

# Model fitting in the presence of nuisance parameters

G. Jogesh Babu  
Center for Astrostatistics  
Department of Statistics, 319 Thomas Building  
The Pennsylvania State University  
University Park, PA 16802, USA  
<http://astrostatistics.psu.edu>  
[babu@stat.psu.edu](mailto:babu@stat.psu.edu)

## Abstract

**A basic problem in any statistical modeling of a scientific dataset is to provide the ‘best’ fit. Such inference is generally based on the empirical distribution function when the underlying process generating the data is not reasonably known. A computationally intensive resampling method called the bootstrap method are presented, to estimate the null distributions of various goodness of fit test statistics, when the underlying process is partially known. These results hold not only in the univariate case but also in the multivariate setting.**

*Keywords: Bootstrap, computationally intensive resampling, Kolmogorov-Smirnov statistic, Gaussian process, Empirical process*

## 1. INTRODUCTION

A vast range of statistical problems arise in modern astronomical and space sciences research, particularly due to the flood of data produced by astronomical surveys at many wavebands. Equally important is the great increase in the complexity of the data sets: some are tabular with many dimensions, some are time series with complex temporal behaviors, and others are linked to highly nonlinear astrophysical models. While the scientific promise is tremendous, it depends critically on the ability to extract useful knowledge from the data. A basic problem in any statistical modeling of a scientific dataset is to provide the most parsimonious ‘best’ fit. Questions such as the following often arise in astrophysics.

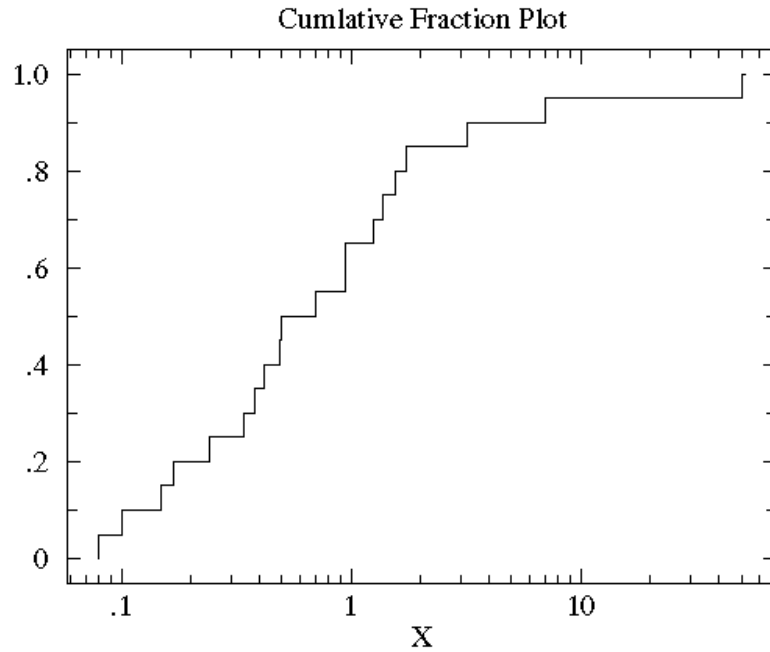
- Is the underlying nature of a quasar spectrum a nonthermal power law or a combination of black bodies?
- Is the topology of a clustered multivariate dataset best modeled as several distinct clusters, as a smooth distribution with voids, or as a stochastic hierarchy of embedded structures?
- Are the fluctuations in the cosmic microwave background best fit by Big Bang models with dark energy or with quintessence?
- Are there interesting correlations among the properties of objects in any given class (e.g. the Fundamental Plane of elliptical galaxies), and what are the optimal analytical expressions of such correlations?
- How do we characterize blips embedded in larger structures?

These issues arise when data are used to repudiate or support astrophysical theories but the underlying processes generating the data are not confidently known.

