Jackknife and Bootstrap

Notes by

G. Jogesh Babu

Center for Astrostatistics
The Pennsylvania State University

1 Introduction

It is often relatively simple to devise a statistic that measures the property of interest, but is almost always difficult or impossible to determine the distribution of that statistic. The classical statistical methods concentrated mainly on the statistical properties of the estimators that have a simple closed form and which can be analyzed mathematically. Except for a few important but simple statistics, these methods involve often unrealistic model assumptions. These limitations have been overcome in the last two decades of the 20th Century with advances in electronic computers. A class of computationally intensive procedures known as *resampling methods* provide inference on a wide range of statistics under very general conditions. Resampling methods involve constructing hypothetical 'populations' derived from the observations, each of which can be analyzed in the same way to see how the statistics depend on plausible random variations in the observations. Resampling the original data preserves whatever distributions are truly present, including selection effects such as truncation and censoring.

Perhaps the *half-sample method* is the oldest resampling method, where one repeatedly chooses at random half of the data points, and estimates the statistic for each resample. The inference on the parameter can be based on the histogram of the resampled statistics. It was used by Mahalanobis in 1946 under the name *interpenetrating samples*. An important variant is the Quenouille–Tukey *jackknife method*. For a dataset with n data points, one constructs exactly n hypothetical datasets each with $n-1$ points, each one omitting a different point. The most important of resampling methods is called the *bootstrap*. Bradley Efron introduced the bootstrap method, also known as resampling with replacement, in 1979. Here one generates a large number of datasets, each with n data points randomly drawn from the original data. The constraint is that each drawing is made from the entire dataset, so a simulated dataset is likely to miss some points and have duplicates or triplicates of others. Thus, bootstrap can be viewed as a *Monte Carlo method* to simulate from an existing data, without any assumption on the underlying population.

2 Jackknife

Jackknife method was introduced by Quenouille (1949) to estimate the bias of an estimator. The method is later shown to be useful in reducing the bias as well as in estimating the variance of an estimator. Let $\hat{\theta}_n$ be an estimator of θ based on n i.i.d. random vectors $X_1, \ldots, X_n$, i.e., $\hat{\theta}_n = f_n(X_1, \ldots, X_n)$, for some function f_n. Let

$$\hat{\theta}_{n,-i} = f_{n-1}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$$

be the corresponding recomputed statistic based on all but the i-th observation. The jackknife estimator of bias $E(\hat{\theta}_n) - \theta$ is given by

$$bias_J = \frac{(n-1)}{n} \sum_{i=1}^{n} \big(\hat{\theta}_{n,-i} - \hat{\theta}_n\big). \tag{1}$$

Jackknife estimator θ_J of θ is given by

$$\theta_J = \hat{\theta}_n - bias_J = \frac{1}{n}\sum_{i=1}^{n}\big(n\hat{\theta}_n - (n-1)\hat{\theta}_{n,-i}\big). \tag{2}$$

Such a bias corrected estimator hopefully reduces the over all bias. The summands above

$$\theta_{n,i} = \hat{\theta}_n - (n-1)\hat{\theta}_{n,-i},\ i = 1,\ldots,n$$

are called *pseudo-values*.

2.1 Bias Reduction

Jackknifing, indeed, helps in reducing bias of an estimator in many cases. Suppose the expected value of the estimator $\hat{\theta}_n$ is of the form

$$E(\hat{\theta}_n) = \theta + \frac{a}{n} + \frac{b}{n^2},$$

then clearly,

$$E(\hat{\theta}_{n,i}) = \theta - \frac{b}{n(n-1)}.$$

Consequently, the bias of the jackknife estimator is $E(\theta_J) - \theta = O(n^{-2})$, which is of lower order than the bias of $\hat{\theta}_n$.

