# On Johnson's asymptotic expansion for a posterior distribution

Patrice Bertail
Laboratoire de Statistique, Crest
3 Ave Pierre Larousse
92000 Malakoff, France
email: Patrice.Bertail@ensae.fr

Albert Y. Lo[1]
Department of Information and Systems Management
The University of Science and Technology
Clear Water Bay, Hong Kong
email: imaylo@ust.hk

**Summary.** For smooth models, the posterior distribution, centered at the MLE, has a second order asymptotic expansion in which the leading term depends on the prior density and its derivative. Recentering the posterior distribution at the posterior mean results in a prior-free second order expansion. Recentering at the posterior mode also leads to a different second order expansion with a leading term depending on the prior. Accuracy of the normal approximations to a posterior distribution based on these centerings are discussed.

[1]This research of this author is supported in part by the Hong Kong RGC grant 674/95P and 6189/98E

AMS 1991 Mathematics subject classification. Primary 62G09; secondary 62G20.

Key words and phrases: Asymptotic expansion, posterior distributions, accelerate normal approximation.

## 1. Introduction.

The classic normal approximation to a posterior distribution states that the posterior distribution of $\theta$ is approximately $N(\hat{\theta}, [-L^{(2)}(\theta|\mathbf{x})]^{-1})$ where $\hat{\theta}$ is a maximum likelihood estimator, and $L^{(j)}(\theta|\mathbf{x})$ the j-order derivative of the log likelihood $L(\theta|\mathbf{x})$. The performance of this approximation is questionable since it does not account for the contribution from the prior distribution. One has to bring in the prior information to the normal approximation in order to have more credible approximations. Another choice for centering the posterior distribution is the posterior mean. The fact that the posterior mean is almost always consistent under natural conditions [Doob (1949)] supports its appropriateness as an estimator for a location parameter. The recently developed MCMC method can be used to computing posterior means and integrals.

To improve on the normal approximation, Johnson (1969, 1970) expanded the posterior distribution for a smooth parametric model at the MLE in an asymptotic series with the standard normal being the initial term. Among the expansions, the second order term expansion provides information on the magnitude of the error of the normal approximation and gives a second order correction term which may be used to correct the normal approximation. Johnson's leading correction term which dictates the magnitude of the normal approximation depends on the prior density. In section 2, we show that centering at the posterior mean leads to a second-order expansion for the posterior distribution which is prior-free. Second order posterior expansion centering at the posterior mode is also discussed. We compare the accuracy of the these approximations to a posterior distribution based on its there second order expansions (Section 3). The expansions are manipulated via two lemmas (Section 2) that may have general interest. The first lemma is a law of large numbers for an average of functions of symmetric statistics in the spirit of Doob (1949) and Blackwell and Dubins (1962). The second lemma discusses how recentering and rescaling by constants alter second order asymptotic expansions.

## 2. Centering an asymptotic expansion of a posterior distribution.

Throughout this note, regularity conditions on the smoothness of the likelihood function $L(\theta|x_1)$ are assumed [see Johnson (1970)]. In particular, we assume that $L(\theta|x_1)$ has four derivatives (in $\theta$) which are dominated by integrable random variables in a compact neighborhood C of the "true" parameter $\theta_0$, and that the prior density has two bounded derivatives and strictly positive on C. Let $T=(\theta-\theta$

)$[-L^{(2)}(\theta\,|\mathbf{x})]^{1/2}$, then from Johnson [1967 and 1970; see also Ghosh, Sinha and Joshi (1982)], we have the following expansion

$$(2.1)\qquad \pi_T\{T\le z|\mathbf{x}\} = \Phi(z)-a_n(2+z^2)\phi(z)-d_\pi\,\phi(z)+o(n^{-1/2})$$
$$= \Phi(z)+a_n(1-z^2)\phi\,(z)-(3a_n+d_\pi\,)\phi(z)+o(n^{-1/2}),$$

where $a_n=(3!)^{-1}L^{(3)}(\theta\,|\mathbf{x})[-L^{(2)}(\theta\,|\mathbf{x})]^{-3/2}$, and

$d_\pi=[-L^{(2)}(\theta\,|\mathbf{x})]^{-1/2}\pi^{(1)}(\hat\theta\,)/\pi\,(\hat\theta\,)$.