Standard penalized likelihood based methods such as the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) can be used, but do not necessarily answer the problem

---

This work is partially supported by NSF grant DMS 0101360



**FIGURE 1:** Empirical Distribution Function. Also known as step function

of the physical scientist. The AIC is optimized for the reduction of prediction error while the BIC is optimized for the maximization of the probability of correct selection. Babu and C. R. Rao have developed nonparametric resampling methods for inference, when the data comes from an unknown distribution which may or may not belong to a specified family ([2, 3]). The asymptotic null distributions of statistics based on empirical distribution function such as Kolmogorov-Smirnov, will not be fixed if nuisance parameters affecting the distribution are present. Computationally intensive bootstrap methods, to estimate these null distributions which hold not only in the univariate case but also in the multivariate setting, are discussed here.

## 2. STATISTICS BASED ON THE EMPIRICAL DISTRIBUTION FUNCTION

Nonparametric goodness of fit tests are generally based on the empirical distribution function (see Figure 1). The problem of goodness of fit tests when parameters are estimated are presented. These results hold not only in the univariate case but also in the multivariate setting. These ideas are taken a step further to develop non-parametric resampling methods for inference, when the data comes from an unknown distribution which may or may not belong to a specified family of distributions.

Many nonparametric goodness of fit tests are based on the empirical distribution function. For example, Kolmogorov-Smirnov:

$$\sup_x |F_n(x) - F(x)|, \sup_x (F_n(x) - F(x))^+, \sup_x (F_n(x) - F(x))^-;$$

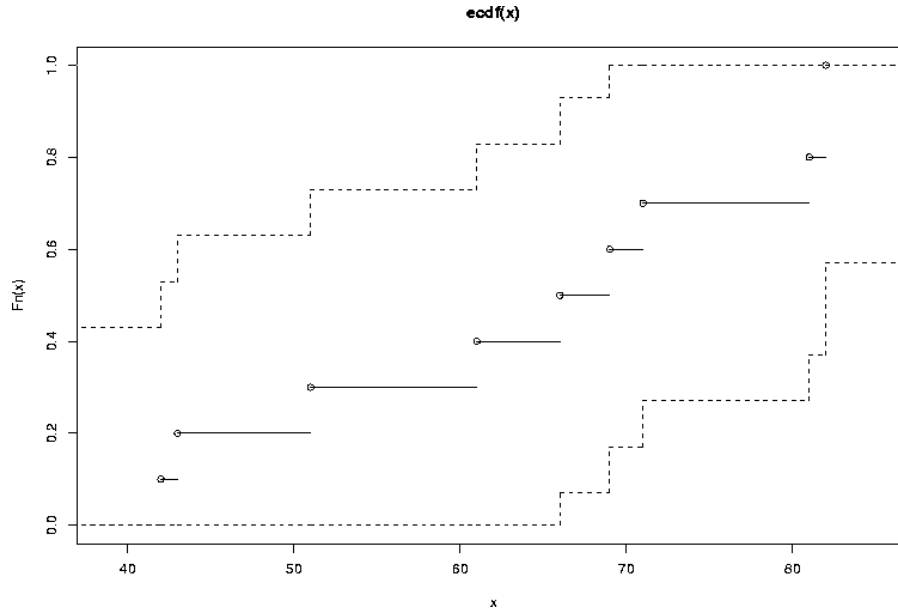
Renyi:

$$\sup_{a \leq F(x) \leq b} \left( \frac{F_n(x) - F(x)}{F(x)} \right)^+, \sup_{a \leq F(x) \leq b} \left( \frac{F_n(x) - F(x)}{F(x)} \right)^-,$$

$$\sup_{a \leq F(x) \leq b} \left| \frac{F_n(x) - F(x)}{F(x)} \right|;$$

Cramér-von Mises:

$$\int (F_n(x) - F(x))^2 dF(x);$$



**FIGURE 2:** K-S Confidence Bands

and Anderson - Darling:

$$\int \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} dF(x).$$

All of these are distribution free. The goodness of fit tests such as Kolmogorov-Smirnov can be used to generate confidence bands for the underlying probability distribution (Figure 2). This figure is generated using the public domain software **R**.

In the presence of nuisance parameters, the tests are generally constructed by first estimating the parameters. In such a case the asymptotic null distribution of the test statistic may depend in a complex way on the unknown parameters. Asymptotic distribution of test statistics based on the empirical distribution function, when parameters are estimated have been extensively studied in [4], [5], [6] and others.