2.2 Estimation of variance

In the case of the sample mean $\hat{\theta}_n = \bar{X}_n$, it is easy to check that the *pseudo-values* are simply,

$$\theta_{n,i}\hat{\theta}_n - (n-1)\hat{\theta}_{n,-i} = X_i,\ i = 1,\ldots,n.$$

This provides motivation for the jackknife estimator of variance of $\hat{\theta}_n$,

$$\begin{aligned} var_J(\hat{\theta}_n) &= \frac{1}{n(n-1)}\sum_{i=1}^{n}(\theta_{n,i} - \theta_J)(\theta_{n,i} - \theta_J)' \\ &= \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{n,-i} - \bar{\theta}_n)(\hat{\theta}_{n,-i} - \bar{\theta}_n)', \end{aligned} \tag{3}$$

where $\bar{\theta}_n = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{n,-i}$. For most statistics, jackknife estimator of variance is consistent, i.e.,

$$Var_J(\hat{\theta}_n)/Var(\hat{\theta}_n) \to 1,$$

as $n \to \infty$ almost surely. This holds in particular for a **smooth functional model**. To describe this, let the statistic of interest $\hat{\theta}_n$ based on n data points be defined by $H(\bar{Z}_n)$, where $\bar{Z}_n$ is the sample mean of multivariate random variables $Z_1,\ldots,Z_n$ and H is continuously differentiable in a neighborhood of $E(\bar{Z}_n)$. Many commonly occurring statistics, including: Sample Means, Sample Variances, Central and Non-central t-statistics (with possibly non-normal populations), Sample Coefficient of

Variation, Maximum Likelihood Estimators, Least Squares Estimators, Correlation Coefficients, Regression Coefficients, Smooth transforms of these statistics, fall under this model.

However, consistency does not always hold; for example the jackknife method fails for non-smooth statistics, such as the sample median. If $\hat{\theta}_n$ denotes the sample median in the univariate case, then in general,

$$Var_J(\hat{\theta}_n)/Var(\hat{\theta}_n) \to \left(\frac{1}{2}\chi_2^2\right)^2$$

in distribution, where χ_2^2 denotes a *chi-square* random variable with 2 degrees of freedom (see Efron 1982, §3.4). So in this case, the jackknife method does not lead to a consistent estimator of the variance. However, a resampling method called *bootstrap* discussed in the next section, would lead to a consistent estimator.

3 Bootstrap

The importance of the bootstrap emerged during the 1980s when mathematical study demonstrated that it gives nearly optimal estimate of the distribution of many statistics under a wide range of circumstances. In several cases, the method yields better results than those obtained by the classical normal approximation theory. However, one should caution that bootstrap is not the solution for all problems. The theory developed in 1980s and 1990s, show that bootstrap fails in some 'non-smooth' situations. Hence, caution should be used and should resist the temptation to use the method inappropriately. Many of these methods work well in the case of **smooth functional model**.

3.1 Description of the bootstrap method

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be data drawn from an unknown population distribution F. Suppose $T_n(\mathbf{X})$ is a good estimator of $T(F)$, a parameter of interest. The interest lies in assessing its accuracy in prediction. Determining the confidence intervals for $T(F)$ requires knowledge of the sampling distribution G_n of $T_n(\mathbf{X}) - T(F)$, *i.e.* $G_n(x) = \mathrm{P}(T_n(\mathbf{X}) - T(F) \leq x)$, for all x.

For example, the sample mean $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ is a good estimator of the population mean μ. To get the confidence interval for μ, we must find the sampling distribution of $\bar{X}_n - \mu$, which depends on the shape and other characteristics of the unknown distribution F.

Classical statistical theory uses the Central Limit Theorem (normal approximation) to the sampling distribution. For example, if $(X_1, Y_1), \ldots, (X_n, Y_n)$ denote observations from a bivariate normal population, then the maximum likelihood estimator of the correlation coefficient ρ is given by Pearson's correlation coefficient,

$$\hat{\rho}_n = \frac{\sum_{i=1}^n (X_i Y_i - \bar{X}_n \bar{Y}_n)}{\sqrt{\left(\sum_{i=1}^n (X_i - \bar{X}_n)^2\right)\left(\sum_{i=1}^n (Y_i - \bar{Y}_n)^2\right)}}.$$

For statistics with asymmetrical distributions, such as that of $\hat{\rho}_n$, the classical theory suggests variable transformations. In this case, *Fisher's Z-transformation Z* given by

$$Z = \frac{\sqrt{(n-3)}}{2}\left(\ln\left(\frac{1+\hat{\rho}_n}{1-\hat{\rho}_n}\right) - \ln\left(\frac{1+\rho}{1-\rho}\right)\right)$$

gives a better normal approximation. is approximately normally distributed. This approximation corrects skewness and is better than the normal approximation of $\sqrt{n}(\hat{\rho}_n - \rho)$. The bootstrap method, when properly used, avoids such individual transformations by taking into account the skewness of the sampling distribution. It automatically corrects for skewness.