What it states is that if the posterior distribution is centered at the MLE $\hat\theta$ and rescaled by $[-L^{(2)}(\theta\,|\mathbf{x})]^{1/2}$ (both of them do not depend on the prior), the effect of the prior distribution appears in $d_\pi$ at the leading correction term. Howewer, according to Johnson (1970) [see also page 51 in Ghosh (1994)],

$$(2.2)\qquad E[\theta\,|\mathbf{x}]-\hat\theta = (3a_n+d_\pi\,)[-L^{(2)}(\hat\theta_\pi\,|\mathbf{x})]^{-1/2}+o(n^{-1}).$$

that is the term $-(3a_n+d_\pi)$ in (2.1) is the result of a "Bayesian bias", defined as the difference between the posterior mean and the centering in question. Eliminating the "Bayesian bias" leads to a one-term expansion (Theorem 1 below) which is free of the effect of the prior distribution. In the following we denote by $SD(\theta\,|\mathbf{x})$ the standard deviation of the posterior distribution of $\theta$. Unless otherwise specified, asymptotic expansions of conditional distributions are expansions with probability one. This improves other previous asymptotic expansion, which are either in probability or L1 expansions (See Bickel and Ghosh (1990)).

**Theorem.** Assume that (2.1) holds then, uniformly in z,

(i) $\pi\,\{\,(\theta-E[\theta\,|\mathbf{x}])\times[-L^{(2)}(\theta\,|\mathbf{x})]^{1/2}\le z|\mathbf{x}\} = \Phi(z)+a_n(1-z^2)\phi(z)+o(n^{-1/2})$

is valid with probability one;

(ii) replacing $[-L^{(2)}(\theta\,|\mathbf{x})]^{1/2}$ by $1/SD(\theta\,|\mathbf{x})$ does not alter the expansion in (i).

The proof of this Theorem is based on the following two lemmas. Lemma 1 is a law of large numbers for an average of functions of symmetric statistics; its proof is based on a backward martingale convergence result due to Blackwell and Dubins (1962). Lemma 2 though obvious is a very usefull tool (which may also be used to understand Bartlett corrections, see Bickel and Ghosh (1990) for results in that direction). It describes how a distribution expanded in terms of $\Phi(z)$, $\phi(z)$, $z\phi(z)$ and $z^2\phi(z)$ up to an $o(1/\sqrt{n})$ error can be standardized, i.e., recenterred and rescaled, to an expansion in terms of $\Phi(z)$ and $(1-z^2)\phi(z)$ with the same $o(1/\sqrt{n})$ error. Through out this paper we denote

$$\text{limsup}_n\delta_n^{-1}\sup_x|h_n(x)-g_n(x)|=0 \text{ by } h_n(x)=g_n(x)+o(\delta_n).$$

**Lemma 1.** Suppose $X_1,...,X_n,...$ are i.i.d. random variables, and for each $n \geq 1$, $T_n$ is a symmetric statistic of $X_1,...,X_n$. Suppose $g(t,x)$ is a function of two variables, and that (i) $|g(t,y)| \leq h(y)$ where $h(X_1)$ is integrable and (ii) for each y, $g(t,y)$ is a continuous function in t.

Then $T_n \rightarrow T_\infty$ with probability one implies

$$n^{-1}\Sigma_{1 \leq i \leq n} g(T_n, X_i) \rightarrow E[g(T_\infty, X_1)] \text{ with probability one.}$$

**Proof.** Write $n^{-1}\Sigma_{1 \leq i \leq n} g(T_n, X_i) = E[g(T_n, X_1) | \mathcal{S}^n]$ where $\mathcal{S}^n$ is the symmetric $\sigma$–field consisting of symmetric functions of $X_1,...,X_n$. By the extended backward martingale convergence theorem of Blackwell and Dubins (1965), with probability one

$$n^{-1}\Sigma_{1 \leq i \leq n} g(T_n, X_i) \rightarrow E[g(T_\infty, X_1) | \mathcal{S}^\infty].$$

The above convergence only requires exchangeability of the $X_i$ s. Here the $X_i$ s are i.i.d., and the Hewitt-Sawage zero-one law then implies that $\mathcal{S}^\infty$ is a trivial $\sigma$–field. Therefore, $E[g(T_\infty, X_1) | \mathcal{S}^\infty]$ is a constant. The constant must be $Eg(T_\infty, X_1)$. ‖

**Lemma 2.** Suppose $\mu_n$, $\varepsilon_n$, $a_n$, $b_n$ and $c_n$ are $O(1/\sqrt{n})$. Expansions (i) and (ii) are equivalent.

