Response to Reviewers
“ On the Wilks phenomenon of Bayes factors and the integrated likelihood ratio test ”

Rui Wang[1,2] and Xingzhong Xu[2,3]

[1] School of Statistics, Renmin University of China, Beijing 100872, China
[2] School of Mathematics and Statistics, Beijing Institute of Technology, Beijing 100081,China
[3] Beijing Key Laboratory on MCAACI, Beijing Institute of Technology, Beijing 100081,China

Monday 7th December, 2020

We are grateful to the AE for informing us a number of useful relevant papers and helpful comments. He/she even provided us an example, namely binomial mixture model, which is very suitable for illustrating the advantages of the proposed method over the LRT. We thank both reviewers for their valuable and helpful comments and critiques. We have carefully revised our paper according to the comments and critiques and have made extensive modifications on the original manuscript. Now the paper is much improved. Below we respond to the AE and each reviewer in turn.

1 Response to the AE

First, the fractional posterior is a special case of a Gibbs posterior, and $\sqrt{n}$ consistency of Gibbs posteriors has been around in the literature for a while; see, for example, the paper ”An MCMC approach to classical estimation” by Chernozhukov and Hong as a representative example. This is particularly not surprising for the fractional posterior since roughly speaking, it keeps the center of the posterior unaffected for large n and reduces the variance by a scale factor depending on the fractional power - hence the classical Bernstein–von Mises phenomenon should go through for the fractional posterior under the same conditions as the exact posterior. I found the integrability condition in Proposition 2 to be somewhat strong in this regard.

Answers: We thank the AE for informing the literature on the Gibbs posterior. The fact that the fractional posterior is a special case of a Gibbs posterior is interesting, and we mention this connection in the revised paper. I found that the results of Chernozhukov and Hong (2003) require certain conditions on the likelihood and only bounded likelihood can satisfy their conditions. Since we would like to include the unbounded likelihood case, we can not use the $\sqrt{n}$ consistency result in their paper. We note that the assumptions of Chernozhukov and Hong (2003) (that is, compact parameter space and continuous prior density) implies that the prior distribution is proper. In fact, we found that most existing results on the consistency of posterior require that the prior is proper. Since improper priors are often used for Bayes factor, existing results are not suitable for our purpose. The integrability condition in Proposition 2 is

$$\int_\Theta \exp\{-c^* D_{1-t}(\theta_0\|\theta)\}\pi(\theta)\,\mathrm{d}\theta < \infty.$$

Note that if $\pi(\theta)$ is proper, then

$$\int_\Theta \exp\{-c^* D_{1-t}(\theta_0\|\theta)\}\pi(\theta)\,\mathrm{d}\theta < \int_\Theta \pi(\theta)\,\mathrm{d}\theta = 1 < \infty.$$

Hence the integrability condition is satisfied for any proper priors. Furthermore, it also allows improper priors provided the tail of the prior density is not too thick. Hence our integrability condition is very weak. In the revised paper, we add a remark on this condition.

The paper doesn't discuss recent advances in the frequentist literature. I would point out the article "Generalized likelihood ratio statistics and Wilks phenomenon" by Fan et al. as an example; this has been cited many times and the authors need to check what is there beyond the usual LRT. Also take a look at the paper "Geometric Understanding of Likelihood Ratio Statistics" by Fan et al. which among other things provides a Bayesian argument for the derivation of the classical Wilk's phenomenon.

Answers: We thank the AE for informing us these references. In the revised paper, we have improved the literature review on the LRT. Although the LRT is a very classical idea, we found that the LRT is an active research topic very recently, with important advances both in theory Sur et al. (2019); Anastasiou and Reinert (2020) and methodology Wasserman et al. (2020). Compared with the LRT, the test method of Wasserman et al. (2020) is very general, and can be used in complex models, such as mixture models. As Wasserman et al. (2020) noted, the generality of their method comes at the price of relatively low power. Theses points are discussed in the revised paper.

2 Response to reviewer 1

Major comments:

1. (Literature review) The paper proposes to use the Bayes factor as a test statistic for a significance test of the composite null hypothesis. I can hardly believe this idea is new. The paper does not mention any papers on this idea. I suggest to review the

literature of significance test using Bayes factors or related issues. I also suggest to review the literature of large sample properties of Bayes factor, and to compare the results of this paper with existing results in the literature.

Answers: Following the reviewer's suggestions, we improved the literature review in the revised paper. The idea of using the Bayes factor as a test statistic is definitely not a new one, and we point out this point explicitly in the revised paper. We add the following comment in the revised paper:

> *The idea is not new. In fact, this methodology dates back at least to Good (1967), and has been considered by many researchers since then. Good (1992) gave a review for some early literature using this idea. See Aerts et al. (2004), Zhou and Guan (2018) and Wang and Xu (2020) for some recent applications of the idea. However, the idea has been mostly used for specific models and has not been systematically and rigorously studied for general models. One reason for this fact may be the Lindley's paradox, that is, the distribution of the Bayes factor depends heavily on the prior density; see, e.g., Shafer (1982). ...*