The bootstrap method presumes that if $\hat{F}_n$ is a good approximation to the unknown population distribution F, then the behavior of the samples from $\hat{F}_n$ closely resemble that of the original data. Here $\hat{F}_n$ can be the *empirical distribution function* (EDF, or a smoothed EDF) of the data $X_1, \ldots, X_n$, or a parametric estimator of the function F. Once $\hat{F}_n$ is provided, datasets $\mathbf{X}^* = (X_1^*, \ldots, X_n^*)$ are resampled from $\hat{F}_n$ and the statistic $T_n(\mathbf{X}^*)$ is computed for each resample. Under very general conditions (see Babu & Singh (1984)), it can be shown that the difference between the sampling distribution G_n of $T_n(\mathbf{X}) - T(F)$ and the 'bootstrap distribution' G_b [ie the distribution of $T_n(\mathbf{X}^*) - T(\hat{F}_n)$] is negligible. G_b can be used to draw inferences about the parameter $T(F)$ in place of the unknown G_n. In principle, Bootstrap distribution (Histogram) G_b is completely known, as it is constructed entirely from the original data. However, to get the complete bootstrap distribution, one needs to compute the statistics for nearly all of $M = n^n$ possible bootstrap samples. For the simple example of sample mean, presumably one needs to compute,

$$\begin{aligned}
&X_1^{*(1)}, \ldots, X_n^{*(1)}, \quad r_1 = \bar{X}^{*(1)} - \bar{X} \\
&X_1^{*(2)}, \ldots, X_n^{*(2)}, \quad r_2 = \bar{X}^{*(2)} - \bar{X} \\
&\qquad \ddots \qquad \ddots\ddots \qquad \ddots \\
&X_1^{*(M)}, \ldots, X_n^{*(M)}, \ r_M = \bar{X}^{*(M)} - \bar{X}.
\end{aligned}$$

The bootstrap distribution is given by the histogram of $r_1, \ldots, r_M$. Even for $n = 10$ data points, M turns out to be ten billion. In practice, the statistic of interest, $T_n(\mathbf{X}^*) - T(\hat{F}_n)$, is computed for a number N (say $N = n(\log n)^2$) of resamples, and its histogram is constructed.

The most popular and simple bootstrap is the *nonparametric bootstrap*, where the resampling with replacement is based on the EDF of the original data. This gives equal weights to each of the original data points. Table 1 gives bootstrap versions of some commonly used statistics. In the case of ratio estimator and the correlation coefficient, the data pairs are resampled from the original data pairs (X_i, Y_i).

3.2 Confidence intervals

The concept of 'pivotalness' is very important for *confidence intervals*. For example, the *scaling factor* $\bar{X} - \mu$ depends on the standard deviation σ or its estimator. It is

Table 1: Statistics and their bootstrap versions

Statistic	Bootstrap Version
Mean, $\bar{X}_n$	$\bar{X}^*_n$
Variance, $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$	$\frac{1}{n}\sum_{i=1}^n (X^*_i - \bar{X}^*_n)^2$
Ratio estimator, $\bar{X}_n/\bar{Y}_n$	$\bar{X}^*_n/\bar{Y}^*_n$
Correlation coefficient, $\frac{\sum_{i=1}^n (X_iY_i - \bar{X}_n\bar{Y}_n)}{\sqrt{\left(\sum_{i=1}^n (X_i-\bar{X}_n)^2\right)\left(\sum_{i=1}^n (Y_i-\bar{Y}_n)^2\right)}}$	$\frac{\sum_{i=1}^n (X^*_iY^*_i - \bar{X}^*_n\bar{Y}^*_n)}{\sqrt{\left(\sum_{i=1}^n (X^*_i-\bar{X}^*_n)^2\right)\left(\sum_{i=1}^n (Y^*_i-\bar{Y}^*_n)^2\right)}}$

possible to obtain confidence intervals or approximate confidence intervals for μ, if $\mathrm{P}((\bar{X} - \mu)/\sigma < x)$ does not depend on μ and σ. Unless the limiting distribution is free from the unknown parameters, once can not invert it to get confidence intervals. It is thus important to focus on pivotal or approximately pivotal quantities in order to get reliable confidence intervals for the parameter of interest.