(i) $P\{S \leq z\} = \Phi(z) + a_n(1-z^2)\phi(z) - b_n z\phi(z) - c_n\phi(z) + o(1/\sqrt{n})$.

(ii) $P\{(S-\mu_n) \times [1+\varepsilon_n]^{-1} \leq z\} = \Phi(z) + a_n(1-z^2)\phi(z) - (b_n - \varepsilon_n)z\phi(z) - (c_n - \mu_n)\phi(z) + o(1/\sqrt{n})$.

**Proof.** Straightforward by Taylor expansions. The uniformity follows from the boundedness of the functions in z (and their derivatives) appearing in (i). This Lemma also holds if $o(1/\sqrt{n})$ is replace by $o(1/n)$. Notice that the linear term $z\phi(z)$ typically leads to a Bartlett type correction (see also Bickel and Ghosh(1990)).

**Proof of Theorem 1.** To show (i), assume (2.1). It suffices to transfer the Bayesian bias $-(3a_n + d_\pi)$ in (2.1) and to control all the terms in the expansion with probability one.

First we show that $a_n$ and $d_\pi$ are $O(n^{-1/2})$ with probability one. Write

$$(1/n)L^{(3)}(\theta | \mathbf{x}) = (1/n)L^{(3)}(\theta | \mathbf{x}) I_{\{\hat{\theta} \ll C\}} + (1/n)L^{(3)}(\theta | \mathbf{x}) I_{\{\hat{\theta} \gg C\}}.$$

Since under the assumptions, the MLE $\hat{\theta}$ converges to $\theta_0$, and C contains an open neighborhood which contains $\theta_0$, $I_{\{\hat{\theta} \gg C\}} = 0$ for all but finitely many n's. Therefore,

(2.3) $(1/n)L^{(3)}(\theta | \mathbf{x}) I_{\{\hat{\theta} \gg C\}} = 0$ for all but finitely many ns.

Since ç ô is a symmetric function of the data $x_1,...,x_n$, Lemma 1 applies to yield,

(2.4) $(1/n)L^{(3)}(\theta | \mathbf{x}) I_{\{\hat{\theta} \ll C\}} \rightarrow E(L^{(3)}(\theta_0 | X_1) \times I_{\{\theta_0 \ll C\}} | \theta_0) = E[L^{(3)}(\theta_0 | X_1) | \theta_0]$.

Hence (2.3) and (2.4) combine to yield that $L^{(3)}(\theta | \mathbf{x})$ is of order O(n) with probability one. The same argument entails that $L^{(2)}(\theta | \mathbf{x})$ is also of order O(n), and we conclude that $a_n$ and $d_\pi$ are $O(n^{-1/2})$ with probability one [$\pi^{(1)}(\theta)$ is continuous on C].

Second, put $\mu_n=\{\ E[\theta\,|\mathbf{x}]-\hat{\theta}\ \}\ [-L_\pi{}^{(2)}(\hat{\theta}_\pi\,|\mathbf{x})]^{1/2}=O(n^{-1/2})$ and $\varepsilon_n=0$ in Lemma 2 to obtain $\pi\,(T-E[\theta\,|\mathbf{x}]\le z|\mathbf{x})=\Phi(z)+a_n(1-z^2)\phi\,(z)+o(n^{-1/2})$; (i) follows.

To show (ii), integrating the expansion in Theorem 1(i), or applying Lemma 2.4 in Johnson (1970), results in

(2.5) $\qquad [-L^{(2)}(\theta\,|\mathbf{x})]^{1/2}\times SD(\theta\,|\mathbf{x})=1+o(n^{-1/2}).$

Lemma 2 states that replacing $[-L^{(2)}(\theta\,|\mathbf{x})^{1/2}$ by $1/SD(\theta\,|\mathbf{x})$ does not change the expansion in (i). ·

**Remark 1.** The uniform errors of the approximations are $o(n^{-1/2})$ with probability one. These error bounds can be upgraded to $O(n^{-1})$, but only in probability. This is due to the fact that the original expansion remain valid up to $O(n^{-1})$, in probability and that the rate of convergence of $(1/n)L^{(3)}(\theta\,|\mathbf{x})-E[L^{(3)}(\theta_0|X_1)|\theta_0]$ to a limiting distribution is $n^{1/2}$.

