Statistics and Its Interface Volume 17 (2024) 9–26

Frequentist Bayesian compound inference

Ao Yuan* and Jinfeng Xu

In practice often either the Bayesian or frequentist method is used, although there are some combined uses of the two methods, a formal unified methodology of the two hasn't been seen. Here we first give a brief review of the two methods and some combination of the two, then propose a procedure using both the frequentist likelihood and the Bayesian posterior loss in parameter estimation and hypothesis testing, as an attempt to unify the two methods. Basic properties of the proposed method are studied, and simulation studies are carried out to evaluate the performance of the method.

MSC 2010 subject classifications: Primary 62F99; secondary 62F10.

Keywords and phrases: Bayesian estimate, Compound inference, High order behavior, Maximum likelihood estimate.

1. INTRODUCTION

The frequentist and the Bayesian methods are the two dominant schools for inference. The two schools mainly favour their own method, and the practitioners often have to choose one of the methods and ignore the other. Each method has its own advantages and dis-advantages. The two methods share some common basic asymptotic properties, which have been studied extensively. The Bernstein-von Mises theorem, for example in Prakasa Rao (1987) and LeCam and Yang (1990), states that under general conditions the Bayes and frequentist inferences are equivalent: the two estimators are asymptotically first order equivalent, and the posterior distribution around its mean is close to the distribution of the maximum likelihood estimate (MLE) around the true parameter – both are efficient, asymptotically normal with mean zero and the inverse Fisher information as the asymptotic covariance matrix, and so it is impossible to improve the two types of estimators in their asymptotic behaviour. However, the two methods are different in second and higher order terms (Gusev, 1975), or in many other aspects, such as finite sample properties, each of them has its own advantage(s) in some situations (Han et al. 2022). For example, if $x_1, ..., x_n$ iid $Bernoulli(\theta)$, the frequentist MLE of the true θ_0 is $\hat{\theta} = \bar{x}$, the data mean. If the sample size n is small, it may happen that $\bar{x} \approx 0$ (or 1), which may not be reasonable. For a Bayes estimator, if we set a prior $\theta \sim U[0, 1]$, then under the squared error loss, the Bayes estimator of θ_0 is $\check{\theta} = (\bar{x} + 1/n)/(1 + 2/n)$, which is a bit further away from 0 or 1, and is more reasonable. On the other hand, if a bad prior is used, the Bayes estimator will be worse than the MLE. Thus in application they have received different appreciations for various reasons. The debate over which method should be preferred has a long history and will continue. However, at the methodological level, both schools appreciate the advantages of each other, and there are some efforts to advocate compromises or combined uses of the two methods. The Bayesian has appreciated steady growth partially due to the development of computation facilities, but in practice the main stream statistical tool is still frequentist. Efron (2005) summarized the main reasons for this as the ease of use, modelling and objectivity. Lindley (1990) gave a broad review of the present position of Bayesian statistics. Below we first give a brief review of each method, and their combined method, then introduce our proposed compound model as a new attempt to unify the two methods.

*Corresponding author.

The frequentist method. The frequentist method is simpler in model formulation and computation, and is the dominating tool in statistics. The commonly used procedures include the parametric, semiparametric and nonparametric MLE, the M-estimator, Z-estimator, order statistic, rank statistic, U-statistic, V-statistic, and various forms of test statistics, such as T-test and chi-squared test statistic. They are convenient to use, and their asymptotic study is relatively easy, often via the empirical process theory. The frequentist's theory of weak convergence on functional space is a powerful tool, which is rarely seen in the Bayesian framework. Advantages including simple in model formulation, objectivity, easier of theoretic stusy, not requiring specification of prior distribution and loss function; and the computation is straight forward and relatively simple. Frequentist measures like p-values and confidence intervals continue to dominate statistical research. For nonparametric/semiparametric models, the frequentist has aboundant methods, such as kernal smoothing, spline, nonparametric maximum likelihood estimate, etc. In this case, the Bernstein-von Mises theorem dose not generally hold, and then the Bayesian method may be inferior to the frequentist's counter part. In hypothesis test, the frequentist developed optimality criteria, such as the most powerful test, and asymptotic efficiency criteria, such as the Pitman, Chernoff, Bahadur and Hodges-Lehmann efficiencies.

The Bayesian method. The Bayesian can incorporate prior knowledge about parameters into the model, and has a broader theoretical framework using the statistic decision theory. Wald's complete class theorem (Wald, 1950. Theorem 2.2, p.183) states that under some general conditions, the set of all Bayesian procedures form a complete class: any admissible procedure to a decision problem can be formulated as a Bayesian procedure, or the limit of a sequence of Bayesian procedures. For example, the frequentist maximum likelihood estimate (MLE) can be formulated as a special Bayesian estimate under a 0-1 loss and uniform prior. Also, the Bayesian model gives uncertainty interpretation of the parameters via the posterior distribution. However, the subjective specification of the prior is often a source of debate, it may vary from one investigator to another. The computation is more difficult than that of the frequentist method. Often the Markov chain Monte Carlo method is need for its computation, and the method itself requires diagnostic tool for convergence of the chain. Asymptotic study of Bayesian procedure is more difficult than the frequentist counter part. In the classic Bayes framework, the prior is asymptotically negligible. Yuan and De Gooijer (2014) considered asymptotically informative prior, whose effect is asympotically non-negligible. For nonparametric Bayesian, the main tool is the Dirichlet process prior, which has limitation in application (Ferguson, 1973). Recently, many other nonparametric Bayesian methods have been proposed for broader applications, but not very convenient to use. For hypothesis test, the Bayesian main tool is the Bayes Factor, decision is based on some subjectively determined values.

Combination of the two methods. It is reasonable to expect that a unified framework of the two methods will give us more flexibility to achieve any given optimality criterion, and provide a comprehensive few in the decision making. There are some combined uses of Bayesian and frequentist methods (Bayarri and Berger, 2004), in which, the authors hope eventually leads to a general methodological unification of the two methods. There are also compromises between the two methods (Good, 1992). A prominent example is the empirical Bayes approach introduced by Robbins (1955), in which some of the parameters in the prior is estimated by the frequentist MLE.

Bayesian frequentist hybrid method. In practice, there are situations in which one of the methods is more favourable than the other by some criteria. Thus in inferring multiparameters, it may happen that on part of the parameters, the frequentist method is preferable, while on the other part a Bayesian procedure is more appropriate. Yuan (2009) proposed a hybrid procedure for Bayesian and frequentist methods, it perform the MLE on part of the parameters and Bayesian procedure on the other part parameters. As in practice, sometimes we have good prior information on part of the parameters, the sound prior knowledge prefers a Bayesian on this subset of parameters, while for the other part of the parameters there is no reliable prior information, even a non-informative prior on this part of parameter is not desirable, and the MLE is favored on this part of parameters. This motivates a joint operation of the two methods on different parts of the parameters in the same model. Such hybrid inference will give us more flexibility than using either methods alone in achieving overall advantage. In this paper we propose a hybrid estimator and study its consistency and asymptotic high order behavior, and we illustrate its application. Also, using the high-order expansions, we considered a new type of second-order matching prior in the objective Bayes context.

The proposed method. Our method here is not such combination nor compromise, not the quasi-Bayesian (which applies Bayesian procedure on quasi-likelihood), not the pseudo-Bayesian (which applies Bayesian procedure on pseudo-likelihood or pseudo-prior), nor the empirical Bayes (which incorporates uncertainty into a frequentist procedure via Bayesian method), nor the Bayesian frequentist hybrid method in the literature. The proposed method here is different from the above, we use both frequentist and Bayesian procedure on the same parameters. Our method optimize both the frequentist criterion of likelihood and the Bayesian criterion of posterior risk simultaneously, and in some sense can be viewed as a unified framework of the two methods. We show that the new estimator is asymptotically equivalent to the MLE, the Bayesian estimator and the hybrid estimator in the first order, also it can optimize a second order criterion, which cannot be achieved by using either the MLE or the Bayes estimator. For hypothesis testing, the frequentist method is based on the likelihood, while in many cases it is desirable that the decision rule be based on both the likelihood and the cost of the problem – the loss function. The proposed method take these factors into consideration simultaneously. We use the word "compound" to distinguish the proposed method from the linear combination of MLE and Bayes estimator, or any other existing methods combining both the frequentist and Bayesian elements.

In Section 2, we describe the proposed method and in Section 3 study its asymptotic behavior. We show that the compound estimator, Bayes estimator, MLE and the hybrid estimator are all first-order equivalent, asymptotic normal and efficient. In Section 4, we study Edgeworth expansions for the density and distribution function of the compound estimator after standardization. In Section 5 we show the compound estimator can achieve a second order optimality criterion. In Section 6 we propose the compound hypothesis testing, and Section 7 give some examples for illustration of the proposed methods. A short Concluding Remarks is given at the end, account for some issues of the method and future research topics. The used regularity conditions and technical proofs of theorems are given in the Appendix.

2. THE PROPOSED METHOD

Let $\mathbf{x}^n = (\mathbf{x}_1, ..., \mathbf{x}_n)$ be a set of i.i.d. observations from $f(\cdot|\theta)$, a given density function of the data distribution and $\theta \in \Theta \subset R^d$ be the parameter of interest, $\pi(\theta)$ be the prior density for θ, and $\pi(\theta|\mathbf{x}^n) = f(\mathbf{x}^n|\theta)\pi(\theta)/m(\mathbf{x}^n)$ the posterior density of θ given the observed data, where $m(x^n) = \int f(\mathbf{x}^n|\theta)\pi(\theta)d\theta$ is the marginal density for X^n. Let $\mathcal{D}$ be the decision space for inferring θ ($\mathcal{D} = \Theta$ for estimation of θ_0, the true parameter value for the observed data), $\mathbf{d}(\mathbf{x}^n) \in \mathcal{D}$ a decision rule, $W(\mathbf{d}(\mathbf{x}^n), \theta)$ the loss function, $R(\mathbf{d}, \theta) = E_\theta[W(\mathbf{d}(\mathbf{x}^n), \theta)]$ the risk of $d(\cdot)$ at θ, $R(\mathbf{d}) = \int R(\mathbf{d}, \theta)\pi(\theta)d\theta$ the Bayes risk of $d(\cdot)$ and

$$R(\mathbf{d}|\mathbf{x}^n) = \int W(\mathbf{d}(\mathbf{x}^n), \theta)\pi(\theta|\mathbf{x}^n)d\theta$$

the posterior risk of $d(\cdot)$ for inferring θ_0. The Bayes decision for θ based on the observed data $\mathbf{x}^n$ is $\check{\theta}_n = \arg\inf_{d\in\Theta} R(\mathbf{d})$, and from Bayes inference theory

$$\check{\theta}_n \stackrel{a.s.}{=} \arg\inf_{d\in\Theta} R(\mathbf{d}|\mathbf{x}^n) = \arg\inf_{\theta\in\Theta} \int W(\theta, \alpha)\pi(\alpha|\mathbf{x}^n)d\alpha.$$

The right-hand side above is called the generalized Bayesian estimator of θ_0.

In contrast, let $l(\theta|\mathbf{x}^n) = \log f(\mathbf{x}^n|\theta)$, the frequentist maximum likelihood estimate (MLE) of θ_0 is

$$\hat{\theta}_n = \arg\sup_{\theta\in\Theta} f(\mathbf{x}^n|\theta) = \arg\sup_{\theta\in\Theta} l(\theta|\mathbf{x}^n).$$

In this compound inference, we infer θ_0 by optimizing both the criteria of the MLE and the generalized Bayesian rule in the same framework. However, $f(\mathbf{x}^n|\theta)$ (or $l(\theta|\mathbf{x}^n)$) and $R(\mathbf{d}|\mathbf{x}^n)$ have different scales, when combining the two parts in the inference, we need to adjust them to the same scale, otherwise one part will be dominant. If we expand $\exp\{\pi(\theta|\mathbf{x}^n)\}$ by Laplace method, the leading term is the density of $N(\hat{\theta}, I^{-1}(\hat{\theta})/n)$, with $I(\theta)$ the Fisher information evaluated at θ, and $n\int W(\theta, \alpha)\pi(\alpha|\mathbf{x}^n)d\alpha = O_p(n)$, which has the same scale as $l(\theta|\mathbf{x}^n)$. This motivates our estimate θ_0 by

$$\theta_n = \arg\sup_{\theta\in\Theta} \Big(f^{1-c}(\mathbf{x}^n|\theta) \exp\big\{-cn\int W(\theta, \alpha)\pi(\alpha|\mathbf{x}^n)d\alpha\big\}\Big)$$

$$(1) \quad = \arg\sup_{\theta\in\Theta} \Big((1-c)l(\theta|\mathbf{x}^n) - cn\int W(\theta, \alpha)\pi(\alpha|\mathbf{x}^n)d\alpha\Big),$$

where $0 \le c \le 1$ is a constant to be determined. It is seen that θ_n optimises the joint criterion of the likelihood and the posterior risk, it is the MLE when $c = 0$, and the Bayesian estimate when $c = 1$. Initially, we just set $c = 1/2$, latter on we realize that put a c in the construction will give us more flexibility to make the compound estimator $\theta_n = \theta_n(c)$ to satisfy some specified optimality criterion. We will see in Section 3 how to choose the constant c for such purpose.

Let $G(\theta)$ be the function inside the outermost bracket of (1), and $G^{(1)}(\theta)$ be its derivative. θ_n can be found as the solution of the 'normal' equation $G^{(1)}(\theta) = \mathbf{0}$. Thus θ_n exists under similar conditions for the MLE and Bayesian estimates.

Let $\hat{\theta}_n$ be the MLE of the true parameter θ_0, and $\check{\theta}_n$ the Bayes estimator. One may wondering why not just simply set the linear combination $\bar{\theta}_n = (1-c)\hat{\theta}_n + c\check{\theta}_n$ as the desired estimator instead of (1). In fact, as $G(\bar{\theta}_n) \le G(\theta_n)$, and generally "=" does not hold, so $\bar{\theta}_n$ is not optimal by the criterion given in (1), nor are the MLE and Bayesian estimator. Also, we'll see in Section 3 that we can choose c so that the compound estimator θ_n achieves a second order optimal criterion, but $\bar{\theta}_n$ cannot.

3. ASYMPTOTIC BEHAVIOR

In this section we study the consistency and high order asymptotic behavior of the compound estimator.

Consistency of the estimator. The study of the consistencies of Bayes estimates, MLE and their relationships has a relatively long history (Bernstein, 1917; Doob, 1949; LeCam, 1953; Strasser, 1981; Wald, 1949; among others). Doob (1949) established strong consistency of Bayes estimators under very general conditions, and there is some speculation that conditions for Bayesian consistency might be found which are weaker than those for the MLE. Under some basic assumptions, Strasser (1981) showed that any conditions for the convergence (a.s.) of MLE assert the concentration (a.s.) of the posterior distribution to the true parametric value. This does not directly imply that conditions for Bayesian consistency are weaker since posterior concentration to the true parameter is not equivalent to the consistency of Bayes estimate. The latter also depends on the loss. Generally, the loss $W(\mathbf{d}, \alpha)$ has the form $W(||\mathbf{d} - \alpha||) = W(\mathbf{d} - \alpha)$. To avoid confusion, we will use W for any of these functional forms.

Theorem 1. *Assume conditions (A1)-(A9) in the Appendix, that $W(\cdot)$ satisfies $W(\mathbf{0}) = 0$, is strictly increasing and continuous in a neighbourhood of $\mathbf{0}$. Then θ_n exists, and as $n \to \infty$ we have*

$$\theta_n \to \theta_0, \quad (a.s.).$$

High-order asymptotic behavior. High-order asymptotic expansions are used to assess estimators when they have similar lower-order behavior. In Linnik and Mitrofanova (1965), Johnson (1970), Chibisov (1973) and Gusev (1975), among others) such expansions of Bayes estimate and MLE were obtained, so were their densities and related quantities in the one-dimensional case. Below we derive high order asymptotic expansion for the compound estimator.

We introduce the following notations. For an integer vector $\mathbf{i} = (i_1, ..., i_d)$ with $i_j \ge 0$ $(j = 1, ..., d)$, denote

$|\mathbf{i}| = \sum_{j=1}^d i_j$, and for any $g(\cdot) \geq 0$, define $\log g(\cdot) = 0$ if $g(\cdot) = 0$. Denote $\mathbf{1} = (1, ..., 1)'$ of length d, $\mathbf{0} = (0, ..., 0)'$ of length d. Let $l(\mathbf{x}|\theta) = \log f(\mathbf{x}|\theta)$ and the score at θ as

$$L(\mathbf{x}|\theta) = (L_1(x|\theta), ..., L_d(x|\theta))' :=$$

$$(\partial l(x|\theta)/\partial\theta_1, ..., \partial l(x|\theta)/\partial\theta_d)', \quad E_{\mathbf{i}}(\theta) = E_\theta L_{\mathbf{i}}(\mathbf{X}|\theta),$$

$$L_{\mathbf{i}}(\mathbf{x}|\theta) = \Big(\frac{\partial^{|\mathbf{i}|}}{\partial\theta_1^{i_1}\cdots\partial\theta_d^{i_d}} L_1(\mathbf{x}|\theta), ..., \frac{\partial^{|\mathbf{i}|}}{\partial\theta_1^{i_1}\cdots\partial\theta_d^{i_d}} L_d(\mathbf{x}|\theta)\Big)',$$

$$S_{\mathbf{i}}(\theta) = \frac{1}{\sqrt{n}}\sum_{j=1}^n L_{\mathbf{i}}(\mathbf{x}_j|\theta), \Delta_{\mathbf{i}}(\theta) = \frac{1}{\sqrt{n}}\sum_{j=1}^n (L_{\mathbf{i}}(\mathbf{x}_j|\theta) - E_{\mathbf{i}}(\theta)),$$

$$\rho_0(\theta) := (\rho_1, ..., \rho_d)' = \Big(\frac{\partial}{\partial\theta_1}\log\pi(\theta), ..., \frac{\partial}{\partial\theta_d}\log\pi(\theta)\Big)',$$

$$\rho_{\mathbf{i}} = (\frac{\partial^{|\mathbf{i}|}}{\partial\theta_1^{i_1}\cdots\partial\theta_d^{i_d}}\rho_1, ..., \frac{\partial^{|\mathbf{i}|}}{\partial\theta_1^{i_1}\partial\theta_d^{i_d}}\rho_d)'$$

and set $S_{\mathbf{i}} = S_{\mathbf{i}}(\theta_0)$, $\Delta_{\mathbf{i}} = \Delta_{\mathbf{i}}(\theta_0)$, $E_{\mathbf{i}} = E_{\mathbf{i}}(\theta_0)$, and $\rho_{\mathbf{i}} = \rho_{\mathbf{i}}(\theta_0)$.