Following the reviewer's suggestion, we also comment on the literature of large sample properties of Bayes factor, as follow:

> *In this work, we treat the Bayes factor and its variants as frequentist test statistics. Hence a closely related area of research is the frequentist properties of the Bayes factor. Most existing results in this area focus on the consistency of the Bayes factor, that is, as the sample size n goes to infinity, the Bayes factor converges to 0 under the null hypothesis, and converges to $+\infty$ under the alternative hypothesis; see Berger et al. (2003), Moreno et al. (2010), Wang and Maruyama (2016), Chatterjee et al. (2018) and the references therein. There are still relatively few researches on the asymptotic distribution of the Bayes factor. Clarke and Barron (1990) derived the asymptotic distribution of the ratio of the marginal likelihood to the likelihood at the true parameter, which can be regarded as the Bayes factor for a point null hypothesis. Gelfand and Dey (1994) derived formal approximations to the Bayas factor and some of its variants. In this work, in order to use the Bayes factor to construct frequentist tests, we give a thorough study of the asymptotic distribution of the Bayes factor and its variants for general models. Thess results are interesting in their own right.*

We add a remark below Theorem 1 which gives a comparison of Theorem 1 and the result in Clarke and Barron (1990):

> *Clarke and Barron (1990) derived the asymptotic null distribution of*
>
> $$\frac{\int_\Theta p_n(\mathbf{X}^n\theta)\pi(\theta)\,\mathrm{d}\theta}{p_n(\mathbf{X}^n\theta_0)},$$
>
> *which can be regarded as the Bayes factor for the point null hypothesis, i.e., $p_0 = 0$. Theorem 1 extends their result and gives the asymptotic distribution of the general Bayes factor under both the null hypothesis and the local alternative hypothesis. Clarke and Barron (1990) imposed some conditions on the likelihood around the true parameter, which may not be easy*

> *to verify for some moderately complex models. In comparison, Theorem 1 only assumes the likelihood can be expanded in a $n^{-1/2}$ neighborhood of the true parameter, which is satisfied by most regular models.*

2. (numerical examples: simulation studies and real applications) The paper does not have any numerical study. I suggest to include extensive simulation studies comparing existing significance tests such as the likelihood ratio test (LRT). If the Bayes factors in the paper are used in hypothesis testing, is there anything that the user needs to be careful about? Please include in simulation study examples in which LRT fails and examples with wide applications.

Answers: We follow the reviewer's suggestion. In the revised paper, we add simulations to examine the performance of the proposed method and compare it with the LRT. The simulation examples include the logistic regression model which is an example with wide applications, and the normal mixture model and the binomial mixture model which are examples that the LRT has irregular behavior. Numerical results show that the proposed method has good performance in terms of both test level and test power.

In the revised paper, we also discussed that the users needs to be careful about the degree of freedom when they deal with irregular models, as follows:

> *The proposed method is quite universal. However, when applied to irregular models, one needs to be careful about the freedom of the asymptotic chi-squared distribution. For example, in Proposition 5 and Proposition 6, the freedom of the asymptotic chi-squared distribution is* 1 *instead of $p - p_0 = 2$. This phenomenon is essentially caused by the loss of identifiability of the mixture models. In general, when the model has a loss of identifiability, the degree of freedom of the asymptotic chi-squred distribution may be less than $p - p_0$. In this case, if the true degree of freedom is not easy to obtain, the user can simply use $p - p_0$ as the degree of freedom, and the resulting test can still preserve the test level. Of course, this simple strategy may lead to decreased power.*

3. (Choice of a and b) According to Theorem 2, there are many a and b such that $\Delta_{a,b}$ satisfies the Wilks phenomenon. In actual testing problem, what values of a and b would you suggest to use?

Answers:

Our results imply that the choice of a and b does not affect the Wilks phenomenon and asymptotic power provided a and b are fixed and $0 < b < a < 1$. In the simulation studies of the revised paper, the default value is $a = 2/3$ and $b = 1/3$. we also add a simulation to examine the effect of a and b on the finite sample performance of the test procedure. It turns out that the proposed methdod is not sensitive to the choice a and b. In practice, if there is no practical evidence for the choice of a and b, one can simply choose $a = 2/3$ and $b = 1/3$.

4. The paper discusses the fractional Bayes factor, the posterior Bayes factor and the integrated Bayes factor, but does not discuss the intrinsic Bayes factor. I

suggest to include some discussion on the intrinsic Bayes factor. Can you get similar asymptotic results for the intrinsic Bayes factor?

Answers: We follow the reviewer's suggestion. In Appendix D of the revised paper, we rigorously derive the asymptotic distribution of the intrinsic Bayes factor. This theoretical result is obtained by applying the law of large numbers of U-statistics. Unfortunately, it turns out that the intrinsic Bayes factor does not have Wilks phenomenon in general. Hence intrinsic Bayes factor can not be directly used as a frequentist test statistic. This point is made clear in our revised paper.

5. The paper gives one example (mixtures of normals) for which the LRT fails. One example for the argument for the Bayes factor as the test statistic is weak. Please give more examples and numerical examples.