In the multivariate case the Kolmogorov-Smirnov function is no longer distribution free. This is even the case when all the parameters are known. The following simple example to illustrate this was given in [7]. Let

$$F(x, y) = \frac{1}{2}ax^2y + \frac{1}{2}(2 - a)y^2x, \quad 0 < x, y < 1$$

denote a bivariate distribution and  $(X_1, Y_1)$  a data from  $F$ . If  $F_1$  denotes the empirical distribution function of  $(X_1, Y_1)$ , then

$$\text{for } a = 0 : F(x, y) = y^2x \text{ and } 0.065 < P(|F_1(x, y) - F(x, y)| < 0.72, \forall x, y) < 0.066,$$

and

$$\text{for } a = 1 : F(x, y) = \frac{1}{2}xy(x + y) \text{ and } 0.057 < P(|F_1(x, y) - F(x, y)| < 0.72, \forall x, y) < 0.058.$$

Thus the distributions and hence critical values differ for K-S statistics for different values of  $a$ .

### 3. THE BOOTSTRAP

The bootstrap resampling scheme will help in obtaining critical values in the testing context, even when the parameters are estimated. Let  $\{F(\cdot; \theta) : \theta \in \Theta\}$  be a family of continuous distribution functions, where  $\Theta$  is an open region in a  $p$ -dimensional Euclidean space. Let  $X_1, \dots, X_n$  be i.i.d.

random variables from a distribution function  $H$ . Statistics based on empirical measures to test  $H = F(\cdot; \theta)$  for some  $\theta = \theta_0$  or if  $\theta$  is partially specified are considered. The statistics such as Kolmogorov-Smirnov and Cramér-von Mises, when  $\theta$  is estimated by  $\hat{\theta}_n = \theta_n(X_1, \dots, X_n)$ , can be viewed as continuous functionals of the process

$$Y_n(x; \hat{\theta}_n) = \sqrt{n}(F_n(x) - F(x; \hat{\theta}_n)),$$

where  $F_n$  denotes the empirical distribution function of  $X_1, \dots, X_n$ . In the case of Gaussian family with  $\theta = (\mu, \sigma^2)$ , we can use  $\hat{\theta} = (\bar{X}_n, s_n^2)$ , where

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{and} \quad s_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Two bootstrap procedures are described here, when the parameters are partially or completely estimated. The procedure will also help in the computation of power under contiguous alternatives  $\lambda_n = \theta_0 + n^{-1/2}\lambda$ . To describe the bootstrap procedures, let  $X_1^*, \dots, X_n^*$  be i.i.d. random variables from  $\hat{F}_n$  and  $\hat{\theta}_n^* = \theta_n(X_1^*, \dots, X_n^*)$ , where  $\hat{F}_n$  is an estimator of the distribution function  $H$ , based on the sample  $X_1, \dots, X_n$ . The resampling method is called nonparametric bootstrap if  $\hat{F}_n = F_n$ , and it is called parametric bootstrap if  $\hat{F}_n = F(\cdot; \hat{\theta}_n)$ . In the Gaussian example,  $\hat{\theta}_n^* = (\bar{X}_n^*, s_n^{*2})$ , where

$$\bar{X}_n^* = \frac{1}{n} \sum_{i=1}^n X_i^* \quad \text{and} \quad s_n^{*2} = \frac{1}{n} \sum_{i=1}^n (X_i^* - \bar{X}_n^*)^2.$$

Under some regularity conditions both parametric and nonparametric procedures lead to correct asymptotic levels. In fact, under very general conditions, the sample process  $Y_n$  given by

$$Y_n(x) = \sqrt{n}(F_n(x) - F(x; \hat{\theta}_n)) = \sqrt{n}(F_n(x) - F(x; \theta) + F(x; \hat{\theta}) - F(x; \hat{\theta}_n))$$

and the parametric bootstrap process  $Y_n^P$  given by

$$Y_n^P(x) = \sqrt{n}(F_n^*(x) - F(x; \hat{\theta}_n^*)) = \sqrt{n}(F_n^*(x) - F(x; \hat{\theta}_n) + F(x; \hat{\theta}_n) - F(x; \hat{\theta}_n^*)),$$

both converge to the same Gaussian process  $Y$ . Hence, in this case,

$$\sqrt{n} \sup_x |F_n(x) - F(x; \hat{\theta}_n)| \quad \text{and} \quad \sqrt{n} \sup_x |F_n^*(x) - F(x; \hat{\theta}_n^*)|$$

have the same limiting distribution.

