\xi) : \|(\theta,\xi)-(\theta_0,\xi_0)\| < \epsilon_0\} \subset \Theta\times\Xi$ where Θ is the p-dimensional parameter space for θ and Ξ is the q-dimensional nuisance parameter space for ξ at the n-th stage. Now, the natural exponential family can be written as

$$f(x|\theta,\xi) = \exp[x^T\eta(\theta\ ,\ \xi) - \psi(\theta,\xi)],$$

where the natural parameter is $\eta = (\eta_1(\theta_1,\xi),\ldots,\eta_p(\theta_p,\xi))$. That is, each component of η consists of one of the θ_j's and possibly all the ξ_j's.

The notation from Section 2 carries over directly. We use $P_{\theta_0,\xi}$ to denote the true distribution of the data, where dependence on n, p, and q is suppressed in the notation. Now, the true mean is p-dimensional and given by $\mu = \mathrm{E}_{\theta_0,\xi}(X) = \psi'(\theta_0,\xi) \in \mathbb{R}^p$ and the $p\times p$ Fisher information matrix for the parameter θ at θ_0 is given by $F_{1,1} = \psi''(\theta_0,\xi)$, where the differentiation is with respect to θ only. The maximum likelihood estimator (MLE) $\hat\theta$ satisfies $\psi'(\hat\theta,\xi) = \bar X = n^{-1}\sum_{i=1}^n X_i$ if ξ is known.

To state our next result, suppose F has been partitioned and write $F_{1,1}$ to mean the upper right $p\times p$ block of F. Suppose ξ is known and the dependence of ψ on ξ is smooth, i.e., has continuous first and second derivatives neither of which are zero on a neighborhood $N(\zeta) = N_\xi(\zeta)$ of radius ζ centered at ξ. Also, assume that the three classes of hypotheses (MCV), (PDB), and (BF) conditional on ξ hold. Then we get a conditional version of Theorem 3.2.

Theorem 3.4. *Suppose that ξ is smooth and the uniform versions of the hypotheses of Theorem 3.2 are satisfied. Then, for each fixed $\xi \in N(\zeta)$, we have*

$$I_\xi(X^n|\Pi) = \frac{p}{2}\log\frac{n}{2\pi e} + \int \Pi(\theta|\xi)\log\frac{\sqrt{\det(F_{1,1}(\theta,\xi))}}{\Pi(\theta|\xi)}d\theta + o(1),$$

and the error term is uniformly small for $\xi \in N(\zeta)$ as n increases.

Proof. This follows from verifying that the uniformized versions of the Lemmas and Theorems continue to hold under the uniformized hypotheses. □

Since the error term in Theorem 3.4 is uniformly small, it is natural to extract a corollary by integrating. We have the following.

Corollary 3.5. *Assume the hypotheses of Theorem 3.4 and that ξ has been assigned a prior $\Pi(\xi)$ on $N(\zeta)$. Then, the conditional SMI satisfies*

$$I(X^n|\Pi) = \frac{p}{2}\log\frac{n}{2\pi e} + \int \Pi(\theta|\xi)\Pi(\xi)\log\frac{\sqrt{\det(F_{1,1}(\theta,\xi))}}{\Pi(\theta|\xi)}d\theta d\xi + o(1). \quad (3.9)$$

Thus, the prior

$$\Pi(\theta|\xi) = \frac{\sqrt{\det(F_{1,1}(\theta,\xi))}}{\int\sqrt{\det(F_{1,1}(\theta,\xi))}d\theta} \quad (3.10)$$

asymptotically maximizes the conditional SMI but the marginal prior for ξ is unconstrained.

4. Examples

In this section we examine three cases in which p is increasing. We can verify that our hypotheses are satisfied and therefore both the asymptotic normality theorem and its consequences for reference priors hold.

4.1. Independent normal model

Consider n IID samples from a p dimensional normal model with mean $\theta = (\theta_1, \ldots, \theta_p)$ and covariance matrix identity, that is, the components of these variables are also independent. Assume also that a nonsingular prior density Π for θ. First we verify Conditions (PDB), (MCV) and (BF) for this case. Then, we see that if $p = \mathcal{O}(n^{1/3-\eta})$ for some $\eta > 0$, the conclusion of Theorem 3.2 holds, and hence Jeffreys' prior, which is the uniform prior on every compact rectangle, is the reference prior.

To begin, observe that $B_n = \mathcal{O}(1)$, $B'_n = \mathcal{O}(1)$ and $B^*_n = \mathcal{O}(1)$ because they are constants. Also, it is easy to see that $\|F\| = \mathcal{O}(1)$, $\|F^{-1}\| = \mathcal{O}(1)$ and $\mathrm{tr}(F) = p$. Note that (PDB1) and (PDB2) are satisfied for any well-behaved product-type prior, in particular the uniform, which is Jeffreys' prior in this case.

Observe that $M_{2r}^{1/r} = \mathcal{O}(1)$, so $m = 0$. Now, it is easy to see that among (MCV1)–(MCV3), the most stringent condition comes from (MCV1) which asks for $p^{3(1+0)/2+3\delta}/\sqrt{n} \to 0$ for some $\delta > 0$. Thus the requirement holds if $p = \mathcal{O}(n^{1/3-\eta})$ for some $\eta > 0$.

It is easy to see that (BF1) (for $\alpha = 0$) is satisfied by the identity covariance matrix. Condition (BF2) is now equivalent to requiring $\sqrt{p}\sqrt{p}\, p^{(1+0)/2+\delta}/\sqrt{n} \to 0$, which is satisfied when $p = \mathcal{O}(n^{1/3-\eta})$. Thus, Theorems 6.1, 3.1 and 3.2 hold whenever $p = \mathcal{O}(n^{1/3-\eta})$ for some $\eta > 0$.

In this example, it is possible to derive a similar expansion under much weaker growth restriction $p/n \to 0$ by direct computation, provided that we use an independence prior for the components of θ. In this case, the posterior is again of product form, so the expected Kullback-Leibler divergence is the sum of Kullback-Leibler divergence for each component. Individually, the Kullback-Leibler divergence between the posterior and the corresponding normal approximation decays like the square of the Hellinger distance, that is as n^{-1}. As there are p components and the Kullback-Leibler divergence is additive in the components, the overall Kullback-Leibler divergence of the product posterior density to the appropriate product normal density decays like $p/n \to 0$.

Indeed, to see why (3.1) in Theorem 3.2 holds, let $\theta = \theta_0 + u/\sqrt{n}$ and write

$$D(\Pi^*||\Pi) = D(\Pi^*||\phi) + \int n^{p/2}\Pi^*(\theta_0 + u/\sqrt{n}) \log \frac{\phi(u)}{n^{p/2}\Pi(\theta_0 + u/\sqrt{n})} du.$$

By asymptotic normality (see Appendix A), the first term is $o(1)$ a.s. for $p/n \to 0$, so it suffices to show that the expectation of the second term with respect to the marginal distribution gives the first two terms of (3.1). Note that

$$\log \frac{\phi(u)}{n^{p/2}\Pi(\theta_0 + u/\sqrt{n})} = \frac{p}{2}\log 2\pi - \frac{\|u\|^2}{2} - \frac{p}{2}\log n - \log \Pi(\theta_0 + \frac{u}{\sqrt{n}}).$$

Since $\mathrm{E}(u_i^2|X_{i,1},\ldots,X_{i,n}) \to 1$ in L_1, the claim follows by integration with respect to Π^* first and then by integrating with respect to the marginal of X. Thus the Jeffreys prior is asymptotically entropy maximizing among all product form priors only under the mild restriction that $p/n \to 0$. Note that, Jeffreys' prior, being the uniform distribution, can be regarded as a product of p constant functions of the components θ_j of θ for $j = 1,\ldots,p$.

Next, to illustrate Corollary 3.5, consider n IID samples from a p-dimensional normal with mean $\theta = (\theta_1,\ldots,\theta_p)$ and covariance matrix $\sigma^2 I_p$, the $p\times p$ diagonal matrix with nonzero entries σ^2 unknown. Treating σ as a nuisance parameter, and the $\theta_i = \mu/\sigma^2$, for $i = 1,\ldots,p$, as the parameters of interest, the three classes of conditions can be verified conditionally on a value of σ in much the same way as in the absence of nuisance parameters. Indeed, it will be apparent that the most stringent condition comes from (MCV1), so our main results will hold when $p = \mathcal{O}(n^{1/3-\eta})$ for some $\eta > 0$.

To see this, start by fixing a value of σ. Observe that the $(p+1)\times(p+1)$ Fisher information matrix as a function of (θ,σ^2) is $F = \mathrm{diag}(\sigma^{-2},\ldots,\sigma^{-2},2\sigma^{-2})$. So, the Fisher information matrix for θ is the $p\times p$ matrix $F_{1,1} = \mathrm{diag}(\sigma^{-2},\ldots,\sigma^{-2}) = \sigma^{-2}I_p$. Now, $J = \sigma^{-1}I_p$ and $V = J^{-1}(X-\theta_0) \sim N(0, I_p)$. Now, conditional on σ, $B_n = \mathcal{O}(1)$, $B_n' = \mathcal{O}(1)$ and $B_n^* = \mathcal{O}(1)$, the same rates as in the absence of a nuisance parameter, so (MCV1)–(MCV4), conditional on σ, are satisfied.

Note that for each fixed σ, $\|F_{1,1}\| = \mathcal{O}(1)$, $\|F_{1,1}^{-1}\| = \mathcal{O}(1)$ and $\mathrm{tr}(F_{1,1}) = \mathcal{O}(p)$. So, conditions (PDB1) and (PDB2) are unchanged apart from conditioning, i.e., by replacing $\Pi(\theta)$ with $\Pi(\theta|\sigma)$. So, $K_n = \mathcal{O}(\sqrt{p})$ for each fixed σ as before and $K_n = K_n(\sigma)$ is continuous as a function of σ. Thus, (PDB1) and (PDB2) are satisfied.

For the third set of conditions, it can be seen that (BF1) is satisfied for $\alpha = 1$ since $\det(F_{1,1}) = \sigma^{-2p}$ and (BF2) is satisfied as in the case when nuisance parameters are not present.

Now, conditional versions (on σ) of Theorems 6.1, 3.1 and 3.2 hold whenever $p = \mathcal{O}(n^{1/3-\eta})$ for some $\eta > 0$. In particular, Theorem 3.4 holds and since conditions (MCV), (PDB), and (BF) hold uniformly for compact sets for which $\sigma > 0$, Corollary 3.5 holds giving that $\sqrt{\det(F_{1,1})} \propto \sigma^{-p}$ is the conditional reference prior. It is seen that this is improper and independent of θ. Indeed, the analysis extends to the case that each of the p components are independent k-dimensional normal random variables all have the same variance matrix $\Sigma(\zeta)$ regarded as a nuisance parameter provided Σ varies over a compact set of non-singular matrices smoothly parametrized (with non-zero derivative) by $\zeta = (\zeta_1,\ldots,\zeta_q)$ for some fixed q and k is fixed as well. If q increases, it is not clear that Theorems 6.1, 3.1, 3.2, 3.4 and Corollary 3.5 hold.

4.2. Multinomial model

Consider a multinomial distribution with $(p+1)$ cells. The distribution is characterized by the probability vector $\pi = (\pi_1, \ldots, \pi_p)$ in which $P(\text{cell } j) = \pi_j$ for $j = 1, \ldots, p$, and the probability of the zeroth cell is $\pi_0 = 1 - \sum_{j=1}^p \pi_j$. The multinomial is an exponential family which has p natural parameters given by $\theta_j = \log(\pi_j/\pi_0)$. This transformation corresponds to $\pi_j = e^{\theta_j}/(1 + \sum_{j=1}^p e^{\theta_j})$ for $j = 1, \ldots, p$, and $\pi_0 = 1/(1+\sum_{j=1}^p e^{\theta_j})$. It will be seen next that it is enough to require that $p = \mathcal{O}(n^{1/9-\eta})$ for some $\eta > 0$ for Theorem 6.1 to hold and for Theorem 3.2 to show that Jeffreys' prior is the reference prior.

To proceed, we verify that the (BF) conditions are satisfied. Suppose for all $j = 1, \ldots, p$, $|\theta_j| \leq M$ for some bound $M > 0$. Essentially, this means that all π_j's are $\mathcal{O}(p^{-1})$. It can be verified that $F = D - \pi\pi^T$, where $D = \text{diag}(\pi_1, \ldots, \pi_p)$. Using standard arguments in matrix algebra and induction on p, it can be shown that $\det(F) = \prod_{j=0}^p \pi_j$. To transform this back into the natural parameters, recall the formula

$$F_\theta(\theta) = U^T F_\pi(\pi(\theta)) U$$

where F_θ is the Fisher information in the θ_j parameters, F_π is the Fisher information in the π_j parameters, and U is the matrix with (i, j) entries

$$u_{ij} = \frac{\partial \pi_i}{\partial \theta_j}(\theta) = \begin{cases} \pi_i \pi_j, & \text{if } i \neq j, \\ \pi_j(1 - \pi_j), & \text{if } i = j, \end{cases} \tag{4.1}$$

that is, $U = U^T = F_\pi(\pi(\theta))$. So, the log determinant of F_θ is

$$\begin{aligned} \log \det F_\theta(\theta) &= \log \left(\Pi_{j=0}^p \pi_j(\theta)\right)^3 \\ &= \sum_{j=1}^p 3\theta_j - 3(p+1)\log(1 + \sum_{j=1}^p e^{\theta_j}). \end{aligned} \tag{4.2}$$

So, for $|\theta_j| \leq M$, (BF1) is satisfied for $\alpha = 1$. This is slightly stronger than applying (2.9).