Answers: We follow the reviewer's suggestion. In the revised paper, we add two examples, the logistic regression model and the binomial mixture model. The logistic regression model is widely used in applications. We note that when using R language to fit a logistic regression, the p-value reported by the method "`summary.glm`" is based on the Wald test. It is known that for logistic regression, the MLE does not exist when the data points are completely separated or quasi-completely separated; see, e.g., Albert and Anderson (1984); Candès and Sur (2020). As a consequence of this phenomenon, the Wald test has undesirable behavior. In contrast, the generalized FBF does not rely on the estimation of the parameter, it turns out that it works well for this model. The binomial mixture model is useful in genetics. It is known that the LRT for this model has complicated behavior. For these two models, we proved that the proposed method has Wilks phenomenon. We also conduct simulations to examine the good performance of the proposed method.

Minor comments:

1. P. 3. L. 19. variantional → variational

Answers: The correction has been made.

2. Assumption 1. The term "inner point" is used but "interior point" is more standard term. Also here Θ and $\tilde{\Theta}$ is assumed open and "interior point" assumption is not necessary.

Answers: The correction has been made. We drop the condition of "interior point".

3. P. 5. L. 36. means → mean

Answers: The correction has been made.

4. Equation (1). It is stated that null distribution of $\mathbf{BF}_t(X_n)$ is free of the nuisance parameter if and only if $\frac{\left|I_{\xi|\nu}(\nu,\xi_0)\right|^{-1/2}\pi(\theta_0)}{\pi_0(\nu)} \equiv c$ for some constant c. But the formula in Theorem 1 is slightly different. It should be

$$\frac{\left|I_{\xi|\nu}(\nu,\xi_0)\right|^{-1/2}\pi(\theta_0)}{\pi_0(\nu_0)} \equiv c$$

for some constant c. If this is the case, the subsequence discussion needs to be modified.

Answers: We think our original statement is correct although our original expression is not good. In the paper, under the null hypothesis, we assume the true parameter is $\theta_0 = (\nu_0^\top, \xi_0^\top)^\top$. Here ξ_0 is specified by the null hypothesis ($H : \xi = \xi_0$). But ν_0 is an unknown nuisance parameter. Hence according to Theorem 1, to make the null distribution of $\mathrm{BF}(\mathbf{X}^n)$ free of the nuisance parameter ν_0, one should require that

$$\frac{\left|I_{\xi|\nu}(\nu_0, \xi_0)\right|^{-1/2} \pi(\theta_0)}{\pi_0(\nu_0)} \equiv c$$

for any nuisance parameter ν_0. But the above statement is equivalent to say that

$$\frac{\left|I_{\xi|\nu}(\nu, \xi_0)\right|^{-1/2} \pi(\theta_0)}{\pi_0(\nu)} \equiv c$$

for any nuisance parameter ν. In the revised paper, we improve the expression to make it easier to read.

5. Please include some discussion on Assumptions 2 and 3. Please explain conditions in these assumptions.

Answers: We follow the reviewer's suggestion. In the revised paper, we add the following comment bellow Assumption 2:

The first condition in Assumption 2 avoids infinite Kullback-Leibler divergence. If the second condition in Assumption 2 does not hold, then there is a $\delta > 0$ and a sequence of parameters $\{\theta_n\}$ such that $\|\theta_n - \theta_0\| \geq \delta$ and $D_t(\theta_0 \| \theta_n) \to 0$. In this case, the model will suffer from loss of identifiability. Hence the second condition assumes that the model is identifiable in a reasonable sense. The third condition in Assumption 2 assumes that $D_t(\theta_0 \| \theta)$ has a reasonable Taylor approximation around $\theta = \theta_0$; see, e.g., van Erven and Harremoes (2014), Section III. H.

We add the following comment bellow Assumption 3:

Assumption 3 assumes that the tail of the prior density is not too thick. To appreciate the conditions, suppose P_θ is the normal distribution $\mathcal{N}(\theta, 1)$ and $\theta_0 = 0$. Then the first two conditions of Assumption 3 becomes

$$\int_\Theta \exp\left\{-c^* \theta^2/2\right\} \pi(\theta)\, \mathrm{d}\theta < \infty, \qquad \int_\Theta \theta^2 \exp\left\{-c^* \theta^2/2\right\} \pi(\theta)\, \mathrm{d}\theta < \infty.$$

The above condition is satisfied for $\pi(\theta) \equiv 1$. This implies that Assumption 3 is weak and it allows improper priors.

6. P. 18. L. 44. It is said "This equality holds for every $M > 0$ and hence also for some $M_n \to \infty$." Please prove this statement.

Answers:

We have obtained that for any fixed $M > 0$,

$$\begin{aligned}&\int_{\{\theta:\|\theta-\theta_0\|\leq M/\sqrt{n}\}} \exp\{-tR_n(\theta_0\|\theta)\}\,\pi(\theta)\,\mathrm{d}\theta\\ =&(1+o_{P^n_{\theta_0}}(1))n^{-p/2}\pi(\theta_0)\exp\left\{\frac{t}{2}\Delta_{n,\theta_0}^\top I(\theta_0)\Delta_{n,\theta_0}\right\}\\ &\cdot\int_{\{h:\|h\|\leq M\}}\exp\left\{-\frac{t}{2}(h-\Delta_{n,\theta_0})^\top I(\theta_0)(h-\Delta_{n,\theta_0})\right\}\,\mathrm{d}h.\end{aligned}$$

We claim that this equality holds for every $M > 0$ and hence also for some $M_n \to \infty$.

