Goodness-of-fit using Nonparametric Full Bayesian Significance Test

Montcho Djidenou[1,2] , Izbicki Rafael[1], Salasar Bueno[1]

[1,2]Universidade Federal de São Carlos-DES, São Carlos, São Paulo, Brasil
[2]Universidade de São Paulo-ICMC, São Carlos, São Paulo, Brasil
[1]hansamos@usp.br, [2]rafaelizbicki@gmail.com, [3]luis.salasar@gmail.com

Introduction

Testing whether a sample comes from a given distribution, referred as one-sample problem or goodness-of-fit test, is one of the most important problem in statistics. Formally stated, given a random sample $X_1, \ldots, X_n \sim \mathcal{P}$, one wishes to test $H_0 : \mathcal{P} = \mathcal{P}_0$ against $H_1 : \mathcal{P} \neq \mathcal{P}_0$. Under the frequentist framework, well known approaches such as t-test, Kolmogorov-Smirnov(KS) and others provide a good answer.

On the other side, in bayesian procedure, one starts assigning a prior probability to H_0 and then computes the change in the relative plausibility of H_0 versus H_1 after gathering data using Bayes theorem. However, bayesian procedure suffers from precise hypothesis testing i.e when H_0 is indexed by a subset having null Lebesgue measure. A solution to this problem in the parametric context, Full Bayesian Significance Test(FBST), was introduced in De Bragança Pereira e Stern (1999).

This work extends the FBST to nonparametric models using pseudo-densities. A Pseudo-density is a surrogate to a probability density over a space of functions(FERRATY; KUDRASZOW; VIEU, 2012). In the rest of this work, we briefly review the parametric FBST, then we introduce our nonparametric FBST. In the sequence, through a simulation study, we compare our proposal to other tests and end with a discussion over possible extensions.

Parametric FBST

Consider $X_1^n := (X_1, \ldots, X_n) \sim P_\theta$ and the hypothesis $H_0 : \theta \in \Theta_0$, $H_1 : \theta \in \Theta - \Theta_0$. Let us assume a prior density f over Θ and, after observing a sample x_1^n, we can obtain the posterior density $f(\cdot \mid x_1^n)$. The evidence for the null H_0 provided by the data x_1^n can be represented by the e-value:

$$ev(H_0 \mid x_1^n) = 1 - \pi\big(\{\theta \in \Theta : f(\theta|x_1^n) > f^*\} \mid x_1^n\big), \quad f^* = \sup_{\theta\in\Theta_0} f(\theta \mid x_1^n)$$

where f^* is the greatest posterior density over H_0. The set $T_{x_1^n}(H_0) = \{\theta \in \Theta : f(\theta|x_1^n) > f^*\}$ is called the tangent set to H_0 and contains all θ that are more plausible than H_0. Thus, a large posterior probability of $T_{x_1^n}(H_0)$ (small value of $ev(H_0 \mid x_1^n)$) indicates evidence against H_0.

Nonparametric FBST

Let $X_1^n := (X_1, \ldots, X_n) \sim P$ and we want to test the hypothesis $H_0 : P \in \mathbb{P}_0$ $H_1 : P \in \mathbb{P} - \mathbb{P}_0$.

Henceforth, we assume that for a given nonparametric prior probability measure π over $\mathbb{P}$ and a sample x_1^n, $\pi(\cdot \mid x_1^n)$ is the posterior probability over $\mathbb{P}$. There is no clear way to define a posterior density of $\pi(\cdot \mid x_1^n)$. In this case, we use a pseudo-density defined by:

$$\tilde{f}(P \mid x_1^n) = \mathbb{E}_{P^*\sim\pi(\cdot\mid x_1^n)}\left[K\left(\frac{d(P,P^*)}{h}\right)\right],$$

where $P \in \mathbb{P}$, K a given kernel over the real line, h a positive bandwidth. Simulating $P_1, \ldots, P_S$ independently from $\pi(\cdot \mid x_1^n)$, we can approximate $\tilde{f}$ by

$$\hat{f}(P \mid x_1^n) \propto \frac{1}{S}\sum_{j=1}^{S} K\left(\frac{d(P,P_j)}{h}\right),$$