For (BF2), we can take $K_n = \mathcal{O}(p^{1/2})$ because (4.2) is a product form prior (in θ) with uniform positivity and as used in (2.7) and (2.8). Now, to get the rate from condition (BF2) we use part A of Lemma 8.1 to get that $F^{-1} = D^{-1} + (1 - \pi' D^{-1}\pi)^{-1} 11^T$. (Note that the denominator is $1 - \pi^T D^{-1}\pi = 1 - \sum_{j=1}^p \pi_j = \pi_0$.) In both the π parametrization and the natural θ parametrization, $\|F^{-1}\| \leq \text{tr}(F^{-1}) = \mathcal{O}(p^2)$, so $\sqrt{\|F^{-1}\|} = \mathcal{O}(p)$. This bound is actually sharp. For instance, when $\theta = 0$, i.e., $\pi_j = 1/(1+p)$ for all $j = 0, \ldots, p$, it can be verified that the largest eigenvalue of F^{-1} is $\mathcal{O}(p^2)$. It will be seen in the (MCV) conditions that, for the multinomial, we can set $m = 1$. Thus, (BF2) will be satisfied if $p^{5/2+\delta}/\sqrt{n} \to 0$.

Next, we examine the (MCV) conditions. We can now use part B of Lemma 8.1 to find $F^{-1/2}$ and verify the (MCV) conditions directly, or we can observe

by Section 3 of (Ghosal, 2000, 60-62), we have $B_n = \mathcal{O}(p^{3/2})$, $B'_n = \mathcal{O}(p^{3/2})$ and $B^*_n = \mathcal{O}(p^2)$. So, we have the following:

(i) $B_n p^{3+3\delta}/\sqrt{n} \to 0$ if and only if $p^{9/2+3\delta}/\sqrt{n} \to 0$, if and only if $p^{9+6\delta}/n \to 0$, so Condition (MCV1) is satisfied if $p_n = \mathcal{O}(n^{1/9-\eta})$ for some $\eta > 0$.

(ii) It can be seen that Condition (MCV2) is equivalent to $B'_n p^{1+\delta}/\sqrt{n} \to 0$ if and only if $p^{5/2+\delta}/\sqrt{n} \to 0$, which holds if and only if $p^{5+2\delta}/n \to 0$.

(iii) Condition (MCV3) is equivalent to $B^*_n p^{4+4\delta}/n \to 0$ if and only if $p^{6+4\delta}/n \to 0$.

(iv) In order to evaluate M_{2r}, note that $V_j = \mathcal{O}(\sqrt{p})$, so $M_{2r}^{1/r} = \mathcal{O}(p)$, that is, Condition (MCV4) holds with $m = 1$.

Thus all the (MCV) conditions are satisfied when $p = \mathcal{O}(n^{1/9-\eta})$ for some $\eta > 0$. Before we can conclude that this is the rate, we must verify (PDB).

Note that conditions (PDB) are written in terms of the natural parametrization, so we must transform from the π_j-parametrization to the θ_j-parametrization, as we did for the (BF) conditions. However, this time we are working with the priors rather than the Fisher information. Let us consider a prior of product form on π, say $\prod_{j=0}^p h_j(\pi_j)$. To find the Jacobian, recall (4.1). where $\pi_j = \pi_j(\theta)$. Since is seen that

$$\det((\frac{\partial \pi_i}{\partial \theta_j})) = \det[\mathrm{diag}(\pi_1, \ldots, \pi_p) - \pi\pi^T] = \pi_0 \cdots \pi_p,$$

we get that $\Pi(\theta) = \prod_{j=0}^p \pi_j h_j(\pi_j)$ where the π_j's are expressed as functions of θ. That is, $\Pi(\theta)$ remains of product form. Note that this does not mean independence since $\pi_0 = 1 - \sum_{j=1}^p \pi_j$.

Now, assuming $|\log h_j(\pi_j)| = \mathcal{O}(\log p)$, which is satisfied for the conjugate Dirichlet class of priors, we have that for θ_0

$$\begin{aligned} -\log \Pi(\theta_0) &= -\sum_{j=0}^p \log \pi_{j,0} - \sum_{j=0}^p \log h_j(\pi_{j,0}) \\ &= \mathcal{O}(p \log p) + C \sum_{j=0}^p |\log \pi_{j,0}| = \mathcal{O}(p \log p), \end{aligned}$$

verifying (PDB1).

For (PDB2), write

$$\left|\log \frac{\Pi(\theta)}{\Pi(\theta')}\right| \leq \sum_{j=0}^p |\log \pi_j - \log \pi'_j| + \sum_{j=0}^p |\log h_j(\pi_j) - \log h_j(\pi'_j)|. \tag{4.3}$$

Now, for θ in a compact set, all π_j are of order $\mathcal{O}(p^{-1})$. So, using the inequality $|\log x - \log y| \leq \max(x^{-1}, y^{-1})|x - y|$, we get that, for $j = 0, \ldots, p$,

$$|\log \pi_j - \log \pi'_j| \leq Cp|\pi_j - \pi'_j|,$$

since $\frac{1}{\pi_j}$ and $\frac{1}{\pi'_j}$ are of order $\mathcal{O}(p)$. Likewise, for all $j = 0, \ldots, p$, we get

$$|\log h_j(\pi_j) - \log h_j(\pi'_j)| \leq Cp|\pi_j - \pi'_j|$$

when the h_j's are Lipschitz with $\mathcal{O}(p)$ constant. Using these in (4.3), observe we bound the $j = 0$ term by the sum of the other terms:

$$|\theta_0 - \theta'_0| = \left| \left(1 - \sum_{j=1}^{p} \pi_j\right) - \left(1 - \sum_{j=1}^{p} \pi'_j\right) \right| \leq \sum_{j=1}^{p} |\pi_j - \pi'_j|.$$

Thus, (4.3) becomes

$$\begin{aligned}
\left|\log \frac{\Pi(\theta)}{\Pi(\theta')}\right| &\leq Cp \sum_{j=1}^{p} |\pi_j - \pi'_j| \\
&= Cp \sum_{j=1}^{p} \left| \frac{e^{\theta_j}}{1 + \sum_{j=1}^{p} e^{\theta_k}} - \frac{e^{\theta'_j}}{1 + \sum_{j=1}^{p} e^{\theta'_k}} \right| \\
&= \frac{Cp}{(1 + \sum_{j=1}^{p} e^{\theta_k})(1 + \sum_{j=1}^{p} e^{\theta_k})} \\
&\quad \times \sum_{j=1}^{p} \left| e^{\theta_j} \left(1 + \sum_{j=1}^{p} e^{\theta'_k}\right) - e^{\theta_j} \left(1 + \sum_{j=1}^{p} e^{\theta_j}\right) \right. \\
&\quad \left. + e^{\theta_j} \left(1 + \sum_{j=1}^{p} e^{\theta_j}\right) - \left(1 + \sum_{j=1}^{p} e^{\theta_j}\right) e^{\theta'_j} \right| \\
&\leq \frac{C}{p} \sum_{j=1}^{p} e^{\theta_j} \left| 1 - \sum_{k=1}^{p} e^{\theta'_k} - 1 + \sum_{k=1}^{p} e^{\theta_k} \right| \\
&\quad + \frac{C}{p} \sum_{j=1}^{p} \left(1 + \sum_{k=1}^{p} e^{\theta_k}\right) |e^{\theta_j} - e^{\theta'_j}|.
\end{aligned}$$

It is seen that e^{θ_j} is bounded on compact sets, $(1 + \sum_{k=1}^{p} e^{\theta_k}) = \mathcal{O}(p)$, and the sum over j in the first term gives p. Thus,

$$\begin{aligned}
\left|\log \frac{\Pi(\theta)}{\Pi(\theta')}\right| &\leq C \sum_{k=1}^{p} |e^{\theta_k} - e^{\theta'_k}| + C \sum_{j=1}^{p} |e^{\theta_j} - e^{\theta'_j}| \\
&\leq C \sum_{k=1}^{p} |\theta_k - \theta'_k| \leq C\sqrt{p}\|\theta - \theta'\|,
\end{aligned}$$

thereby verifying (PDB2).

Note that growth restrictions on p are more stringent for the multinomial than for the normal to identify Jeffreys' prior. This is because in the normal

case, components are independent, giving diagonal Fisher information matrix and moment bounds which do not grow with p. These features let us treat the components nearly separately, leading to a weaker growth restriction on p. This is consistent with the growth restrictions required for asymptotic normality studied in Ghosal (2000).

It is worth contrasting the priors obtained here via Theorem 3.2, Theorem 3.4 and Corollary 3.5 with other priors for the multinomial. Aside from Jeffreys' prior, the earliest seems to have been developed by Sono (1983) on the basis of transforming the π_j's so that the standardized highest posterior density (HPD) regions in the transformed parameters match those of the likelihood by itself. Essentially, this is a sort of invariance and frequentist matching approach. Sono's method gives Jeffrey's prior for $p = 1$ but not for $p \geq 2$. Sono (1983) noted that the resulting priors depend on the ordering of the parameters and that the Bayes test for a point hypothesis on π based on the standardized HPD regions is equivalent to the likelihood ratio test (independent of the ordering of the π_j's).

Berger and Bernardo (1992b) studied the same problem from the standpoint of reference priors developing ordered group reference priors. This technique is helpful when there is a natural way to partition a finite dimensional parameter of interest into several subvectors that can be ranked in order of inferential importance. They observe that product form priors, such as the Jeffreys prior in the multinomial case, allow inference about the groups of parameters to be decoupled in the sense that the product form of the prior leads to a sort of product form for the posterior. Moreover, the case of the Jeffreys prior for the multinomial is quite special in that it is proper and marginalizes, i.e., integrating out the last subvector of parameters in the prior leads to the reference prior for all but the last subvector of parameters. Berger and Bernardo (1992b) verifies that there are often numerous ways to partition a parameter vector into subvectors and that the results are typically not equivalent. Most recently, Bernardo (2010) derived that a Dirichlet$(p^{-1}, \ldots, p^{-1})$ is obtained when the p parameters are partitioned into p groups of one parameter each, i.e., the prior assignment is done one-at-a-time treating the remaining parameters at each stage as a nuisance. In the present context of increasing p, however, this "converges" to a Dirichlet$(0, \ldots, 0)$ which is improper and does not satisfy the requirement that all α_j^{-1} be bounded.

Beyond tractable cases like the multinomial, the ordered group reference prior method may not work as conveniently because it rests on assigning an objective prior at each conditioning step. In general, the prior they used comes out of their paper Berger and Bernardo (1989) and follows from an asymptotic expansion in Ghosh and Mukerjee (1992) that requires independence assumptions that are often not satisfied. Thus, while this prior may be a sensible choice in general and may be regarded as an approximation to the reference prior identified in Corollary 3.5, it is not in general a reference prior. Nevertheless, the method of building up objective priors by ordering the parameters and choosing priors in m stages does provide a way to find non-informative priors when there are many parameters.

From a more heuristic perspective, the Dirichlet$(\alpha_1, \ldots, \alpha_p)$ distribution is conjugate to the multinomial π and choosing $\alpha_j = 1$ for all j gives a uniform distribution on the $(p-1)$-dimensional simplex. This prior is objective in the sense of the Principle of Insufficient Reason. This differs from Jeffreys' prior, which is Dirichlet$(\frac{1}{2}, \ldots, \frac{1}{2})$ and proper but is not uniform. By contrast, setting all the α_j's to zero results in the limiting improper prior resulting from one-parameter-at-a-time prior assignment. The Dirichlet$(0, \ldots, 0)$, however, can be regarded as uniform on the $\log \pi_j$'s, see Chap. 3, Sec. 5 in Gelman et al. (2004). Heo and Kim (2007) examined the behavior of the posterior from a multinomial likelihood using the Dirichlet prior for a variety of choices of the α_j's.

4.3. Dirichlet model

As a third example, consider a Dirichlet distribution which we write in the form

$$\frac{\Gamma(\theta_1+1)\cdots\Gamma(\theta_p+1)}{\Gamma(\sum_{j=1}^p \theta_j + p)} y_1^{\theta_1} \cdots y_p^{\theta_p}$$

for $\theta_1, \ldots, \theta_p > -1$, $0 < y_j < 1$ and $\sum_{j=1}^p y_j = 1$. Putting this in the form of a natural exponential family enables us to recognize the sufficient statistic $X = (\log Y_1, \ldots, \log Y_p)$, the natural parameter $\theta = (\theta_1, \ldots, \theta_p)$ and

$$\psi(\theta) = \log \Gamma(\sum_{j=1}^p \theta_j + p) - \sum_{j=1}^p \log \Gamma(\theta_j + 1).$$

The second partial derivatives of ψ with respect to the θ_j's give the Fisher information. Let Ψ denote the digamma function, that is, the derivative of the logarithm of the gamma function and let Ψ' denote the derivative of the digamma function, called the trigamma function. So, we can write the Fisher information in terms of the trigamma function Ψ':

$$\frac{\partial^2}{\partial\theta_j \partial\theta_k}\psi(\theta) = \begin{cases} \Psi'(\sum_{j=1}^p \theta_j + p) - \Psi'(\theta_j + 1), & \text{if } j = k, \\ \Psi'(\sum_{j=1}^p \theta_j + p), & \text{if } j \neq k. \end{cases} \tag{4.4}$$

and hence

$$F = D - a11^T,$$

where 1 indicates the 1-vector $1 = (1, \ldots, 1)$, $D = \text{diag}(a_1, \ldots, a_p)$, $a_j = a_j(\theta) = -\Psi'(\theta_j + 1)$ and $a = -\Psi'(\sum_{j=1}^p \theta_j + p)$. On the real line, the trigamma function is given by $\Psi'(x) = \sum_{k=0}^\infty (x+k)^{-2}$. It is positive, decreasing in x, and has an asymptote at zero.