For vector $\mathbf{H} = (H_1, ..., H_d)'$ and integer vector $\mathbf{i} = (i_1, ..., i_d)$, define $\mathbf{H}^{\mathbf{i}} = (H_1^{i_1}, ..., H_d^{i_d})'$, $\langle\mathbf{H}^{\mathbf{i}}\rangle = \prod_{j=1}^d H_j^{i_j}$, $\mathbf{i}! = \prod_{j=1}^d i_j!$, and $[\mathbf{i}]! = (i_1!, ..., i_d!)'$. For $\mathbf{a} = (a_1, ..., a_d)'$ and $\mathbf{b} = (b_1, ..., b_d)'$, define $\mathbf{a}+\mathbf{b} = (a_1+b_1, ..., a_d+b_d)'$, $\mathbf{ab} = (a_1b_1, ..., a_db_d)'$, $\mathbf{a}/\mathbf{b} = (a_1/b_1, ..., a_d/b_d)$ and $\langle\mathbf{ab}\rangle = \prod_{i=1}^d a_ib_i$. Denote $\mathbf{e}_j = (0, .., 0, 1, 0, ..., 0)'$, the d-vector with jth element be 1 and the others be zeros.

As in Gusev (1975), for two real or random vector functions $\xi_1 = \xi_1(n, \delta, \theta_0, y)$ and $\xi_2 = \xi_2(n, \delta, \theta_0, y)$, we define $\xi_1 \overset{k}{\sim} \xi_2$, if there exist finite positive constants C_{1k}, C_{2k} and C such that for any compact $K \subset \Theta$ and for any $0 < \delta < C_{1k}$,

$$\sup_{\theta_0\in K}\sup_y ||\xi_1(n, \delta, \theta_0, y) - \xi_2(n, \delta, \theta_0, y)|| \leq Cn^{-k/2}n^{\delta C_{2k}}$$

for all $n > C$, and we write $\xi_1 - \xi_2 = O_p(n^{-k/2})$.

To simplify notation, denote $\mathbf{I} = \mathbf{I}(\theta_0)$ the Fisher information, and $\mathbf{I}^{-1}$ its inverse.

Theorem 2. *Under conditions (B1)-(B10) in the Appendix, we have*

$$\sqrt{n}(\theta_n - \theta_0) = \sum_{r=0}^{k-1} n^{-r/2}\mathbf{B}_r + O_p(n^{-k/2}),$$

where the $\mathbf{B}_r$'s are d-vectors given at the end of the proof, they are polynomials in the $\Delta_{\mathbf{i}}$'s of degree $r+1$, their coefficients are polynomials in the $\mathbf{E}_{\mathbf{i}}$'s, $|\mathbf{i}| = 2, ..., r+1$, in $\mathbf{I}^{-1}$ and the $\rho_{\mathbf{i}}$'s; the $\mathbf{B}_r$'s are given by the recursive formula $(0 \leq r \leq k-1)$, with $\mathbf{D}$ given in condition (B9), which will involve $\langle\mu\rangle_{\mathbf{i}} = E[\langle\theta^{\mathbf{i}}\rangle]$ and $\mu_{\mathbf{1},\mathbf{j}} = E[\theta\langle\theta^{\mathbf{j}}\rangle]$, with $\theta \sim N(\mathbf{0}, \mathbf{I}^{-1})$.

Remark 1. In Theorem 2, the expressions of the $\mathbf{B}_r$'s are involved and given at the end of the proof. They are computed recursively. i.e., first set $\mathbf{B}_0 = \mathbf{I}^{-1}\Delta_0$, then $\mathbf{B}_1$ can be computed; given $\mathbf{B}_0$ and $\mathbf{B}_1$, $\mathbf{B}_2$ can be obtained;..., given $\mathbf{B}_0, ...\mathbf{B}_r$. $\mathbf{B}_{r+1}$ can be obtained. Also, in the expression for $\mathbf{B}_r$, we note the fact that $\langle\mu\rangle_{\mathbf{j}} = 0$ for $|\mathbf{j}|$ odd, and $\mu_{\mathbf{1},\mathbf{j}} = 0$ for $|\mathbf{j}|$ even. This will simplify the evaluations.

Remark 2. By the results in Yuan (2009) and the above Theorem, the MLE, Bayes estimator, hybrid estimator and the compound estimator all have the same first order term $\mathbf{I}^{-1}\Delta_0 \overset{D}{\to} N(\mathbf{0}, \mathbf{I}^{-1})$, thus all these estimators are first order equivalent and efficient.

In Theorem 2, if we set $c = 0$, then we get the asymptotic expansion of the MLE, and the $\mathbf{B}_r$'s are exactly the same as the $\mathbf{H}_r^o$'s for MLE given in Yuan (2009). If we set $c = 1$, then we get the asymptotic expansion of the Bayes estimator, in this case the $\mathbf{H}_r$'s should be the same as the $\mathbf{G}_r$'s for Bayes estimator given in Yuan (2009), though they have seemingly different expressions. The general expressions for the $\mathbf{B}_r$'s are complicated, but in application the second term $\mathbf{B}_1$ is of more interest than the higher order terms, which has a much simpler expression as given by the following

Corollary 1. *The second order term $\mathbf{B}_1$ in the expansion of $\sqrt{n}(\theta_n - \theta_0)$ is*

$$\mathbf{B}_1 = \mathbf{I}^{-1}\Big(\sum_{|\mathbf{i}|=2} E_{\mathbf{i}}\frac{\langle\mathbf{B}_0^{\mathbf{i}}\rangle}{\mathbf{i}!} + \sum_{|\mathbf{i}|=1}\Delta_{\mathbf{i}}\langle\mathbf{B}_0^{\mathbf{i}}\rangle\Big)$$

$$+[(1-c)\mathbf{I} + 2c\mathbf{D}]^{-1}2c\mathbf{D}\mathbf{I}^{-1}\rho_0,$$

where $\mathbf{B}_0 = \mathbf{I}^{-1}\Delta_0$ and $\mathbf{D}$ is given in condition (B9).

Note. For the compound estimator, its second order term in the expansion is not a linear combination of those of the MLE and of the Bayesian estimator. i.e., $\mathbf{B}_1$ is not of the form $a(c)\mathbf{H}_1^o + (1-a(c))\mathbf{G}_1$ for some $0 \leq a(c) \leq 1$, where $\mathbf{H}_1^o$ and $\mathbf{G}_1$ are the second order term in the expansions of the MLE and of the Bayesian estimator respectively, as given in Yuan (2009). In particular, $\mathbf{H}_1^0$ is the second term in the expansion of Theorem 2 corresponds to $c = 0$ (MLE), and $\mathbf{G}_1$ is that term corresponds to the case $c = 1$ (Bayes estimator). Their expressions are somewhat involved and not show here.

4. EDGEWORTH EXPANSION

Edgeworth expansion (EE) has been extensively studied in the literature. EEs for density and distribution functions of normalized i.i.d. random vector summation can be found in Bhattacharya and Rao (1986).

Edgeworth expansion for the density $f_n(\cdot)$ and distribution function $F_n(\cdot)$ of the sample mean or smooth function of the sample mean often have the form,

$$f_n(x) = \phi_\Lambda(x) + \sum_{j=1}^{k} n^{-j/2} p_j(x)\phi_\Lambda(x) + O_p(n^{-(k+1)})R(x),$$

where $p_j(x)$ is a homogeneous polynomial in x of degree $3j$, it is an odd or even function depends on j is an odd or even number.; Λ is the covariance matrix of the sample, $\phi_\Lambda(\cdot)$ and $\Phi_\Lambda(\cdot)$ are the density function and distribution function of the normal vector $N(\mathbf{0}, \Lambda)$; and

$$F_n(x) = \Phi_\Lambda(x) + \sum_{j=1}^{k} n^{-j/2} q_j(x)\phi_\Lambda(x) + O_p(n^{-(k+1)}),$$

where $q_j(x)$ is a polynomial in x of degree no more than $3j-1$, given by the relationship $q_j(x)\phi_\Lambda(x) = \int^x p_j(y)\phi_\Lambda(y)dy$ (or equivalently $p_j(x)\phi_\Lambda(x) = \frac{d}{dx}[q_j(x)\phi_\Lambda(x)]$); it is an odd or even function depends on j is an even or odd number.

Hall (1992) contains EE for statistics of the form $H(S_n)$. Bickel (1974) surveyed works of EE on some asymptotic normal nonparametric statistics. EEs of distribution functions of the maximum likelihood estimate (MLE) of a location parameter were derived by Linnik and Mitrofanova (1963, 1965). Mitrofanova (1967) obtained EEs of distribution functions of the MLE in general. Chibisov (1972, 1973) obtained EEs of distribution functions of statistics with the form $\sum_{i=0}^{k} n^{-j/2} H_j(S_n)$ and $S_n = \sum_{i=1}^{n} X_i$. Gusev (1975, 1976) obtained EEs for the MLE, Bayes estimator and the moments. Lahiri (2010) derived EEs for studentized statistics with mixing weak dependent observations without plug-in estimators of nuisance parameters.

In this section we derive the k-th order EE for the distribution function $F_n(\mathbf{x}) = P_{\theta_0}\big(\sqrt{n}(\theta_n - \theta_0) \leq \mathbf{x}\big)$ and its density function $f_n(\mathbf{x})$, $\mathbf{x} \in R^d$. Let $\mathbf{V} = (\Delta_\mathbf{i} : |\mathbf{i}| = 0, 1, ..., k-1)$, and $\Omega = Cov_{\theta_0}(\mathbf{V})$. $\mathbf{V}$ has dimension $d_k = k$ when $d = 1$, and $d_k = d\sum_{r=0}^{k-1} d^r = d(d^k - 1)/(d-1)$ when $d > 1$.

Theorem 3. *Assume conditions of Theorem 2, that* $\mathbf{V}$ *has finite* $(k+1)$*-th moments and an integrable characteristic function. Then, uniformly over* $\mathbf{x}$*,*

$$f_n(\mathbf{x}) = \phi_{\mathbf{I}^{-1}}(\mathbf{x}) + \sum_{r=1}^{k-1} n^{-r/2} g_r(\mathbf{x})\phi_{\mathbf{I}^{-1}}(\mathbf{x}) + O(n^{-k/2}),$$

$$F_n(\mathbf{x}) = \Phi_{\mathbf{I}^{-1}}(\mathbf{x}) + \sum_{r=1}^{k-1} n^{-r/2} G_r(\mathbf{x})\phi_{\mathbf{I}^{-1}}(\mathbf{x}) + O(n^{-k/2}),$$

where for each r*,* $g_r(\mathbf{x}) = E[T_r(\mathbf{IX}, \mathbf{Y})|X = x]$*, and* $T_r(\mathbf{IX}, \mathbf{Y})$ *is a polynomial in* (X, Y) *(its construction is given in the proof in the Appendix); the expectation is with respect to* $\mathbf{Y} \sim N(\mathbf{0}, \Omega_2)$*,* Ω_2 *is the sub-block matrix in* Ω *corresponding to* $(\Delta_\mathbf{i} : |\mathbf{i}| = 1, ..., k-1)$*;* $G_r(\cdot)$ *is given by the relationship* $G_r(\mathbf{x})\phi_{\mathbf{I}^{-1}}(\mathbf{x}) = \int^{\mathbf{x}} g_r(\mathbf{y})\phi_{\mathbf{I}^{-1}}(\mathbf{y})d\mathbf{y}$.

Note. In Theorem 3, the g_r's and G_r's are computed from $\Omega = \Omega_k$ of dimension d_k, so we may denote them as $g_{k,r}$'s and $G_{k,r}$'s. A natural question is: are they really depend on k? The answer is Yes, g_r's computed using Ω_k are correct up to order $O(n^{k/2})$, while those computed using Ω_l are correct up to order $O(n^{-l/2})$.

5. CHOICE OF C

The constant c in $\theta_n(c)$ allow us to control the behavior of the compound estimator to satisfy some specified optimality criterion. Since the MLE, Bayes estimator, hybrid estimator and the compound estimator are first order equivalent, regardless of c in $\theta_n(c)$, a natural consideration is to minimize the second order instability of $\theta_n(c)$. Although $E(\mathbf{B}_0) = \mathbf{0}$, where the expectation is obtained under θ_0, the second order term $\mathbf{B}_1$ has non-zero expectation, and a way to measure the amount of stability is its expectation. We want it to be as small as possible in absolute value componentwise. As the exact expectation is generally not easy to compute, we compute that for the weak limit $\tilde{\mathbf{B}}_1(c)$ of $\mathbf{B}_1(c)$. Note the weak limits of the $\Delta_\mathbf{i}$'s and the $\delta_\mathbf{i}$'s are normal, so the weak limit $\tilde{\mathbf{B}}_1(c)$ of $\mathbf{B}_1(c)$ is easily obtained.

To compute $E[\tilde{\mathbf{B}}_1(c)]$, let 1_j be the column d-vector with j-th component be 1 and others be zero, $\mathbf{A}_j = E_{\theta_0}[L_j(\mathbf{X}|\theta_0)\mathbf{L}(\mathbf{X}|\theta_0]$, ${}_j\mathbf{I}^{-1}$ and $\mathbf{I}_j^{-1}$ be the j-th row and j-th column of $\mathbf{I}^{-1}$. Proposition 1 below gives a clear expression of $E[\tilde{\mathbf{B}}_1(c)]$ in terms of c, so that we can optimize $g(c) = ||E[\tilde{\mathbf{B}}_1(c)]||$ over c.

Proposition 1. *We have*

$$E[\tilde{\mathbf{B}}_1(c)] = [(1-c)\mathbf{I} + 2c\mathbf{D}]^{-1}\Big([(1-c)I + 2c\mathbf{D}\mathbf{I}^{-1}]\big[\sum_{j=1}^{d}\mathbf{A}_j\mathbf{I}_j^{-1}$$

$$+ \sum_{i,j=1}^{d}\mathbf{E}_{1_i+1_j}\ {}_i\mathbf{I}^{-1}\mathbf{I}^{-1}\mathbf{I}_j^{-1}/(1_i+1_j)!\big] + 2c\mathbf{D}\mathbf{I}^{-1}\rho_\mathbf{0}\Big).$$

Re-write $E[\tilde{\mathbf{B}}_1(c)]$ as

$$E[\tilde{\mathbf{B}}_1(c)] = [(1-c)\mathbf{I} + 2c\mathbf{D}]^{-1}\Big((1-c)\big[\sum_{j=1}^{d}\mathbf{A}_j\mathbf{I}_j^{-1}$$

$$+ \sum_{i,j=1}^{d}\mathbf{E}_{1_i+1_j}\ {}_i\mathbf{I}^{-1}\mathbf{I}^{-1}\mathbf{I}_j^{-1}/(1_i+1_j)!\big] + 2c\Big(\mathbf{D}\mathbf{I}^{-1}\big[\sum_{j=1}^{d}\mathbf{A}_j\mathbf{I}_j^{-1}$$

$$+ \sum_{i,j=1}^{d}\mathbf{E}_{1_i+1_j}\ {}_i\mathbf{I}^{-1}\mathbf{I}^{-1}\mathbf{I}_j^{-1}/(1_i+1_j)!\big] + \mathbf{D}\mathbf{I}^{-1}\rho_\mathbf{0}\Big)\Big)$$

$$:= [(1-c)\mathbf{I} + 2c\mathbf{D}]^{-1}\Big((1-c)\mathbf{a} + 2c\mathbf{b}\Big).$$

Note that $E[\tilde{\mathbf{B}}_1(c)]$ is a vector, $\mathbf{a}$ is from the likelihood component, and $\mathbf{b}$ is from the Bayesian component. Initially we define the asymptotic instability of $\mathbf{B}_1(c)$ as

$$g(c) = ||E[\tilde{\mathbf{B}}_1(c)]|| = ||\big((1-c)\mathbf{I} + 2c\mathbf{D}\big)^{-1}\big((1-c)\mathbf{a} + 2c\mathbf{b}\big)||.$$

We are to select c to minimize $g(c)$. Since $g(c)$ is a single variable function, $\arg\min_c g(c)$ can be easily found by plotting $g(c)$ on $[0,1]$.

After investigating the function $g(c)$ for a variety of choices of ($\mathbf{a}$ and $\mathbf{b}$, we found $g(c)$ is nearly linear in c, and will end up $arg\min_c g(c)$ to be 0 or 1. So we redefine the function $g(\cdot)$ and the instability as the following convex function in c,

$$g(c) = (1-c)^2||\mathbf{a}|| + c^2||\mathbf{b}||$$

and choose c to minimize the above $g(\cdot)$ on [0,1], and it gives $c = ||a||/(||a|| + ||b||)$.

For the linear combination estimator $\bar{\theta}_n = (1-c)\hat{\theta}_n + c\check{\theta}_n$, its second order term is linear in c, and linear in c for the above criterion, its minimal value is attained at either $c = 0$ or $c = 1$, with the corresponding estimator being the MLE or the Bayesian. So it cannot achieve the above second order optimal criterion, nor is any other combination estimators such as $\bar{\theta}_n = \frac{1}{1+c}\hat{\theta}_n + \frac{c}{1+c}\check{\theta}_n$ with $c \in [0,\infty)$, as c with $c \in [0,1]$ and $1/(1+c)$ with $c \in [0,\infty)$ has a 1-1 correspondence; nor any estimator of the form $a(c)\hat{\theta}_n + (1-a(c))\check{\theta}_n$ with $0 \le a(c) \le 1$.

6. HYPOTHESIS TESTING

Consider testing the hypothesis $H_0 : \theta \in \Theta_0 = \{\theta_j = \theta_{j,0}, j = 1, ..., d-r\}$ $(1 \le r < d)$. A commonly used frequentist procedure is to use the likelihood ratio test statistic

$$\Lambda_n = \frac{\sup_{\theta\in\Theta_0} f(|\mathbf{x}^n|\theta)}{\sup_{\theta\in\Theta} f(\mathbf{x}^n|\theta)}, \quad \text{or} \quad \lambda_n = -2\log\Lambda_n.$$

Under H_0, asymptotically and suitable conditions, one has (Wilks,1938; Serfling 1980)

$$\lambda_n \xrightarrow{D} \chi^2_r.$$

However this test procedure does not use the loss function, which in some cases will be desirable to be a factor in the decision making. Recall the definition of $G(\theta) = G(\theta|\mathbf{x}^n)$ in Section 1. Now we consider compound test procedure with both the likelihood and Bayesian criteria, of the form

$$\lambda_{c,n} = 2\bigg(\sup_{\theta\in\Theta} G(\theta|\mathbf{x}^n) - \sup_{\theta\in\Theta_0} G(\theta|\mathbf{x}^n)\bigg)$$

One important part of a test statistic is its power at the alternative, or the asymptotic distribution of the test statistic under the alternative. Often the alternative can be some fixed point (or region) of θ outside the null, or it can be locally around θ_0 with a shrinking rate of $n^{-1/2}$. The fixed alternative has the problem that the power always tends to 1 as the sample size n increases without bound, and subject to the criticizing that a test is not the testing procedure itself, rather its the game of sample size. So we consider the local alternative $\mathbf{H}_{0,n} : \theta = \theta_0 + n^{-1/2}\delta$, for some fixed δ.