In practice, one takes a bootstrap sample  $X_1^*, \dots, X_n^*$  from  $X_1, \dots, X_n$  and computes

$$\sqrt{n} \sup_x |F_n^*(x) - F(x; \hat{\theta}_n^*)|.$$

This procedure is repeated a large number of times. The histogram of the resulting values approximate the distribution of

$$\sqrt{n} \sup_x |F_n(x) - F(x; \hat{\theta}_n)|.$$

Hence these values can be used to obtain the critical values for testing or to get confidence bands.

In the case of nonparametric bootstrap, a bias correction

$$B_n(x) = \sqrt{n}(F_n(x) - F(x; \hat{\theta}_n))$$

is needed. Under very general regularity conditions, the process  $Y_n$  and the process  $Y_n^N$  given by

$$\begin{aligned} Y_n^N(x) &= \sqrt{n}(F_n^*(x) - F(x; \hat{\theta}_n^*)) - B_n(x) \\ &= \sqrt{n}(F_n^*(x) - F_n(x) + F(x; \hat{\theta}_n) - F(x; \hat{\theta}_n^*)), \end{aligned}$$

both converge to the same Gaussian process  $Y$ . Hence

$$\sqrt{n} \sup_x |F_n(x) - F(x; \hat{\theta}_n)| \quad \text{and} \quad \sup_x |\sqrt{n}(F_n^*(x) - F(x; \hat{\theta}_n^*)) - B_n(x)|$$

have the same limiting distribution. Thus the bootstrap method consistently estimates the null distributions of various goodness of fit tests. So both parametric and nonparametric procedures lead to correct asymptotic levels. The results hold also in the multivariate setting.

#### 4. LOCATION FAMILIES

A family  $\{F(\cdot; \theta) : \theta \in \Theta\}$  is called location family if  $F(x; \theta) = F(x - \theta)$ . An estimator  $\hat{\theta}_n$  is called location invariant, if  $\hat{\theta}_n(X_1 + a, \dots, X_n + a) = \hat{\theta}_n(X_1, \dots, X_n) + a$ . Computations become much simpler under location invariance.

**Example:** Gaussian family.

Let  $V_1, \dots, V_n$  be data from normal distribution with  $\theta = (\mu, \sigma^2)$ . Let  $F_{n,V}$  and  $\hat{F}_{n,V}$  denote the empirical distribution functions of  $V_1, \dots, V_n$  and  $((V_1 - \bar{V})/s_V), \dots, ((V_n - \bar{V})/s_V)$ , where  $\bar{V}$  and  $s_V^2$  are the sample mean and sample variance of  $V_1, \dots, V_n$ . Then

$$\sup_y |\hat{F}_{n,V}(y) - \Phi(y)| = \sup_x |F_{n,V}(x) - \Phi_{\bar{V}, s_V}(x)|.$$

To implement bootstrap, draw a sample  $R_1, \dots, R_n$  from normal distribution with mean  $\bar{V}$  and variance  $s_V^2$ . Clearly,

$$\sup_y |\hat{F}_{n,R}(y) - \Phi(y)| = \sup_x |F_{n,R}(x) - \Phi_{\bar{R}, s_R}(x)|,$$

where  $\hat{F}_{n,R}$  is the empirical distribution function of  $((R_1 - \bar{R})/s_R), \dots, ((R_n - \bar{R})/s_R)$ . Thus

$$\sup_y \sqrt{n} |\hat{F}_{n,R}(y) - \Phi(y)| \quad \text{and} \quad \sup_y \sqrt{n} |\hat{F}_{n,V}(y) - \Phi(y)|$$

have the same limiting distributions.