We need to find F^{-1} and J which we then use to find J^{-1}. Using part A of Lemma 8.1 we get

$$F^{-1} = D^{-1} - \frac{a}{1 + a1^T D^{-1} 1} D^{-1} 11^T D^{-1},$$

in which the factor simplifies to $a(1 - a\sum_{j=1}^p a_j^{-1})^{-1}$ and the (j,k)-th entry in $D^{-1}11^TD^{-1}$ is $a_j^{-1}a_k^{-1}$ for $j,k = 1,\ldots p$. Note that F^{-1} is well defined provided $a^{-1} \neq \sum_{j=1}^p a_j^{-1}$. It is seen that, for any ℓ, a_j is bounded above for $\theta_j > \ell > -1$ and a_j goes to zero as θ increases. Likewise, a is bounded above for $\sum_{j=1}^p \theta_j + p > \ell > -1$ and goes to zero as $\sum_{j=1}^p \theta_j$ increases. Since F is the variance-covariance matrix of a nonsingular distribution, F is positive definite for any set of θ's satisfying $\theta_j > -1$. Thus, $a^{-1} \neq \sum_{j=1}^p a_j^{-1}$. Further, because of the continuous dependence of the a_j's on the θ_j's, it follows that $b = 1 - a\sum_{j=1}^p a_j^{-1}$ remains boinded away from zero if $\theta_j > \ell > -1$ for all j, and any fixed ℓ. It follows that $\|F^{-1}\| \leq \mathrm{tr}(F^{-1}) = \mathcal{O}(p)$.

Now, we use part B of Lemma 8.1 to find J, a square root of F. Letting $u = \sqrt{a}1$, we find

$$J = D^{1/2} + vw^T,$$

where $v = -a(1 + \sqrt{1 - a\sum_{j=1}^p a_j^{-1}})^{-1}1$ and $w^T = (a_1^{-1/2}, \ldots, a_p^{-1/2})$.

So, using part A of Lemma 8.1 again, we find the inverse is

$$J^{-1} = D^{-1/2} - \frac{D^{-1/2}vw^TD^{-1/2}}{1 + w^TD^{-1/2}v} = D^{-1/2} + \frac{c}{1 - c\sum_{j=1}^p a_j^{-1}}M$$

where $c = a(1 + \sqrt{1 - a\sum_{j=1}^p a_j^{-1}})^{-1}$ and M is the matrix with (j,k)-th entry $a_j^{-1}a_k^{-1/2}$. By simple algebra $1 - c\sum_{j=1}^p a_j^{-1} = \sqrt{b}$, which is bounded away from zero and $c \leq a$ is bounded uniformly in θ as long as $\theta_j > \ell$ for any j, for any fixed $\ell > -1$. Note that the entries in M are bounded as well.

To find the rates required for the (MCV) conditions we must examine the moments of $V = J^{-1}(X - \mu)$. So, consider a p-dimensional unit vector $u^T = (u_1, \ldots, u_p)$ and observe that, from the form of V, we have

$$u^TV = \sum_{j=1}^p u_j \frac{(X_j - \mu_j)}{\sqrt{a_j}} + \frac{c}{\sqrt{b}}\sum_{j=1}^p \frac{(X_j - \mu_j)}{\sqrt{a_j}}\sum_{j=1}^p \frac{u_j}{a_j}. \tag{4.5}$$

Recall that $X_j = \log Y_j$ and that $Y_j = W_j/W$ in distribution, where the W_j's are independent Gamma$(\theta_j + 1, 1)$ random variables and $W = \sum_{j=1}^p W_j$ is a Gamma$(\sum_{j=1}^p \theta_j + p, 1)$ random variable. Consequently, $X_j = \log W_j - \log W$ and $\mu_j = \mathrm{E}(\log W_j) - \mathrm{E}(\log W)$. So, denoting $\widetilde{\log W}_j = \log W_j - \mathrm{E}(\log W_j) = \log W_j - \mu_j$ and $\widetilde{\log W} = \log W - \mathrm{E}(\log W) = \log W - \sum_{j=1}^p \mu_j$ respectively, we can rewrite u^TV in (4.5) as

$$\begin{aligned} u^TV &= \sum_{j=1}^p u_j \left(\frac{\widetilde{\log W}_j - \widetilde{\log W}}{\sqrt{a_j}}\right) \\ &\quad + \frac{c}{\sqrt{b}}\sum_{j=1}^p \left(\frac{\widetilde{\log W}_j - \widetilde{\log W}}{\sqrt{a_j}}\right)\sum_{j=1}^p \frac{u_j}{a_j} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{p} u_j \frac{\widetilde{\log W}_j}{\sqrt{a_j}} - \widetilde{\log W} \sum_{j=1}^{p} \frac{u_j}{\sqrt{a_j}} \\
&\quad + \frac{c}{\sqrt{b}} \left(\sum_{j=1}^{p} \frac{\widetilde{\log W}_j}{\sqrt{a_j}} \right) \left(\sum_{j=1}^{p} \frac{u_j}{a_j} \right) \\
&\quad - \frac{c}{\sqrt{b}} \left(\widetilde{\log W} \sum_{j=1}^{p} \frac{1}{\sqrt{a_j}} \right) \left(\sum_{j=1}^{p} \frac{u_j}{a_j} \right).
\end{aligned}
$$

It is the fourth moment of the last upper bound that we must control for (MCV3). So, using $(a+b+c+d)^2 \leq 4a^2 + 4b^2 + 4c^2 + 4d^2$ gives four terms

$$
\begin{aligned}
\mathrm{E}|u^T V|^4 \quad \leq \quad & 4\mathrm{E} \left(\sum_{j=1}^{p} u_j \frac{\widetilde{\log W}_j}{\sqrt{a_j}} \right)^4 && (4.6) \\
& +4 \left(\sum_{j=1}^{p} \frac{u_j}{\sqrt{a_j}} \right)^4 \mathrm{E}|\widetilde{\log W}|^4. && (4.7) \\
& +4 \left(\frac{c}{\sqrt{b}} \right)^4 \left(\sum_{j=1}^{p} \frac{u_j}{a_j} \right)^4 \mathrm{E} \left(\sum_{j=1}^{p} \frac{\widetilde{\log W}_j}{\sqrt{a_j}} \right)^4 && (4.8) \\
& +4 \left(\frac{c}{\sqrt{b}} \right)^4 \left(\sum_{j=1}^{p} \frac{u_j}{a_j} \right)^4 \left(\sum_{j=1}^{p} \frac{1}{\sqrt{a_j}} \right)^4 \mathrm{E}(\widetilde{\log W})^4. && (4.9)
\end{aligned}
$$

We bound (4.6) and (4.8) by the Marcinkiewicz-Zygmund inequality for centered random variables with finite $2r$-th moments:

$$
\mathrm{E} \left(\sum_{j=1}^{p} Z_j \right)^{2r} \leq C_r p^{r-1} \sum_{j=1}^{p} \mathrm{E}|Z_j|^{2r}.
$$

Thus, for $r = 2$ (4.6) is bounded by

$$
4C_2 p \sum_{j=1}^{p} \frac{u_j^4}{a_j^2} \mathrm{E}|\widetilde{\log W_j}|^4 \tag{4.10}
$$

and (4.8) is bounded by

$$
4 \left(\frac{c}{\sqrt{b}} \right)^4 \left(\sum_{j=1}^{p} \frac{u_j}{a_j} \right)^4 C_2 p \sum_{j=1}^{p} \frac{\mathrm{E}|\log \widetilde{W}_j|^4}{a_j^2}. \tag{4.11}
$$

Now, to work out rates for the (MCV) conditions, we start by finding the orders of (4.7), (4.9), (4.10), and (4.11).

Consider the expression in (4.10). As before, when the θ's are bounded, $a_j^{-2} < C'$ and there is a C so that $\mathrm{E}|\widetilde{\log W_j}|^4 < C$. Since $1 = \|u\|^4 \geq u_1^4 + \cdots + u_p^4$, (4.10) is bounded by

$$4C_2 p C' C \sum_{j=1}^{p} u_j^4 = \mathcal{O}(p).$$

Next, for (4.7), recall that $W \sim \mathrm{Gamma}(\sum_{j=1}^{p} \theta_j + p, 1)$. So $\mathrm{E}(\widetilde{\log W})^4 = \Psi'''(\sum_{j=1}^{p} \theta_j + p) \leq c$ since all polygamma functions are bounded above as long as the argument stays away from zero. This is due to the series representation

$$\Psi^{(m)}(t) = (-1)^{m+1} m! \sum_{k=0}^{\infty} \frac{1}{(t+k)^{-(m+1)}}.$$

Indeed,

$$|\Psi^{(m)}(t)| \leq c \int_t^{\infty} \frac{1}{x^{-(m+1)}} dx = \mathcal{O}(t^{-m})$$

as $t \to \infty$. In the present case, all $\theta_j > \ell > -1$, so $\sum_{j=1}^{p} \theta_j + p$ grows like p, i.e., $\mathrm{E}(\widetilde{\log W})^4 = \mathcal{O}(p^{-3})$. For bounded θ, there is a ℓ so that $a_j > \ell > 0$, i.e., $a_j^{-1/2}$ has a finite bound C'. So, (4.7) is bounded by

$$4C'^4 C \left(\sum_{j=1}^{p} u_j \right)^4 \mathcal{O}(p^{-3}) \leq 4C'^4 C \left[\sqrt{p} \sqrt{\sum_{j=1}^{p} u_j^2} \right]^4 \mathcal{O}(p^{-3}) = \mathcal{O}(p^{-1}).$$

Now consider the expression in (4.11). Bounding $\mathrm{E}|\widetilde{W}_j|^4$'s and a_j^{-2}'s by a constant, (4.11) is bounded from above by

$$Cp^2 \left(\frac{c}{\sqrt{b}} \right)^4 \left(\sum_{j=1}^{p} \frac{u_j}{a_j} \right)^4.$$

We now argue that $c^2(\sum_{j=1}^{p}(u_j/a_j))^2$ is bounded. Then, since b is bounded away from zero, it will follow that the above expression is $\mathcal{O}(p^2)$. To this end, observe that $c = a(1+\sqrt{b})^{-1}$ and note that

$$\begin{aligned} c^2 \left(\sum_{j=1}^{p} \frac{u_j}{a_j} \right)^2 &\leq \frac{a^2 \sum_{j=1}^{p} u_j^2 \sum_{j=1}^{p} a_j^{-2}}{1+\sqrt{b})^2} \leq a^2 \sum_{j=1}^{p} a_j^{-2} \\ &\leq a \sum_{j=1}^{p} a_j^{-1} = 1 - b \leq 1, \end{aligned}$$

since $a a_j^{-1} \leq 1$. Alternatively, one can use the fact that $c = \mathcal{O}(p^{-1})$.

Finally, we bound (4.9). As $\mathrm{E}(\widetilde{\log W})^4 = \mathcal{O}(p^{-3})$, letting C' be a bound on all a_j^{-1}'s and C'' be a bound on $a_j^{-1/2}$'s for bounded θ, (4.8) is bounded by

$$Cp^{-3}\left(\sum_{j=1}^{p}\frac{1}{\sqrt{a_j}}\right)^4 \frac{c^4}{b^2}\left(\sum_{j=1}^{p}\frac{u_j}{a_j}\right)^4 = \mathcal{O}(p),$$

since, as shown above, $c\sum_{j=1}^{p}(u_j/a_j) = \mathcal{O}(1)$ and $\sum_{j=1}^{p} a_j^{-1/2} = \mathcal{O}(p)$. Thus $\mathrm{E}|u^TV|^4 = \mathcal{O}(p^2)$ for any u with $\|u\| = 1$ for θ bounded away from zero and uniformly bounded above. That is, $B_n^* = \mathcal{O}(p^2)$ since the domain of θ's includes local neighborhoods of the sort used in the definition of B_n^*. To get rates for B_n and B_n', note that $\mathrm{E}|X|^3 \le (\mathrm{E}|X|^4)^{3/4}$. Thus, $\mathrm{E}|u^TV|^3 \le (\mathcal{O}(p^2))^{3/4} = \mathcal{O}(p^{3/2})$ and $B_n = B_n' = \mathcal{O}(p^{3/2})$.