Let $\mathbf{I}_0$ be the $r \times r$ Fisher information under H_0 (its the lower right $r \times r$ block in $\mathbf{I}$), and $\tilde{\mathbf{I}}$ be the $d \times d$ matrix with the lower right $r \times r$ block be $\mathbf{I}_0^{-1}$ and other elements be 0's.

Theorem 4. *Assume conditions of Theorem 2 with $k = 1$. Then*

$$\text{Under} \quad H_0, \qquad \lambda_{c,n} \xrightarrow{D} \sum_{j=1}^{r} \gamma_j \chi^2_j;$$

$$\text{Under} \quad H_{0,n}, \qquad \lambda_{c,n} \xrightarrow{D} \mathbf{Z}'[(1-c)\mathbf{I} + 2c\mathbf{D}]\mathbf{Z},$$

where $\gamma_1, ..., \gamma_r$ are all the non-zero eigenvalues of $\mathbf{B} = (\mathbf{I}^{-1/2})'(\mathbf{I}^{-1} - \tilde{\mathbf{I}})[(1-c)\mathbf{I} + 2c\mathbf{D}](\mathbf{I}^{-1} - \tilde{\mathbf{I}})\mathbf{I}^{-1/2}$, $(\mathbf{I}^{-1/2})'\mathbf{I}^{-1/2} = \mathbf{I}^{-1}$, and the χ^2_j's are independent chisquared distributions with one degree of freedom; $\mathbf{Z} \sim N(\delta, (\mathbf{I}^{-1} - \tilde{\mathbf{I}})\mathbf{I}(\mathbf{I}^{-1} - \tilde{\mathbf{I}}))$.

A well known property of the log-likelihood ratio is the locally asymptotic normality (LAN) property (LeCam, 1960), which states that, under suitable conditions,

$$l(\theta_0 + n^{-1/2}\delta|\mathbf{x}^n) - l(\theta_0|\mathbf{x}^n) = \delta'\Delta_0 - \frac{1}{2}\delta'\mathbf{I}\delta + o_p(1)$$

The following LAN property for the compound log-likelihood posterior risk $G(\theta|\mathbf{x}^n)$ parallel that of the log-likelihood ratio, and may have its own interest.

Proposition 2. *Assume conditions of Theorem 1 in Walker (1969), and condition (B9) in the Appendix, then for $\delta \in R^d$,*

$$G(\theta_0 + n^{-1/2}\delta|\mathbf{x}^n) - G(\theta_0|\mathbf{x}^n)$$

$$= \delta'[(1-c)I + 2c\mathbf{D}\mathbf{I}^{-1}]\Delta_0 - \frac{1}{2}\delta'\mathbf{I}\delta + o_p(1).$$

7. EXAMPLES

Example 1. Take $W(\theta, \alpha) = ||\theta - \alpha||^2$ to the commonly used squared error loss, then θ_n is the solution of the following equation

$$(1-c)L(\theta_n|\mathbf{x}^n) - 2cn\theta_n + 2cnE(\theta|\mathbf{x}^n) = \mathbf{0},$$

where $L(\theta_n|\mathbf{x}^n) = \partial l(\theta_n|\mathbf{x}^n)/\partial\theta$, and $E(\theta|\mathbf{x}^n) = \int \theta\pi(\theta|\mathbf{x}^n)d\theta$ is the posterior mean of θ. Consider the example where $X \sim N(\theta, I)$ with I be the identity matrix, if we use the squared error loss, then the above equation leads to

$$(1-c)\sum_{i=1}^{n}(\mathbf{x}_i - \theta_n) - 2cn\theta_n + 2cnE(\theta|\mathbf{x}^n) = \mathbf{0}$$

and so with $\hat{\theta}_n = n^{-1}\sum_{i=1}^n \mathbf{x}_i$ be the MLE and $\check{\theta}_n = E(\theta|\mathbf{x}^n)$ be the Bayesian estimate,

$$\theta_n = \frac{1}{1+c}\big((1-c)\hat{\theta}_n + 2c\check{\theta}_n\big)$$

when $c = 1/3$, we get $\theta_n = \frac{1}{2}(\hat{\theta}_n + \check{\theta}_n)$. But generally θ_n is not a linear combination of, nor any function of $\hat{\theta}_n$ and $\check{\theta}_n$, even for the squared error loss.

Example 2. Take $W(\theta,\alpha) = \sum_{j=1}^d |\theta_j - \alpha_j|$ be the absolute error loss, $\theta_n = (\theta_{n,1},...,\theta_{n,d})'$ is the solution of the following equation

$$\mathbf{0} = G^{(1)}(\theta) = (1-c)L(\theta_n|\mathbf{x}^n)$$

$$-cn\bigg(\int_{\theta\le\theta_n}\pi(\theta|\mathbf{x}^n)d\theta - \int_{\theta>\theta_n}\pi(\theta|\mathbf{x}^n)d\theta\bigg),$$

where $\theta \le (>)\theta_n$ is in the componentwise sense.

Example 3. For $W(\theta,\alpha)$ to be the 0-1 loss, the corresponding Bayesian estimator is $\check{\theta}_n = \sup_{\theta\in\Theta} f(\mathbf{x}^n|\theta)\pi(\theta)$, thus in the compound estimation case,

$$\theta_n = \sup_{\theta\in\Theta}\bigg((1-c)l(\theta|\mathbf{x}^n)m(\mathbf{x}^n) + cnf(\mathbf{x}^n|\theta)\pi(\theta)\bigg).$$

8. SIMULATION STUDIES

Two scenarios are considered. In Scenario 1, we assume that X follows an exponential distribution with probability density function (PDF) $f_X(x) = \lambda\exp(-\lambda x), x \ge 0$. To estimate λ, we adopt its conjugate prior, that is, $\lambda \sim Gamma(\alpha_1,\beta_1)$ with PDF $f_\lambda(t) = \frac{\beta_1^{\alpha_1}t^{\alpha_1-1}\exp(-t\beta_1)}{\Gamma(\alpha_1)}$, $t \ge 0$. We replicate 1,000 times to make inference. For each replicate, we generate n *i.i.d.* observations. The first three columns of Table 1 show the parameter settings. Let the sample be $x_1, x_2, \cdots, x_n$. Then the MLE, Bayisan estimator and the compound estimator are, respectively, $\hat{\lambda}_n = \frac{n}{\sum x_i}, \check{\lambda}_n = \frac{n+\alpha_1}{\beta_1+\sum x_i}, \lambda_n = \frac{-u+\sqrt{u^2-4wv}}{2w}$, where $u = (1-c)\frac{1}{n}\sum x_i - 2c\frac{\alpha_1}{\beta_1}$, $v = c-1$ and $w = 2c$.

In Scenario 2, we assume that X follows a binomial distribution with $\Pr(X=1) = p, \Pr(X=0) = 1-p$. Denote the prior, $p \sim Beta(\alpha_2,\beta_2)$. We replicate 1,000 times and generate n *i.i.d.* observations in each replicate. The first three columns of Table 2 gives the details of the parameter setting. Denote the sample as $x_1, x_2, \cdots, x_n$. Then the MLE, Bayesian estimator and the compound estimator are, respectively, $\hat{p}_n = \frac{\sum x_i}{n}$, $\check{p}_n = \frac{\sum x_i + \alpha_2 - 1}{n+\alpha_2+\beta_2-2}$, and p_n, which is the solution on p of the equation:

$$2cp^3 - 2c(1+\check{p}_n)p^2 + (2c\check{p}_n - 1 + c)p + (1-c)\hat{p}_n = 0,\ p \in [0,1]$$

We define the relative efficiency (RE) of the two estimators to be the inverse of the ratio of the mean-squared errors and the REs are summarized in Tables 1 and 2. We can see, under the considered scenarios, the empirical mean-squared errors (MSEs) are smaller than those of the MLE and the Bayesian estimator. For example, as for the exponential distribution, when $1/\lambda = 0.7$ and $n = 100$, two parameters of the prior are 4 and 10, the MSE of the proposed method is 0.0176, which is by far smaller than those of MLE with 0.0222 and Bayesian estimator with 0.0265. The improvement rate over the Bayes estimator is 34% (=(0.0265-0.0176)/0.0265), and that over the MLE is 21%(=(0.0222-0.0176)/0.0222). However, when the sample size increases,

Table 1. The relative efficiency (RE) of the compound estimate (λ_n) compared to the MLE ($\hat{\lambda}_n$) and the Bayesian estimator ($\check{\lambda}_n$) when the data are generated from an exponential distribution. The results are based on one thousand replications.

n	α_1	β_1	$1/\lambda$	$\mathrm{RE}(\lambda_n,\hat{\lambda}_n)$	$\mathrm{RE}(\lambda_n,\check{\lambda}_n)$
100	3	8	0.8	1.24	1.28
	3	9	0.8	1.28	1.39
	4	9	0.7	1.23	1.40
	4	10	0.7	1.26	1.51
	5	9	0.7	1.27	1.27
	5	10	0.8	1.22	1.24
150	3	8	0.8	1.07	1.13
	3	9	0.8	1.11	1.23
	4	9	0.7	1.08	1.24
	4	10	0.7	1.11	1.30
	5	9	0.7	1.12	1.12
	5	10	0.8	1.06	1.07
200	3	8	0.8	1.02	1.05
	3	9	0.8	1.05	1.14
	4	9	0.7	1.01	1.15
	4	10	0.7	1.04	1.24
	5	9	0.7	1.05	1.05
	5	10	0.8	1.01	1.02

Table 2. The relative efficiency (RE) of the compound estimate (p_n) compared to the MLE ($\hat{p}_n$) and the Bayesian estimator ($\check{p}_n$) when the data are generated from a binomial distribution. The results are based on one thousand replications.

n	p	α_2	β_2	$\mathrm{Re}(p_n,\hat{p}_n)$	$\mathrm{Re}(p_n,\check{p}_n)$
100	0.2	7	7	1.21	1.24
	0.2	7	10	1.24	1.35
	0.2	8	8.5	1.20	1.36
	0.2	8	9.5	1.22	1.47
150	0.2	7	7	1.15	1.18
	0.2	7	10	1.18	1.29
	0.2	8	8.5	1.14	1.30
	0.2	8	9.5	1.17	1.40
200	0.2	7	7	1.00	1.04
	0.2	7	10	1.04	1.13
	0.2	8	8.5	1.00	1.13
	0.2	8	9.5	1.02	1.22

Table 3. Empirical power and type-I error rate of the tests based on the compound estimate (CE) and the MLE when the data are generated from a binomial distribution. The results are based on one thousand replications.

n	$1/\lambda$	α_1	β_1	CE	MLE
100	0.8	3	8	0.051	0.055
		3	9	0.049	0.055
	0.7	3	8	0.460	0.353
		3	9	0.453	0.353
150	0.8	3	8	0.051	0.052
		3	9	0.049	0.052
	0.7	3	8	0.511	0.436
		3	9	0.516	0.436
200	0.8	3	8	0.049	0.056
		3	9	0.048	0.056
	0.7	3	8	0.568	0.538
		3	9	0.571	0.538

the improvement decreases and the effect will diminish or become less pronounced. For Scenario 1, we test the hypothesis $H_0 : \lambda = 1/0.8$. Letting $1/\lambda = 0.8$ or $1/\lambda = 0.7$, we conduct simulations to compare the type-I error or power of the tests based on the MLE and Compound estimator. The results are summarized in Table 3.

9. AN APPLICATION TO DEEP GROOVE BALL HEARINGS DATA

The data set is from Lawless (1982). The data given arose in tests on endurance of deep groove ball bearings. The data are the number of million revolutions before failure (Y) for each of the 23 ball bearings in the life tests and they are:

17.88, 28.92, 33.00, 41.52, 42.12, 45.60, 48.80, 51.84, 51.96, 54.12, 55.56, 67.80, 68.44, 68.64, 68.88, 84.12, 93.12, 98.64,

105.12, 105.84, 127.92, 128.04, 173.40. We use the exponential distribution to model Y and the MLE is 0.014 and the compound estimator (CE) is 0.012. The 95% normal-based confidence intervals from the bootstrap samples ($B = 1000$) are respectively $(0.011, 0.017)$, and $(0.009, 0.015)$.

10. CONCLUDING REMARKS

We proposed a frequentist and Bayesian compound method, as an attempt to unify the two methods. Basic properties are studied, like each of the two methods there are still lots of issues to be tackled, such as bias reduction in both point estimation and hypothesis testing, choice of prior, possibility of extension to nonparametric problems, and difficulties in high-dimensional settings, etc. These issues will be topics of our future research. Currently only the second order term in the expansion is taken into account, the usefulness of the third and higher order terms is also a research topic.

APPENDIX A. APPENDIX SECTION

Regularity conditions. Throughout this paper we assume the densities are with respect to the Lebesgue measure. In the following, conditions (A1)–(A3) are A 2.1, A 2.6 and A 2.7 in Bickel and Yahav (1969).

(A1) θ belongs to an open subset of R^d.

(A2) Let $l(\mathbf{x}|\theta)$ be the log-likelihood. Assume $\partial l(\mathbf{x}|\theta)/\partial\theta$ and $\partial^2 l(\mathbf{x}|\theta)/(\partial\theta\partial\theta')$exist and are continuous in θ for almost all $\mathbf{x}$.

(A3) $E_\theta\left(\sup_{\eta\in\Theta}||\partial^2 l(\mathbf{x}|\eta)/(\partial\theta\partial\theta')|| : ||\eta-\theta|| < \epsilon(\theta)\right) < \infty$ for some $\epsilon(\theta) > 0$ and all $\theta \in \Theta$.

Let P_θ be the data distribution given $\theta \in \Theta$, and $l_n(\mathbf{x}^n|\theta) = \frac{1}{n}\sum_{i=1}^n l(\mathbf{x}_i|\theta)$. Conditions (A4)-(A9) below are those of (1)-(6) in Strasser (1981).

(A4) The metric space (Θ, d) is separable, where $d(\theta, \eta) = ||P_\theta - P_\eta||$.

(A5) The functions $(l_n(\cdot|\theta))_{\theta\in\Theta}, n \in N$, are separable and measurable.

(A6) $f(\cdot|\theta)$, $\theta \in \Theta$, are lower semicontinuous, that is, $\limsup_{n\to\infty} f(\cdot|\theta_n) \le$
$f(\cdot|\theta)$ (a.e) if $d(\theta_n, \theta) \to 0$.

(A7) For every $\theta, \eta \in \Theta$, there is an open neighborhood $U_{\theta,\eta}$ of η such that
$E_\theta(\inf_{\theta'\in U_{\theta,\eta}} l_n(\mathbf{x}^n|\theta')) > -\infty$ for at least one n.

(A8) For every $\theta \in \Theta$ and $\epsilon > 0$, $\Pi\big(\eta \in \Theta : E_\theta l(\mathbf{x}|\eta) < E_\theta l(\mathbf{x}|\theta) + \epsilon\big) > 0$,
where $\Pi(\cdot)$ is the distribution for $\pi(\cdot)$.

(A9) For every $\theta \in \Theta$ there is some n_θ such that $P_\theta^n(\mathbf{x}^n : \int\prod_{i=1}^n f(\mathbf{x}_i|\eta)\Pi(d\eta)$
$< \infty) = 1$ if $n \ge n_\theta$.

Conditions (B1)–(B10) are multivariate versions of those of 1–10 in Gusev (1975).

(B1) For $\theta \ne \eta$, $\int |f(\mathbf{x}|\theta) - f(\mathbf{x}|\eta)|d\mathbf{x} > 0$.

(B2) For some $p_1 > 0$ and some compact set $\mathbf{K} \in \Theta$, $\sup_{\theta\in K,\eta\in\Theta}||\theta-\eta||^{p_1}$
$\times \int \sqrt{f(\mathbf{x}|\theta)f(\mathbf{x}|\eta)}d\mathbf{x} < \infty$.

(B3) $f(\mathbf{x}|\cdot)$ is continuous on Θ^c, the closure of Θ on R^d and has $k + 2$ $(k \ge 1)$
continuous derivatives on Θ.

(B4) (a) For some $b > 0$, and every compact $\mathbf{K} \in \Theta$,
$\sup_{\theta\in K} E_\theta||\mathbf{L}(\mathbf{x}|\theta)||^{3\vee(k+1+b)} < \infty$.
(b) For every compact $\mathbf{K} \in \Theta$,
$\max_{1\le|\mathbf{i}|\le k}\sup_{\theta\in K} E_\theta||\mathbf{L_i}(\mathbf{x}|\theta)||^{k+1} < \infty$.
(c) For every compact $\mathbf{K} \in \Theta$, and for some $\epsilon_1(\mathbf{K}) > 0$,
$\max_{|\mathbf{i}|=k+1}\sup_{\theta\in K,||\theta-\eta||\le\epsilon_1(\mathbf{K})} E_\theta||\mathbf{L_i}(\mathbf{x}|\eta)||^{(k+1)/2} < \infty$.

(B5) (a) For some $p_2 \ge 0$, $\sup_{\theta\in\Theta}(1 + ||\theta||^{p_2})^{-1}||\mathbf{I}(\theta)|| < \infty$.

(b) $\mathbf{I}(\theta)$ is positive definite for $\theta \in \Theta$.

(B6) $\pi(\cdot)$ has k continuous derivatives on Θ.

(B7) For some $p_3 > 0$, $\sup_{\theta\in\Theta}(1 + ||\theta||^{p_3})^{-1}\pi(\theta) < \infty$.

(B8) $W(\cdot) \geq 0$, is convex, that is, for any $t \in [0,1]$ and $\mathbf{u}_1$ and $\mathbf{u}_2$,
$W(t\mathbf{u}_1 + (1-t)\mathbf{u}_2) \leq tW(\mathbf{u}_1) + (1-t)W(\mathbf{u}_2)$.