In practice, one generates data from normal distribution, computes

$$\sup_y \sqrt{n} |\hat{F}_{n,R}(y) - \Phi(y)|,$$

and repeats this process a large number of times. The histogram of the resultant values approximate the distribution of

$$\sup_y \sqrt{n} |\hat{F}_{n,V}(y) - \Phi(y)|.$$

#### 5. CONFIDENCE LIMITS UNDER MISSPECIFICATION

So far the bootstrap approach for testing the null hypothesis that a sample comes from a specified parametric family of distributions, is considered. If the hypothesis is rejected, it would be of interest to examine how close the actual alternative distribution is to the specified family of distributions. A non-parametric bootstrap method is used to obtain confidence limits to the difference between the true distribution function  $H$  and a member of the specified family closest to it in the sense of Kullback-Leibler measure.

Let  $H$  be closest to  $F(\cdot, \theta_0)$  so that an estimator, such as a maximum likelihood estimator,  $\hat{\theta}_n$  converges to  $\theta_0$ . The distribution function  $H$  may or may not belong to the family  $\{F(\cdot; \theta) : \theta \in \Theta\}$ . Then under some regularity conditions, both  $Y_n^N$  and the process  $U_n$  given by

$$U_n(x; \hat{\theta}_n) = \sqrt{n}(F_n(x) - F(x; \hat{\theta}_n)) - \sqrt{n}(H(x) - F(x; \theta_0)),$$

converge weakly to the same Gaussian process under very general conditions.

If  $m \rightarrow \infty$  and  $m/n \rightarrow 0$ , then both  $U_n$  and the process  $Y_n^m$  given by

$$Y_n^m(x) = \sqrt{m}(F_m^*(x) - F(x; \hat{\theta}_m^*)),$$

converge to the same limiting Gaussian process. This can be used to obtain confidence bands for  $H - F(\cdot; \theta_0)$ .

By a result of [1], for any  $0 < \alpha < 1$ ,

$$P\left(\sqrt{n} \sup_x |F_n(x) - F(x; \hat{\theta}_n) - (H(x) - F(x; \theta_0))| \leq C_\alpha^*\right) \rightarrow \alpha,$$

where  $C_\alpha^*$  is the  $\alpha$ -th quantile of

$$\sup_x |\sqrt{n} (F_n^*(x) - F(x; \hat{\theta}_n^*)) - \sqrt{n} (F_n(x) - F(x; \hat{\theta}_n))|.$$

This provides an estimate of the distance between the true distribution and the family of distributions under consideration.

Similar conclusions can be drawn for von Mises-type distance, for example,

$$\int (F_n(x) - F(x; \hat{\theta}_n) - (H(x) - F(x; \theta_0)))^2 dF(x; \theta_0)$$

or

$$\int (F_n(x) - F(x; \hat{\theta}_n) - (H(x) - F(x; \theta_0)))^2 dF(x; \hat{\theta}_n).$$

The details of the results presented here appear in [2] and [3].

## REFERENCES

- [1] BABU, G. J., AND BOSE, A. Bootstrap confidence intervals. *Statistics & Probability Letters* 7 (1988), 151–160.
- [2] BABU, G. J., AND RAO, C. R. Confidence limits to the distance of the true distribution from a misspecified family by bootstrap. *Journal of Statistical Planning and Inference* (2003), **115**, no. 2, 471–478.
- [3] BABU, G. J., AND RAO, C. R. Goodness-of-fit tests when parameters are estimated. *Sankhyā* (2003), **66**, 1–12.
- [4] DARLING, D. A. The Cramér-Smirnov test in the parametric case. *Annals of Mathematical Statistics* 26 (1955), 1–20.
- [5] DURBIN, J. Weak convergence of the sample distribution function when parameters are estimated. *The Annals of Statistics* 1 (1973), 279–290.
- [6] KAC, M., KIEFER, J., AND WOLFOWITZ, J. On tests of normality and other tests of goodness of fit based on distance methods. *Annals of Mathematical Statistics* 26 (1955), 189–211.
- [7] SIMPSON, P. B. Note on the estimation of a bivariate distribution function. *Annals of Mathematical Statistics* 22 (1951), 476–478.