Note that the preceding extends to give $M_{2r} = \max_j \mathrm{E}_{\theta_0}|V_j|^{2r} = \mathcal{O}(1)$. We can now identify rates because we can choose $m = 0$ to satisfy (MCV4). Write

$$\begin{aligned} V_j &= \frac{X_j-\mu_j}{\sqrt{a_j}} + \frac{c}{\sqrt{b}a_j}\sum_{k=1}^{p}\frac{X_k-\mu_k}{\sqrt{a_k}} \\ &= \frac{\widetilde{\log W_j} - \widetilde{\log W}}{\sqrt{a_j}} + \frac{c}{\sqrt{b}a_j}\sum_{k=1}^{p}\frac{\widetilde{\log W_k}}{\sqrt{a_k}} - \frac{c\,\widetilde{\log W}}{\sqrt{b}a_j}\sum_{k=1}^{p}\frac{1}{\sqrt{a_k}}. \end{aligned}$$

Then, for any $r \ge 1$,

$$\begin{aligned} |V_j|^{2r} \le\ & C\left[\frac{|\widetilde{\log W_j}|^{2r}}{a_j^r} + \frac{|\widetilde{\log W}|^{2r}}{a_j^r}\right. \\ & \left. + \frac{c^{2r}}{b^r a_j^{2r}}\left|\sum_{k=1}^{p}\frac{\widetilde{\log W_k}}{\sqrt{a_k}}\right|^{2r} + \frac{c^{2r}|\widetilde{\log W}|^{2r}}{b^r a_j^{2r}}\left(\sum_{k=1}^{p}\frac{1}{\sqrt{a_k}}\right)^{2r}\right]. \end{aligned}$$

Taking expectations gives

$$\begin{aligned} \mathrm{E}|V_j|^{2r} \le\ & C'\left[\frac{\mathrm{E}|\widetilde{\log W_j}|^{2r}}{a_j^r} + \frac{\mathrm{E}|\widetilde{\log W}|^{2r}}{a_j^r}\right. \\ & \left. + \frac{c^{2r}p^{r-1}}{b^r a_j^{2r}}\sum_{k=1}^{p}\frac{\mathrm{E}|\widetilde{\log W_k}|^{2r}}{a_k^r} + \frac{c^{2r}\mathrm{E}|\widetilde{\log W}|^{2r}}{b^r a_j^{2r}}\left(\sum_{k=1}^{p}\frac{1}{\sqrt{a_j}}\right)^{2r}\right]. \\ \le\ & C'\left[\mathcal{O}(1) + \mathcal{O}(p^{-(2r-1)}) + \mathcal{O}(c^{2r}p^r) + \mathcal{O}(c^{2r}p^{-(2r-1)}p^{2r})\right]. \end{aligned}$$

Now,

$$c = \frac{a}{1+\sqrt{b}} \le a = \Psi'(\sum_{j=1}^{p}\theta_j + p) = \mathcal{O}(p^{-1}),$$

uniformly for all θ_j's bounded and $\theta_j > \ell > -1$. This gives $\mathrm{E}|V_j|^{2r} = \mathcal{O}(1)$ for any $r \ge 1$ and $j = 1, \ldots, p$. So, we can choose $m = 0$ in (MCV4) as claimed.

Now, (MCV1)–(MCV3) give the following.

1. For (MCV1), p must satisfy $p^{3/2}p^{3(1+0)/2+2\delta}/\sqrt{n} \to \infty$. That is, $p = \mathcal{O}(n^{1/6-\eta})$ for some $\eta > 0$.
2. For (MCV2), p must satisfy $p^{3/2}p^{(1+0)/2+2\delta}/\sqrt{n} \to \infty$. That is, $p = \mathcal{O}(n^{1/4-\eta})$ for some $\eta > 0$.
3. For (MCV3), p must satisfy $p^2 p^{2(1+0)+4\delta}/n \to 0$. That is, $p = \mathcal{O}(n^{1/4-\eta})$ for some $\eta > 0$.

It remains to verify (BF1), (BF2), (PDB1) and (PDB2). Verification of (BF1) is easy: apply (2.9) to see that any $\alpha > 1$ will suffice. By contrast, Conditions (BF2), (PDB1) and (PDB2) involve properties of the prior and so are related to each other. We consider two cases: Π is a conjugate prior and Π is Jeffreys' prior.

Beginning with conjugate priors, recall that a regular exponential family such as the Dirichlet has natural form $\exp\left[\sum_{j=1}^p \eta_j x_j - \psi(\theta)\right]$. So, its conjugate family is is of the form $\Pi(\theta) \propto \exp\left[\sum_{j=1}^p \eta_j \alpha_j - \lambda\psi(\eta)\right]$, where $\lambda > 0$ and $\alpha_j/\lambda < 0$ on bounded sets of θ, see Chap. 4, p. 113, Brown (1986). In the present setting, θ ranges over $\{\theta : \text{ for all } j, -1 < \ell < \theta_j < M\}$, and the Dirichlet is regular in the natural parameter θ. Note that the conjugate prior is not of the product form, so we cannot use (2.7) or (2.8) and must proceed with a direct verification of (PDB1) and (PDB2).

For any conjugate prior, (PDB1) is

$$|\log \Pi(\theta)| = |\sum_{j=1}^p \theta_j \alpha_j - \lambda\psi(\theta)| \le \mathcal{O}(p) + |\lambda||\Psi'(\sum_{j=1}^p \theta_j + p)|.$$

The last term on the right is bounded when the θ_j's are in a compact set so (PDB1) is satisfied for conjugate priors with rate $\mathcal{O}(p)$.

For (PDB2), the difference of logarithms for conjugate priors is

$$\begin{aligned} |\log \Pi(\theta) - \log \Pi(\theta_0)| &\le \sum_{j=1}^p |\theta_j - \theta_j'|\alpha_j + \lambda|\Psi'(\sum_{j=1}^p \theta_j + p) - \Psi'(\sum_{j=1}^p \theta_j' + p)| \\ &\le C\sum_{j=1}^p |\theta_j - \theta_j'| + \lambda|\Psi''(\zeta^*)|\sum_{j=1}^p |\theta_j - \theta_j'| \\ &\le C\sqrt{p}\|\theta - \theta'\|. \end{aligned}$$

Thus, it is seen that the rate K_n in (BF2) is $K_n(p) = \mathcal{O}(\sqrt{p})$.

Now, verification of (BF2) is easy. Since $\sqrt{\|F^{-1}\|} = \mathcal{O}(\sqrt{p})$, the rate from (BF2) becomes $\sqrt{p}\sqrt{p}p^{(1+0)/2+\delta}/\sqrt{n} \to 0$, i.e., $p^{(3/2)+\delta}/\sqrt{n} \to 0$ giving $p = \mathcal{O}(n^{1/3-\eta})$, a weaker constraint that the (MCV) conditions did. So, the overall rate is $p = \mathcal{O}(p^{1/6-\eta})$.

Next, we turn to the verification of (BF2), (PDB1) and (PDB2) for Jeffreys' prior

$$\Pi(\theta) \propto \sqrt{\det[\text{diag}(a_1, \ldots, a_p)] - a11^T},$$

where $a_j = -\Psi'(\theta_j + 1)$ and $a = \Psi'(\sum_{j=1}^p \theta_j + p)$. Since Jeffreys' prior is based on the Fisher information, it is enough to observe that as long as the entries in F are polynomially bounded in p, $\log\det(F)$ will be $\mathcal{O}(p\log p)$ using (2.9) thereby satisfying (PDB1).

To begin the verification of (PDB2), write

$$|\log\sqrt{\det F(\theta)} - \log\sqrt{\det F(\theta')}| \quad = \quad \frac{1}{2}|\log\det[I_p + F(\theta)^{-1}(F(\theta') - F(\theta))]|.$$

We first show that the "error term"

$$F(\theta)^{-1}(F(\theta') - F(\theta)) \leq Cp\|\theta - \theta'\|I_p \tag{4.12}$$

in matrix ordering. To do this, we use the form of $F(\theta)$ and calculate directly. We have

$$\begin{aligned} F^{-1}(\theta)(F(\theta') - F(\theta)) \quad = \quad & \operatorname{diag}\left(\frac{a_1' - a_1}{a_1}, \ldots, \frac{a_p' - a_p}{a_p}\right) \\ & + \frac{a}{b}\begin{pmatrix} a_1^{-1} \\ \vdots \\ a_p^{-1} \end{pmatrix} (\frac{a_1' - a_1}{a_1}, \ldots, \frac{a_p' - a_p}{a_p}) \\ & -(a' - a)\begin{pmatrix} a_1^{-1} \\ \vdots \\ a_p^{-1} \end{pmatrix} 1 \\ & -\frac{a(a' - a)}{b}\left(\sum_{j=1}^p \frac{1}{a_j}\right)\begin{pmatrix} a_1^{-1} \\ \vdots \\ a_p^{-1} \end{pmatrix} 1, \end{aligned} \tag{4.13}$$

where 1 is the vector $(1, \ldots, 1)^T$ of p ones, $a' = \Psi'(\sum_{j=1}^p \theta_j' + p)$ and $a_j' = \Psi'(\theta_j' + 1)$. It is sufficient to show that all entries are bounded by a constant multiple of $\|\theta - \theta'\|$. Note that $\max_j |\theta_j - \theta'| \leq \|\theta - \theta'\|$.

For the first term in (4.13), note $|a_j - a_j'|$ is bounded by $|\theta_j - \theta_j'|$, so each entry is bounded by $\max_j |\theta_j - \theta_j'|$. The second term is similar because all the entries in the matrix are of the form $(a_j - a_j')$ times a bound which is finite when all the θ_j's are bounded and that $|a_j - a_j'| \leq C|\theta_j - \theta_j'|$ by the mean value theorem. Further, $a = \mathcal{O}(p^{-1})$, so the entries in the second term are bounded by $Cp^{-1}\|\theta - \theta'\|$. For the third term, note that

$$\begin{aligned} |a' - a| \quad &= \quad \left|\Psi'\left(\sum_{j=1}^p \theta_j + p\right) - \Psi'\left(\sum_{j=1}^p \theta_j' + p\right)\right| \\ &\leq \quad Cp^{-1}\sum_{j=1}^p |\theta_j - \theta_j'| \leq Cp^{-1/2}\|\theta - \theta'\| \end{aligned}$$

by the mean value theorem, since $|\Psi'(t)| = \mathcal{O}(t^{-1})$ as $t \to \infty$. The other terms are of the form a_j^{-1} and hence bounded above since the a_j's are bounded below

when the θ_j's are bounded. Finally, as observed above, $(a'-a)$ in the first factor in the fourth term in (8.2) is bounded by $Cp^{-1/2}\|\theta-\theta'\|$ and $a=\mathcal{O}(p^{-1})$. The second factor is $\mathcal{O}(p)$ since the a_j^{-1}'s are bounded by a constant when the θ_j's are bounded. The remaining two factors are entry-wise bounded as well, so all entries in this term are bounded by the overall bound $Cp^{-1/2}\|\theta-\theta'\|$. Collecting these bounds together, all entries of $F(\theta)^{-1}(F(\theta')-F(\theta))$ are bounded by a constant multiple of $\|\theta-\theta'\|$. An application of Lemma 8.2 now gives (4.12). This leads to

$$\begin{aligned}\log\det(I_p+F^{-1}(\theta)[F'(\theta)-F(\theta)]) &\leq \log\det(I_p+Cp\|\theta-\theta'\|I_p)\\ &= \log(1+Cp\|\theta-\theta'\|)^p\\ &\leq Cp^2\|\theta-\theta'\|. \end{aligned} \tag{4.14}$$

So, we get (PDB2) and can take $K_n=\mathcal{O}(p^2)$ in (BF2). Using this and $\|F^{-1}\|=\mathcal{O}(p)$, (BF2) becomes $p^2p^{1/2}p^{(1+0)/2+\delta}/\sqrt{n}\to 0$, i.e., $p=\mathcal{O}(n^{1/6-\eta})$, as in (MCV1).

5. Discussion

Here we have established an asymptotic expansion for the relative entropy $D(p_{\theta_0}^n\|m_n)$ between an n-fold product of an i.i.d. model and the mixture over such models with respect to the prior. The error term is $o(1)$ and the dimension p of θ is increasing with n. We observe that our expansion is uniform under appropriate assumptions. This leads to an expansion for $I(X^n|\Pi)$, the SMI between a parameter and a sample of size n for a general class of priors. The term involving the prior can be maximized so that the corresponding reference prior is seen to be Jeffreys' prior, even when p is increasing with n. We have verified that in three examples, the normal, the multinomial, and the Dirichlet, that our hypotheses are satisfied when $p=\mathcal{O}(n^{1/3-\eta})$, $p=\mathcal{O}(n^{1/9-\eta})$, and $p=\mathcal{O}(p^{1/6-\eta})$, respectively, for some $\eta>0$. An analogous treatment can be given when the model depends on a nuisance parameter. In particular, one can integrate the asymptotic expression for the SMI given a specific value of the nuisance parameter over a range of nuisance parameters to obtain the conditional SMI which can be optimized as well to give a conditional version of Jeffreys' prior, although the prior on the nuisance parameter is indeterminate. We comment that the treatment given for the more general setting of Clarke and Yuan (2004) is also expected to generalize. Moreover, other measures of distance may be amenable to the same sort of treatment, parallel to Ghosh et al. (2010).

Our main results, like other reference prior derivations, rest on an asymptotic normality result in Appendix A. This key feature of this result, in contrast to other asymptotic normality results, is that the error of approximation admits an explicit bound in the increasing parameter case. The approximation is in L_1-distance and the set on which the bound fails has probability decaying polynomially in p.

From Bernardo (1979), Clarke and Yuan (2004), the present results, and numerous other authors, it can be seen that, typically, when SMI can be optimized

in its conditional or unconditional form, the result is a prior based on the normalized square roots of asymptotic variances. The reference prior obtained in Section 5 is also based on Jeffreys' prior, but conditionally on the nuisance parameter (whose distribution is unconstrained). In Clarke and Yuan (2004), all the priors obtained are based on square roots of asymptotic variances, typically of an asymptotically normal statistic, or on ratios of asymptotic variances from asymptotically normal statistics. This general form is consistent with those derived under invariance considerations by George and McCulloch (1993).

The merit of Jeffreys' prior, and variants such as ratios of asymptotic variances, remains somewhat inconclusive for high dimensional problems. Obviously, if the information-theoretic assumptions are satisfied, then Jeffreys' prior, or its similarly derived variants are ineluctable. Even so, using the Jeffreys prior directly can be cumbersome when the Fisher information is far from diagonal, e.g., the Dirichlet example. One way around this (suggested originally by Jeffreys and since studied extensively) is to use a product of Jeffreys priors for individual parameters, or groups of parameters, see Berger and Bernardo (1992b) for one instance of this. There is evidence that this is a viable solution in some cases. There are also cases in which truncating the parameter space to get propriety leads to nontrivial dependence on the truncation. This can be examined via robustness to cut-off specification and some researchers have put a hyper prior on the point of truncation to good effect.