(B9) For some small $\epsilon_2 > 0$ and positive definite $\mathbf{D}$,
$W(\theta) = \theta'\mathbf{D}\theta$ for $||\theta|| \leq \epsilon_2$.

(B10) For some $p_4 > 0$, $\sup_{\theta\in\Theta}(1 + ||\theta||^{p_4})^{-1}W(\theta) < \infty$.

Proof of Theorem 1. Let $\tilde{\theta}_n$ be the MLE and $\check{\theta}_n$ be the Bayesian estimate of θ, and write $\theta_n = \theta_n(c)$ to stress its dependence on c. Then $\theta_n(1) = \tilde{\theta}_n$, $\theta_n(0) = \check{\theta}_n$.

With the given conditions, $\tilde{\theta}_n$ and $\check{\theta}_n$ exist and consistent (a.s.), and so for large n, in a small neighborhood of θ_0, $(1-c)l(\theta|\mathbf{x}^n)$ and $cn\int W(\theta,\alpha)f(\mathbf{x}^n|\alpha)\pi(\alpha)d\alpha$ are peaked at $\tilde{\theta}_n$ and $\check{\theta}_n$ respectively, both are convex as a function of θ.

Recall the definition of $G(\theta)$ after expression (1). First consider the case $d = 1$. It is seen that if $\theta < \tilde{\theta}_n \wedge \check{\theta}_n$ or $\theta > \tilde{\theta}_n \vee \check{\theta}_n$, then $G(\theta) < \max\{G(\tilde{\theta}_n), G(\check{\theta}_n)\} \leq G(\theta_n)$, so we must have $\tilde{\theta}_n \wedge \check{\theta}_n \leq \theta_n \leq \tilde{\theta}_n \vee \check{\theta}_n$, i.e., θ_n exists. Since under the given conditions $\tilde{\theta}_n \to \theta_0$ (a.s.) and $\check{\theta}_n \to \theta_0$ (a.s.), we must have $\theta_n \to \theta_0$ (a.s.).

For $d > 1$, let $\tilde{\theta}_n \wedge \check{\theta}_n$ and $\tilde{\theta}_n \vee \check{\theta}_n$ be in the component-wise sense. Similarly, we cannot have $\theta_n \in [\tilde{\theta}_n \wedge \check{\theta}_n, \tilde{\theta}_n \vee \check{\theta}_n]^c$. So at lease one component $\theta_{j,n}$ of θ_n satisfies $\theta_{j,n} \in [\tilde{\theta}_{j,n} \wedge \check{\theta}_{j,n}, \tilde{\theta}_{j,n} \vee \check{\theta}_{j,n}]$. Fix this component and let the new parameter be θ_{-j} which is θ with component j removed, and define $\tilde{\theta}_{-j,n}$, $\check{\theta}_{-j,n}$ and $\theta_{-j,n}$ accordingly. Then we see that at least one component of $\theta_{-j,n}$ will be in the corresponding component of $[\tilde{\theta}_{-j,n} \wedge \check{\theta}_{-j,n}, \tilde{\theta}_{-j,n} \vee \check{\theta}_{-j,n}]$;... continue this way we have $\theta_n \in [\tilde{\theta}_n \wedge \check{\theta}_n, \tilde{\theta}_n \vee \check{\theta}_n]$, so the existence of θ_n, and the a.s. consistency of it. □

Let

$$Z_n(\theta) = \left(\prod_{i=1}^{n} \frac{f(x_i|\theta_0 + \theta n^{-1/2})}{f(x_i|\theta_0)}\right)\frac{\pi(\theta_0 + \theta n^{-1/2})}{\pi(\theta_0)}.$$

Let $\hat{\theta}_n$ be the posterior mode, i.e.

$$\hat{\theta}_n = \sup_{\theta}\{f(\mathbf{x}^n|\theta)\pi(\theta)\} \quad \text{and} \quad \hat{\theta}_n' = \sqrt{n}(\hat{\theta}_n - \theta_0).$$

The following Lemma is a multivariate modification of Gusev (1975, Theorem 2).

Lemma 1. *Under the conditions of Theorem 2,*

$$\frac{Z_n(\theta + \hat{\theta}_n')}{Z_n(\hat{\theta}_n')} = exp\left(-\frac{1}{2}\theta'\mathbf{I}\theta\right)\left(1 + \sum_{|\mathbf{r}|=1}^{k-1} n^{-|\mathbf{r}|/2} \sum_{|\mathbf{i}|=2}^{3|\mathbf{r}|} \langle\theta^{\mathbf{i}}\rangle N_{\mathbf{i},\mathbf{r}}\right)$$

$$+ n^{-k/2}R_{k,n}(\theta),$$

where $R_{k,n}(\theta)$ is such that for some C_1 and C_2 and for every δ with $0 < \delta < C_1$, for every compact K,

$$\sup_{\theta_0\in K} P_{\theta_0}(\sup_{\theta} |R_{k,n}(\theta)| > n^{\delta}) = O(n^{(k-1+C_2\delta)/2}),$$

$$N_{\mathbf{i},\mathbf{r}} = \sum_{I_2(\mathbf{r},\mathbf{i})} \prod_{v=1}^{|\mathbf{r}|} \sum_{I_1(2,v+2,\mathbf{k}_v,\mathbf{i}_v)} \prod_{s=2}^{v+2} \frac{F_{s,v}^{|\mathbf{j}_s|}}{\mathbf{j}_s!(s!)^{|\mathbf{j}_s|}},$$

in the above the summations are for $(\mathbf{k}_v, \mathbf{i}_v) \in I_2(\mathbf{r},\mathbf{i})$; for each given pair $(\mathbf{k}_v, \mathbf{i}_v)$, $\mathbf{j}_s (s = 2, ..., v+2) \in I_1(2, v+2, \mathbf{k}_v, \mathbf{i}_v)$, $I_2(\mathbf{r},\mathbf{i})$ and $I_1(m,r,\mathbf{k}_v,\mathbf{i}_v)$ given below, and

$$F_{s,v} = \sum_{|\mathbf{j}|=v,|\mathbf{k}|=0}^{v-s} \rho_{\mathbf{j}-\mathbf{k}} \sum_{|\mathbf{i}|=s,\ I_1(0,|\mathbf{k}|,\mathbf{k},\mathbf{j}-\mathbf{i}-\mathbf{k})} \prod_{v=0}^{|\mathbf{k}|} \langle\mathbf{H}_v^{\mathbf{i}_v}\rangle/\mathbf{i}_v!$$

$$+ \sum_{|\mathbf{j}|=v+1,|\mathbf{k}|=0}^{v+1-s} \Delta_{\mathbf{j}-\mathbf{k}} \sum_{|\mathbf{i}|=s,\ I_1(0,|\mathbf{k}|,\mathbf{k},\mathbf{j}-\mathbf{i}-\mathbf{k})} \prod_{v=0}^{|\mathbf{k}|} \frac{1}{\mathbf{i}_v!}$$

$$+ \sum_{|\mathbf{j}|=v+2,|\mathbf{k}|=0}^{v+2-s} E_{\mathbf{j}-\mathbf{k}} \sum_{|\mathbf{i}|=s,\ I_1(0,|\mathbf{k}|,\mathbf{k},\mathbf{j}-\mathbf{i}-\mathbf{k})} \prod_{v=0}^{|\mathbf{k}|} \langle\mathbf{H}_v^{\mathbf{i}_v}\rangle/\mathbf{i}_v!.$$

The $\mathbf{H}_s$'s parallel the same notations as in Theorem 2 in Yuan (2009), and

$$I_1(a,r,\mathbf{m},\mathbf{l}) = \{(\mathbf{i}_a, ..., \mathbf{i}_r) : \sum_{v=a}^{r} v\mathbf{i}_v = \mathbf{m}, \quad \sum_{v=a}^{r} \mathbf{i}_v = \mathbf{l}\}.$$

$$I_2(\mathbf{r},\mathbf{i}) = \Big\{(\mathbf{k}_1, ..., \mathbf{k}_{|\mathbf{r}|}), (\mathbf{m}_1, ..., \mathbf{m}_{|\mathbf{r}|}) :$$

$$\sum_{v=1}^{|\mathbf{r}|} \mathbf{k}_v = \mathbf{r}, \quad \frac{\mathbf{i}-\mathbf{r}}{2} \leq \sum_{v=1}^{|\mathbf{r}|} \mathbf{k}_v \leq \frac{\mathbf{i}}{2},$$

$$\sum_{v=1}^{|\mathbf{r}|} \mathbf{m}_v = \mathbf{i}, \quad 2\mathbf{k}_v \leq |\mathbf{m}_v| \leq (v+2)\mathbf{k}_v, \quad v = 1, ..., |\mathbf{r}|\Big\}.$$

Proof of Theorem 2. The given conditions ensure the validity of the results as in Gusev (1975) and Yuan (2009). Below we derive the details of the results. The proof is involved, we break it into steps.

Step 1. Let $\theta_n' = \sqrt{n}(\theta_n - \theta_0)$, recall the definition of $\hat{\theta}_n'$ given before the Lemma, $m(\mathbf{x}^n) = \int f(\mathbf{x}^n|\theta)\pi(\theta)d\theta$, and define $\mathbf{d}_n' = \theta_n' - \hat{\theta}_n'$. By definition of θ_n, we have, with $\mathbf{S}_0$ be the score,

$$\mathbf{0} = (1-c)\mathbf{S_0}(\theta_n)m(\mathbf{x}^n)$$

$$-n^{1/2}c\int \mathbf{W}^{(1)}(\theta_n - \theta)f(\mathbf{x}^n|\theta)\pi(\theta)d\theta$$

$$= \int \Big((1-c)\mathbf{S_0}(\theta_n) - n^{1/2} c \mathbf{W}^{(1)}(\theta_0 - \theta + n^{-1/2}\theta_n') \Big)$$

$$f(\mathbf{x}^n|\theta)\pi(\theta)d\theta$$

$$= n^{-d/2} \int \Big((1-c)\mathbf{S_0}(\theta_n) - n^{1/2} c \mathbf{W}^{(1)}(\frac{\theta_n' - \eta}{\sqrt{n}}) \Big)$$

$$f(\mathbf{x}^n|\theta_0 + n^{-1/2}\eta)\pi(\theta_0 + n^{-1/2}\eta)d\eta$$

$$= n^{-d/2} \int \Big((1-c)\mathbf{S_0}(\theta_n) - n^{1/2} c \mathbf{W}^{(1)}(\frac{\mathbf{d}_n' - \theta}{\sqrt{n}}) \Big)$$

$$f(\mathbf{x}^n|\theta_0 + n^{-1/2}(\theta + \hat{\theta}_n'))\pi(\theta_0 + n^{-1/2}(\theta + \hat{\theta}_n'))d\theta,$$

and we get
(A.1)

$$\mathbf{0} = \int \Big((1-c)\mathbf{S_0}(\theta_n) - n^{1/2} c \mathbf{W}^{(1)}(\frac{\mathbf{d}_n' - \theta}{\sqrt{n}}) \Big) \frac{Z_n(\theta + \hat{\theta}_n')}{Z_n(\hat{\theta}_n')} d\theta.$$

Step 2. Solving out $\theta_n' = \sqrt{n}(\theta_n - \theta_0)$ from the above equation to get the desired result. Expanding $\mathbf{S_0}(\theta_n) = \mathbf{S_0}(\theta_0 + n^{-1/2}\theta_n')$ we get

$$\mathbf{S_0}(\theta_0 + n^{-1/2}\theta_n') = \sum_{r=0}^{k} n^{-r/2} \sum_{|\mathbf{i}|=r} \mathbf{S_i} \frac{\langle (\theta_n')^{\mathbf{i}} \rangle}{\mathbf{i}!} + r_{1,k,n}(\theta_n').$$

Using results of Ibragimov and Khas'minski (1973a, 1973b), Gusev (1975) proved, in the case of $d = 1$, $r_{1,k,n}(\hat{\theta}_n') \overset{k}{\sim} 0$ and $r_{1,k,n}(\tilde{\theta}_n') \overset{k}{\sim} 0$, where $\tilde{\theta}_n' = \sqrt{n}(\tilde{\theta}_n - \theta_0)$, and $\tilde{\theta}_n$ is the MLE (the original result is for the posterior mode $\hat{\theta}_n$, which is the MLE if we choose the prior to be constant). The same results hold for $d > 1$, and as in the proof of Theorem 1, since $\theta_n \in [\tilde{\theta}_n \wedge \check{\theta}_n, \tilde{\theta}_n \vee \check{\theta}_n]$, and thus the a.s. consistency of θ_n, we have $\theta_n' \in [\tilde{\theta}_n' \wedge \check{\theta}_n', \tilde{\theta}_n' \vee \check{\theta}_n']$, and consequently, $r_{1,k,n}(\theta_n') \overset{k}{\sim} \mathbf{0}$.

Step 3. Write the integral on the right hand side of (A,1) as $\int = \int_{\|\theta\| \le n^\delta} + \int_{\|\theta\| > n^\delta}$, and justify that $\int_{\|\theta\| > n^\delta}$ is negligibly small. Note by definition $Z_n(\hat{\theta}_n') \ge 1$. As in Gusev (1975, p.485-486), condition (B8) implies that, for some $0 < C < \infty$ and $0 < \alpha < 1/2$, $||W^{(1)}(\frac{\mathbf{d}_n' - \theta}{\sqrt{n}})|| \le C||\theta + \hat{\theta}_n'||^\alpha$, also, since $\theta_n \to \theta_0$ (a.s.), and it is known that $\hat{\theta}_n' = \mathbf{I}^{-1}\mathbf{S_0}(\theta_0) + o_p(1)$, so for all large n, $||\mathbf{S_0}(\theta_n)|| \le C||\theta + \hat{\theta}_n'||^\alpha$ (a.s.) for $||\theta||$ large, also, $Z_n(\hat{\theta}_n') \ge 1$ for large n, consequently

$$\|\int_{||\theta||>n^\delta} \Big((1-c)\mathbf{S_0}(\theta_n) - n^{1/2} c \mathbf{W}^{(1)}(\frac{\mathbf{d}_n' - \theta}{\sqrt{n}}) \Big) \frac{Z_n(\theta + \hat{\theta}_n')}{Z_n(\hat{\theta}_n')} d\theta\|$$

$$\le \|\int_{||\theta||>n^\delta} \Big((1-c)\mathbf{S_0}(\theta_n) - n^{1/2} c \mathbf{W}^{(1)}(\frac{\mathbf{d}_n' - \theta}{\sqrt{n}}) \Big) Z_n(\theta + \hat{\theta}_n') d\theta\|$$

$$= O_p(n^{-(k+|\mathbf{a}|-1)/2}).$$

Step 4. The above, with condition (B9), on the set $||\theta|| \le n^\delta/2$ we can replace $n^{1/2} c \mathbf{W}^{(1)}(n^{-1/2}(\mathbf{d}_n' - \theta))$ by $2c\mathbf{D}(\mathbf{d}_n' - \theta)$ and using Lemma 1, recall the definition of $\overset{k}{\sim}$, (A.1) is now

$$\mathbf{0} \overset{k}{\sim} (\int_{||\theta|| \le n^\delta/2} + \int_{||\theta|| > n^\delta/2}) \Big((1-c)\mathbf{S_0}(\theta_n)$$

$$-n^{1/2} c \mathbf{W}^{(1)}(\frac{\mathbf{d}_n' - \theta}{\sqrt{n}}) \Big) \frac{Z_n(\theta + \hat{\theta}_n')}{Z_n(\hat{\theta}_n')} d\theta$$

$$= \int_{||\theta|| \le n^\delta/2} \Big((1-c)\mathbf{S_0}(\theta_n) - n^{1/2} c \mathbf{W}^{(1)}(\frac{\mathbf{d}_n' - \theta}{\sqrt{n}}) \Big)$$

$$\frac{Z_n(\theta + \hat{\theta}_n')}{Z_n(\hat{\theta}_n')} d\theta + O_p(n^{-(k+|\mathbf{a}|-1)/2})$$

$$= \int_{||\theta|| \le n^\delta/2} \Big((1-c)\mathbf{S_0}(\theta_n) - 2c\mathbf{D}(\mathbf{d}_n' - \theta) \Big) \frac{Z_n(\theta + \hat{\theta}_n')}{Z_n(\hat{\theta}_n')} d\theta$$

$$+ O_p(n^{-(k+|\mathbf{a}|-1)/2})$$

$$\overset{k}{\sim} \int_{||\theta|| \le n^\delta/2} \Big((1-c)\mathbf{S_0}(\theta_n) - 2c\mathbf{D}(\mathbf{d}_n' - \theta) \Big)$$

$$\exp\Big(-\frac{1}{2}\theta'\mathbf{I}\theta \Big) \Big(1 + \sum_{|\mathbf{r}|=1}^{k-1} n^{-|\mathbf{r}|/2} \sum_{|\mathbf{j}|=2}^{3r} \langle \theta^{\mathbf{j}} \rangle N_{\mathbf{j},\mathbf{r}} \Big) d\theta$$

$$\overset{k}{\sim} \int \Big((1-c)\mathbf{S_0}(\theta_n) - 2c\mathbf{D}(\mathbf{d}_n' - \theta) \Big)$$

$$\exp\Big(-\frac{1}{2}\theta'\mathbf{I}\theta \Big) \Big(1 + \sum_{|\mathbf{r}|=1}^{k-1} n^{-|\mathbf{r}|/2} \sum_{|\mathbf{j}|=2}^{3|\mathbf{r}|} \langle \theta^{\mathbf{j}} \rangle N_{\mathbf{j},\mathbf{r}} \Big) d\theta$$

$$\overset{k}{\sim} \int \Big((1-c) \sum_{r=0}^{k} n^{-r/2} \sum_{|\mathbf{i}|=r} \mathbf{S_i} \frac{\langle (\theta_n')^{\mathbf{i}} \rangle}{\mathbf{i}!} - 2c\mathbf{D}(\mathbf{d}_n' - \theta) \Big)$$

$$\times \Big[\exp\Big(-\frac{1}{2}\theta'\mathbf{I}\theta \Big) \Big(1 + \sum_{|\mathbf{r}|=1}^{k-1} n^{-|\mathbf{r}|/2} \sum_{|\mathbf{j}|=2}^{3|\mathbf{r}|} \langle \theta^{\mathbf{j}} \rangle N_{\mathbf{j},\mathbf{r}} \Big) \Big] d\theta$$

$$\overset{k}{\sim} E_\theta \Big[(1-c) \Big(\sum_{r=0}^{k} n^{-r/2} \sum_{|\mathbf{i}|=r} \mathbf{S_i} \frac{\langle (\theta_n')^{\mathbf{i}} \rangle}{\mathbf{i}!} - 2c\mathbf{D}(\theta_n' - \hat{\theta}_n' - \theta) \Big)$$

$$\Big(1+\sum_{|\mathbf{r}|=1}^{k-1} n^{-|\mathbf{r}|/2}\sum_{|\mathbf{j}|=2}^{3|\mathbf{r}|}\langle\theta^{\mathbf{j}}\rangle N_{\mathbf{j},\mathbf{r}}\Big)\Big],$$

where $\theta\sim N(\mathbf{0},\mathbf{I}^{-1})$, and in the above we used the fact that for all $0<\delta<\infty$, $\int_{||\theta||>n^{\delta}/2}\Big(\mathbf{S_0}(\theta_n)-2c\mathbf{D}(\mathbf{d}'_n-\theta)\Big)\exp\Big(-\frac{1}{2}\theta'\mathbf{I}\theta\Big)\Big(1+\sum_{|\mathbf{r}|=1}^{k-1} n^{-|\mathbf{r}|/2}\sum_{|\mathbf{j}|=2}^{3|\mathbf{r}|}\langle\theta^{\mathbf{j}}\rangle N_{\mathbf{j},\mathbf{r}}\Big)d\theta = o_p(n^{-k/2})$ for all finite k.