A function $T_n(\mathbf{X}; F)$ is pivotal, if its distribution is free from the unknown parameters of F. In case, $X_i \sim N(\mu, \sigma^2)$, then $T_n(\mathbf{X}; F) = \sqrt{n}(\bar{X} - \mu)/s_n$ is pivotal. In non-normal case, it is approximately pivotal. To obtain bootstrap confidence interval for μ, we compute $\sqrt{n}(\bar{X}^{*(j)} - \bar{X})/s_n$ for N bootstrap samples, arrange the values in increasing order

$$h_1 < h_2 < \cdots < h_N.$$

One can then read off from the histogram (say) the 90% confidence interval of the parameter. That is, the 90% confidence interval for μ is given by

$$\bar{X} - h_m \frac{s_n}{\sqrt{n}} \le \mu < \bar{X} - h_k \frac{s_n}{\sqrt{n}},$$

where $k = [0.5N]$ and $m = [0.95N]$. Babu & Singh (1983) have shown that $N(\log n)^2$ bootstrap iterations would be sufficient.

It is important to note that even when σ is known the bootstrap version of $\sqrt{n}(\bar{X} - \mu)/\sigma$ is $\sqrt{n}(\bar{X}^* - \bar{X})/s_n$. One should not replace $\sqrt{n}(\bar{X}^* - \bar{X})/s_n$ by $\sqrt{n}(\bar{X}^* - \bar{X})/\sigma$.

3.3 Bootstrap at its best: *Smooth function model*

It is well established using Edgeworth expansions that the bootstrap provides a good approximation for a 'Studentized smooth functional model'. A broad class of commonly used statistics, including least squares estimators and some maximum likelihood estimators, can be expressed as smooth function of multivariate means. The model is illustrated using Pearson's well known estimator $\hat{\rho}_n$ of correlation

coefficient ρ. The sample correlation coefficient $\hat{\rho}_n$ based on the data $(X_i, Y_i), i = 1, \ldots, n$, can be expressed as $\hat{\rho}_n = H(\bar{\mathbf{Z}}_n)$, and $\rho^* = H(\bar{\mathbf{Z}}^*_n)$, where

$$\mathbf{Z}_i = (X_iY_i, X_i^2, Y_i^2, X_i, Y_i), \quad \mathbf{Z}^*_i = (X^*_iY^*_i, X^{*2}_i, Y^{*2}_i, X^*_i, Y^*_i)$$

and

$$H(a_1, a_2, a_3, a_4, a_5) = \frac{(a_1 - a_4a_5)}{\sqrt{((a_2 - a_4^2)(a_3 - a_5^2))}}.$$

Note that H is a differentiable function.

In general, if the standard deviation of $T_n(\mathbf{X}; F)$ is not known (which is often the case), the function may be divided by a good estimator of the standard deviation of the statistic. This makes it an 'approximate pivotal' quantity. Such a correction by a special type of estimator of standard deviation for the smooth function model refers to Studentization, as it is similar to the Student's t-statistic. The empirical distribution of the data is used to estimate the standard deviation of the statistic in a special way, making it an 'approximate pivotal' quantity. For the smooth function model, a good estimator of the variance of $\sqrt{n}H(\bar{\mathbf{Z}}_n)$ is given by $\hat{\sigma}^2 = \ell^T(\bar{\mathbf{Z}}_n)\Sigma_n\ell(\bar{\mathbf{Z}}_n)$, where $\ell(\mathbf{x})$ denotes the vector of first order partial derivatives of H at $\mathbf{x}$, ^{T} denotes transpose, and Σ_n denotes the variance-covariance matrix computed from the $\{\mathbf{Z}_i\}$. That is,

$$\Sigma_n = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{Z}_i - \bar{\mathbf{Z}}_n)(\mathbf{Z}_i - \bar{\mathbf{Z}}_n)^T. \tag{4}$$

This leads to Studentization or approximate pivotal function

$$t_n = \sqrt{n}(H(\bar{\mathbf{Z}}_n) - H(\mathrm{E}(\mathbf{Z}_1))/\hat{\sigma} \tag{5}$$