Nevertheless, some investigators argue that relative to ideal inference, Jeffreys prior can put too little or too much weight on tail regions of the parameter space: Chen et al. (2009) noted that in many binomial regression problems Jeffreys prior has tails that are lighter than any multivariate t-distribution. By contrast, Jeffreys prior for the mean and variance, (μ, Σ), in a multivariate normal problem is $\Pi(\mu, \Sigma) \propto |\Sigma|^{-(p+2)/2}$ but the exact frequentist matching prior is $\Pi_{FM}(\mu, \Sigma) \propto |\Sigma|^{-p}$, see Geisser and Cornfield (1963), indicating the tails of Jeffreys prior may be heavier than desirable. Note that this means Jeffreys' prior can put too little or too much mass around some points such as zero. Even in the simplest setting of a Bernoulli(π) where Jeffreys' prior is $\Pi_J(\pi) \propto (\pi(1-\pi))^{-1/2}$ and puts relatively high mass around $\frac{1}{2}$ making some values of π more reasonable a priori than others. Zhu and Lu (2004) explain this by an estimator matching argument. Roughly, they looked for priors that make the MLE equal to the posterior mean and argue that the uniform is not always least informative, deriving the Haldane prior $\Pi_H(\pi) \propto (\pi(1-\pi))^{-1}$ (when one wants a uniform distribution on $\log(\pi/(1-\pi))$) and a prior that concentrates near $\pi = 0$ or 1.

On the other hand, for some sparse problems involving dimension reduction via principal components, see Guan and Dy (2009), Jeffreys' prior seems to work well. Also, various modifications of the Jeffreys' prior such as Berger and Bernardo (1992b) and Yang and Berger (1994) give good performance, even in certain regression problems; see Chen et al. (2009). Overall, it seems rare that Jeffreys' prior, or some modification of it, will fail to give good results.

6. Appendix A: Posterior normality

The proof of Theorem 3.1 rests on the asymptotic normality of the posterior in the L_1-sense. This posterior normality is of a particularly strong form because we obtain an explicit bound on the L_1-distance on a "good" set

$$W = W_{n,p,\theta_0,\delta} = \{\|\Delta_n\| \leq \frac{1}{4}p^{(1+m)/2+\delta}\}$$

such that $P(W)$ decays to zero at a polynomial rate in p^{-1} (see Lemma 6.5). The first step of the proof of posterior normality is to use an instance of an inequality that can be stated informally as follows. Let a and b be positive integrable functions of u. Let $\{N, N^c\}$ be a partition of the domain such that informally N stands for the central region, where $|a-b|$ is small, whereas N^c stands for the tail region, where a and b are individually small. Then, we can estimate the L_1-distance between the normalized functions as follows. By adding and subtracting $a/\int b$, bounding the first term by the second, and partitioning the domain of integration we get

$$\begin{aligned}
&\int \left|\frac{a}{\int adu} - \frac{b}{\int bdu}\right| du \\
&\quad \leq \left|\frac{1}{\int adu} - \frac{1}{\int bdu}\right| \int adu + \frac{1}{\int bdu}\int |a-b|du \\
&\quad \leq 2\left(\frac{1}{\int bdu}\int_N |a-b|du + \int_{N^c}\frac{a}{\int bdu}du + \int_{N^c}\frac{b}{\int bdu}du\right). \qquad (6.1)
\end{aligned}$$

Now, we can state our bound on the L_1-distance between the posterior and its normal approximation.

Theorem 6.1. *Assume Conditions (MCV), (PDB) and (BF). Then on W, we have*

$$\begin{aligned}
&\int \left|\Pi_n^*(u) - \phi_p(u|\Delta_n, I_p)\right| du \\
&\quad \lesssim K_n\sqrt{\|F^{-1}\|}\frac{p^{(1+m)/2+\delta}}{\sqrt{n}} + \frac{p^{3(1+m)/2+3\delta}}{\sqrt{n}}B_n + \frac{p^{2+2m+4\delta}}{n}B_n^* \\
&\quad + e^{-p^{1+m+2\delta}/16}. \qquad (6.2)
\end{aligned}$$

Proof. The proof is very similar to that of Theorem 2.3 of Ghosal (2000). Start by using (6.1) with $a(u) = \Pi(\theta_0 + n^{-1/2}Hu)Z_n(u)$, $b(u) = \Pi(\theta_0)\tilde{Z}_n(u)$, recognizing that ϕ_p is $b(u)/\int b(v)dv$. This gives that the left hand side of (6.2) is bounded by two times

$$\begin{aligned}
&\left(\int \Pi(\theta_0)\tilde{Z}_n(u)du\right)^{-1} \\
&\quad \times \int_{\|u\|\leq p^{(1+m)/2+\delta}} |\Pi(\theta_0) + \frac{1}{\sqrt{n}}Hu)Z_n(u) - \Pi(\theta_0)\tilde{Z}_n(u)|du
\end{aligned}$$

$$+\left(\int \Pi(\theta_0)\tilde{Z}_n(u)du\right)^{-1}$$
$$\times \int_{\|u\|>p^{(1+m)/2+\delta}} \Pi(\theta_0+\frac{1}{\sqrt{n}}Hu)Z_n(u)du$$
$$+\int_{\|u\|>p^{(1+m)/2+\delta}} \phi_p(u|\Delta_n, I_p)du. \tag{6.3}$$

The first term in (6.3) can be bounded by adding and subtracting $\Pi(\theta_0)Z_n(u)$ and using the triangle inequality, namely by

$$\sup_{\|u\|\le p^{(1+m)/2+\delta}} \left|\frac{\Pi(\theta_0+\frac{1}{\sqrt{n}}Hu)}{\Pi(\theta_0)}-1\right| \frac{\int_{\|u\|\le p^{(1+m)/2+\delta}} Z_n(u)du}{\int \tilde{Z}_n(u)du}$$
$$+\frac{\int |Z_n(u)-\tilde{Z}_n(u)|du}{\int \tilde{Z}_n(u)du}$$
$$\le \frac{2K_n\sqrt{\|F^{-1}\|}p^{(1+m)/2+\delta}}{\sqrt{n}}\mathcal{O}(1)+\frac{B_np^{3(1+m)/2+3\delta}}{6\sqrt{n}}+\frac{B_n^*p^{2+2m+4\delta}}{n}$$

in view of Lemmas 6.8 and 6.12 below, respectively. The bound on the second term of (6.3) follows from Lemma 6.10 and is $e^{-p^{1+m+2\delta}/16}$. The bound on the third term of (6.3) follows directly from Lemma 6.11 below and is $e^{-c_1p^{1+m+2\delta}}$. □

To complement Theorem 6.1, we extract a corollary that bounds the probability of W^c. It is this result that is used in Theorem 3.1.

Corollary 6.2. *On* $W=\{\Delta_n\le p^{(1+m)/2+\delta}/4\}$, $\Pi_n(\sqrt{n}\|J(\theta-\theta_0)\|>p^{(1+m)/2+\delta})$ *is bounded by a multiple of*

$$\lambda_n^* := K_n\sqrt{\|F^{-1}\|}\frac{p^{(1+m)/2+\delta}}{\sqrt{n}}+\frac{p^{3(1+m)/2+3\delta}}{\sqrt{n}}B_n+\frac{p^{2+2m+4\delta}}{n}B_n^*$$
$$+e^{-p^{1+m+2\delta}/16}+e^{-4p^{1+m+2\delta}}.$$

Proof. By adding and subtracting the limiting normal and using the triangle inequality,

$$\Pi_n^*(\|u\|>p^{(1+m)/2+\delta}) \lesssim \int |\Pi_n^*(u)-\phi_p(u|\Delta_n, I_p)|du$$
$$+\int_{\|u\|>p^{(1+m)/2+\delta}} \phi_p(u|\Delta_n, I_p)du. \tag{6.4}$$

The corollary now follows from Theorem 6.1 and Lemma 6.11 below. □

Next, we turn to the formal proof of Theorem 6.1 which we have broken up into a series of Lemmas. We begin with a result on local expansions for ψ and ψ'. It is a restatement, in terms of the local parameter, of an approximation result due to Portnoy (1988).

Lemma 6.3. *The normalizing function in the exponential family has the local expansion for every u,*

$$\psi(\theta_0 + n^{-1/2}Hu) = \psi(\theta_0) + \frac{1}{\sqrt{n}}u^T J^{-1}\mu + \frac{1}{2n}\|u\|^2 + R_{1n}, \tag{6.5}$$

where, for some $\tilde{\theta}$ lies between θ_0 and $\theta_0 + n^{-1/2}Hu$,

$$\begin{aligned} |R_{1n}| &= \left| \frac{1}{6n^{3/2}}\mathrm{E}_{\theta_0}(u^T V)^3 + \frac{1}{24n^2}\{\mathrm{E}_{\tilde{\theta}}(u^T V)^4 - 3\left[\mathrm{E}_{\tilde{\theta}}(u^T V)^2\right]^2 \right| \\ &\leq \frac{\|u\|^3}{6n^{3/2}}B_n + \frac{\|u\|^4}{24n^2}B_n^*. \end{aligned} \tag{6.6}$$

Further,

$$\psi'(\theta_0 + n^{-1/2}Hu) = \mu + \frac{1}{\sqrt{n}}Ju + R_{2n} \tag{6.7}$$

where $R_{2n} = \frac{1}{2n}\mathrm{E}_{\tilde{\theta}}\left[(u^T V)^2 JV\right]$.

The following lemma bounds the moments of $\|\Delta_n\|$ and hence controls probabilities of deviation of it.

Lemma 6.4. *Let $r \geq 1$. Then there exist universal constants C_{2r}, depending only on r, so that $\mathrm{E}\|\Delta_n\|^{2r} \leq C_{2r}M_{2r}p^r$.*

Proof. For $i = 1, \ldots, n$, let $(V_{1,i}, \ldots, V_{p,i})^T$ be i.i.d. outcomes of the random variable $V = (V_1, \ldots, V_p)^T$. Then

$$\begin{aligned} \mathrm{E}\|\Delta_n\|^{2r} &= \mathrm{E}\left\| \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n V_{i1}, \ldots, \frac{1}{\sqrt{n}}\sum_{i=1}^n V_{ip} \right) \right\|^{2r} \\ &= \mathrm{E}\left\{ \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n V_{i1} \right)^2 + \cdots + \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n V_{ip} \right)^2 \right\}^r \\ &\leq p^{r-1}\left\{ \mathrm{E}\left| \frac{1}{\sqrt{n}}\sum_{i=1}^n V_{i1} \right|^{2r} + \cdots + \mathrm{E}\left| \frac{1}{\sqrt{n}}\sum_{i=1}^n V_{ip} \right|^{2r} \right\} \qquad (6.8) \\ &\leq C_{2r}p^{r-1}\{\mathrm{E}|V_1|^{2r} + \cdots + \mathrm{E}|V_p|^{2r}\} \qquad (6.9) \\ &\leq C_{2r}p^r \max_{1\leq j\leq p} \mathrm{E}|V_j|^{2r} \\ &= C_{2r}p^r M_{2r}. \end{aligned}$$

Note that (6.8) follows from the standard inequality $(a_1 + \cdots + a_k)^r \leq k^{r-1}(a_1^r + \cdots + a_k^r)$ and (6.9) follows from the Marcinkiewicz-Zygmund inequality — for mean-zero IID random variables, $\mathrm{E}|n^{-1/2}\sum_{j=1}^n V_j|^r \leq C_r\mathrm{E}|V|^r, r \geq 1$ for some constant C_r independent of the summands. □

The next lemma bounds the probability of W^c, ensuring it is unlikely that $\|\Delta_n\|$ is too large.

Lemma 6.5. *Let $\delta > 0$, $r \geq 1$, and assume Condition (MCV4). Then there exists a constant $C = C(r, m)$ but independent of δ, n, and p, so that $P_{\theta_0}(W^c) \leq C/p^{2r\delta}$.*

Proof. The proof follows from Markov's inequality, Lemma 6.4 and the use of Condition (MCV4) to control M_{2r}. We have from Condition (MCV4) that

$$P_{\theta_0}\left(\|\Delta_n\| > \frac{p^{(1+m)/2+\delta}}{4}\right) \leq 4^{2r}\frac{\mathrm{E}_{\theta_0}\|\Delta_n\|^{2r}}{p^{r+mr+2r\delta}} \lesssim \frac{p^r M_{2r}}{p^{r+mr+2r\delta}} \lesssim \frac{1}{p^{2r\delta}}. \tag{6.10}$$

□

Note that the main role of (MCV4) appears in Lemma 6.5 above to control the probability of W^c. This kind of condition is not needed for posterior normality because posterior normality is a local property, i.e., depends only on a shrinking neighborhood of θ_0, and on the increase in number of data points. It is only when we want to aggregate over data sequences by taking a probability that we must control moments as in (MCV4).

On W, the set where $\|\Delta_n\|$ is relatively small, a bound on the maximum likelihood estimator for standardized parameter can be given. The probability that this bound can be violated can also be controlled at the $\mathcal{O}(p^{-2r\delta})$ rate. This is formalized in the following.