In essence, this claim follows from a result in mathematical analysis and the fact that convergence in probability is metrizable. To make the proof more readable, we add a lemma in the revised paper (Lemma 2 in the revised paper).

Lemma. *Let $T_{m,n}$, $m = 1, 2, \ldots$, $n = 1, 2, \ldots$, be random variables such that for fixed m, $T_{m,n}$ converges in probability to 0 as $n \to \infty$. Then there exists a sequence $\{h(n)\}$ such that $h(n) \to \infty$ and $T_{h(n),n}$ converges in probability to 0 as $n \to \infty$.*

Using this lemma, we can prove our claim as follow:

Note that for every fixed $M > 0$, the term $o_{P^n_{\theta_0}}(1)$ in the above equality converges in probability to 0 as $n \to \infty$. Hence by Lemma 2, this term also converges in probability to 0 for some $M_n \to \infty$. Therefore the above equality still holds if we replace M by M_n.

In the revised paper, we also simplify some other proofs to improve the readability of the paper.

7. P. 24. L. 4. A typo. $L_t(\Theta; \mathbf{X}_n) \leq L_1^{1/t}(\Theta; \mathbf{X}_n) \to L_t(\Theta; \mathbf{X}_n) \leq L_1^t(\Theta; \mathbf{X}_n)$**.**

Answers: The correction has been made.

8. P. 25. In the integral of the displayed formula. $\pi(\theta|\mathbf{X}_n) \to \pi_t(\theta|\mathbf{X}_n)$**.**

Answers: The correction has been made. During the revision of the paper, we also corrected many other typos and mistakes. We are sorry for such typos and mistakes.

3 Response to reviewer 2

1. Some related reference is missing such as Zhou and Guan (2018) JASA paper On the Null Distribution of Bayes Factors in Linear Regression. In this paper, they showed by considering the Bayes factor as a test statistics (like this submitted paper), they derived the null distribution of the Bayes factor for linear models, which is a weighted sum of chi-squared random variables. Even though this work is restricted to linear models under simple assumptions, it is worth to cite. Also, I believe that you can find more references that consider a Bayes factors as a test statistics in frequentist settings.

Answers: We thank the reviewer for informing us this useful paper. In the revised paper, we reviewed the literature that consider a Bayes factors as a frequentist test statistic, and add the following comment in the revised paper:

> *. . . In fact, this methodology dates back at least to Good (1967), and has been considered by many researchers since then. Good (1992) gave a review for some early literature using this idea. See Aerts et al. (2004), Zhou and Guan (2018) and Wang and Xu (2020) for some recent applications of the idea. However, the idea has been mostly used for specific models and has not been systematically and rigorously studied for general models. One reason for this fact may be the Lindley's paradox, that is, the distribution of the Bayes factor depends heavily on the prior density; see, e.g., Shafer (1982). . . .*

Following another reviewer's suggestion, we also comment on the literature of large sample properties of Bayes factor, as follow:

> *In this work, we treat the Bayes factor and its variants as frequentist test statistics. Hence a closely related area of research is the frequentist properties of the Bayes factor. Most existing results in this area focus on the consistency of the Bayes factor, that is, as the sample size n goes to infinity, the Bayes factor converges to* 0 *under the null hypothesis, and converges to* $+\infty$ *under the alternative hypothesis; see Berger et al. (2003), Moreno et al. (2010), Wang and Maruyama (2016), Chatterjee et al. (2018) and the references therein. There are still relatively few researches on the asymptotic distribution of the Bayes factor. Clarke and Barron (1990) derived the asymptotic distribution of the ratio of the marginal likelihood to the likelihood at the true parameter, which can be regarded as the Bayes factor for a point null hypothesis. Gelfand and Dey (1994) derived formal approximations to the Bayas factor and some of its variants. In this work, in order to use the Bayes factor to construct frequentist tests, we give a thorough study of the asymptotic distribution of the Bayes factor and its variants for general models. Thess results are interesting in their own right.*

2. On page 4, "The prior $\pi(\theta)$ and $\pi_0(\nu)$ may be improper. . . ". This is wrong in general. If they are improper, we cannot calculate the Bayes factor. Of course, I understand that even when the priors are improper, the fractional Bayes factor, introduced later, circumvents this issue. But, I think that it would be better to briefly note this point at the place of the sentence "The prior $\pi(\theta)$ and $\pi_0(\nu)$ may be improper. . . ", to avoid a confusion.

Answers: We are not pretty sure if the statement "*If they are improper, we cannot calculate the Bayes factor*" means that the Bayes factor is not well-defined or the Bayes factor is hard to compute. Below we discuss these two points separately.