Its bootstrap version is

$$t^*_n = \sqrt{n}(H(\bar{\mathbf{Z}}^*_n) - H(\bar{\mathbf{Z}}_n))/\sqrt{\ell^T(\bar{\mathbf{Z}}^*_n)\Sigma^*_n\ell(\bar{\mathbf{Z}}^*_n)}, \tag{6}$$

where Σ^*_n denotes the variance-covariance matrix computed from the bootstrap sample $\{\mathbf{Z}^*_i\}$, *i.e.*

$$\Sigma^*_n = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{Z}^*_i - \bar{\mathbf{Z}}^*_n)(\mathbf{Z}^*_i - \bar{\mathbf{Z}}^*_n)^T. \tag{7}$$

If $H(\bar{\mathbf{Z}}_n)$ represents the sample mean $\bar{X}_n$, then $\hat{\sigma}^2 = s_n^2$, and if $H(\bar{\mathbf{Z}}_n)$ represents the ratio statistic $\hat{\theta} = \bar{X}_n/\bar{Y}_n$, then $\hat{\sigma}^2 = \bar{Y}^{-2}n^{-1}\sum_{i=1}^{n}(X_i - \hat{\theta}Y_i)^2$.

Under very general conditions, if $\ell(\mathrm{E}(\mathbf{Z}_1)) \neq 0$, then the approximation of the sampling distribution of t_n by the bootstrap distribution (the distribution of t^*_n) is better than the classical normal approximation. This is mainly because the bootstrap automatically corrects for the skewness factor. This is established using Edgeworth expansion (see Babu & Singh (1983), and Babu & Singh (1984)):

$$P(t_n \leq x) = \Phi(x) + \frac{1}{\sqrt{n}}p(x)\phi(x) + \text{ error}$$
$$P^*(t^*_n \leq x) = \Phi(x) + \frac{1}{\sqrt{n}}p_n(x)\phi(x) + \text{ error}.$$

The 'error' terms are so small that

$$\sqrt{n}|\mathrm{P}(t_n \le x) - \mathrm{P}^*(t_n^* \le x)| \to 0.$$

The theory above is applicable in very general set up that includes the statistics: *Sample Means, Sample Variances, Central and Non-central t-statistics (with possibly non-normal populations), Sample Coefficient of Variation, Maximum Likelihood Estimators, Least Squares Estimators, Correlation Coefficients, Regression Coefficients*, and Smooth transforms of these statistics.

Thus the sampling distribution of several commonly occurring statistics are closer to the corresponding bootstrap distribution than the normal distribution. These conditional approximations are suggestive of the unconditional ones, though one cannot be derived from the other by elementary methods. Babu & Bose (1988) provide theoretical justification for the accuracy of the bootstrap confidence intervals both in terms of the actual coverage probability achieved and also the limits of the confidence interval.

In spite of these positive results, one should use caution in using bootstrap methods. It is not a 'cure all' solution. There are cases where bootstrap method fails. These include, non-smooth statistics such as $\hat{\theta} = \max_{1\le i\le n} X_i$ (see Bickel & Freedman (1981)), heavy tailed distributions, $\hat{\theta} = \bar{X}$ and $\mathrm{E}X_1^2 = \infty$ (see Babu (1984) and Athreya (1987)), and asymptotically non-linear statistics such as, $\hat{\theta} - \theta = H(\bar{\mathbf{Z}}_n) - H(\mathrm{E}(\mathbf{Z}_1)$ when $\partial H(\mathrm{E}(\mathbf{Z}_1)) = 0$. In the last case the limiting distribution is like that of linear combinations of Chi-squares, but here a modified version works (Babu (1984)).

3.4 Linear regression

Consider the simple linear regression model, where the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ satisfies

$$Y_i = \alpha + \beta X_i + e_i, \tag{8}$$

where α and β are unknown parameters, $X_1, \ldots, X_n$ are often called the design points. The error variables e_i need not be Gaussian, but are assumed to be independent with zero mean and standard deviation σ_i. This model is called homoscedastic if all the σ_i are identical. Otherwise, the model is known as heteroscedastic. In what follows, for any sequence of pairs $\{(U_1, V_1), \ldots, (U_n, V_n)\}$ of numbers, we use the notation

$$S_{UV} = \sum_{i=1}^{n} (U_i - \bar{U}_n)(V_i - \bar{V}_n) \quad \text{and} \quad \bar{U}_n = \frac{1}{n}\sum_{i=1}^{n} U_i. \tag{9}$$