Lemma 6.6. *Suppose that Conditions (MCV2) and (MCV4) hold. Then on W, we have the bound*

$$\|\sqrt{n}J^T(\hat{\theta} - \theta_0)\| \leq \frac{p^{(1+m)/2+\delta}}{2}. \tag{6.11}$$

Consequently,

$$P_{\theta_0}\left(\|\sqrt{n}J^T(\hat{\theta} - \theta_0)\| > \frac{p^{(1+m)/2+\delta}}{2}\right) \leq \mathcal{O}(p^{-2r\delta}). \tag{6.12}$$

Proof. The proof is similar to that of Theorem 2.1 of Ghosal (2000). Define $G(u) = \sqrt{n}J^{-1}\{\psi'(\theta_0 + \frac{1}{\sqrt{n}}Hu) - \overline{X}\}$. Since ψ is convex, in view of Theorem 6.3.4 in Ortega and Rheinboldt (1970), $G(u) = 0$ has a unique root in the set $\{u : \|u\| \leq p^{(1+m)/2+\delta}/2\}$ provided we show that $u^T G(u) \geq 0$ on $\{u : \|u\| = p^{(1+m)/2+\delta}/2\}$. Since $\sqrt{n}J^T(\hat{\theta} - \theta_0)$ is clearly a root, in order to verify (6.11), it suffices to show $u^T G(u) \geq 0$ on W for all u with $\|u\| = p^{(1+m)/2+\delta}/2$.

First observe that by (6.7),

$$\begin{aligned} u^T G(u) &= \sqrt{n}u^T J^{-1}\{\psi'(\theta_0 + \frac{1}{\sqrt{n}}Hu) - \overline{X}\} \\ &= \sqrt{n}u^T J^{-1}\{\mu + \frac{1}{\sqrt{n}}Ju + R_{2n} - \overline{X}\} \\ &= -u^T\Delta_n + \|u\|^2 + \frac{u^T J^{-1}}{2\sqrt{n}}\mathrm{E}_{\tilde{\theta}}[(u^T V)^2 JV]. \end{aligned} \tag{6.13}$$

When $\|u\| \le p^{(1+m)/2+\delta}/2$, the last term in (6.13) is bounded as

$$u^T J^{-1} \mathrm{E}_{\tilde{\theta}}[(u^T V)^2 JV] \le \mathrm{E}_{\tilde{\theta}}|u^T V|^3 \le p^{1+m+2\delta} B_n' \|u\|/4. \tag{6.14}$$

Thus on W, the triangle inequality gives

$$\begin{aligned} u^T G(u) &\ge \|u\|^2 - \|u\|\|\Delta_n\| - \frac{\|u\|}{8\sqrt{n}} p^{(1+m)+2\delta} B_n' \\ &= \|u\| \left\{ \frac{p^{(1+m)/2+\delta}}{2} - \frac{p^{(1+m)/2+\delta}}{4} - \frac{p^{(1+m)+2\delta}}{8\sqrt{n}} B_n' \right\} \\ &= \|u\| \frac{p^{(1+m)/2+\delta}}{4} \left(1 - \frac{p^{(1+m)/2+\delta}}{2\sqrt{n}} B_n' \right) \ge 0 \end{aligned} \tag{6.15}$$

for all sufficiently large n by Condition (MCV2). Now, (6.12) follows from Lemma 6.5. □

The control on Δ_n helps us study the local likelihood ratio process Z_n

$$\tilde{Z}_n(u) = \exp[u^T \Delta_n - \frac{1}{2}\|u\|^2]. \tag{6.16}$$

Clearly, $\int \tilde{Z}_n(u)\, du = (2\pi)^{p/2} e^{\|\Delta_n\|^2/2}$.

Lemma 6.7. *For all $\|u\| \le p^{(1+m)/2+\delta}$, we have that*

$$|\log Z_n(u) - \log \tilde{Z}_n(u)| \le \lambda_n \|u\|^2, \tag{6.17}$$

where $\lambda_n = \dfrac{p^{(1+m)/2+\delta}}{\sqrt{n}} B_n + \dfrac{p^{1+m+2\delta}}{n} B_n^*$. Further, we have the one sided bounds

$$\log Z_n(u) \le u^T \Delta_n - \frac{1}{2}\|u\|^2 (1 - 2\lambda_n), \tag{6.18}$$

$$\log Z_n(u) \ge u^T \Delta_n - \frac{1}{2}\|u\|^2 (1 + 2\lambda_n). \tag{6.19}$$

Proof. By the definitions of Z_n and $\tilde{Z}_n$, and (6.5) in Lemma 6.3,

$$\begin{aligned} &\log Z_n(u) - \log \tilde{Z}_n(u) \\ &\quad = \sqrt{n} u^T J^{-1} \overline{X} - n \left[\psi(\theta_0 + \frac{1}{\sqrt{n}} Hu) - \psi(\theta_0) \right] - u^T \Delta_n + \frac{1}{2}\|u\|^2 \\ &\quad = \sqrt{n} u^T J^{-1} \overline{X} - \sqrt{n} u^T J^{-1} \mu - \frac{1}{2}\|u\|^2 + n R_{1n} - u^T \Delta_n + \frac{1}{2}\|u\|^2 \\ &\quad \le n\|u\|^2 \{ \frac{p^{(1+m)/2+\delta}}{6n^{3/2}} B_n + \frac{p^{1+m+2\delta}}{24 n^2} B_n^* \} \le \lambda_n \|u\|^2, \end{aligned} \tag{6.20}$$

which proves (6.17).

Relations (6.18) and (6.19) clearly follow from the triangle inequality and the definition of $\tilde{Z}_n(u)$. □

Lemma 6.8. *Assume Conditions (MCV1) and (MCV3). Then on W, we have*

$$\frac{\int_{\|u\|\le p^{(1+m)/2+\delta}} \mid Z_n(u) - \tilde{Z}_n(u) \mid du}{\int \tilde{Z}_n(u)du} \lesssim p^{1+m+2\delta}\lambda_n = o(1), \tag{6.21}$$

$$\frac{\int_{\|u\|\le p^{(1+m)/2+\delta}} Z_n(u)du}{\int \tilde{Z}_n(u)du} = \mathcal{O}(1). \tag{6.22}$$

Proof. The proof is based on the proofs of Lemma 2.1 and Lemma 2.3 of Ghosal (2000). Using $|e^x - e^y| \le |x-y|\max(e^x, e^y)$ gives

$$\begin{aligned}|Z_n(u) - \tilde{Z}_n(u)| &\le |\log Z_n(u) - \log \tilde{Z}_n(u)| \times \exp[u^T\Delta_n - \frac{1}{2}(1-2\lambda_n)\|u\|^2] \\ &\le \lambda_n\|u\|^2 \exp[u^T\Delta_n - \frac{1}{2}(1-2\lambda_n)\|u\|^2],\end{aligned} \tag{6.23}$$

from Lemma 6.7, (6.17), and (6.18). By a completing-the-square argument,

$$u^T\Delta_n - \frac{1}{2}(1-2\lambda_n)\|u\|^2 = -\frac{1}{2}(1-2\lambda_n)\left\|u - \frac{\Delta_n}{1-2\lambda_n}\right\|^2 + \frac{\|\Delta_n\|^2}{2(1-2\lambda_n)}$$

and hence

$$\begin{aligned}&\int \|u\|^2 \exp[u^T\Delta_n - \frac{1}{2}(1-2\lambda_n)\|u\|^2]du \\ &\quad= (2\pi)^{p/2}(1-2\lambda_n)^{-p/2}\exp\left[\frac{\|\Delta_n\|^2}{2(1-2\lambda_n)}\right] \\ &\qquad\times \int \|u\|^2(2\pi)^{-p/2}(1-2\lambda_n)^{p/2}\exp\left[-\frac{(1-2\lambda_n)}{2}\|u - \frac{\Delta_n}{1-2\lambda_n}\|^2\right]du \\ &\quad= (2\pi)^{p/2}(1-2\lambda_n)^{-p/2+1}\exp\left[\frac{\|\Delta_n\|^2}{2(1-2\lambda_n)}\right]\left(p + \frac{\frac{\|\Delta_n\|^2}{(1-2\lambda_n)^2}}{(1-2\lambda_n)^{-1}}\right) \\ &\quad= (2\pi)^{p/2}(1-2\lambda_n)^{-p/2+1}\exp\left[\frac{\|\Delta_n\|^2}{2(1-2\lambda_n)}\right]\left(p + \|\Delta_n\|^2(1-2\lambda_n)^{-1}\right).\end{aligned}$$

Using this and (6.23),

$$\begin{aligned}&\left(\int \tilde{Z}_n(u)du\right)^{-1}\int_{\|u\|\le p^{(1+m)/2+\delta}} |Z_n(u) - \tilde{Z}_n(u)|du \\ &\quad\le \lambda_n\left(p + \frac{\|\Delta_n\|^2}{1-2\lambda_n}\right)(1-2\lambda_n)^{-p/2+1}\exp\left[\frac{\|\Delta_n\|^2}{2}((1-2\lambda_n)^{-1} - 1)\right] \\ &\quad\le \lambda_n\left(p + \frac{p^{1+m+2\delta}}{4(1-2\lambda_n)}\right)(1-2\lambda_n)^{-p/2+1}\exp\left[\frac{\lambda_n p^{1+m+2\delta}}{1-2\lambda_n}\right] \\ &\quad\lesssim p^{1+m+2\delta}\lambda_n,\end{aligned}$$

in which the restriction to W was used at the second inequality. Clearly, conditions (MCV1) and (MCV3) give $p^{1+m+2\delta}\lambda_n \to 0$. This also implies that $(1-2\lambda_n)^{-p/2} \to 1$. In particular, that $1-2\lambda_n > \frac{1}{2}$ for sufficiently large n. These assertions together imply (6.21).

For (6.22), note that

$$\frac{\int_{\|u\|\le p^{(1+m)/2+\delta}} Z_n(u)du}{\int \tilde{Z}_n(u)du} \le 1 + \frac{\int_{\|u\|\le p^{(1+m)/2+\delta}} |Z_n(u)-\tilde{Z}_n(u)|du}{\int \tilde{Z}_n(u)du}$$
$$= 1 + \mathcal{O}(p^{1+m+2\delta}\lambda_n) = \mathcal{O}(1)$$

□

Lemma 6.9. *Under Conditions (MCV1)–(MCV3),* $\log Z_n(u) \le -p^{1+m+2\delta}/8$ *on* W.

Proof. The proof is similar to that of Lemma 2.2 of Ghosal (2000). Let $\hat{u} = \sqrt{n}J^T(\hat{\theta}-\theta_0)$, where $\hat{\theta}$ is the MLE of θ. On W, we have $\|\hat{u}\| \le p^{(1+m)/2+\delta}/2$ by Lemma 6.6. By the convexity of $\psi(\theta)$, the log-likelihood function is log-concave, and hence the likelihood decreases if θ moves away from $\hat{\theta}$ along any line passing through $\hat{\theta}$.

Let u, $\|u\| > p^{(1+m)/2+\delta}$, be given and let ξ be the point of intersection of the line passing through the origin and u with the sphere $\|u\| = p^{(1+m)/2+\delta}$. Thus, $\|\xi\| = p^{(1+m)/2+\delta}$. Now, by Lemma 6.7 and (6.18), we have

$$\begin{aligned} Z_n(u) \le Z_n(\xi) &\le \exp[\xi^T\Delta_n - \frac{1}{2}(1-2\lambda_n)\|\xi\|^2] \\ &\le \exp[\|\xi\|\|\Delta_n\| - \frac{1}{2}(1-2\lambda_n)\|\xi\|] \\ &= \exp[p^{(1+m)/2+\delta}\frac{1}{4}p^{(1+m)/2+\delta} - \frac{1}{2}(1-2\lambda_n)p^{1+m+2\delta}] \\ &\le \exp[-\frac{1}{8}p^{1+m+2\delta}]. \end{aligned}$$

□

Lemma 6.10. *Assume Conditions (MCV1)–(MCV3), (BF1), and (PDB1). Then on* W*, we have*

$$\frac{\int_{\|u\|>p^{(1+m)/2+\delta}} \Pi(\theta_0 + \frac{1}{\sqrt{n}}Hu)Z_n(u)du}{\int \Pi(\theta_0)\tilde{Z}_n(u)du} \le e^{-p^{1+m+2\delta}/16}. \tag{6.24}$$

Proof. By Lemma 6.9 and a change of variables, it is seen that the second integral in (6.24) is bounded from above by

$$\int_{\|u\|>p^{(1+m)/2+\delta}} \Pi(\theta_0 + \frac{1}{\sqrt{n}}Hu)Z_n(u)du$$
$$\le e^{-p^{1+m+2\delta}/8} n^{p/2}\sqrt{\det(F)}\int \Pi(\theta)d\theta$$

$$= \exp[-\frac{1}{8}p^{1+m+2\delta} + \frac{p}{2}\log n + \frac{1}{2}\log\det(F)].$$
$$\leq \exp[-\frac{1}{8}p^{1+m+2\delta} + Cp\log p]$$

using $\log n = \mathcal{O}(\log p)$ and the fact that $\log\det F = O(p\log p)$, in view of (BF1).