Then, we define the pseudo e-value of $H_0 : P \in \mathbb{P}_0$ as:

$$\tilde{ev}(H_0 \mid x_1^n) = \pi\big(\{P : \tilde{f}(P \mid x_1^n) \le f^*\} \mid x_1^n\big); \quad f^* = \sup_{P\in\mathbb{P}_0} \tilde{f}(P \mid x_1^n) \qquad (1)$$

Here, we consider a Dirichlet process prior over $\mathbb{P}$, ie $P \sim DP(G, \alpha)$, where G is the base distribution function and $\alpha > 0$ the concentration parameter. From the conjugacy property of the DP, one can easily obtain posterior samples and compute 1. Moreover, we use a gaussian kernel $K = exp\{\frac{-d(\cdot,\cdot)^2}{h}\}$ and the $KS = Sup_n\,|F_n - F^*|$ and $L_2 = \left[\int (F_n - F^*)^2 dx\right]^{\frac{1}{2}}$ as distances between the posteriors.

Acknowledgements

The authors are thankful for the support from the following organizations:

Simulations

Setup:Let $X_1, X_2, \cdots, X_n \sim P$ Gaussian Kernel, h=1, S=500, 2000 replicates.

1. $H_0 : P = N(0{,}1)$; $H_1 : P \neq N(0{,}1)$, $G_0 = \mathcal{N}(\prime,\infty)$, $\alpha = 1$, prior-FBST: N(0,25)

Tabela 1 – Power comparison between t-test, AD, CVM, KS, fbst, and our tests fbst.ks, fbst.l2

θ	0	0.25	0.5	0.75	1	1.25
fbst.ks	0.062	0.212	0.631	0.936	0.992	1.000
fbst.l2	0.047	0.191	0.640	0.939	0.995	1.000
fbst	0.046	0.224	0.721	0.968	0.997	1.000
AD	0.042	0.214	0.688	0.962	0.996	1.000
t	0.043	0.202	0.678	0.952	0.996	1.000
CVM	0.050	0.200	0.655	0.947	0.995	1.000
KS	0.054	0.176	0.577	0.907	0.991	0.999

2. $H_0 : P = Exp(2)$; $H_1 : P \neq Exp(2)$, $G_0 = Exp(1)$, $\alpha = 1$, prior-FBST: Gamma(0.16,0.08)

Tabela 2 – Power comparison between AD, CVM, KS, fbst, and our test fbst.l2

	0.7	0.85	1	1.15	1.3	1.45	1.6	1.75	1.9	2
fbst.ks	0.993	0.966	0.897	0.726	0.556	0.424	0.157	0.101	0.059	0.052
fbst.l2	0.981	0.931	0.817	0.630	0.411	0.254	0.142	0.094	0.056	0.047
fbst	0.999	0.990	0.970	0.862	0.699	0.496	0.298	0.193	0.072	0.041
AD	0.993	0.976	0.900	0.711	0.517	0.335	0.176	0.114	0.065	0.037
CVM	0.982	0.935	0.807	0.613	0.441	0.279	0.156	0.105	0.065	0.040
KS	0.974	0.902	0.766	0.559	0.393	0.243	0.136	0.096	0.069	0.048

Induced Posterior: Instead of sampling F from the DP,

1. $\mu \sim N(0{,}25)$
2. From $f(\mu|X_1^n)$, sample $P_\mu = N(\mu,1)$ induced by μ
3. $\hat{f}(P \mid x_1^n) \propto \frac{1}{S}\sum_{j=1}^{S} K\left(\frac{d(P,P_j)}{h}\right)$