There are two conceptually separate models to consider, random and fixed design models. In the first case, the pairs $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ are assumed to be random data points and the conditional mean and variance of e_i given X_i are assumed to be zero and σ_i^2. In the latter case, $X_1, \ldots, X_n$ are assumed to be fixed numbers (fixed design). In both the cases, the least squares estimators $\hat{\alpha}$ and $\hat{\beta}$ of α and β are given by

$$\hat{\beta} = S_{XY}/S_{XX} \quad \text{and} \quad \hat{\alpha} = \bar{Y}_n - \hat{\beta}\bar{X}_n. \tag{10}$$

However the variances of these estimators are different for a random and fixed designs, though the difference is very small for large n. We shall concentrate on the fixed design case here.

The variance of the slope $\hat{\beta}$ is given by

$$\text{var}(\hat{\beta}) = \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \sigma_i^2 / S_{XX}^2, \tag{11}$$

and depends on the individual error deviations σ_i, which may or may not be known. Knowledge of $\text{var}(\hat{\beta})$ provides the confidence intervals for β. Several resampling methods are available in the literature to estimate the sampling distribution and $\text{var}(\hat{\beta})$. We consider three bootstrap procedures: a) the classical bootstrap, b) weighted bootstrap, c) the paired bootstrap.

The classical bootstrap

Let $\hat{e}_i$ denote the residual of the i-th element of $\hat{e}_i = Y_i - \hat{\alpha} - \hat{\beta} X_i$ and define $\tilde{e}_i$ to be

$$\tilde{e}_i = \hat{e}_i - \frac{1}{n} \sum_{j=1}^{n} \hat{e}_j. \tag{12}$$

A bootstrap sample is obtained by randomly drawing $e_1^*, \ldots, e_n^*$ with replacement from $\tilde{e}_1, \ldots, \tilde{e}_n$. The bootstrap estimators β^* and α^* of the slope and the intercept are given by

$$\beta^* - \hat{\beta} = S_{Xe^*}/S_{XX} \quad \text{and} \quad \alpha^* - \hat{\alpha} = (\hat{\beta} - \beta^*)\bar{X}_n + \bar{e}_n^*. \tag{13}$$

To estimate the sampling distribution and variance, the procedure is repeated N times to obtain

$$\beta_1^*, \ldots, \beta_N^* \quad \text{where} \quad N \sim n(\log n)^2. \tag{14}$$

The histogram of these β^*s give a good approximation to the sampling distribution of $\hat{\beta}$ and the estimate of the variance $\hat{\beta}$ is given by

$$\text{var}_{\text{Boot}} = \frac{1}{N} \sum_{j=1}^{N} (\beta_j^* - \hat{\beta})^2. \tag{15}$$

This variance estimator is the best among all the three methods proposed here, if the residuals are homoscedastic; *i.e.* if the variances of the residuals $\text{E}(\epsilon_i^2) = \sigma_i^2 = \sigma^2$ are all the same. However if they are not, then the bootstrap estimator of the variance is an inconsistent estimator, and does not approach the actual variance. The *paired bootstrap* (§3.4) is robust against heteroscedasticity, giving consistent estimator of variance when the residuals have different standard deviations.

The paired bootstrap

The paired bootstrap are useful to handle heteroscedastic data. The paired bootstrap method treats the design points as random quantities. A simple random

sample $(\tilde{X}_1, \tilde{Y}_1), \ldots, (\tilde{X}_n, \tilde{Y}_n)$ is drawn from (X_1, Y_1), ..., (X_n, Y_n) and the paired bootstrap estimators of slope and intercept are constructed as

$$\tilde{\beta} = \frac{\sum_{i=1}^{n}(\tilde{X}_i - \bar{\tilde{X}})(\tilde{Y}_i - \bar{\tilde{Y}})}{\sum_{i=1}^{n}(\tilde{X}_i - \bar{\tilde{X}})^2}, \quad \text{and} \quad \tilde{\alpha} = \bar{\tilde{Y}} - \tilde{\beta}\bar{\tilde{X}}$$

The variance is obtained by repeating the resampling scheme N times and applying equation (15).