On the other hand,

$$\int \Pi(\theta_0)\tilde{Z}_n(u)du = \Pi(\theta_0)e^{\|\Delta_n\|^2/2}(2\pi)^{p/2} \geq \Pi(\theta_0) \geq e^{-cp\log p},$$

by Condition (PDB1). Thus the ratio in (6.24) is bounded above by

$$\exp[-\frac{1}{8}p^{1+m+2\delta} + \frac{p}{2}\log n + \frac{1}{2}\log\det(F) + Cp\log p] \leq e^{-p^{1+m+2\delta}/16}.$$

□

Lemma 6.11. *On W, we have that for some $c_1 > 0$,*

$$\int_{\|u\|>p^{(1+m)/2+\delta}} \phi_p(u|\Delta_n, I_p)du \leq e^{-c_1 p^{1+m+2\delta}}.$$

More generally, for a and b remaining in a fixed bounded set D, there exists $c_2, c_3 > 0$ depending on D only such that on W,

$$\int_{\|u\|>c_3 p^{(1+m)/2+\delta}} \phi_p(u|a\Delta_n, bI_p)du \leq e^{-c_2 p^{1+m+2\delta}}.$$

Proof. Let $\xi \sim N_p(\Delta_n, I_p)$. Then $\xi - \Delta_n \sim N_p(0, I_p)$ and $\|\xi - \Delta_n\|^2 \sim \chi^2_p$. Since $\|\Delta_n\| \leq \frac{1}{4}p^{(1+m)/2+\delta}$ on W,

$$\int_{\|u\|>p^{(1+m)/2+\delta}} \phi_p(u|\Delta_n, I_p)du \leq \mathrm{P}(\|\xi - \Delta_n\| > \frac{3}{4}p^{(1+m)/2+\delta}) \leq e^{-c_1 p^{1+m+2\delta}}.$$

More generally, with $\xi \sim N(a\Delta_n, bI_p)$, we have

$$\int_{\|u\|>c_3 p^{(1+m)/2+\delta}} \phi_p(u|a\Delta_n, bI_p)du = \mathrm{P}\left(\frac{\|\xi - a\Delta_n\|}{\sqrt{b}}) > \frac{c_3 - a}{\sqrt{b}}p^{(1+m)/2+\delta}\right)$$
$$\leq e^{-\frac{c_1(c_3-a)^2}{b}p^{(1+m)/2+\delta}},$$

so it is enough to choose $c_2 = c_1(c_3 - a)^2/b$. □

Lemma 6.12. *Assume Conditions (BF2) and (PDB2). Then*

$$\sup_{\|u\|\leq p^{(1+m)/2+\delta}} \left|\frac{\Pi(\theta_0 + \frac{1}{\sqrt{n}}Hu)}{\Pi(\theta_0)} - 1\right| \leq 2K_n\sqrt{\|F^{-1}\|}\frac{p^{(1+m)/2+\delta}}{\sqrt{n}}. \tag{6.25}$$

Moreover, the following bounds hold for all u with $\|u\| \le p^{(1+m)/2+\delta}$:

$$\exp\left[-K_n\sqrt{\|F^{-1}\|}\frac{p^{(1+m)/2+\delta}}{\sqrt{n}}\right] \le \frac{\Pi(\theta_0+\frac{1}{\sqrt{n}}Hu)}{\Pi(\theta_0)} \le \exp\left[K_n\sqrt{\|F^{-1}\|}\frac{p^{(1+m)/2+\delta}}{\sqrt{n}}\right].$$

Proof. First, notice that $n^{-1/2}\|Hu\| \le n^{-1/2}\sqrt{\|F^{-1}\|}\|u\|$. Therefore, by Condition (PDB2),

$$\sup_{\|u\|\le p^{(1+m)/2+\delta}}\left|\log\frac{\Pi(\theta_0+\frac{1}{\sqrt{n}}Hu)}{\Pi(\theta_0)}\right| \le 2K_n\sqrt{\|F^{-1}\|}\frac{p^{(1+m)/2+\delta}}{\sqrt{n}},$$

from which the two one-sided bounds follow.

Using $|e^x - e^y| \le |x-y|\max(e^x, e^y)$, Condition (PDB2) gives that

$$\begin{aligned}\left|\frac{\Pi(\theta_0+n^{-1/2}Hu)}{\Pi(\theta_0)}-1\right| &\le \left|\log\frac{\Pi(\theta_0+\frac{1}{\sqrt{n}}Hu)}{\Pi(\theta_0)}\right|\exp\left[\left|\log\frac{\Pi(\theta_0+\frac{1}{\sqrt{n}}Hu)}{\Pi(\theta_0)}\right|\right]\\ &\le K_n\sqrt{\|F^{-1}\|}\frac{p^{(1+m)/2+\delta}}{\sqrt{n}}\exp\left[K_n\sqrt{\|F^{-1}\|}\frac{p^{(1+m)/2+\delta}}{\sqrt{n}}\right],\end{aligned}$$

for $\|u\| \le p^{(1+m)/2+\delta}$, so that under Condition (BF2), (6.25) follows by noting that the last factor is bounded by 2. □

This completes the proof of Theorem 6.1.

7. Appendix B: Detailed proof of Lemma 3.3 and Theorem 3.1

To see how asymptotic normality of the posterior as in Corollary 6.2 applies to the derivation of reference priors, let $N = \{u : \|u\| \le p^{(1+m)/2+\delta}\}$ be the local neighborhood (at 0) corresponding to $\tilde{N} = \{\theta = \theta_0 + n^{-1/2}Hu : \|u\| \le p^{(1+m)/2+\delta}\}$ at θ_0 in the original parametrization and recall the bound for $\Pi_n^*(N^c) \le \lambda_n^*$ on W by use of (6.4) in Corollary 6.2. Then

$$\begin{aligned}\lambda_n^* &\ge \Pi_n^*(\|u\| > p^{(1+m)/2+\delta})\\ &= \frac{\int_{N^c}\Pi^*(u)p(X^n|\theta_0+n^{-1/2}Hu)du}{\int_N\Pi^*(u)p(X^n|\theta_0+n^{-1/2}Hu)du+\int_{N^c}\Pi^*(u)p(X^n|\theta_0+n^{-1/2}Hu)du}\end{aligned}$$

can be re-arranged to give

$$\int_{N^c}\Pi^*(u)p(X^n|\theta_0+n^{-1/2}Hu)du \le \frac{\lambda_n^*}{1-\lambda_n^*}\int_N\Pi^*(u)p(X^n|\theta_0+n^{-1/2}Hu)du.$$

We begin with the proof of Lemma 3.3.

Proof of Lemma 3.3: For the upper bound, we have

$$\begin{aligned}
\frac{m_n(X^n)}{p(X^n|\theta_0)} &= \int_{\tilde{N}} \Pi(\theta)\frac{p(X^n|\theta)}{p(X^n|\theta_0)}d\theta + \int_{\tilde{N}^c} \Pi(\theta)\frac{p(X^n|\theta)}{p(X^n|\theta_0)}d\theta \\
&\leq \left(1+\frac{\lambda_n^*}{1-\lambda_m^*}\right)\int_{\tilde{N}} \Pi(\theta)\frac{p(X^n|\theta)}{p(X^n|\theta_0)}d\theta \\
&= (1-\lambda_n^*)^{-1}n^{-p/2}(\det(F))^{-1/2}\int_N \Pi(\theta_0+n^{-1/2}Hu)Z_n(u)du \\
&\leq (1-\lambda_n^*)^{-1}\exp\left[K_n\sqrt{\|F^{-1}\|}\frac{p^{(1+m)/2+\delta}}{\sqrt{n}}\right]\Pi(\theta_0)n^{-p/2}(\det(F))^{-1/2} \\
&\quad \times \int \exp\left[u^T\Delta_n - \frac{1}{2}(1-2\lambda_n)\|u\|^2\right]du \\
&= (1-\lambda_n^*)^{-1}\exp\left[K_n\sqrt{\|F^{-1}\|}\frac{p^{(1+m)/2+\delta}}{\sqrt{n}}\right]\Pi(\theta_0)n^{-p/2}(\det(F))^{-1/2} \\
&\quad \times (2\pi)^{p/2}(1-2\lambda_n)^{-p/2}\exp\left[\frac{\|\Delta_n\|^2}{2(1-2\lambda_n)}\right].
\end{aligned}$$

Note that at the second inequality, estimate (6.18) and the upper bound of the prior ratio in Lemma 6.12 were used.

The proof of the lower bound is actually simpler because the first inequality in the proof of the upper bound is not needed; the nonlocal term can be dropped. Otherwise, the same reasoning can be followed. Thus

$$\begin{aligned}
&n^{-p/2}(\det(F))^{-1/2}\int_N \Pi(\theta_0+n^{-1/2}Hu)Z_n(u)du \\
&\quad \geq \exp\left[-K_n\sqrt{\|F^{-1}\|}\frac{p^{(1+m)/2+\delta}}{\sqrt{n}}\right]\Pi(\theta_0)n^{-p/2}(\det(F))^{-1/2} \\
&\qquad \times \int_N \exp[u^T\Delta_n - \frac{1}{2}(1+2\lambda_n)\|u\|^2]du \\
&\quad = \exp\left[-K_n\sqrt{\|F^{-1}\|}\frac{p^{(1+m)/2+\delta}}{\sqrt{n}}\right]\Pi(\theta_0)n^{-p/2}(\det(F))^{-1/2}(2\pi)^{p/2} \\
&\qquad \times \exp\left[\frac{\|\Delta_n\|^2}{2(1+2\lambda_n)}\right](1+2\lambda_n)^{-p/2} \\
&\qquad \times \left(1-\int_{N^c}\phi_p\left(u|\frac{\|\Delta_n\|}{1+2\lambda_n},(1+2\lambda_n)I_p\right)du\right).
\end{aligned}$$

Here, in the second step, estimate (6.19) and the lower bound of the prior ratio in Lemma 6.12 were used. Now, the statement follows from the second part of Lemma 6.11. □

Proof of Theorem 3.1: Let us start with the first error term (3.4).

By restricting the domain of integration, we have

$$\begin{aligned} m_n(X^n) &= \int p(X^n|\theta)\Pi(\theta)d\theta \\ &\geq \int_{\|\theta-\theta_0\|<\frac{1}{n}} p(X^n|\theta)\Pi(\theta)d\theta \\ &= \Pi\left(\|\theta-\theta_0\| < \frac{1}{n}\right)\left\{\frac{1}{\Pi(\|\theta-\theta_0\|<\frac{1}{n})}\int_{\|\theta-\theta_0\|<\frac{1}{n}} p(X^n|\theta)\Pi(\theta)d\theta\right\}; \end{aligned}$$

here and afterwards, we shall slightly abuse notation to denote prior probabilities by the same symbol Π. Therefore, by Jensen's inequality ($-\log \mathrm{E}(X) \leq -\mathrm{E}(\log X)$), Condition (PDB1), and the form of exponential families for n independent random variables, $\log(p(X^n|\theta_0)/m_n(X^n))$ is

$$\begin{aligned} &-\log\Pi\left(\|\theta-\theta_0\| < \frac{1}{n}\right) - \frac{\int_{\|\theta-\theta_0\|<\frac{1}{n}}\left\{\log\frac{p(X^n|\theta)}{p(X^n|\theta_0)}\right\}\Pi(\theta)d\theta}{\Pi(\|\theta-\theta_0\|<\frac{1}{n})} \\ &\quad\leq -\log\left(\frac{e^{-cp\log p}}{n^p}\right) \\ &\qquad -\frac{n}{\Pi(\|\theta-\theta_0\|<\frac{1}{n})}\int_{\|\theta-\theta_0\|<\frac{1}{n}}\{(\theta-\theta_0)^T\bar{X} - \psi(\theta) + \psi(\theta_0)\}\Pi(\theta)d\theta. \end{aligned}$$

Taken together, we now can bound (3.4) using the following:

$$\begin{aligned} &\mathrm{E}_{\theta_0}\left\{\log_+\left(\frac{p(X^n|\theta_0)}{m_n(X^n)}\right)\mathbb{1}_{W^c}\right\} \\ &\quad\leq (p\log n + cp\log p)P_{\theta_0}(W^c) + \frac{n}{\Pi(\|\theta-\theta_0\|<\frac{1}{n})} \\ &\qquad\times\int_{\|\theta-\theta_0\|<\frac{1}{n}}\mathrm{E}_{\theta_0}\left\{|(\theta-\theta_0)^T\overline{X} - \psi(\theta) + \psi(\theta_0)|\mathbb{1}_{W^c}\right\}\Pi(\theta)d\theta \\ &\quad\leq \frac{C(p\log n + p\log p)}{p^{2r\delta}} + \frac{n}{\Pi(\|\theta-\theta_0\|<\frac{1}{n})} \\ &\qquad\times\int_{\|\theta-\theta_0\|<\frac{1}{n}}\mathrm{E}_{\theta_0}\left\{|(\theta-\theta_0)^T\bar{X} - \psi(\theta) + \psi(\theta_0)|\mathbb{1}_{W^c}\right\}\Pi(\theta)d\theta. \quad (7.1) \end{aligned}$$

The first term in (7.1) tends to zero when n is polynomial in p provided we choose $r > 1/(2\delta)$. Thus, it is enough to use expression (6.5) in the second term of (7.1) to see it is bounded by

$$\begin{aligned} &\frac{Cn}{p^{r\delta}\Pi(\|\theta-\theta_0\|<\frac{1}{n})}\int_{\|\theta-\theta_0\|<\frac{1}{n}}\Big(\mathrm{E}_{\theta_0}|(\theta-\theta_0)^T(\overline{X}-\mu) \\ &\qquad\qquad -\frac{1}{2n}(\theta-\theta_0)^T F(\theta-\theta_0) + R_{1,n}|^2\Big)^{1/2}\Pi(\theta)d\theta \end{aligned}$$

$$\leq \frac{Cn}{p^{r\delta}\Pi(\|\theta-\theta_0\|<\frac{1}{n})}\int_{\|\theta-\theta_0\|<\frac{1}{n}}\Bigg(\mathrm{E}_{\theta_0}[(\theta-\theta_0)^T(\bar{X}-\mu)]^2 + \left[\frac{1}{2n}(\theta-\theta_0)^T F(\theta-\theta_0)\right]^2 + \mathrm{E}_{\theta_0}(R_{1,n}^2)\Bigg)^{1/2}\Pi(\theta)d\theta. \tag{7.2}$$

To deal with the three terms under the square root in (7.2), note that $\|u\| \leq \sqrt{n\|F\|}\|\theta-\theta_0\|$ and that the domain of integration is $\|\theta-\theta_0\| \leq n^{-1}$. So, the first term under the square root is

$$\mathrm{E}_{\theta_0}[(\theta-\theta_0)^T(\overline{X}-\mu)]^2 = \frac{1}{n}(\theta-\theta_0)^T F(\theta-\theta_0) \leq \frac{\|F\|}{n^3} \tag{7.3}$$

Similarly, the second term under the square root is bounded above by $(4n^6)^{-1}\|F\|^2$. To deal with the third term under the square root, use Lemma 6.3 and $(a+b)^2 \leq 2a^2+2b^2$ to obtain

$$\mathrm{E}_{\theta_0}(R_{1,n}^2) \leq \left(\frac{\|u\|^3}{6n^{3/2}}B_n + \frac{\|u\|^4}{24n^2}B_n^*\right)^2 \leq \frac{\|F\|^3 B_n^2}{3n^6} + \frac{\|F\|^4(B_n^*)^2}{12n^8}.$$

Using these three bounds and the fact that the resulting integral cancels the prior probability, (7.2) is bounded above by

$$\frac{Cn}{p^{r\delta}}\sqrt{\frac{\|F\|}{n^3} + \frac{\|F\|^2}{4n^6} + \frac{\|F\|^3 B_n^2}{3n^6} + \frac{\|F\|^4(B_n^*)^2}{12n^8}}. \tag{7.4}$$

Since $\sqrt{a_1+\cdots+a_k} \leq \sqrt{a_1}+\cdots+\sqrt{a_k}$ when $a_j \geq 0$, (7.4) is bounded by

$$\frac{C}{p^{r\delta}}\left(\sqrt{\frac{\|F\|}{n}} + \frac{\|F\|}{2n^2} + \frac{\|F\|^{3/2}B_n}{\sqrt{3}n^2} + \frac{\|F\|^2 B_n^*}{2\sqrt{3}n^3}\right),$$

which goes to zero as a consequence of Condition (BF0) and the (MCV) conditions. So, (3.4) holds.