Indeed, if the prior $\pi(\theta)$ is improper, then the marginal density $\int_\Theta p_n(\mathbf{X}^n|\theta)\pi(\theta)\,\mathrm{d}\theta$ may be infinite for certain models, especially for small sample size n. That is, the Bayes factors with improper priors are not universally well-defined. In fact, for some irregular models, the improper priors can not be used for any n. For example, consider the normal mixture model $p(x|\mu) = 0.5\phi(x) + 0.5\phi(x-\mu)$. Then if the prior $\pi(\mu)$ is improper, then for any n,

$$\int_{-\infty}^{+\infty}\prod_{i=1}^n p(x_i|\mu)\pi(\mu)\,\mathrm{d}\mu \geq \int_{-\infty}^{+\infty}\prod_{i=1}^n(0.5\phi(x_i))\pi(\mu)\,\mathrm{d}\mu = \prod_{i=1}^n(0.5\phi(x_i))\int_{-\infty}^{+\infty}\pi(\mu)\,\mathrm{d}\mu = +\infty.$$

In this sence, the reviewer's comment is correct. However, we think that including the case for improper priors is important for at least two reasons. First, in the context of statistical decision theory, Bayes rules with improper priors are important. In particular, Farrell (1968) Theorem 5.1 implies that, under very weak conditions, if $\int_{\tilde{\Theta}_0} \pi_0(\nu)\,\mathrm{d}\nu = 1$, $\int_{\Theta} \pi(\theta)\,\mathrm{d}\theta = +\infty$, then the test based on Bayes factor is *admissible.* Second, in Bayesian literature, many, if not most, popular priors for Bayes factors are improper. Take the linear model for example, improper priors are often adopted for the variance parameter; see, e.g., Liang et al. (2008). Hence we think that it is worthwhile to include the case of improper priors in our main theoretical results. In contrast, most existing results on the frequentist properties of Bayes methods assume proper priors.

Now we turn to the computation issue. If the priors are proper, a brute force method to compute the marginal likelihood is to sample the parameter from the prior $\pi(\theta)$ and then use the sampled parameters $\theta_1^*, \ldots, \theta_B^*$ to calculate the average of the likelihood

$$\int_{\Theta} p(\mathbf{X}^n|\theta)\pi(\theta)\,\mathrm{d}\theta \approx \frac{1}{B}\sum_{i=1}^{n} p(\mathbf{X}^n|\theta_i^*).$$

If the priors are improper, then one can not sample from the prior and the above method can not be directly applied. In this sence, the reviewer's comment is correct. Nevertheless, one can use the importance sampling method to remedy this problem. Note that

$$\int_{\Theta} p(\mathbf{X}^n|\theta)\pi(\theta)\,\mathrm{d}\theta = \int_{\Theta} p(\mathbf{X}^n|\theta)\frac{\pi(\theta)}{\tilde{\pi}(\theta)}\tilde{\pi}(\theta)\,\mathrm{d}\theta$$

where $\tilde{\pi}(\theta)$ is an appropriate proper prior. Then one can sample from $\tilde{\pi}(\theta)$ and then compute the average of $p(\mathbf{X}^n|\theta)\pi(\theta)/\tilde{\pi}(\theta)$. Hence one can compute the Bayes factor even if the prior is improper. The computation of the Bayes factor has been actively studied, see Friel and Wyse (2012) for a review.

We add the following comment in the revised paper:

> *In Bayesian literature, conventional Bayes factors often use improper priors for certain parameters, and consequently,* $\pi(\theta)$ *and* $\pi_0(\nu)$ *may be improper, that is,* $\int_{\Theta} \pi(\theta)\,\mathrm{d}\theta = +\infty$, $\int_{\tilde{\Theta}_0} \pi_0(\nu)\,\mathrm{d}\nu = +\infty$*; see, e. g., Berger and Pericchi (2001), Section 2.1. To accomodate these cases, throughout the paper,* $\pi(\theta)$ *and* $\pi_0(\nu)$ *are allowed to be improper unless otherwise stated.*

3. On line 31 on page 6, what is f?

Answers: Here f is a probability density function. We are sorry that we did not explain the meaning of f in the original paper.

4. On line 38 on page 6, you may want to note an extra weakness of the traditional approach. When the prior density is proportional to a function of the determinant of the Fisher information, its prior normalizing constant is difficult to evaluate, and the normalizing constant is critical for the Bayes factor. Although you noted "...Fisher

information matrix has a complicated form. . . undesirable. . . ", it would be better to be more specific.

Answers: We agree that the original statement is not specific enough. Following the reviewer's suggestion, we the following comment to the revised paper:

> *Second, in order to construct priors satisfying (2), the determinant of the Fisher information matrix should be evaluated, which is a difficult task for many models. Hence it may not be easy to construct priors satisfying (2), especially for complex models.*

5. This theoretical work is investigated under assumed that the posterior achieves $\sqrt{n}$-consistency (or the posterior contraction rate is $n^{1/2}$). In many practical statistical models, the optimal posterior contraction rate would be slower than $n^{-1/2}$. These examples include high-dimensional sparse problem like $(\log p/n)^{1/2}$ or nonparametric function estimation $n^{-\beta/(2\beta+p)}$, where β is a smoothing factor for the true function. Under these interesting models, can you extend this result to more general models?