Step 5. We make some notation clear-up. Note $\mathbf{S_0}=\Delta_{\mathbf{0}}$, $\mathbf{E_0}=\mathbf{0}$; $\mathbf{S_i}=\Delta_{\mathbf{i}}+n^{1/2}\mathbf{E_i}$ for $|\mathbf{i}|\geq 1$, $\mathbf{I}=-(\mathbf{E_i}:|\mathbf{i}|=1)$, $\sum_{|\mathbf{i}|=1}\mathbf{E_i}\langle(\theta'_n)^{\mathbf{i}}\rangle=-\mathbf{I}\theta'_n$, and $n^{-k/2}\sum_{|\mathbf{i}|=k}\Delta_{\mathbf{i}}\langle(\theta'_n)^{\mathbf{i}}\rangle/\mathbf{i}!=O_p(n^{-k/2})$. Let $\langle\mu\rangle_{\mathbf{i}}=E[\langle\theta^{\mathbf{i}}\rangle]$ and $\mu_{\mathbf{1},\mathbf{j}}=E[\theta\langle\theta^{\mathbf{j}}\rangle]$, with $\theta\sim N(\mathbf{0},\mathbf{I}^{-1})$. We have

$$\mathbf{0}\overset{k}{\sim}E_\theta\Big[\Big((1-c)(\Delta_0-\mathbf{I}\theta'_n)-2c\mathbf{D}(\theta'_n-\hat{\theta}'_n-\theta)$$

$$+\sum_{r=1}^{k-1}n^{-r/2}(1-c)\big(\sum_{|\mathbf{i}|=r}\Delta_{\mathbf{i}}\frac{\langle(\theta'_n)^{\mathbf{i}}\rangle}{\mathbf{i}!}+\sum_{|\mathbf{i}|=r+1}\mathbf{E_i}\frac{\langle(\theta'_n)^{\mathbf{i}}\rangle}{\mathbf{i}!}\big)\Big)$$

$$\Big(1+\sum_{|\mathbf{r}|=1}^{k-1}n^{-|\mathbf{r}|/2}\sum_{|\mathbf{j}|=2}^{3|\mathbf{r}|}\langle\theta^{\mathbf{j}}\rangle N_{\mathbf{j},\mathbf{r}}\Big)\Big]$$

$$=(1-c)(\Delta_0-\mathbf{I}\theta'_n)-2c\mathbf{D}(\theta'_n-\hat{\theta}'_n)$$

$$+\sum_{r=1}^{k-1}n^{-r/2}(1-c)\big(\sum_{|\mathbf{i}|=r}\Delta_{\mathbf{i}}\frac{\langle(\theta'_n)^{\mathbf{i}}\rangle}{\mathbf{i}!}+\sum_{|\mathbf{i}|=r+1}\mathbf{E_i}\frac{\langle(\theta'_n)^{\mathbf{i}}\rangle}{\mathbf{i}!}\big)$$

$$+\sum_{|\mathbf{r}|=1}^{k-1}n^{-|\mathbf{r}|/2}\sum_{|\mathbf{j}|=2}^{3|\mathbf{r}|}N_{\mathbf{j},\mathbf{r}}\Big((1-c)\langle\mu\rangle_{\mathbf{j}}(\Delta_0-\mathbf{I}\theta'_n)$$

$$-2c\mathbf{D}[(\theta'_n-\hat{\theta}'_n)\langle\mu\rangle_{\mathbf{j}}-\mu_{\mathbf{1},\mathbf{j}}]\Big)$$

$$+\sum_{r=2}^{k-1}n^{-r/2}(1-c)\sum_{l+|\mathbf{m}|=r,|\mathbf{m}|\geq 1}\sum_{|\mathbf{j}|=2}^{3|\mathbf{m}|}\langle\mu\rangle_{\mathbf{j}}N_{\mathbf{j},\mathbf{m}}\Big(\sum_{|\mathbf{i}|=l}\Delta_{\mathbf{i}}\frac{\langle(\theta'_n)^{\mathbf{i}}\rangle}{\mathbf{i}!}$$

$$+\sum_{|\mathbf{i}|=l+1}\mathbf{E_i}\frac{\langle(\theta'_n)^{\mathbf{i}}\rangle}{\mathbf{i}!}\Big).$$

Also, the same way as in the proof in Yuan (2009,p2480–2481) and Theorem 2.2 there, we have, with $\Delta_0=\sum_{i=1}^d e_ie'_i\Delta_{\mathbf{e}_i}$ be the score,

$$\theta'_n\overset{k}{\sim}\sum_{r=0}^{k-1}n^{-r/2}\mathbf{B}_r,\quad \hat{\theta}'_n\overset{k}{\sim}\sum_{r=0}^{k-1}n^{-r/2}\mathbf{H}_r,\quad\text{and}\quad \mathbf{H}_0=\mathbf{I}^{-1}\Delta_0.$$

Step 6. Further simplifications. These give, with I being the d-dimensional identity matrix (without confusing it with the Fisher information matrix $\mathbf{I}$)

$$[(1-c)\mathbf{I}+2c\mathbf{D}]\theta'_n\overset{k}{\sim}[(1-c)I+2c\mathbf{D}\mathbf{I}^{-1}]\Delta_0$$

$$+\sum_{r=1}^{k-1}n^{-r/2}\Big(2c\mathbf{D}\mathbf{H}_r+(1-c)\big(\sum_{|\mathbf{i}|=r}\Delta_{\mathbf{i}}\frac{\langle(\theta'_n)^{\mathbf{i}}\rangle}{\mathbf{i}!}+\sum_{|\mathbf{i}|=r+1}\mathbf{E_i}\frac{\langle(\theta'_n)^{\mathbf{i}}\rangle}{\mathbf{i}!}\big)\Big)$$

$$+\sum_{|\mathbf{r}|=1}^{k-1}n^{-|\mathbf{r}|/2}\sum_{|\mathbf{j}|=2}^{3|\mathbf{r}|}N_{\mathbf{j},\mathbf{r}}\Big((1-c)\langle\mu\rangle_{\mathbf{j}}(\Delta_0-\mathbf{I}\theta'_n)$$

$$-2c\mathbf{D}[(\theta'_n-\hat{\theta}'_n)\langle\mu\rangle_{\mathbf{j}}-\mu_{\mathbf{1},\mathbf{j}}]\Big)$$

$$+\sum_{r=2}^{k-1}n^{-r/2}(1-c)\sum_{l+|\mathbf{m}|=r,|\mathbf{m}|\geq 1}\sum_{|\mathbf{j}|=2}^{3|\mathbf{m}|}\langle\mu\rangle_{\mathbf{j}}N_{\mathbf{j},\mathbf{m}}\Big(\sum_{|\mathbf{i}|=l}\Delta_{\mathbf{i}}\frac{\langle(\theta'_n)^{\mathbf{i}}\rangle}{\mathbf{i}!}$$

$$+\sum_{|\mathbf{i}|=l+1}\mathbf{E_i}\frac{\langle(\theta'_n)^{\mathbf{i}}\rangle}{\mathbf{i}!}\Big).$$

Recall $\sum_{r=t}^{s}(\cdots)=0$ for $t>s$, so for $k=1$, we get

$$[(1-c)\mathbf{I}+2c\mathbf{D}]\theta'_n\overset{1}{\sim}[(1-c)I+2c\mathbf{D}\mathbf{I}^{-1}]\Delta_0$$

or

$$\sqrt{n}(\theta_n-\theta_0)=\theta'_n\overset{1}{\sim}[(1-c)\mathbf{I}+2c\mathbf{D}]^{-1}[(1-c)I+2c\mathbf{D}\mathbf{I}^{-1}]\Delta_0.$$

Step 7. Expand $[(1-c)\mathbf{I}+2c\mathbf{D}]^{-1}$ in the above to get more clear result. Note $[(1-c)\mathbf{I}+2c\mathbf{D}]$ is positive definite and hence invertible. Using the formula $(A-BD^{-1}C)^{-1}=A^{-1}+A^{-1}B(D-CA^{-1}B)^{-1}CA^{-1}$, with $A=(1-c)\mathbf{I}$, $B=-2c\mathbf{D}$, $C=D=I$, we get $[(1-c)\mathbf{I}+2c\mathbf{D}]^{-1}=(1-c)^{-1}\mathbf{I}^{-1}-2c(1-c)^{-1}\mathbf{I}^{-1}\mathbf{D}[(1-c)I+2c\mathbf{I}^{-1}\mathbf{D}]^{-1}\mathbf{I}^{-1}$, note $\mathbf{I}^{-1}[(1-c)I+2c\mathbf{D}\mathbf{I}^{-1}]=[(1-c)I+2c\mathbf{I}^{-1}\mathbf{D}]\mathbf{I}^{-1}$, and get

$$[(1-c)\mathbf{I}+2c\mathbf{D}]^{-1}[(1-c)I+2c\mathbf{D}\mathbf{I}^{-1}]=\mathbf{I}^{-1}+2c(1-c)^{-1}\mathbf{I}^{-1}\mathbf{D}\mathbf{I}^{-1}$$

$$-2c(1-c)^{-1}\mathbf{I}^{-1}\mathbf{D}[(1-c)I+2c\mathbf{I}^{-1}\mathbf{D}]^{-1}[(1-c)I+2c\mathbf{I}^{-1}\mathbf{D}]\mathbf{I}^{-1}$$

$$=\mathbf{I}^{-1},$$

consequently,

$$\sqrt{n}(\theta_n-\theta_0)\overset{1}{\sim}\mathbf{I}^{-1}\Delta_0$$

and, with $\mathbf{B}_0=\mathbf{I}^{-1}\Delta_0$,

$$\sqrt{n}(\theta_n-\theta_0)\overset{k}{\sim}\mathbf{B}_0+\sum_{r=1}^{k-1}n^{-r/2}[(1-c)\mathbf{I}+2c\mathbf{D}]^{-1}\Big(2c\mathbf{D}\mathbf{H}_r$$

$$+(1-c)\Big(\sum_{|\mathbf{i}|=r}\Delta_{\mathbf{i}}\frac{\langle(\theta_n')^{\mathbf{i}}\rangle}{\mathbf{i}!}+\sum_{|\mathbf{i}|=r}\mathbf{E}_{\mathbf{i}}\frac{\langle(\theta_n')^{\mathbf{i}}\rangle}{\mathbf{i}!}\Big)\Big)$$

$$+\sum_{|\mathbf{r}|=1}^{k-1}n^{-|\mathbf{r}|/2}\sum_{|\mathbf{j}|=2}^{3|\mathbf{r}|}N_{\mathbf{j},\mathbf{r}}[(1-c)\mathbf{I}+2c\mathbf{D}]^{-1}$$

$$\Big((1-c)\langle\mu\rangle_{\mathbf{j}}(\Delta_0-\mathbf{I}\theta_n')-2c\mathbf{D}[(\theta_n'-\hat{\theta}_n')\langle\mu\rangle_{\mathbf{j}}-\mu_{\mathbf{1},\mathbf{j}}]\Big)$$

$$+\sum_{r=2}^{k-1}n^{-r/2}\sum_{l+|\mathbf{m}|=r,|\mathbf{m}|\geq 1}\sum_{|\mathbf{j}|=2}^{3|\mathbf{m}|}(1-c)\langle\mu\rangle_{\mathbf{j}}N_{\mathbf{j},\mathbf{m}}[(1-c)\mathbf{I}+2c\mathbf{D}]^{-1}$$

$$\Big(\sum_{|\mathbf{i}|=l+1}\Delta_{\mathbf{i}}\frac{\langle(\theta_n')^{\mathbf{i}}\rangle}{\mathbf{i}!}+\sum_{|\mathbf{i}|=l+1}\mathbf{E}_{\mathbf{i}}\frac{\langle(\theta_n')^{\mathbf{i}}\rangle}{\mathbf{i}!}\Big)$$

$$\overset{k}{\sim}\mathbf{B}_0+\sum_{r=1}^{k-1}n^{-r/2}[(1-c)\mathbf{I}+2c\mathbf{D}]^{-1}\Big(2c\mathbf{D}\mathbf{H}_r$$

$$+(1-c)\Big(\sum_{|\mathbf{i}|=r}\Delta_{\mathbf{i}}\frac{\langle(\sum_{s=0}^{k-2}n^{-s/2}\mathbf{B}_s)^{\mathbf{i}}\rangle}{\mathbf{i}!}$$

$$+\sum_{|\mathbf{i}|=r+1}\mathbf{E}_{\mathbf{i}}\frac{\langle(\sum_{s=0}^{k-2}n^{-s/2}\mathbf{B}_s)^{\mathbf{i}}\rangle}{\mathbf{i}!}\Big)\Big)$$

$$+\sum_{|\mathbf{r}|=1}^{k-1}n^{-|\mathbf{r}|/2}\sum_{|\mathbf{j}|=2}^{3|\mathbf{r}|}N_{\mathbf{j},\mathbf{r}}[(1-c)\mathbf{I}+2c\mathbf{D}]^{-1}$$

$$\Big((1-c)\langle\mu\rangle_{\mathbf{j}}(\Delta_0-\mathbf{I}\sum_{s=0}^{k-2}n^{-s/2}\mathbf{B}_s)$$

$$-2c\mathbf{D}[\sum_{s=0}^{k-2}n^{-s/2}(\mathbf{B}_s-\mathbf{H}_s)\langle\mu\rangle_{\mathbf{j}}-\mu_{\mathbf{1},\mathbf{j}}]\Big)$$

$$+\sum_{r=2}^{k-1}n^{-r/2}\sum_{l+|\mathbf{m}|=r,|\mathbf{m}|\geq 1}(1-c)\sum_{|\mathbf{j}|=2}^{3|\mathbf{m}|}\langle\mu\rangle_{\mathbf{j}}N_{\mathbf{j},\mathbf{m}}[(1-c)\mathbf{I}+2c\mathbf{D}]^{-1}$$

$$\times\Big(\sum_{|\mathbf{i}|=l}\Delta_{\mathbf{i}}\frac{\langle(\sum_{s=0}^{k-2}n^{-s/2}\mathbf{B}_s)^{\mathbf{i}}\rangle}{\mathbf{i}!}+\sum_{|\mathbf{i}|=l+1}\mathbf{E}_{\mathbf{i}}\frac{\langle(\sum_{s=0}^{k-2}n^{-s/2}\mathbf{B}_s)^{\mathbf{i}}\rangle}{\mathbf{i}!}\Big).$$

Step 8. Collecting terms to get the final result. Similarly, as in Yuan (2009, p.2480),

$$\langle(\sum_{s=0}^{k-2}n^{-s/2}\mathbf{B}_s)^{\mathbf{i}}\rangle=\mathbf{i}!\sum_{s=0}^{k-2}n^{-s/2}\sum_{|\mathbf{l}|=s}\sum_{I_1(0,s,\mathbf{l},\mathbf{i})}\prod_{\nu=0}^{s}\frac{\langle\mathbf{B}_\nu^{\mathbf{i}_\nu}\rangle}{\mathbf{i}_\nu!},$$

plugging in the above into the previous expression, collecting terms in powers of $n^{-r/2}$, and notice that for $l\neq\mathbf{0}$, $I_1(0,0,l,i)$ is an empty set, we have

$$\sqrt{n}(\theta_n-\theta_0)\overset{k}{\sim}\mathbf{B}_0+\sum_{r=1}^{k-1}n^{-r/2}[(1-c)\mathbf{I}+2c\mathbf{D}]^{-1}\Big(2c\mathbf{D}\mathbf{H}_r$$

$$+(1-c)\sum_{t+s=r,t\geq 1}\Big[\sum_{|\mathbf{i}|=t}\Delta_{\mathbf{i}}\sum_{|\mathbf{l}|=s}\sum_{I_1(0,s,\mathbf{l},\mathbf{i})}\prod_{\nu=0}^{s}\frac{\langle\mathbf{B}_\nu^{\mathbf{i}_\nu}\rangle}{\mathbf{i}_\nu!}$$

$$+\sum_{|\mathbf{i}|=t+1}\mathbf{E}_{\mathbf{i}}\sum_{|\mathbf{l}|=s}\sum_{I_1(0,s,\mathbf{l},\mathbf{i})}\prod_{\nu=0}^{s}\frac{\langle\mathbf{B}_\nu^{\mathbf{i}_\nu}\rangle}{\mathbf{i}_\nu!}\Big]$$

$$+\sum_{|\mathbf{m}|=r}\sum_{|\mathbf{j}|=2}^{3r}N_{\mathbf{j},\mathbf{m}}\big[(1-c)\langle\mu\rangle_{\mathbf{j}}\Delta_0+2c\mathbf{D}\mu_{\mathbf{1},\mathbf{j}}\big]$$

$$-\sum_{t+s=r,t\geq 1}\sum_{|\mathbf{m}|=t}\sum_{|\mathbf{j}|=2}^{3t}N_{\mathbf{j},\mathbf{m}}\langle\mu\rangle_{\mathbf{j}}\big([(1-c)\mathbf{I}+2c\mathbf{D}]\mathbf{B}_s-2c\mathbf{D}\mathbf{H}_s\big)\Big)$$