To bound the second error term (3.5), let $G_n = W^c \cap \{m_n(X^n) \geq p(X^n|\theta_0)\}$. Then, for ν_n the n-fold product of ν,

$$\begin{aligned}
&\mathrm{E}_{\theta_0}\left[\mathbb{1}_{W^c}\log_-\frac{p(X^n|\theta_0)}{m_n(X^n)}\right] \\
&= P_{\theta_0}(G_n)\mathrm{E}_{\theta_0}\left[\frac{\mathbb{1}_{G_n}}{P_{\theta_0}(G_n)}\log\frac{\int\Pi(\theta)p(X^n|\theta)d\theta}{p(X^n|\theta_0)}\right] \\
&\leq P_{\theta_0}(G_n)\log\left\{\int\frac{\mathbb{1}_{G_n}}{P_{\theta_0}(G_n)}\frac{\int\Pi(\theta)p(X^n|\theta)d\theta}{p(X^n|\theta_0)}p(X^n|\theta_0)d\nu_n\right\} \\
&= -P_{\theta_0}(G_n)\log P_{\theta_0}(G_n) + P_{\theta_0}(G_n)\log\int P_\theta(G_n)\Pi(\theta)d\theta \\
&\leq -P_{\theta_0}(G_n)\log P_{\theta_0}(G_n) + P_{\theta_0}(G_n)\log 1 \\
&= -P_{\theta_0}(G_n)\log P_{\theta_0}(G_n).
\end{aligned}$$

Now $-x\log x \to 0$ as $x \to 0$ and $P_{\theta_0}(G_n) \le P_{\theta_0}(W^c) \le C^* p^{-2r\delta} \to 0$ for any choice of r, so (3.5) follows.

The third error term (3.6) is bounded by

$$\left(\frac{p}{2}\log\frac{n}{2\pi}\right)P_{\theta_0}(W^c) + \left(\log\frac{\Pi(\theta_0)}{(\det(F))^{1/2}}\right)P_{\theta_0}(W^c) \lesssim \frac{\log p}{p^{2r\delta-1}} + \frac{\log p}{p^{2r\delta}} \tag{7.5}$$

using $\log n = \mathcal{O}(\log p)$, (BF1), and (PDB1). Thus, the bound goes to zero if $r > 1/(2\delta)$.

To bound the fourth and last error term (3.7), let $r > 1$ be chosen and s be the conjugate index $r/(r-1)$. Then,

$$\begin{aligned}
&\mathrm{E}_{\theta_0}\left(\|\Delta_n\|^2 \cdot \mathbb{1}\{\|\Delta_n\| > \frac{1}{4}p^{(1+m)/2+\delta}\}\right)\\
&\le \left[\mathrm{E}_{\theta_0}\|\Delta_n\|^{2r}\right]^{1/r}\left[P_{\theta_0}(\|\Delta_n\| > \frac{1}{4}p^{(1+m)/2+\delta})\right]^{1/s}\\
&\lesssim (p^r M_{2r})^{1/r}(Cp^{-2r\delta})^{1/s}\\
&\lesssim p^{1+m-2r\delta/s},
\end{aligned}$$

by using Hölder's inequality, Lemma 6.5 and Condition (MCV4). Thus, if $r > 1 + (m+1)/2\delta$, the upper bound (3.7) goes to zero. Finally, recalling that $\mathrm{E}_{\theta_0}\|\Delta_n\|^2 = p$, the limiting form for R_n is obtained. □

8. Appendix C: Useful lemmas

Lemma 8.1. *A) Let A be a real $p \times p$ matrix and u and v be vectors of length p. If A is nonsingular then*

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u},$$

provided that $v^TA^{-1}u \neq -1$.

B) Let A be a real, symmetric $p \times p$ matrix and u be a vector of length p. If there exists a symmetric square root B for A i.e., $B = B^T$ and $A = BB^T = B^2$, then

$$C = B - \frac{uu^TB^{-1}}{1 + \sqrt{1 - u^TA^{-1}u}}$$

is a square root for $A - uu^T$, provided that $u^TA^{-1}u < 1$.

Proof. Part A follows from direct verification. For part B, we need to verify that $CC^T = A - uu^T$. Write $d = 1 + \sqrt{1 - u^TA^{-1}u}$, $C = B - (uu^TB^{-1})/d$ and $C^T = B - (B^{-1}uu^T)/d$. Now,

$$CC^T = A - \frac{2uu^T}{d} + \frac{u^TA^{-1}u}{d^2}uu^T.$$

To make $CC^T = A - uu^T$, observe that d satisfies the quadratic equation $d^2 - 2d + u^T A^{-1} u = 0$. □

Here, A is diagonal, so finding inverses or symmetric square roots is easy.

Lemma 8.2. *Let $A = ((a_{ij}))$ be a real $p \times p$ matrix with all entries bounded as $|a_{ij}| < \eta$. Then $A \leq \eta I_p$ in matrix ordering, or equivalently, for any vector x of length p, $x^T A x \leq \eta p x^T x$.*

In particular, $\det(A) \leq \det(\eta p I_p) = (\eta p)^p$.

Proof. Observe that

$$\begin{aligned} x^T A x &= \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij} x_i x_j \leq \sum_{i=1}^{p}\sum_{j=1}^{p} |a_{ij}||x_i||x_j| \\ &\leq \eta \sum_{i=1}^{p}\sum_{j=1}^{p} |x_i||x_j| = \eta \left(\sum_{i=1}^{p} |x_i|\right)^2 \\ &\leq \eta(\sqrt{p}\|x\|)^2 = \eta p \|x\|^2. \end{aligned}$$

□

References

Berger, J. O. and J. M. Bernardo (1989). Estimating a product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc. 84*, 200–207. MR0999679

Berger, J. O. and J. M. Bernardo (1991). Reference priors in a variance components problem. In P. Goel and N. Iyengar (Eds.), *Bayesian Inference in Statistics and Econometrics*, pp. 177–194. New York: Springer. MR1194392

Berger, J. O. and J. M. Bernardo (1992a). On the development of reference priors. In J. M. Bernardo, J. O. Berger, A. Dawid, and A. Smith (Eds.), *Bayesian Statistics IV*, pp. 36–60. Oxford: Clarendon Press. MR1380269

Berger, J. O. and J. M. Bernardo (1992b). Ordered group reference priors with application to the multinomial. *Biometrika 25*, 25–37. MR1158515

Berger, J. O., J. M. Bernardo, and M. Mendoza (1991). On priors that maximize expected information. In J. Klein and J. Lee (Eds.), *Recent Developments in Statistics and Their Applications*, pp. 1–20. Seoul: Freedom Academy.

Berger, J. O., J. M. Bernardo, and D. Sun (2009). The formal definition of reference priors. *Ann. Statist. 37*, 905–938. MR2502655

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B 41*, 113–147. MR0547240

Bernardo, J. M. (2010). Integrated objective Bayesian estimation and hypothesis testing. In J. M. Bernardo, J. O. Berger, A. P. D. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics IX*, Oxford. Clarendon Press.

Boucheron, S. and E. Gassiat (2009). A Bernstein-von Mises theorem for discrete probability distributions. *Elec. J. Statist. 3*, 114–148. MR2471588

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families*. Vol. 9, Lecture Notes –Monograph Series. Hayward, CA: Institute of Mathematical Statistics. MR0882001

Chen, M.-H., J. Ibrahim, and S. Kim (2009). Properties and implementation of Jeffreys' prior in binomial regression models. *J. Amer. Stat. Assoc. 103*, 1659–1664.

Clarke, B. and A. Barron (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inform. Theory 36*, 453–471. MR1053841

Clarke, B. and A. Barron (1994). Jeffreys' prior is the reference prior under entropy loss. *J. Stat. Planning and Inference 41*, 37–60. MR1292146

Clarke, B. and D. Sun (1997). Reference priors under the chi-square distance. *Sankhya 59*, 215–231. MR1665703

Clarke, B. and A. Yuan (2004). Partial information reference priors: derivation and interpretations. *J. Stat. Plann. Inf. 123*, 313–345. MR2062985

Geisser, S. and J. Cornfield (1963). Posterior distributions for multivariate normal parameters. *J. Roy. Stat. Soc. Ser. B 25*, 368–376. MR0171354

Gelman, A., J. Carlin, S. Stern, and D. Rubin (2004). *Bayesian Data Analysis*. Boca Raton, FL: Chapman and Hall. MR2027492

George, E. and R. McCulloch (1993). On obtaining invariant prior distributions. *J. Statist. Plann. Inf. 37*, 169–179. MR1243795

Ghosal, S. (1997). Normal approximation to the posterior distribution for generalized linear models with many covariates. *Math. Methods Statist. 6*, 332–348. MR1475901

Ghosal, S. (1999). Asymptotic normality of posterior distributions in high dimensional linear models. *Bernoulli 5*, 315–331. MR1681701

Ghosal, S. (2000). Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *J. Multivariate Anal. 74*, 49–68. MR1790613

Ghosal, S., J. K. Ghosh, and R. V. Ramamoorthi (1997). Non-informative priors via sieves and packing numbers. In S. Panchapakesan and N. Balakrishnan (Eds.), *Advances in Statistical Decision Theory and Applications*, pp. 119–132. New York: Birkhauser. MR1479180

Ghosal, S., J. K. Ghosh, and A. W. van der Vaart (2000). Convergence rates of posterior distributions. *Ann. Statist. 30*(2), 500–531. MR1790007

Ghosh, J. K. and R. Mukerjee (1992). Noninformative priors. In J. M. Bernardo, J. O. Berger, A. P. D. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics IV*, Oxford, pp. 195–210. Clarendon Press. MR1380277

Ghosh, J. K. and R. V. Ramamoorthi (2003). *Bayesian Nonparametrics*. New York, NY: Springer. MR1992245

Ghosh, M., V. Mergel, and R. Liu (2010). A general divergence criterion for prior selection. *To appear: Ann. Inst. Stat. Math.*.

Guan, Y. and J. Dy (2009). Sparse probabilistic principal component analysis. In *JMLR Workshop and Conference Proceedings Vol. 5: AISTATS*, pp. 185–192.

Heo, T. and J. Kim (2007). Bayesian inference for multinomial group testing. *Korean Communications in Statistics 14*, 81–92.

Ibragimov, I. and R. Hasminsky (1973). On the information in a sample about a parameter. In *Proc. 2nd Internat. Symp. on Information Theory*, Budapest, pp. 295–309. Akademiai, Kiado. MR0356948

Lindley, D. (1956). On a measure of the information provided by an experiment. *Ann. Math. Statist. 27*, 986–1005. MR0083936

Ortega, J. and W. Rheinboldt (1970). *Iterative Solution of Nonlinear Equations in Several Variables.* New York, NY: Academic Press. MR0273810

Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist. 16*, 356–366. MR0924876

Shannon, C. (1948a). A mathematical theory of communication, part i. *Bell Syst. Tech. J. 27*, 379 – 423. MR0026286

Shannon, C. (1948b). A mathematical theory of communication, part ii. *Bell Syst. Tech. J 27*, 623 – 656. MR0026286

Sono, S. (1983). On a non-informative prior distribution for Bayesian inference of multinomial distribution parameters. *Ann. Inst. Statist. Math. 35*(Part A), 167–174. MR0716027

Sun, D. and J. O. Berger (1998). Reference priors with partial information. *Biometrika 85*, 55–71. MR1627242

Yang, R. and J. O. Berger (1994). Estimation of a covariance matrix using a reference prior. *Ann. Statist. 22*, 1195–1211. MR1311972

Zhang, Z. (1994). *Discrete Noninformative Priors.* Ph. D. thesis, Department of Statistics, Yale. MR2636766

Zhu, M. and A. Lu (2004). The counter-intuitive non-informative prior for the Bernoulli family. *J. Stat. Ed. 12*, 1–10.