**Remark 2.** Another possible location for centering is the posterior mode which captures a part of the prior information. Assume in addition that $\pi$ has four bounded derivatives. Put $\qquad L_\pi(\theta|\mathbf{x})=L(\theta|\mathbf{x})+\log\pi(\theta),$
and denote $a_\pi$ analogously as $a_n$:

$$a_\pi=(3!)^{-1}L_\pi{}^{(3)}(\hat{\theta}_\pi\,|\mathbf{x})[-L_\pi{}^{(2)}(\hat{\theta}_\pi\,|\mathbf{x})]^{-3/2}.$$

Let $\hat{\theta}_\pi$ be a posterior mode such that $L_\pi{}^{(1)}(\hat{\theta}_\pi\,|\mathbf{x})=0$. Johnson's argument gives

(2.6)
$$\begin{aligned}\pi\{\ (\theta-\hat{\theta}_\pi)\times[-L_\pi{}^{(2)}(\hat{\theta}_\pi\,|\mathbf{x})]^{1/2}\le z|\mathbf{x}\}\\ &=\Phi(z)+a_\pi\,(1-z^2)\phi(z)-3a_\pi\,\phi(z)+o(n^{-1/2})\\ &=\Phi(z)+a_n(1-z^2)\phi(z)-3a_n\phi(z)+o(n^{-1/2}).\end{aligned}$$

Comparing with Theorem 1 (i), the "Bayesian bias" results in an extra term $-3a_n\phi(z)$ in the expansion which may explain by the fact [integrate (2.6) to see it] that

(2.7) $\qquad E[\theta\,|\mathbf{x}]=\hat{\theta}_\pi+3a_\pi\,[-L_\pi{}^{(2)}(\hat{\theta}_\pi\,|\mathbf{x})]^{-1/2}+o(n^{-3/2}).$

**Remark 3.** Because the second order terms in the posterior expansions centered only depends on the data, it is easy to invert these expansions without any further assumptions to get second order valid uniform approximation of posterior probabilities. This is in great contrast to inverting an Edgeworth expansion for a sampling distribution which typically depends on unknown parameters [See for instance Hall (1983)].

**3. Comparison of approximations**.

Let $\pi_1(z)$ and $\pi_2(z)$ be two sequences of distributions which have N(0,1) limit with the ratio of two variances tending to unity (the subscript n is suppressed). The second-order efficiency of $\pi_1$ relative to $\pi_2$ is

(3.1) $$e(\pi_1,\pi_2)=\lim_{n\to\infty} n^{1/2}\sup_z|\pi_1(z)-\Phi(z)|/\lim_{n\to\infty} n^{1/2}\sup_z|\pi_2(z)-\Phi(z)|,$$

and the criterion for performance is "$\pi_1$ beats $\pi_2$ if $e(\pi_1,\pi_2)\le 1$; $\pi_1$ ties with $\pi_2$ in case of equality." This criterion was used in Singh (1981) to compare the accuracy of normal approximation and Efron's bootstrap approximation to a sampling distribution. This criterion can be applied to compare performance of approximations to posterior distributions, and Bertail and Lo (1996) apply this criterion to compare the performance of normal and the weighted bootstrap approximations [Newton and Rafetery (1994)]. Here we consider the comparison of posterior distributions centered at the posterior mean, at the posterior mode, or at the MLE. Let $S=(\theta-E[\theta|\mathbf{x}])\times[-L^{(2)}(\theta|\mathbf{x})]^{1/2}$ (i.e., centered at the posterior mean) and $V=(\theta-\hat{\theta}_\pi)\times[-L_\pi{}^{(2)}(\theta_\pi|\mathbf{x})]^{1/2}$ (i.e., centered at the posterior mode). From the Edgeworth expansion for the distributions for $S$ (i.e., $\pi_S$) and $V$ (i.e., $\pi_V$),

(3.2) $$\lim_{n\to\infty} n^{1/2}\sup_z|\pi_S(z)-\Phi(z)|=\lim_{n\to\infty} n^{1/2}(2\pi)^{-1/2}|a_n|,$$

and $$\lim_{n\to\infty} n^{1/2}\sup_z|\pi_V(z)-\Phi(z)|=\lim_{n\to\infty} n^{1/2}(2\pi)^{-1/2}|2a_\pi|.$$

Hence, $e(\pi_S,\pi_V)=1/2$. That is, centering at the posterior mean is always better than centering at the posterior mode in terms of second-order efficiency.