$$+\sum_{r=2}^{k-1}n^{-r/2}[(1-c)\mathbf{I}+2c\mathbf{D}]^{-1}$$

$$\sum_{b+t+s=r,t\geq 2,}\ \sum_{b+a=t}\ \sum_{|\mathbf{m}|=a,|\mathbf{m}|\geq 1}\ \sum_{|\mathbf{j}|=2}^{3a}N_{\mathbf{j},\mathbf{m}}(1-c)\langle\mu\rangle_{\mathbf{j}}$$

$$\times\Big(\sum_{|\mathbf{i}|=b}\Delta_{\mathbf{i}}\sum_{|\mathbf{l}|=s}\sum_{I_1(0,s,\mathbf{l},\mathbf{i})}\prod_{\nu=0}^{s}\frac{\langle\mathbf{B}_\nu^{\mathbf{i}_\nu}\rangle}{\mathbf{i}_\nu!}$$

$$+\sum_{|\mathbf{i}|=b+1}\mathbf{E}_{\mathbf{i}}\sum_{|\mathbf{l}|=s}\sum_{I_1(0,s,\mathbf{l},\mathbf{i})}\prod_{\nu=0}^{s}\frac{\langle\mathbf{B}_\nu^{\mathbf{i}_\nu}\rangle}{\mathbf{i}_\nu!}\Big)$$

$$:=\sum_{r=0}^{k-1}n^{-r/2}\mathbf{B}_r.\qquad\square$$

Proof of Corollary 1. Recall $\mathbf{B}_0=\mathbf{H}_0=\mathbf{I}^{-1}\Delta_0$, and so Theorem 2 gives, for $r=1$,

$$\mathbf{B}_1=[(1-c)\mathbf{I}+2c\mathbf{D}]^{-1}\Big(2c\mathbf{D}\mathbf{H}_1$$

$$+(1-c)\sum_{|\mathbf{m}|=1}\sum_{|\mathbf{j}|=2}N_{\mathbf{j},\mathbf{m}}\langle\mu\rangle_{\mathbf{j}}\Delta_0+2c\sum_{|\mathbf{m}|=1}\sum_{|\mathbf{j}|=2}N_{\mathbf{j},\mathbf{m}}\mathbf{D}\mu_{\mathbf{1},\mathbf{j}}$$

$$-\sum_{|\mathbf{m}|=1}\sum_{|\mathbf{j}|=2}N_{\mathbf{j},\mathbf{m}}\langle\mu\rangle_{\mathbf{j}}\big([(1-c)\mathbf{I}+2c\mathbf{D}]\mathbf{B}_0-2c\mathbf{D}\mathbf{H}_0\big)$$

$$+(1-c)[\sum_{|\mathbf{i}|=1}\Delta_{\mathbf{i}}\langle\mathbf{B}_0^{\mathbf{i}}\rangle+\sum_{|\mathbf{i}|=2}\mathbf{E}_{\mathbf{i}}\frac{\langle\mathbf{B}_0^{\mathbf{i}}\rangle}{\mathbf{i}!}]\Big)$$

$$=[(1-c)\mathbf{I}+2c\mathbf{D}]^{-1}\Big(2c\mathbf{D}\mathbf{H}_1+2c\sum_{|\mathbf{m}|=1}\sum_{|\mathbf{j}|=2}N_{\mathbf{j},\mathbf{m}}\mathbf{D}\mu_{\mathbf{1},\mathbf{j}}$$

$$+(1-c)[\sum_{|\mathbf{i}|=1}\Delta_{\mathbf{i}}\langle\mathbf{B}_0^{\mathbf{i}}\rangle+\sum_{|\mathbf{i}|=2}\mathbf{E}_{\mathbf{i}}\frac{\langle\mathbf{B}_0^{\mathbf{i}}\rangle}{\mathbf{i}!}]\Big).$$

By the Lemma, for $|\mathbf{m}|=1$ and $|\mathbf{j}|=2$, $N_{\mathbf{j},\mathbf{m}}=\sum_{I_2(\mathbf{m},\mathbf{j})}\sum_{I_1(2,3,\mathbf{k},\mathbf{i})}\prod_{s=2}^{3}\frac{F_{s,1}^{|\mathbf{j}_s|}}{\mathbf{j}_s!(s!)^{|\mathbf{j}_2|}}$. Recall that $\mathbf{e}_i$ is the d-vector with i-th element be 1 and others be 0's. With $|\mathbf{m}|=1$ and $|\mathbf{j}|=2$, $I_2(\mathbf{m},\mathbf{j})$ is non-empty only when $(\mathbf{m},\mathbf{j})=(\mathbf{e}_i,2\mathbf{e}_i)$ for $i=1,...,d$. The corresponding $I_1(2,3,\mathbf{m},\mathbf{j})=\{(\mathbf{i}_2,\mathbf{i}_3)=\{(\mathbf{e}_i,\mathbf{0})\}$. So for given $|\mathbf{m}|=1$ and $|\mathbf{j}|=2$, $N_{\mathbf{j},\mathbf{m}}=\frac{F_{2,1}}{\mathbf{j}!}$. In the formula for $F_{2,1}$, the first summation is over an empty set; the second summation is $\sum_{|\mathbf{j}|=2}\Delta_{\mathbf{j}}\sum_{|\mathbf{i}|=2}\sum_{I_1(0,0,\mathbf{0},\mathbf{0})}1/\mathbf{i}_0!$. Since $I_1(0,0,\mathbf{0},\mathbf{0})=\{\mathbf{0}\}$, this term is $\sum_{|\mathbf{j}|=2}\Delta_{\mathbf{j}}$. The third summation is $\sum_{|\mathbf{j}|=3}\mathbf{E}_{\mathbf{j}}\sum_{|\mathbf{i}|=2}\sum_{I_1(0,0,\mathbf{0},\mathbf{j}-\mathbf{i})}\langle\mathbf{H}_0^{\mathbf{i}_0}\rangle/\mathbf{i}_0!=\sum_{|\mathbf{j}|=3}\mathbf{E}_{\mathbf{j}}\sum_{|\mathbf{i}|=2}\langle\mathbf{H}_0^{\mathbf{j}-\mathbf{i}}\rangle/(\mathbf{j}-\mathbf{i})!$. This gives, with $\mathbf{i}\le\mathbf{j}$,

$$N_{\mathbf{j},\mathbf{m}}=\frac{1}{2}\sum_{|\mathbf{j}|=1}\Delta_{\mathbf{j}}+\frac{1}{2}\sum_{|\mathbf{j}|=2}\mathbf{E}_{\mathbf{j}}\sum_{|\mathbf{i}|=2}\langle(\mathbf{I}^{-1}\Delta_0)^{\mathbf{j}-\mathbf{i}}\rangle/(\mathbf{j}-\mathbf{i})!.$$

However, it is easy to see that for $|\mathbf{j}|=2$, $\mu_{1,\mathbf{j}}=\mathbf{0}$, so we get

$$\mathbf{B}_1=[(1-c)\mathbf{I}+2c\mathbf{D}]^{-1}\Big(2c\mathbf{D}\mathbf{H}_1$$

$$+(1-c)[\sum_{|\mathbf{i}|=1}\Delta_{\mathbf{i}}\langle\mathbf{B}_0^{\mathbf{i}}\rangle+\sum_{|\mathbf{i}|=2}\mathbf{E}_{\mathbf{i}}\frac{\langle\mathbf{B}_0^{\mathbf{i}}\rangle}{\mathbf{i}!}]\Big),$$

where

$$\mathbf{H}_1=\mathbf{I}^{-1}\Big(\sum_{|\mathbf{i}|=2}E_{\mathbf{i}}\frac{\langle\mathbf{B}_0^{\mathbf{i}}\rangle}{\mathbf{i}!}+\sum_{|\mathbf{i}|=1}\Delta_{\mathbf{i}}\langle\mathbf{B}_0^{\mathbf{i}}\rangle+\rho_0\Big)$$

is the multivariate version of Gusev (1975, p.496). This gives

$$\mathbf{H}_1=\mathbf{I}^{-1}\Big(\sum_{|\mathbf{i}|=2}E_{\mathbf{i}}\frac{\langle\mathbf{B}_0^{\mathbf{i}}\rangle}{\mathbf{i}!}+\sum_{|\mathbf{i}|=1}\Delta_{\mathbf{i}}\langle\mathbf{B}_0^{\mathbf{i}}\rangle\Big)$$

$$+[(1-c)\mathbf{I}+2c\mathbf{D}]^{-1}2c\mathbf{D}\mathbf{I}^{-1}\rho_0.$$

Note that, when $d=1$, our notation $(E_2,\Delta_1,\Delta_0,\rho_0)$ here correspond to $(E_3,\Delta_2,\Delta_1,\rho_{0,1})$ in Gusev (1975). □

Proof of Proposition 1. By the Proposition in Yuan (2009), asymptotically we have

$$E_{\theta_0}[\sum_{|\mathbf{i}|=1}\Delta_{\mathbf{i}}\langle(\mathbf{I}^{-1}\Delta_0)^{\mathbf{i}}\rangle+\sum_{|\mathbf{i}|=2}\mathbf{E}_{\mathbf{i}}\frac{\langle(\mathbf{I}^{-1}\Delta_0)^{\mathbf{i}}\rangle}{\mathbf{i}!}]$$

$$=\sum_{j=1}^{d}\mathbf{A}_j\mathbf{I}_j^{-1}+\sum_{i,j=1}^{d}\mathbf{E}_{1_i+1_j}\ _i\mathbf{I}^{-1}\mathbf{I}^{-1}\mathbf{I}_j^{-1}/(1_i+1_j),$$

here our notations $\mathbf{A}_j$ and 1_j correspond to $\mathbf{D}_j$ and $\mathbf{e}_j$ there, this gives the result. □

Proof of Theorem 3. Under the given conditions the validity of the results are justified in Chibisov (1973) and Gusev (1976). Below we give such detailed constructions of the terms in the EE. Recall by Theorem 2,

$$\sqrt{n}(\theta_n-\theta_0)=\sum_{r=0}^{k-1}n^{-r/2}\mathbf{B}_r+O_p(n^{-k/2}),$$

where $\mathbf{B}_r$ is a polynomial of degree r, of the random vectors $\Delta_{\mathbf{i}}$ ($|\mathbf{i}|=0,1,...,r$) and of the random variables $\delta_{\mathbf{i}}$'s ($|\mathbf{i}|=r,r+1$), but the $\delta_{\mathbf{i}}$'s are components of some $\Delta_{\mathbf{j}}$'s with $|\mathbf{j}|<|\mathbf{i}|$. The coefficients of the $\mathbf{B}_r$'s depend on the $\mathbf{E}_{\mathbf{i}}$'s, $e_{\mathbf{i}}$s, $\langle\mu\rangle_{\mathbf{i}}$'s, $\mu_{\mathbf{1},\mathbf{i}}$ ($|\mathbf{i}|=0,1,...,r+1$) and $\mathbf{D}$. In summary, $\sum_{r=0}^{k-1}n^{-r/2}\mathbf{B}_r$ is a d_k-variate polynomial of degree k, of the random vector

$$\mathbf{V}=(\Delta_{\mathbf{i}}:|\mathbf{i}|=0,1,...,k-1),\quad\text{and let}\quad\Omega=Cov_{\theta_0}(\mathbf{V}).$$

Thus each $\mathbf{B}_r$ is a function of $\mathbf{V}$ and denoted as $\mathbf{B}_r(\mathbf{V})$ ($r=0,1,...,k-1$), although $\mathbf{B}_0=\mathbf{I}^{-1}\Delta_0$ is a function only of Δ_0, and $\mathbf{B}_r$ is only a function of ($\Delta_{\mathbf{i}}:|\mathbf{i}|\le r$). For each fixed $\mathbf{i}$, $\Delta_{\mathbf{i}}$ is of dimension d; for $|\mathbf{i}|=r$, there are d^r different $\Delta_{\mathbf{i}}$'s. Thus $\mathbf{V}$ has dimension $d_k=k$ when $d=1$, and $d_k=d\sum_{r=0}^{k-1}d^r=d(d^k-1)/(d-1)$ when $d>1$. Note that $E_{\theta_0}(\mathbf{V})=\mathbf{0}$. Let $\phi_\Omega(\cdot)$ and $\Phi_\Omega(\cdot)$ be the density and distribution function of the normal random vector with mean zero and covariance matrix Ω, and

$$\mathbf{X}=\sum_{r=0}^{k-1}n^{-r/2}\mathbf{B}_r(\mathbf{V}).$$

Let $g_n(\mathbf{x})$ and $G_n(\mathbf{x})$ be the density and distribution functions of $\mathbf{X}$. By results in Chibisov (1972,1973) and Gusev (1975), we have

$$\sup_{\mathbf{x}}|f_n(\mathbf{x})-g_n(\mathbf{x})|=o(n^{-k/2}),$$

$$\sup_{\mathbf{x}}|F_n(\mathbf{x})-G_n(\mathbf{x})|=o(n^{-k/2}).$$

So we only need to compute the EE for $g_n(\mathbf{x})$ and $G_n(\mathbf{x})$ of $\mathbf{X}$.

We first compute the EE for $g_n(\mathbf{x})$, and in turn we need the EE for the density function $q_n(\mathbf{v})$ of $\mathbf{V}$. Denote $\mathbf{Y}=(\Delta_{\mathbf{i}}:|\mathbf{i}|=1,...,k-1)$, thus $\mathbf{V}=(\Delta_0,\mathbf{Y})$. By Theorem 19.2 in Battacharya and Rao (1986, p.192), with the given conditions we have, uniformly over $\mathbf{v}=(\Delta_0,\mathbf{y})$,

$$q_n(\mathbf{v})=\phi_\Omega(\Delta_0,\mathbf{y})\big[1+\sum_{r=1}^{k-1}n^{-r/2}P_r(\Delta_0,\mathbf{y})\big]$$

$$(A.1) \qquad + \frac{O(n^{-k/2})}{1+||(\Delta_0,\mathbf{y})||^{k+1}},$$

where

$$P_r(\Delta_0,\mathbf{y}) = -\sum_{|\nu|=r+2}\frac{\chi_\nu}{\nu!}p_v(\Delta_0,\mathbf{y}),$$

$p_v(\Delta_0,\mathbf{y})$ is the r-th Hermit polynomial for $\phi_\Omega(\cdot)$, given by $\phi_\Omega^{(\nu)}(\Delta_0,\mathbf{y}) = p_v(\Delta_0,\mathbf{y})\phi_\Omega(\Delta_0,\mathbf{y})$, and χ_ν is the ν-th cummunant of $\mathbf{V}$, ν is a d_k-dimensional non-negative integer vector.

Our strategy is to compute the EE of $g_n(\cdot)$ from that of $q_n(\cdot)$ as given in (A.1), via the following transformation

$$\begin{cases} \mathbf{x} = \mathbf{I}^{-1}\Delta_0 + \sum_{r=1}^{k-1} n^{-r/2}\mathbf{B}_r(\Delta_0,\mathbf{y}), \\ \mathbf{y} = \mathbf{y}. \end{cases}$$

$$\text{or} \quad \begin{cases} \Delta_0 = \mathbf{I}\mathbf{x} - \sum_{r=1}^{k-1} n^{-r/2}\mathbf{I}\mathbf{B}_r(\Delta_0,\mathbf{y}), \\ \mathbf{y} = \mathbf{y}. \end{cases}$$

However in the above transformation, Δ_0 cannot be solved out as function of $(\mathbf{x},\mathbf{y})$ explicitly, and the transformation Jacobin is difficult to work with, as the $\mathbf{B}_r$'s are polynomials of Δ_0 and $\mathbf{y}$. Thus we cannot get the density of $\mathbf{X}$ directly from the above transformation.

To overcome these difficulties, Chibisov (1972) used the following method and proved its validity. First replace Δ_0 in $\mathbf{B}_1,...,\mathbf{B}_{k-1}$ by a random vector $\mathbf{u}=\mathbf{u}(\delta)$ different from Δ_0, $\mathbf{u}$ and Δ_0 have the same dimension. Then in the transformation we have

$$\begin{cases} \Delta_0 = \mathbf{I}\mathbf{x} - \mathbb{B}(\mathbf{u},\mathbf{y}), \\ \mathbf{y}=\mathbf{y}, \quad \mathbf{u}=\mathbf{u}. \end{cases} \qquad \text{with } \mathbb{B}(\mathbf{u},\mathbf{y}) = \sum_{r=1}^{k-1} n^{-r/2}\mathbf{I}\mathbf{B}_r(\mathbf{u},\mathbf{y}),$$

in which Δ_0 is solved out explicitly, the transformation Jacobin is $|\mathbf{I}|$, and the EE of $g_n(\cdot)$ on $\mathbf{x}$ will be easily obtained by plugging in the above relationships into (A.1) and integrate with respect to $(\mathbf{u},\mathbf{y})$, expanding and collecting terms in powers of $n^{-1/2}$ will give EE for $\mathbf{x}$ from that of $(\Delta_0,\mathbf{u},\mathbf{y})$. Then let $\mathbf{u}=\mathbf{u}(\delta)\to\Delta_0$ as $\delta\to 0$, to get the EE on $\mathbf{x}$ from that of $(\Delta_0,\mathbf{y})$, the limit does not depend on $\pi(\cdot)$, and gives us the desired result.

Specifically, let $\mathbf{u}=\mathbf{u}(\delta)=\Delta_0+\delta\mathbf{w}$, with $\mathbf{w}$ independent of $\mathbf{V}$ and having density $\pi(\cdot)$ with various order of partial derivatives, and with support on some compact set, say on $||\mathbf{w}||\le 1$. By (A.1), the above transformation and Lemma 2 in Chibisov (1972) we have

$$g_n(\mathbf{x}) = \lim_{\delta\to 0} J_{n,k}(\delta) + O(n^{-k/2}),$$

i.e., the limit $\lim_{\delta\to 0} J_{n,k}(\delta)$ is the k-th order EE for $g_n(\cdot)$. Since $f_n(\cdot) = g_n(\cdot) + o(n^{-k/2})$, the limit $\lim_{\delta\to 0} J_{n,k}(\delta)$ is also the k-th order EE for $f_n(\cdot)$, where

$$J_{n,k}(\delta) = |\mathbf{I}|\int\int\Big(1+\sum_{r=1}^{k-1} n^{-r/2}P_r\big(\mathbf{I}\mathbf{x}-\mathbb{B}(\mathbf{u},\mathbf{y}),\mathbf{y}\big)\Big)$$

$$\times\phi_\Omega(\mathbf{I}\mathbf{x}-\mathbb{B}(\mathbf{u},\mathbf{y}),\mathbf{y})\pi(\frac{\mathbf{u}-\mathbf{I}\mathbf{x}+\mathbb{B}(\mathbf{u},\mathbf{y})}{\delta})\frac{1}{\delta^d}d\mathbf{u}d\mathbf{y}.$$

Note Lemma 2 of Chibisov is for integration over $\mathbf{x}$ on any set A, take $A=(-\infty,\mathbf{x}]$ we get the result for distribution function, and then take derivative with respect to $\mathbf{x}$ to get result for density function.