Figure 1 provides a simple FORTRAN code for jackknife and paired bootstrap resampling.

```
C       PAIRED BOOTSTRAP RESAMPLING
        NSIM = INT(N * ALOG(FLOAT(N))**2)
        DO 20 ISIM = 1,NSIM
        DO 10 I = 1,N
           J = INT(RANDOM * N + 1.0)
           XBOOT(I) = X(J)
  10       YBOOT(I) = Y(J)
  20    CONTINUE

C       JACKKNIFE RESAMPLING
        DO 40 NSIM = 1,N
        DO 30 I = 1,N-1
           IF(I.LT.NSIM)
             XJACK(I) = X(I)
             YJACK(I) = Y(I)
           ELSE
             XJACK(I) = X(I+1)
             YJACK(I) = Y(I+1)
           ENDELSE
  30    CONTINUE
  40    CONTINUE
```

Figure 1: FORTRAN code illustrating the paired bootstrap and jackknife resampling for a two dimensional dataset $(x_i, y_i), i = 1, \ldots, N$.

The bootstrap methodology, mathematics and second order properties are reviewed in Babu & Rao (1993). A detailed account of second order asymptotics can be found in Hall (1992). A less mathematical overview of the bootstrap is presented in Efron and Tibshirani (1993). The book by Zoubir & Iskander (2004) serves as a handbook on 'bootstrap' for engineers, to analyze complicated data with little or no model assumptions. Bootstrap has found many applications in engineering field including, artificial neural networks, biomedical engineering, environmental engineering, image processing, and Radar and sonar signal processing. Majority of the applications in the book are taken from signal processing literature.

References

[1] Athreya, K. B. (1987). Bootstrap of the mean in the infinite variance case. *Ann. Statist.,* **15**, no. 2, 724-731.

[2] Babu, G. J. (1984). Bootstrapping statistics with linear combinations of chi-squares as weak limit. *Sankhya*, Series A, **46**, 85-93.

[3] Babu, G. J. (1986). A note on bootstrapping the variance of the sample quantile. *Ann. Inst. Statist. Math.*, **38**, Part A, 439-443.

[4] Babu, G. J., and Bose, A. (1988). Bootstrap confidence intervals. *Statistics and Probability Letters*, **7**, 151-160.

[5] Babu, G. J., and Rao, C. R. (1993). Bootstrap methodology. In *Handbook of Statistics,* **Vol. 9** "Computational Statistics." C. R. Rao (Ed.), Elsevier Science Publishers B. V., Amsterdam, 627-659.

[6] Babu, G. J., and Singh, K. (1983). Inference on means using the bootstrap. *Annals of Statistics*, **11**, 999-1003.

[7] Babu, G. J., and Singh, K. (1984a). On one term Edgeworth correction by Efron's Bootstrap. *Sankhya*, Series A, **46**, 219-232.

[8] Babu, G. J., and Singh, K. (1984b). Asymptotic representations related to jackknifing and bootstrapping L-statistics. *Sankhyā*, Series A, **46**, 195-206.

[9] Bickel, P. J., and Freedman, D. A. Some asymptotic theory for the bootstrap. *Ann. Statist.*, **9**, no. 6, 1196–1217.

[10] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.,* **7**, no. 1, 1-26.

[11] Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans.* CBMS-NSF Regional Conference Series in Applied Mathematics, 38. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa.

[12] Efron, B., and Tibshirani, R. J. (1993). *An introduction to the bootstrap.* Monographs on Statistics and Applied Probability, 57. Chapman and Hall, New York.

[13] Hall, P. (1992). *The bootstrap and Edgeworth expansion.* Springer Series in Statistics. Springer-Verlag, New York.

[14] Quenouille, M. (1949). Approximate tests of correlation in time series. *J. Roy. Statist. Soc.*, Ser. B, **11**, pp. 1884.

[15] Zoubir, A. M., and Iskander, D. R. (2004). *Bootstrap Techniques for Signal Processing.* Cambridge University Press, Cambridge, U.K.