The efficiency of $\pi_T$ (centered at the MLE) relative to $\pi_V$ (centered at the posterior mode) has also been investigated. In that case we have

(3.3) $$\lim_{n\to\infty}\sup_z|\pi(T\le z|\mathbf{x})-\Phi(z)|=(2\pi)^{-1/2}|2a_n|\times\max\{|1+d_\pi/(2a_n)|,\exp\{d_\pi/(2a_n)\}$$

Notice that $\lim_{n\to\infty} n^{-1}L^{(3)}(\hat{\theta}|\mathbf{x})=-I^{(1)}(\theta_0)$ (by Lemma 1), and denote

$$\rho(\theta)=\lim d_\pi/a_n=[\pi^{(1)}(\theta)/\pi(\theta)]/[-I^{(1)}(\theta)/I(\theta)]$$
$$=[(d/d\theta)\log\pi(\theta)]/[(d/d\theta)\log(I(\theta)^{-1})]$$

(3.1), (3.2), and (3.3) state that centering at the MLE beats centering at the posterior mode if $\rho(\theta_0)\ll(-4,0)$ (otherwise centering at the posterior mode is better), and centering at the MLE is better than centering at the posterior mean if $\rho(\theta_0)\ll(-3,-2\log(2))$. The following table summarizes the ranking of the three centerings.

Table 3.1

| $\rho(\theta_0) \ll$ | $(-\infty,-4)$ | $(-4,-3)$ | $(-3,-2\log(2))$ | $(-2\log(2),0)$ | $(0,\infty)$ |
|---|---|---|---|---|---|
| Posterior mean | 1 | 1 | 2 | 1 | 1 |
| Posterior mode | 2 | 3 | 3 | 3 | 2 |
| MLE | 3 | 2 | 1 | 2 | 3 |

For Jeffreys type prior i.e. $\pi(\theta) \propto I(\theta)^{-\alpha}$, and $\rho(\theta)=\alpha>0$ for all $\theta$. Centering at the posterior mean beats centering at the posterior mode, which beats centering at the MLE for all "true" parameter value $\theta_0$.

**References**


Bickel, P.J. and Ghosh, J.K. (1990). A decomposition of the likelihood Ratio Statistics and the Bartlett Correction--A bayesian argument. ***Ann. Statist.,* 18**, 1070-1090.

Blackwell, D. and Dubins, L. (1962). Merging of opinion with increasing information, ***Ann. Math. Statist.,* 33**, 882-886.

Bertail,P. and Lo, A.Y. (1996). Accurate approximate posterior inference. Preprint paper Inra-Corela, n 9607.

Blackwell, D. and Dubins, L. (1962). Merging of opinion with increasing information, ***Ann. Math. Statist.,* 33**, 882-886.

Doob, J. (l949) An application of the theory of martingales, ***Coll. Int. du CNRS,*** Paris, 22-28.

Ghosh, J. K. (1994) Higher Order Asymptotics NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 4.

Ghosh, J. K., Sinha, B. K., and Joshi, S. N. (1982) Expansions for posterior probability and integrated Bayes risk, ***Statistical Decision Theory and Related Topics*** 3 **1**, 403-456, Academic, New York.

Hall, P. (1983). Inverting an Edgeworth Expansion, ***Ann. Statist.,* 11**, 569-576.

Johnson, R. A. (1967). An asymptotic expansion for posterior distributions. ***Ann. Math. Statist.,* 38**, 1899-1906.

Johnson, R. A. (1970) Asymptotic expansions associated with posterior distributions, ***Ann. Math. Statist.,* 41**, 851-864.

Newton, M. A. and Rafetery. A. E. (1994) Approximate Bayesian inference with the weighted likelihood bootstrap (with discussions), ***JRSS B,* 56**, 3-48.

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. ***Ann. Statist.,* 9**, 1187-1195.