Expanding $P_r\big(\mathbf{I}\mathbf{x}-\mathbb{B}(\mathbf{u},\mathbf{y}),\mathbf{y}\big)$ and $\phi_\Omega(\mathbf{I}\mathbf{x}-\mathbb{B}(\mathbf{u},\mathbf{y}),\mathbf{y})$ at $\mathbf{I}\mathbf{x}$ in powers of $n^{-1/2}$, and expanding $\pi(\frac{\mathbf{u}-\mathbf{I}\mathbf{x}+\mathbb{B}(\mathbf{u},\mathbf{y})}{\delta})$ at $(\mathbf{u}-\mathbf{I}\mathbf{x})/\delta$ in powers of $n^{-1/2}$, i.e.

$$P_r\big(\mathbf{I}\mathbf{x}-\mathbb{B}(\mathbf{u},\mathbf{y}),\mathbf{y}\big) = \sum_{s=0}^{k-1} n^{-s/2}P_{r,s}(\mathbf{I}\mathbf{x},\mathbf{u},\mathbf{y}) + O(n^{-k/2}),$$

with $P_{r,0}(\mathbf{I}\mathbf{x},\mathbf{u},\mathbf{y}) = P_r\big(\mathbf{I}\mathbf{x},\mathbf{y}\big)$ for $r>0$, $P_{0,0}(\mathbf{I}\mathbf{x},\mathbf{u},\mathbf{y})\equiv 1$, and $P_{0,l}(\mathbf{I}\mathbf{x},\mathbf{u},\mathbf{y})\equiv 0$ for $l>0$. In the above expansion the powers on $\mathbf{u}$ comes from $\mathbf{B}_r(\mathbf{u},\mathbf{y})$, not from $P_r(\mathbf{I}\mathbf{x},\mathbf{y})$ nor its derivatives; and $\mathbf{B}_r(\mathbf{u},\mathbf{y})$ is a polynomial of degree $r+1$, so the power of $\mathbf{u}$ in $Q_{r,s}(\mathbf{I}\mathbf{x},\mathbf{u},\mathbf{y})$ is at most some finite number;

$$\phi_\Omega(\mathbf{I}\mathbf{x}-\mathbb{B}(\mathbf{u},\mathbf{y}),\mathbf{y}) = \phi_\Omega(\mathbf{I}\mathbf{x},\mathbf{y})\Bigg(1+$$

$$\sum_{|\mathbf{i}|=1}^{k-1} p_{\mathbf{i}}(\mathbf{I}\mathbf{x},\mathbf{y})\frac{\langle\mathbb{B}^{\mathbf{i}}(\mathbf{u},\mathbf{y})\rangle}{\mathbf{i}!} + O(n^{-k/2})\Bigg)$$

$$= \phi_\Omega(\mathbf{I}\mathbf{x},\mathbf{y})\Bigg(1+\sum_{|\mathbf{i}|=1}^{k-1} p_{\mathbf{i}}(\mathbf{I}\mathbf{x},\mathbf{y})\sum_{s=1}^{k-1} n^{-s/2}$$

$$\sum_{|\mathbf{l}|=s}\sum_{(1,s,\mathbf{l},\mathbf{i})}\prod_{v=1}^{s}\frac{\langle\big(\mathbf{I}\mathbf{B}_v(\mathbf{u},\mathbf{y})\big)^{\mathbf{i}_v}\rangle}{\mathbf{i}_v!} + O(n^{-k/2})\Bigg)$$

$$= \phi_\Omega(\mathbf{I}\mathbf{x},\mathbf{y})\Bigg(1+\sum_{s=1}^{k-1} n^{-s/2}$$

$$\sum_{|\mathbf{i}|=1}^{k-1} p_{\mathbf{i}}(\mathbf{I}\mathbf{x},\mathbf{y})\sum_{|\mathbf{l}|=s}\sum_{(1,s,\mathbf{l},\mathbf{i})}\prod_{v=1}^{s}\frac{\langle\big(\mathbf{I}\mathbf{B}_v(\mathbf{u},\mathbf{y})\big)^{\mathbf{i}_v}\rangle}{\mathbf{i}_v!} + O(n^{-k/2})\Bigg)$$

$$:= \phi_\Omega(\mathbf{I}\mathbf{x},\mathbf{y})\Bigg(1+\sum_{s=1}^{k-1} n^{-s/2}L_s(\mathbf{I}\mathbf{x},\mathbf{u},\mathbf{y}) + O(n^{-k/2})\Bigg),$$

in the above $\mathbf{i}$ is a d-dimensional non-negative integer vector, and $p_{\mathbf{i}}$ is p_ν given in (A.1) with $\nu=(\mathbf{i},\mathbf{0})$; Since $B_v(\mathbf{u},\mathbf{y})$ is a polynomial in $(\mathbf{u},\mathbf{y})$ of degree $v+1$, so the power of $\mathbf{u}$ in $L_s(\mathbf{I}\mathbf{x},\mathbf{u},\mathbf{y})$ is at most $(s^2+3s)/2$; Also,

$$\pi(\frac{\mathbf{u}-\mathbf{I}\mathbf{x}+\mathbb{B}(\mathbf{u},\mathbf{y})}{\delta}) = \pi(\frac{\mathbf{u}-\mathbf{I}\mathbf{x}}{\delta})+$$

$$\sum_{|\mathbf{i}|=1}^{k-1}\frac{1}{\mathbf{i}!}\pi^{(\mathbf{i})}(\frac{\mathbf{u}-\mathbf{Ix}}{\delta})\frac{1}{\delta^{|\mathbf{i}|}}\langle\mathbb{B}^{\mathbf{i}}(\mathbf{u},\mathbf{y})\rangle+O(n^{-k/2})$$

$$=\pi(\frac{\mathbf{u}-\mathbf{Ix}}{\delta})+\sum_{|\mathbf{i}|=1}^{k-1}\pi^{(\mathbf{i})}(\frac{\mathbf{u}-\mathbf{Ix}}{\delta})\frac{1}{\delta^{|\mathbf{i}|}}\sum_{s=|\mathbf{i}|}^{k-1}n^{-s/2}$$

$$\sum_{|\mathbf{l}|=s}\sum_{(1,s,\mathbf{l},\mathbf{i})}\prod_{v=1}^{s}\frac{\langle\left(\mathbf{IB}_v(\mathbf{u},\mathbf{y})\right)^{\mathbf{i}_v}\rangle}{\mathbf{i}_v!}+O(n^{-k/2})$$

$$=\pi(\frac{\mathbf{u}-\mathbf{Ix}}{\delta})+\sum_{s=1}^{k-1}n^{-s/2}\sum_{|\mathbf{i}|=1}^{s}\pi^{(\mathbf{i})}(\frac{\mathbf{u}-\mathbf{Ix}}{\delta})\frac{1}{\delta^{|\mathbf{i}|}}$$

$$\sum_{|\mathbf{l}|=s}\sum_{(1,s,\mathbf{l},\mathbf{i})}\prod_{v=1}^{s}\frac{\langle\left(\mathbf{IB}_v(\mathbf{u},\mathbf{y})\right)^{\mathbf{i}_v}\rangle}{\mathbf{i}_v!}+O(n^{-k/2})$$

$$:=\pi(\frac{\mathbf{u}-\mathbf{Ix}}{\delta})+\sum_{s=1}^{k-1}n^{-s/2}\sum_{|\mathbf{i}|=1}^{s}\pi^{(\mathbf{i})}(\frac{\mathbf{u}-\mathbf{Ix}}{\delta})\delta^{-|\mathbf{i}|}M_{\mathbf{i},s}(\mathbf{u},\mathbf{y})$$

$$+O(n^{-k/2}),$$

where the $Q_{r,s}(\cdot,\cdot,\cdot)$'s and $L_s(\cdot,\cdot,\cdot)$'s are polynomials in $(\mathbf{Ix},\mathbf{u},\mathbf{y})$; and the $M_{\mathbf{i},s}(\cdot,\cdot)$'s are polynomials in $(\mathbf{u},\mathbf{y})$; and $\pi^{(\mathbf{i})}(\cdot)$ is the $\mathbf{i}$-th derivative of $\pi(\cdot)$. Define $L_0(\cdot,\cdot,\cdot)=M_{\mathbf{0},0}(\cdot,\cdot)\equiv 1$.

Now $J_{n,k}(\delta)$ can be written in the form, omit term of order $O(n^{-k/2})$,

$$J_{n,k}(\delta)=|\mathbf{I}|\int\int\left(\sum_{r=0}^{k-1}n^{-r/2}\sum_{l+s=r}^{l,s\geq 0}P_{l,s}(\mathbf{Ix},\mathbf{u},\mathbf{y})\right)$$

$$\left(\sum_{r=0}^{k-1}n^{-r/2}L_r(\mathbf{Ix},\mathbf{u},\mathbf{y})\right)$$

$$\times\left(\sum_{r=0}^{k-1}n^{-r/2}\sum_{|\mathbf{i}|=0}^{r}\pi^{(\mathbf{i})}(\frac{\mathbf{u}-\mathbf{Ix}}{\delta})\delta^{-|\mathbf{i}|}M_{\mathbf{i},r}(\mathbf{u},\mathbf{y})\right)\phi_\Omega(\mathbf{Ix},\mathbf{y})d\mathbf{u}d\mathbf{y}$$

$$=|\mathbf{I}|\int\int\left(\sum_{r=0}^{k-1}n^{-r/2}\sum_{l+s+j+t=r}^{l,s,j,t\geq 0}P_{l,s}(\mathbf{Ix},\mathbf{u},\mathbf{y})L_j(\mathbf{u},\mathbf{y})\right.$$

$$\left.\times\sum_{|\mathbf{i}|=0}^{t}M_{\mathbf{i},t}(\mathbf{u},\mathbf{y})\pi^{(\mathbf{i})}(\frac{\mathbf{u}-\mathbf{Ix}}{\delta})\frac{1}{\delta^{|\mathbf{i}|+d}}\right)\phi_\Omega(\mathbf{Ix},\mathbf{y})d\mathbf{u}d\mathbf{y}$$

$$=|\mathbf{I}|\iint\left(\pi(\frac{\mathbf{u}-\mathbf{Ix}}{\delta})\frac{1}{\delta^d}+\sum_{r=1}^{k-1}n^{-r/2}\sum_{l+s+j+t=r}^{l,s,j,t\geq 0}P_{l,s}(\mathbf{Ix},\mathbf{u},\mathbf{y})L_j(\mathbf{u},\mathbf{y})\right.$$

$$\left.\times\sum_{|\mathbf{i}|=1}^{t}M_{\mathbf{i},t}(\mathbf{u},\mathbf{y})\pi^{(\mathbf{i})}(\frac{\mathbf{u}-\mathbf{Ix}}{\delta})\frac{1}{\delta^{|\mathbf{i}|+d}}\right)\phi_\Omega(\mathbf{Ix},\mathbf{y})d\mathbf{u}d\mathbf{y}$$

$$=|\mathbf{I}|\iint\left(\pi(\frac{\mathbf{u}-\mathbf{Ix}}{\delta})\frac{1}{\delta^d}+\sum_{r=1}^{k-1}n^{-r/2}S_r(\mathbf{Ix},\mathbf{u},\mathbf{y})\right)\phi_\Omega(\mathbf{Ix},\mathbf{y})d\mathbf{u}d\mathbf{y},$$

where

$$S_r(\mathbf{Ix},\mathbf{u},\mathbf{y})=\sum_{l+s+j+t=r}^{l,s,j,t\geq 0}P_{l,s}(\mathbf{Ix},\mathbf{u},\mathbf{y})L_j(\mathbf{u},\mathbf{y})$$

$$\sum_{|\mathbf{i}|=1}^{t}M_{\mathbf{i},t}(\mathbf{u},\mathbf{y})\pi^{(\mathbf{i})}(\frac{\mathbf{u}-\mathbf{Ix}}{\delta})\frac{1}{\delta^{|\mathbf{i}|+d}}.$$

Since $P_{l,s}(\mathbf{Ix},\mathbf{u},\mathbf{y})$, $L_s(\mathbf{u},\mathbf{y})$ and $M_{\mathbf{i},t}(\mathbf{u},\mathbf{y})$ are polynomials, $S_{r,\mathbf{i}}(\mathbf{Ix},\mathbf{u},\mathbf{y})$ is a summation of terms of the form $C(\mathbf{Ix})^{\mathbf{a}}\mathbf{u}^{\mathbf{b}}\mathbf{y}^{\mathbf{c}}$, and the above integration over $\mathbf{u}$ is a summation of terms of the form

$$C\langle(\mathbf{Ix})^{\mathbf{a}}\rangle\langle\mathbf{y}^{\mathbf{c}}\rangle\int\langle(\mathbf{I}^{-1}\mathbf{u})^{\mathbf{b}}\rangle\pi^{(\mathbf{j})}(\frac{\mathbf{u}-\mathbf{Ix}}{\delta})\frac{1}{\delta^{|\mathbf{j}|+d}}d\mathbf{u}.$$

Using integration by parts, note $\pi(\cdot)$ has compact support $\{\mathbf{u}:||\mathbf{u}||\leq 1\}$, and for small δ, $(\pm\mathbf{1}-\mathbf{Ix})/\delta$ lies outside of this support, thus $\langle(\mathbf{I}^{-1}\mathbf{u})^{\mathbf{b}-\mathbf{l}}\rangle\pi^{(\mathbf{j}-\mathbf{l})}(\frac{\mathbf{u}-\mathbf{Ix}}{\delta})\big|_{||\mathbf{u}||=1}=0$, and that for $\mathbf{b}\geq\mathbf{j}$ (means componentwise greater or equal)

$$\lim_{\delta\to 0}\int\langle(\mathbf{I}^{-1}\mathbf{u})^{\mathbf{b}-\mathbf{j}}\rangle\pi(\frac{\mathbf{u}-\mathbf{Ix}}{\delta})\frac{1}{\delta^d}d\mathbf{u}=\langle\mathbf{x}^{\mathbf{b}-\mathbf{j}}\rangle,$$

we have

$$\int\langle(\mathbf{I}^{-1}\mathbf{u})^{\mathbf{b}}\rangle\pi^{(\mathbf{j})}(\frac{\mathbf{u}-\mathbf{Ix}}{\delta})\frac{1}{\delta^{|\mathbf{j}|+d}}d\mathbf{u}$$

$$=\begin{cases}(-1)^{|\mathbf{j}|}\frac{\mathbf{b}!}{(\mathbf{b}-\mathbf{j})!}\langle\mathbf{x}^{\mathbf{b}-\mathbf{j}}\rangle, & \text{if } \mathbf{b}\geq\mathbf{j},\\ 0, & \text{else.}\end{cases}$$

Thus, we get, with $T_r(\mathbf{Ix},\mathbf{y})$ being $S_r(\mathbf{Ix},\mathbf{u},\mathbf{y})$ with factors of the form $\langle(\mathbf{I}^{-1}\mathbf{u})^{\mathbf{b}}\rangle$ in $Q_{l,s}(\mathbf{Ix},\mathbf{u},\mathbf{y})L_s(\mathbf{u},\mathbf{y})M_{\mathbf{i},t}(\mathbf{u},\mathbf{y})$ replaced by $(-1)^{|\mathbf{i}|}\frac{\mathbf{b}!}{(\mathbf{b}-\mathbf{i})!}\langle\mathbf{x}^{\mathbf{b}-\mathbf{i}}\rangle$ if $\mathbf{b}\geq\mathbf{i}$, and 0 otherwise (especially, $Q_{l,s}(\mathbf{Ix},\mathbf{u},\mathbf{y})L_s(\mathbf{u},\mathbf{y})$ is a special case with $\mathbf{i}=\mathbf{0}$),

$$\lim_{\delta\to 0}J_{n,k}(\delta)=|\mathbf{I}|\int\left(1+\sum_{r=1}^{k-1}n^{-r/2}T_r(\mathbf{Ix},\mathbf{y})\right)\phi_\Omega(\mathbf{Ix},\mathbf{y})d\mathbf{y},$$

and consequently, uniformly over $\mathbf{x}$,

$$f_n(\mathbf{x})=|\mathbf{I}|\phi_{\Omega_1}(\mathbf{Ix})+\sum_{r=1}^{k-1}n^{-r/2}E[T_r(\mathbf{Ix},\mathbf{Y})]|\mathbf{I}|\phi_{\Omega_1}(\mathbf{Ix})$$

$$+O(n^{-k/2}),$$

where Ω_1 is the sub-block d-dimensional matrix in Ω corresponding to Δ_0, the expectation is for $\mathbf{Y}\sim N(\mathbf{0},\Omega_2)$, Ω_2

is the sub-block matrix in Ω corresponding to $(\Delta_{\mathbf{i}} : |\mathbf{i}| = 1, ..., k-1)$, and for each r, $g_r(\cdot)$ is a polynomial. Note that $\Omega_1 = Cov_{\theta_0}(\Delta_0) = \mathbf{I}$, and that $|\mathbf{I}|\phi_{\mathbf{I}}(\mathbf{I}\mathbf{x}) = \phi_{\mathbf{I}^{-1}}(\mathbf{x})$, so we have

$$f_n(\mathbf{x}) = \phi_{\mathbf{I}^{-1}}(\mathbf{x}) + \sum_{r=1}^{k-1} n^{-r/2} E[T_r(\mathbf{I}\mathbf{x}, \mathbf{Y})]\phi_{\mathbf{I}^{-1}}(\mathbf{x}) + O(n^{-k/2})$$

$$:= \phi_{\mathbf{I}^{-1}}(\mathbf{x}) + \sum_{r=1}^{k-1} n^{-r/2} g_r(\mathbf{x})\phi_{\mathbf{I}^{-1}}(\mathbf{x}) + O(n^{-k/2}),$$

The EE for $F_n(\cdot)$ is obtained by integrating that for $f_n(\cdot)$ and its validity is given in Theorem 1 of Chibisov (1972). □