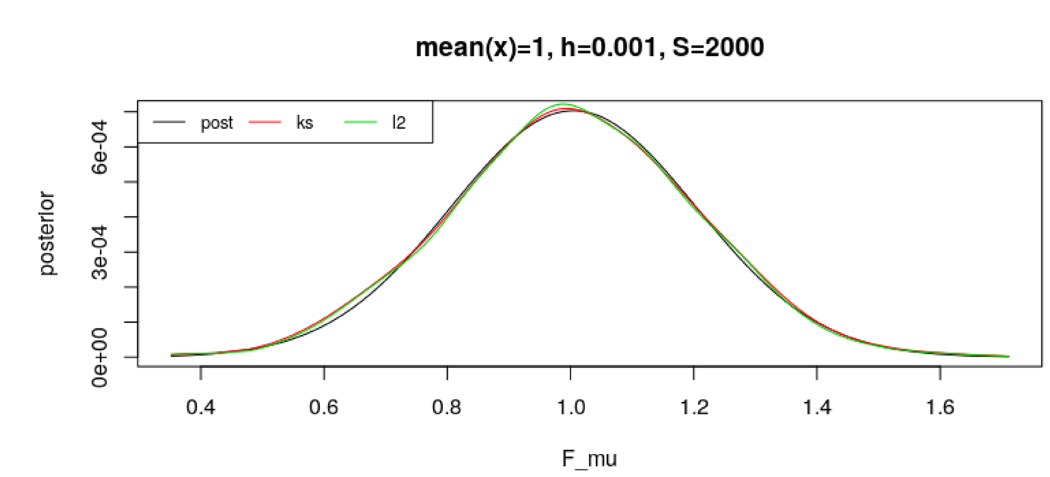

Figura 1 – Posterior density Vs Pseudo-density using KS distance

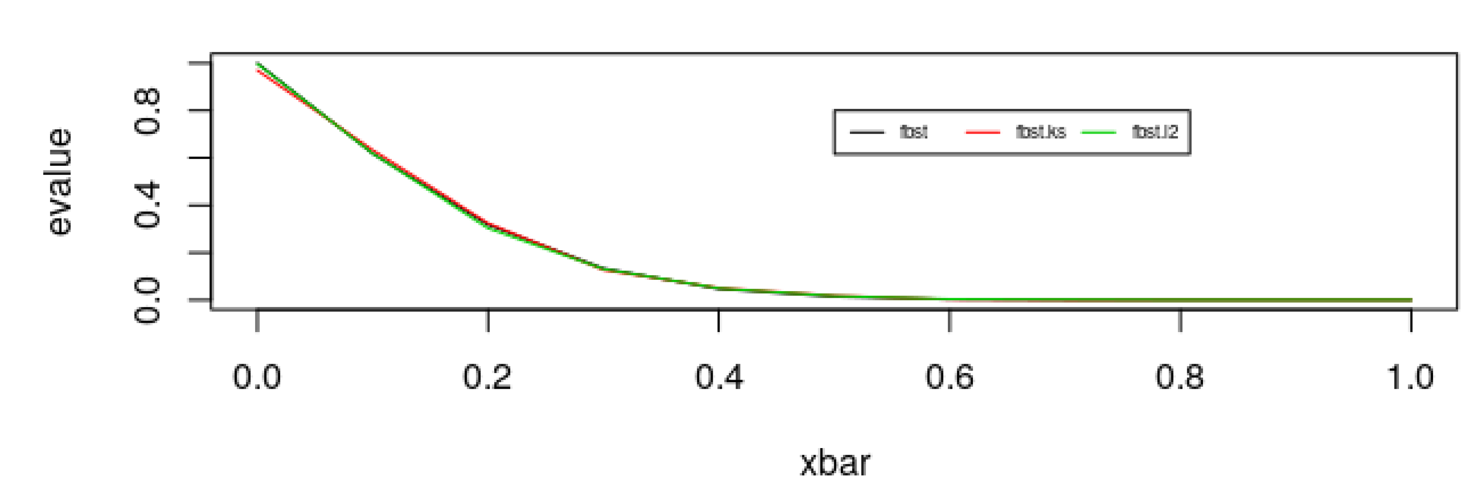

Figura 2 – Pseudo evalue Vs parametric evalue

Conclusion

In this work we propose the nonparametric FBST using pseudo-density and show its performances in situations where Θ_0 has only one element. The next step is its extension to larger spaces and also develop a version for the two sample problem. Other interesting questions of theoretical aspects such as convergence to parametric evalue, effect of the pseudo-density and kernel are worth evaluating.

References

- DE BRAGANÇA PEREIRA, C.; STERN, J. Evidence and credibility: full Bayesian significance test for precise hypotheses. **Entropy**, Molecular Diversity Preservation International, v. 1, n. 4, p. 99–110, 1999.
- FERRATY, F.; KUDRASZOW, N.; VIEU, P. Nonparametric estimation of a surrogate density function in infinite-dimensional spaces. **Journal of Nonparametric Statistics**, Taylor & Francis, v. 24, n. 2, p. 447–464, 2012.