Answers:

As the reviewer noted, our general theoretical results are obtained with the assumption that the posterior contraction rate of the parameter is $n^{1/2}$. The reviewer also noted that for many models, the optimal posteior contraction rate is slower than $n^{1/2}$. We note that this is indeed the case for some of our examples, i.e. normal mixture models and binomial mixture models, where the null model suffers from loss of identifiability and consequently, the posterior does not contract to a single parameter value.

In our paper, the general results are for the low-dimensional parametric models. We note that for exponential family (Section 4.1) and logistic model (Section 4.2), it is straightforward to extend our proof to the case that p slowly increases as $n \to \infty$. However, our proof can not be extended to the real interesting setting that $p \asymp n$. In fact, in the setting of $p \asymp n$, even the asymptotic distribution of the classical LRT was not well understood until recently Sur and Candès (2019); Sur et al. (2019). The application of the Bayes factors to the high-dimensional hypothesis testing problems is very interesting, and we have made some simple exploration in this direction for linear model Wang and Xu (2020).

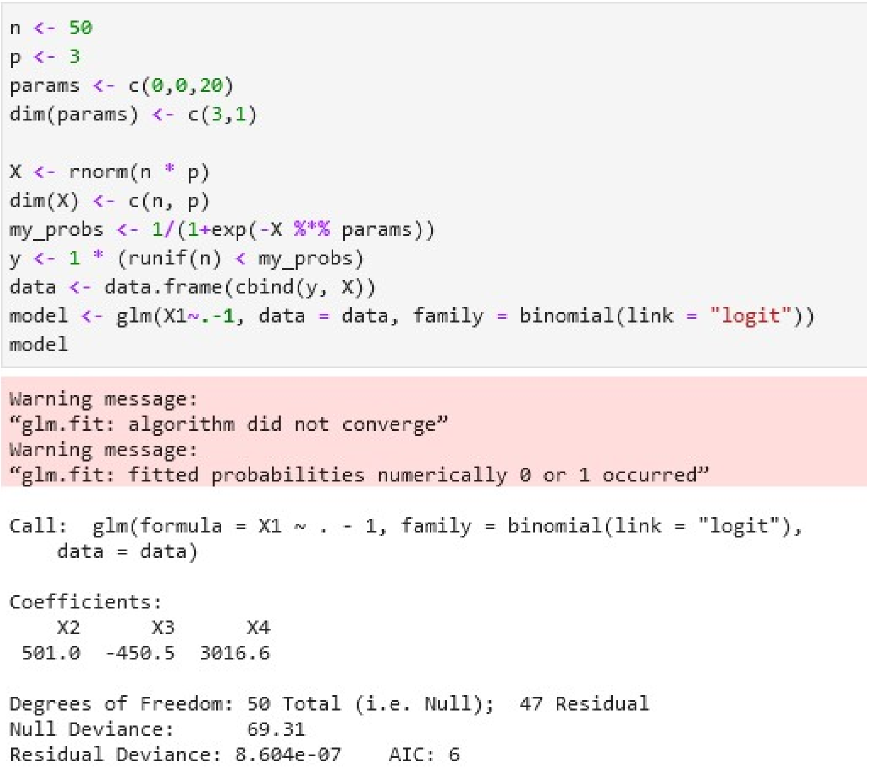
The reviewer mentioned the high-dimensional sparse problem and nonparametric problem, and asked if our results can be extended to include such models. Recently, we found that the work of Gao et al. (2020) provided a general framework which can be used to derive optimal posterior contraction rate of both Bayes high-dimensional statistics and Bayes nonparametrics. It is very interesting to study the asymptotic distributin of the Bayes factor in such general framework. However, it may be very challenging going from posterior contraction result to the asymptotic distribution of the Bayes factor. In fact, in frequentist literatures, the LASSO estimator is desparsified to do hypothesis testing; see, e.g., Zhang and Cheng (2017) and the references therein. Perhaps Bayes factors also need to be "desparsified" to obtain a tractable distribution. We have to admit that currently we can not extend our results to such framework. The main reason is, of course, the limitation of our current ability. We would like to leave this interesting problem for future research.

6. You proposed theoretical results on asymptotic null distribution of Bayes factor which is a linear transformation of a chi-square distribution. But, you have not considered any simulation works to examine finite sample behavior of the approximated distribution. In practice, the accuracy of the asymptotic null distribution with finite samples is of interest. You may want to show that this asymptotic null distribution is useful in practice, especially for complicated and practical models. The examples you considered (without simulation studies) are too simple and far from a practical point of view.

Answers: We follow the reviewer's suggestion. In the revised paper, we conduct simulations to examine the finite sample performance of the proposed method. The simulation results verify our asymptotic theory.

In the original paper, the examples include the exponential family models and two submodels of the normal mixture model. The reviwer noted that our original examples are far from a practical point of view. In the revised paper, to make the examples more practical, we add the logistic regression model which is widely used in practice, and the binomial mixture model which arises in genetics. For these two commonly used models, the commonly used tests still have undesired properties. We prove that the generalized FBF has Wilks phenomenon in these models. We also use simulation studies to verify our theory.