Proof of Theorem 4. Let $\theta_{0,n}$ be the compound estimator under H_0 (i.e., its first r components are fixed as given in Θ_0), $G^{(\mathbf{1})}(\theta|\mathbf{x}^n) = \partial G(\theta|\mathbf{x}^n)/\partial\theta$, $G^{(\mathbf{2})}(\theta|\mathbf{x}^n) = \partial G(\theta|\mathbf{x}^n)/(\partial\theta\partial\theta')$, $W^{(\mathbf{2})}(\cdot)$ be the matrix of second partial derivatives of $W(\cdot)$, and $\tilde{\theta}_n$ be between θ_n and $\theta_{0,n}$. Note by definition of θ_n, $G^{(\mathbf{1})}(\theta_n|\mathbf{x}^n) = \mathbf{0}$, so we have

$$\lambda_{c,n} = -2\Big(G(\theta_{0,n}|\mathbf{x}^n) - G(\theta_n|\mathbf{x}^n)\Big)$$

$$= -2G^{(\mathbf{1})'}(\theta_n|\mathbf{x}^n)(\theta_{0,n} - \theta_n)$$

$$-(\theta_n - \theta_{0,n})' G^{(\mathbf{2})}(\tilde{\theta}_n|\mathbf{x}^n)(\theta_n - \theta_{0,n})$$

$$= -(\theta_n - \theta_{0,n})' G^{(\mathbf{2})}(\tilde{\theta}_n|\mathbf{x}^n)(\theta_n - \theta_{0,n})$$

$$= n(1-c)(\theta_n - \theta_{0,n})'[-n^{-1} l^{(\mathbf{2})}(\tilde{\theta}_n|\mathbf{x}^n)](\theta_n - \theta_{0,n})$$

$$+cn \int (\theta_n - \theta_{0,n})' W^{(\mathbf{2})}(\tilde{\theta}_n - \alpha)(\theta_n - \theta_{0,n})\pi(\alpha|\mathbf{x}^n) d\alpha.$$

Note $\tilde{\theta}_n \to \theta_0$ (a.s.), $\sqrt{n}(\theta_n - \theta_{0,n}) = O_p(1)$, and $\pi(\alpha|\mathbf{x}^n)$ is asymptotically $N(\hat{\theta}_n, n^{-1}\mathbf{I}^{-1})$ in probability, for some small $\delta > 0$, on $||\tilde{\theta}_n - \alpha|| \leq n^{-\delta}$, we can replace $W^{(\mathbf{2})}(\tilde{\theta}_n - \alpha)$ by $2\mathbf{D}$, so we have

$$\lambda_{c,n} = n(1-c)(\theta_n - \theta_{0,n})'\mathbf{I}(\theta_n - \theta_{0,n})$$

$$+cn \int_{||\tilde{\theta}_n - \alpha|| \leq n^{-\delta}} (\theta_n - \theta_{0,n})' W^{(\mathbf{2})}(\tilde{\theta}_n - \alpha)(\theta_n - \theta_{0,n})\pi(\alpha|\mathbf{x}^n) d\alpha$$

$$+o_p(1)$$

$$= n(1-c)(\theta_n - \theta_{0,n})'\mathbf{I}(\theta_n - \theta_{0,n})$$

$$+2cn \int_{||\tilde{\theta}_n - \alpha|| \leq n^{-\delta}} (\theta_n - \theta_{0,n})' \mathbf{D}(\theta_n - \theta_{0,n})\pi(\alpha|\mathbf{x}^n) d\alpha + o_p(1)$$

$$= n(1-c)(\theta_n - \theta_{0,n})'\mathbf{I}(\theta_n - \theta_{0,n})$$

$$+2cn \int (\theta_n - \theta_{0,n})' \mathbf{D}(\theta_n - \theta_{0,n})\pi(\alpha|\mathbf{x}^n) d\alpha + o_p(1)$$

$$= n(\theta_n - \theta_{0,n})'[(1-c)\mathbf{I} + 2c\mathbf{D}](\theta_n - \theta_{0,n}) + o_p(1).$$

By Theorem 2, $\sqrt{n}(\theta_n - \theta_0) = \mathbf{I}^{-1}\Delta_0 + o_p(1)$. Let $\mathbf{I}_0$ be the $r \times r$ Fisher information under H_0, and $\tilde{\mathbf{I}}$ be the $d \times d$ matrix with the lower right $r \times r$ block be $\mathbf{I}_0^{-1}$ and other elements be 0's. Then we have $\sqrt{n}(\theta_{0,n} - \theta_0) = \tilde{\mathbf{I}}\Delta_0 + o_p(1)$. Note $\Delta_0 \xrightarrow{D} N(\mathbf{0}, \mathbf{I})$. Thus we have

$$\lambda_{c,n} = \Delta_0'(\mathbf{I}^{-1} - \tilde{\mathbf{I}})[(1-c)\mathbf{I} + 2c\mathbf{D}](\mathbf{I}^{-1} - \tilde{\mathbf{I}})\Delta_0 + o_p(1)$$

$$= \mathbf{Z}'\mathbf{B}\mathbf{Z} + o_p(1),$$

where $\mathbf{Z} \sim N(\mathbf{0}, I)$, $\mathbf{B} = (\mathbf{I}^{-1/2})'(\mathbf{I}^{-1} - \tilde{\mathbf{I}})[(1-c)\mathbf{I} + 2c\mathbf{D}](\mathbf{I}^{-1} - \tilde{\mathbf{I}})\mathbf{I}^{-1/2}$, $(\mathbf{I}^{-1/2})'\mathbf{I}^{-1/2} = \mathbf{I}^{-1}$. Especially, if we take $c = 0$, then $\mathbf{B} = (\mathbf{I}^{-1/2})'(\mathbf{I}^{-1} - \tilde{\mathbf{I}})\mathbf{I}(\mathbf{I}^{-1} - \tilde{\mathbf{I}})\mathbf{I}^{-1/2}$, and $\lambda_{c,n}$ is the likelihood ratio statistic, and $\lambda_{c,n} \xrightarrow{D} \chi_r^2$, the chisquared distribution with r degrees of freedom. Thus $rank((\mathbf{I}^{-1/2})'(\mathbf{I}^{-1} - \tilde{\mathbf{I}})\mathbf{I}(\mathbf{I}^{-1} - \tilde{\mathbf{I}})\mathbf{I}^{-1/2}) = r$. Since $\mathbf{I}$ and $\mathbf{I}^{-1/2}$ are of full rank, we must have $rank(\mathbf{I}^{-1} - \tilde{\mathbf{I}}) = r$, and so $rank(\mathbf{B}) = r$ for all $0 \leq c \leq 1$. Finally, we have

$$\lambda_{c,n} = \mathbf{Z}'\mathbf{B}\mathbf{Z} + o_p(1) \xrightarrow{D} \sum_{j=1}^{r} \gamma_j \chi_j^2,$$

where $\gamma_1, ..., \gamma_r$ are all the non-zero eigenvalues of $\mathbf{B}$, and the χ_j^2's are independent chisquared distributions with one degree of freedom.

Under $H_{0,n}$, we have $\sqrt{n}(\theta_n - \theta_0) = \sqrt{n}(\theta_n - (\theta_0 + n^{-1/2}\delta)) + \delta \xrightarrow{D} N(\delta, \mathbf{I}^{-1})$. Similarly, under $H_{0,n}$, $\sqrt{n}(\theta_{0,n} - \theta_0) = \tilde{\mathbf{I}}\Delta_0 + \delta$, so under $H_{0,n}$,

$$\lambda_{c,n} = [\delta' + \Delta_0'(\mathbf{I}^{-1} - \tilde{\mathbf{I}})][(1-c)\mathbf{I} + 2c\mathbf{D}][(\mathbf{I}^{-1} - \tilde{\mathbf{I}})\Delta_0 + \delta]$$

$$+o_p(1) = \mathbf{Z}'[(1-c)\mathbf{I} + 2c\mathbf{D}]\mathbf{Z} + o_p(1),$$

with $\mathbf{Z} = (\mathbf{I}^{-1} - \tilde{\mathbf{I}})\Delta_0 + \delta \sim N(\delta, (\mathbf{I}^{-1} - \tilde{\mathbf{I}})\mathbf{I}(\mathbf{I}^{-1} - \tilde{\mathbf{I}}))$, and the conclusion follows. □

Proof of Proposition. Recall that $\hat{\theta}_n$ is the MLE. Under the given conditions, we have

$$G(\theta_0 + n^{-1/2}\delta|\mathbf{x}^n) - G(\theta_0|\mathbf{x}^n)$$

$$= (1-c)\big(l(\theta_0 + n^{-1/2}\delta|\mathbf{x}^n) - l(\theta_0|\mathbf{x}^n)\big)$$

$$-cn \int \Big(W(\theta_0 + n^{-1/2}\delta - \alpha) - W(\theta_0 - \alpha)\Big)\pi(\alpha|\mathbf{x}^n) d\alpha$$

$$= (1-c)\big(\Delta_0'\delta - \frac{1}{2}\delta'\mathbf{I}\delta + O_p(n^{-1/2})\big)$$

$$-cn \int \big(n^{-1/2} W^{(\mathbf{1})'}(\theta_0 - \alpha)\delta$$

$$+\frac{1}{2}n^{-1}\delta' W^{(\mathbf{2})}(\tilde\theta-\alpha)\delta\big)\pi(\alpha|\mathbf{x}^n)d\alpha$$

$$=(1-c)\big(\Delta_0'\delta-\frac{1}{2}\delta'\mathbf{I}\delta+O_p(n^{-1/2})\big)$$

$$-c\int\big[W^{(\mathbf{1})'}(\theta_0-\hat\theta_n-n^{-1/2}\beta)\delta$$

$$+\frac{1}{2}n^{-1/2}\delta' W^{(\mathbf{2})}(\tilde\theta-\hat\theta_n-n^{-1/2}\beta)\delta\big]\pi(\hat\theta_n+n^{-1/2}\beta|\mathbf{x}^n)d\beta$$

$$=(1-c)\big(\Delta_0'\delta-\frac{1}{2}\delta'\mathbf{I}\delta\big)-c\int_{||\beta||\le n^\delta}\big[W^{(\mathbf{1})'}(\theta_0-\hat\theta_n-n^{-1/2}\beta)\delta$$

$$+\frac{1}{2}n^{-1/2}\delta' W^{(\mathbf{2})}(\tilde\theta-\hat\theta_n-n^{-1/2}\beta)\delta\big]\pi(\hat\theta_n+n^{-1/2}\beta|\mathbf{x}^n)d\beta$$

$$+o_p(1),$$

where $0<\delta<1/2$, $\tilde\theta$ is an intermediate point between θ_0 and $\theta_0+n^{-1/2}\delta$. By condition (B9), and that $\theta_0-\hat\theta_n=O_p(n^{-1/2})$, so on the set $||\beta||\le n^\delta$, we can replace $W^{(\mathbf{1})'}(\theta_0-\hat\theta_n-n^{-1/2}\beta)$ by $2(\theta_0-\hat\theta_n-n^{-1/2}\beta)'\mathbf{D}$, and $W^{(\mathbf{2})}(\tilde\theta-\hat\theta_n-n^{-1/2}\beta)$ by $2\mathbf{D}$. By Theorem 1 in Walker (1969), for any $\mathbf{a},\mathbf{b}\in R^d$,

$$\int_{\mathbf{a}}^{\mathbf{b}}\pi(\hat\theta_n+n^{-1/2}\beta|\mathbf{x}^n)d\beta=\int_{\mathbf{a}}^{\mathbf{b}}\phi_{\mathbf{I}^{-1}}(\beta)d\beta+o_p(1),$$

now we have

$$G(\theta_0+n^{-1/2}\delta|\mathbf{x}^n)-G(\theta_0|\mathbf{x}^n)$$

$$=(1-c)\big(\Delta_0'\delta-\frac{1}{2}\delta'\mathbf{I}\delta\big)-2c\int_{||\beta||\le n^\delta}\big[(\theta_0-\hat\theta_n-n^{-1/2}\beta)'\mathbf{D}\delta$$

$$+\frac{1}{2}n^{-1/2}\delta'\mathbf{D}\delta\big]\pi(\hat\theta_n+n^{-1/2}\beta|\mathbf{x}^n)d\beta+o_p(1)$$

$$=(1-c)\big(\Delta_0'\delta-\frac{1}{2}\delta'\mathbf{I}\delta\big)-2c\int\big[(\theta_0-\hat\theta_n-n^{-1/2}\beta)'\mathbf{D}\delta$$

$$+\frac{1}{2}n^{-1/2}\delta'\mathbf{D}\delta\big]\pi(\hat\theta_n+n^{-1/2}\beta|\mathbf{x}^n)d\beta+o_p(1)$$

$$=(1-c)\big(\Delta_0'\delta-\frac{1}{2}\delta'\mathbf{I}\delta\big)-2c\int\big[(\theta_0-\hat\theta_n-n^{-1/2}\beta)'\mathbf{D}\delta$$

$$+\frac{1}{2}n^{-1/2}\delta'\mathbf{D}\delta\big]\phi_{\mathbf{I}^{-1}}(\beta)d\beta+o_p(1)$$

$$=(1-c)\big(\Delta_0'\delta-\frac{1}{2}\delta'\mathbf{I}\delta\big)+2c\int\Delta_0'\mathbf{I}^{-1}\mathbf{D}\delta\phi_{\mathbf{I}^{-1}}(\beta)d\beta+o_p(1)$$

$$=\Delta_0'[(1-c)I+2c\mathbf{I}^{-1}\mathbf{D}]\delta-\frac{1}{2}\delta'\mathbf{I}\delta+o_p(1).\qquad\square$$

ACKNOWLEDGEMENTS

The authors are very grateful to the Co-editors Professor Ming-Hui Chen and Professor Yuedong Wang, an associate editor, and two reviewers for their very helpful and constructive comments. Xu's work was supported in part by General Research Fund (17308820) of Hong Kong.

Received 17 October 2022

REFERENCES

[1] Bayarri, M. J. and Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science*, **19**, 58–80. MR2082147

[2] Bernstein, S. (1917). *Theory of Probability* (in Russian).

[3] Bhattacharya, R. N. and Rao, R. R. (1986). *Normal Approximation and Asymptotic Expansions*, Robert, E. Krieder Publishing Company, Malabar, Florida. MR0855460

[4] Bickel, P. J. (1974). Edgeworth expansions in nonparametric statistics. *Annals of Statistics*, **2**, 1–20. MR0350952

[5] Chibisov, D. M. (1972). Asymptotic expansions for statistics distributions admitting asymptotic expansions. *Theory of Probability and Its Applications*, **17**, No.4, 620–630. MR0323007

[6] Chibisov, D. M. (1973). An asymptotic expansion for the distribution of sums of a special form with an application to minimum-contrast estimates. *Theory of Probability and Its Applications*, **18**, No.4, 649–661. MR0329092

[7] Efron, B. (2005). Bayesian, frequentists, and the scientists. *Journal of the American Statistical Association*, **100**, 1–5. MR2166064

[8] Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**(2), 209–230. MR0350949

[9] Good, I. J. (1992). The Bayes/non-Bayes compromise: A brief review. *Journal of the American Statistical Association*, **87**, 597–606. MR1185188

[10] Gusev, S. I. (1975). Asymptotic expansions associated with some statistical estimators in the smooth case I. Expansions of random variables. *Theory of Probability and its Applications*, **20**, 470–498. MR0381082

[11] Gusev, S. I. (1976). Asymptotic expansions associated with some statistical estimators in the smooth case, II. Expansions of moments and distribution. *Theory of Probability and Its Applications*, **21**, No.1, 14–33. MR0413327

[12] Hall, P. (1992). *Bootstrap and Edgeworth Expansion*, Springer, New York. MR1145237

[13] Han, G., Santner, T.J., Lin, H., Yuan, A. (2022). Bayesian-frequentist hybrid inference in applications with small sample sizes, *The American Statistician*, https://doi.org/10.1080/00031305.2022.2127897.

[14] Ibragimov, I.A. and Khas'minski, R.Z. (1973a). Asymptotic behavior of some statistical estimators. II. limit theorems for the a posterior density and Bayes estimators. *Theory of Probability and its Applications*, **18**, 76–91.

[15] Ibragimov, I.A. and Khas'minski, R.Z. (1973b). On moments of generalized Bayes estimators and maximum likelihood estimators. *Theory of Probability and its Applications*, **18**, 508–520.

[16] Lahiri, S.N. (2010). Edgeworth expansions for studentized statistics under weak dependence. *Annals of Statistics*, **38**, 388–434. MR2589326

[17] LeCam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *Univ. California Publ. Statist.*, **1**, 277–330. MR0054913

[18] LeCam, L. (1960). Locally asymptotically normal families of distributions. *Univ. California Publ. Statist.*, **3**, 37–98. MR0126903

[19] LeCam, L. M. and Yang, G. (1990). *Asymptotics in Statistics: Some Basic Concepts*. Springer, New York. MR1066869

[20] Linnik, Y.V. and Mitrofanova, N.M. (1963). On the asymptotic distribution of the maximum likelihood estimate. *Dokl. Akad. Nauk, SSSR.* **149**, 518–520. MR0150877

[21] Linnik, Y.V. and Mitrofanova, N.M. (1965). Some asymptotic expansions for the distribution of the maximum likelihood estimate. *Sankhya*, Ser. A **27**, 73–82. MR0189172

[22] Mitrofanova, N.M. (1967). On the asymptotic distribution of the maximum likelihood estimate of a vector parameter, *Theory of Probability and Its Applications*, **12**, No.3, 364–372. MR0219162

[23] Robbins, H. (1955). An empirical Bayes approach to statistics. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1**, 157–164. MR0084919

[24] Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*, Wiley. MR0595165

[25] Strasser, H. (1981). Consistency of Maximum Likelihood and Bayes Estimates. *Annals of Statistics*, **9**, 1107–1113. MR0628766

[26] Wald, A. (1950). Statistical Decision Function. *Annals of Mathematical Statistics*, **20**(2), 165–205. MR0044802

[27] Walker, A.M. (1969). On the asymptotic behavior of posterior distributions, *Journal of the Royal Statistical Society*, Ser. B, **1**, 80–88. MR0269000

[28] Wilks, S.S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses, *Annals of Mathematical Statistics* **9**, 60–62.

[29] Yuan, A. (2009). Bayesian frequentist hybrid inference. *Annals of Statistics*, **37**, 2458–2501. MR2543699

[30] Yuan, A. and De Gooijer, J. (2014). Asymptotically informative prior for Bayesian analysis, *Communications in Statistics: Theory and Methods*, **43**, 3080–3094. MR3225047

Ao Yuan
Georgetown University
Washington DC
20057
USA
E-mail address: ay312@georgetown.edu

Jinfeng Xu
City University of Hong Kong
Hong Kong
China
E-mail address: jinfengxu@gmail.com