Here we would like to elaborate some problems associated with the logistic regression from a practical point of view. A notable phenomenon for the logistic regression model is that the MLE does not exist when the data points are completely separated or quasi-completely separated; see, e.g., Albert and Anderson (1984); Candès and Sur (2020). We would like to examine the behavior of software when data are completely separated. For R language (R Core Team, 2020), when fitting the logistic regression with "`glm`", it can successfully fit a model although there are two warnings indicating that the algorithm did not converge and the fitted probabilities numerically 0 or 1 occurred; see Figure 1. In comparison, for Python module *statsmodels* Seabold and Perktold (2010), version 0.12.0, it throws an error "`Perfect separation detected, results not available`", and does not provide a fitted model. Although R language can provide a fitted model, the p-value provided by "`summary.glm`" is close to 1; see Figure 2. Since the true parameter is largely deviated from 0, this phenomenon is unreasonable. We note that the p-value returned by "`summary.glm`" is based on the Wald test rather than the LRT. Since the Wald test relies on the MLE, when the MLE does not exist, it is not surprising that the Wald test has bad numerical behavior. Note that although the MLE does not exist, the likelihood is bounded by 1. Hence the LRT always exists. Our simulation results show that the LRT performs better than the Wald test. On the other hand, the generalized FBF does not rely on the estimation of the parameter, and is always well defined. Our theoretical and simulation results show that the generalized FBF has similar performance as the LRT. When ξ is largely deviated from 0, the empirical power of the LRT exhibits slight instability



```
n <- 50
p <- 3
params <- c(0,0,20)
dim(params) <- c(3,1)

X <- rnorm(n * p)
dim(X) <- c(n, p)
my_probs <- 1/(1+exp(-X %*% params))
y <- 1 * (runif(n) < my_probs)
data <- data.frame(cbind(y, X))
model <- glm(X1~.-1, data = data, family = binomial(link = "logit"))
model
```

```
Warning message:
“glm.fit: algorithm did not converge”
Warning message:
“glm.fit: fitted probabilities numerically 0 or 1 occurred”
```

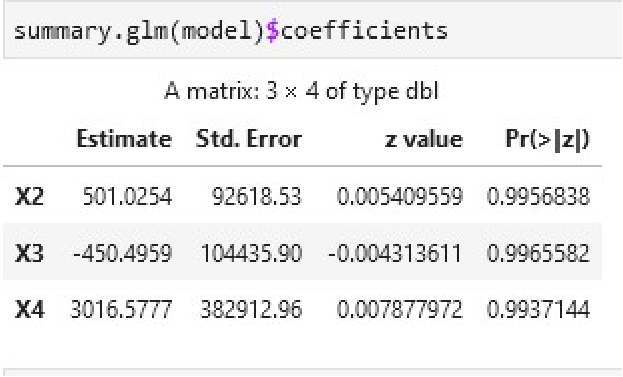
```
Call:  glm(formula = X1 ~ . - 1, family = binomial(link = "logit"), 
    data = data)

Coefficients:
    X2      X3      X4  
 501.0  -450.5  3016.6  

Degrees of Freedom: 50 Total (i.e. Null);  47 Residual
Null Deviance:      69.31 
Residual Deviance: 8.604e-07    AIC: 6
```

Figure 1: R code for using “`glm`” to fit a logistic regression model with completely separated data.

which is due to the instability of the numerical optimization. In comparison, the computation of the generalized FBF does not involve numerical optimization, and its empirical power of is exact 1 for large ξ.

The reviwer noted that our original examples are too simple. Indeed, the original normal mixture models have simple forms. For these models, the proposed method is simple to compute and has regular behavior. However, we would like to point out that the two normal mixture models and the binomial mixture model are irregular models and the LRT for these models are not very simple to use. Take the second normal mixture model in the paper as an example. For this model, the asymptotic distribution of the LRT is a difficult problem. Bickel and Chernoff (1993) studied this problem, but they did not fully solve it. Liu and Shao (2004) gave the asymptotic null distribution of the LRT. Hall and Stewart (2005) derived the asymptotic power of the LRT. According to the result of Hall and Stewart (2005), the LRT has trivial power under $n^{-1/2}$ local alternative hypothesis. For this irreagular model, the generalized FBF still has Wilks phenomenon and has better performance than the LRT. We think these examples can illustrate the advantages of the proposed method over the LRT.

It is interesing to use the proposed method to solve more complex and practical problems. We would like to leave it for future work.



```
summary.glm(model)$coefficients
```

```
             A matrix: 3 × 4 of type dbl
      Estimate   Std. Error      z value    Pr(>|z|)
X2    501.0254     92618.53  0.005409559   0.9956838
X3   -450.4959    104435.90 -0.004313611   0.9965582
X4   3016.5777    382912.96  0.007877972   0.9937144
```

Figure 2: R code for using "`summary.glm`" to obtain the p-value of the coefficients.

7. This paper is lack of tuning parameter selection which will be critical in the hypothesis testing result. Theoretically, it can satisfy some simple conditions, but in practice how to choose the tuning parameters a and b is very important.

Answers: This question was raised by both reviewers. Our Answer is as follow:

Our results imply that the choice of a and b does not affect the Wilks phenomenon and asymptotic power provided a and b are fixed and $0 < b < a < 1$. In the simulation studies of the revised paper, the default value is $a = 2/3$ and $b = 1/3$. we also add a simulation to examine the effect of a and b on the finite sample performance of the test procedure. It turns out that the proposed methdod is not sensitive to the choice a and b. In practice, if there is no practical evidence for the choice of a and b, one can simply choose $a = 2/3$ and $b = 1/3$.

4 List of major changes

- We add an example of logistic regression model.
- We add an example of binomial mixture model.
- We add simulations studies for the logistic regression model, normal mixture model and the binomial mixture model.
- We add the asymptotic distribution of the intrinsic Bayes factor.
- The literature review is improved.
- We add many remarks to explain our results and assumptions.
- We simplify some of the theoretical proofs.

References

Aerts, M., Claeskens, G., and Hart, J. D. (2004). Bayesian-motivated tests of function fit and their asymptotic frequentist properties. *The Annals of Statistics*, 32(6):2580–2615.

Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10.

Anastasiou, A. and Reinert, G. (2020). Bounds for the asymptotic distribution of the likelihood ratio. *The Annals of Applied Probability*, 30(2):608–643.

Berger, J. O., Ghosh, J. K., and Mukhopadhyay, N. (2003). Approximations and consistency of bayes factors as model dimension grows. *Journal of Statistical Planning and Inference*, 112:241–258.

Berger, J. O. and Pericchi, L. R. (2001). *Objective Bayesian Methods for Model Selection: Introduction and Comparison*, volume Volume 38 of *Lecture Notes–Monograph Series*, pages 135–207. Institute of Mathematical Statistics, Beachwood, OH.

Bickel, P. J. and Chernoff, . (1993). Asymptotic distribution of the likelihood ratio statistic in a prototypical non regular problem. In Ghosh, e. a., editor, *Statistics and Probability: A Raghu Raj Bahadur Festschcrift*, pages 83–96. Eastern Limited.

Candès, E. J. and Sur, P. (2020). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics*, 48(1):27–42.

Chatterjee, D., Maitra, T., and Bhattacharya, S. (2018). A short note on almost sure convergence of bayes factors in the general set-up. *The American Statistician*. Online.

Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. *Journal of Econometrics*, 115(2):293–346.

Clarke, B. S. and Barron, A. R. (1990). Information-theoretic asymptotics of bayes methods. *IEEE Transactions on Information Theory*, 36(3):453–471.

Farrell, R. H. (1968). Towards a theory of generalized Bayes tests. *Annals of Mathematical Statistics*, 39:1–22.

Friel, N. and Wyse, J. (2012). Estimating the evidence - a review. *Statistica Neerlandica*, 66(3):288–308.

Gao, C., van der Vaart, A. W., and Zhou, H. H. (2020). A general framework for bayes structured linear models. *Annals of Statistics*, 48(5):2848–2878.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(3):501–514.

Good, I. J. (1967). A Bayesian significance test for multinomial distributions. (With discussion). *Journal of the Royal Statistical Society. Series B. Methodological*, 29:399–431.

Good, I. J. (1992). The Bayes/non-Bayes compromise: a brief review. *Journal of the American Statistical Association*, 87(419):597–606.

Hall, P. and Stewart, M. (2005). Theoretical analysis of power in a two-component normal mixture model. *Journal of Statistical Planning and Inference*, 134(1):158 – 179.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423.

Liu, X. and Shao, Y. (2004). Asymptotics for the likelihood ratio test in a two-component normal mixture model. *Journal of Statistical Planning and Inference*, 123(1):61 – 81.

Moreno, E., Girn, F. J., and Casella, G. (2010). Consistency of objective bayes factors as the model dimension grows. *The Annals of Statistics*, 38(4):1937–1952.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Seabold, S. and Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 92 – 96.

Shafer, G. (1982). Lindley's paradox. *Journal of the American Statistical Association*, 77(378):325–351. With discussion and with a reply by the author.

Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525.

Sur, P., Chen, Y., and Candès, E. J. (2019). The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability Theory and Related Fields*, 175(1-2):487–558.

van Erven, T. and Harremoes, P. (2014). Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.

Wang, M. and Maruyama, Y. (2016). Consistency of bayes factor for nonnested model selection when the model dimension grows. *Bernoulli*, 22(4):2080–2100.

Wang, R. and Xu, X. (2020). A bayesian-motivated test for high-dimensional linear regression models with fixed design matrix. *Statistical Papers*.

Wasserman, L., Ramdas, A., and Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890.

Zhang, X. and Cheng, G. (2017). Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association*, 112(518):757–768.

Zhou, Q. and Guan, Y. (2018). On the null distribution of bayes factors in linear regression. *Journal of the American Statistical Association*, 113(523):1362–1371.