Goodness-of-Fit using bootstrap

Notes by

G. Jogesh Babu

Center for Astrostatistics
The Pennsylvania State University

1 The Problem of Model Selection and Fitting

The aim of model fitting is to provide most parsimonious 'best' fit of a parametric model to data. It might be a simple, heuristic model to phenomenological relationships between observed properties in a sample of astronomical objects. Examples include characterizing the Fundamental Plane of elliptical galaxies or the power law index of solar flare energies. Perhaps more important are complex nonlinear models based on our astrophysical understanding of the observed phenomenon. Here, if the model family truly represents the underlying phenomenon, the fitted parameters give insights into sizes, masses, compositions, temperatures, geometries, and evolution of astronomical objects. Examples of astrophysical modeling include:

- Interpreting the spectrum of an accreting black hole such as a quasar. Is it a nonthermal power law, a sum of featureless blackbodies, and/or a thermal gas with atomic emission and absorption lines?
- Interpreting the radial velocity variations of a large sample of solar-like stars. This can lead to discovery of orbiting systems such as binary stars and exoplanets, giving insights into star and planet formation.
- Interpreting the spatial fluctuations in the cosmic microwave background radiation. What are the best fit combinations of baryonic, Dark Matter and Dark Energy components? Are Big Bang models with quintessence or cosmic strings excluded?

The mathematical procedures used to link data with astrophysical models fall into the realm of statistics. The relevant methods fall under the rubrics of statistical model selection, regression, and goodness-of-fit. Astronomers' understanding of such methods are often rather simplistic, and we seek here to develop increased sophistication in some aspects of the methodological issues. We discuss the advantages and limitations of some traditional model fitting methods and suggest new procedures when these methods are inadequate. In particular, we discuss some recently developed procedures based on nonparametric resampling designed for model selection and goodness-of-fit when the astronomer not only seeks the best parameters of the model, but wishes to consider entirely different families of parametric models.

2 Challenges of Model Selection and Fitting

Consider the astronomical spectrum illustrated in Figure 1a where flux from a source is plotted against energy of light received by an X-ray telescope. Here the photons are shown collected into constant-width bins, and the measured flux value F is accompanied by an error bar σ representing the uncertainty of the intensity at each energy based on the square-root of the number of counts in the bin. The dataset shown happens to be a low-resolution spectrum from the *Chandra* Orion Ultradeep Project (COUP) where NASA's *Chandra* X-ray Observatory observed about 1400

pre-main sequence stars in the Orion Nebula region for 13 days (Getman *et al.* 2005). But it could easily be an optical spectrum of a high-redshift starburst galaxy, or a millimeter spectrum of a collapsing molecular cloud core, or the spectrum of a gamma-ray burst at the birth of a black hole.

The histogram in Figure 1a shows the best-fit astrophysical model assuming a plausible emission mechanism: a single-temperature thermal plasma with solar abundances of elements. This model M has three free parameters – plasma temperature, line-of-sight absorption, and normalization – which we denote by the vector θ. The astrophysical model has been convolved with complicated functions representing the sensitivity of the telescope and detector. The model is fitted by minimizing $\chi^2(\theta) = \sum[F_i - M_i(\theta)]^2/\sigma_i^2$ with an iterative procedure. Confidence intervals on best-fit parameter values are obtained using a χ^2_{min}-plus-constant criterion. These procedures are familiar in the astronomical community (*e.g.* Bevington 1969).

There are important limitations to χ^2 minimization for use in modern astronomical model selection and fitting. It fails when the errors are non-Gaussian (*e.g.* small-N problems with Poissonian errors). It does not provide clear procedures for adjudicating between models with different numbers of parameters (*e.g.* one- vs. two-temperature models) or between different acceptable models (*e.g.* local minima in $\chi^2(\theta)$ space). It can be difficult to obtain confidence intervals on parameters when complex correlations between the parameters are present (*e.g.* non-parabolic shape near the minimum in $\chi^2(\theta)$ space).

Figure 1b shows an important alternative approach to the model fitting and goodness-of-fit problem. Here the energies of photons of observed spectrum are shown individually rather than in a binned histogram. In statistical parlance, this is called the empirical distribution function (EDF), and is advantageous over the binned histogram because the exact measured values are used. This avoids the often arbitrary choices of bin width(s) and starting point in histograms, and the sometimes-inaccurate assumption of $\sqrt{n}$ error bars on binned values. There is a large statistical literature on the difficulty of choosing bin widths, and indeed on choosing between histograms and other data smoothing procedures. Narrow bins or smoothing kernels can be dominated by noise while wide bins can miss physically important structure.

Among astronomers, the Kolmogorov-Smirnov (K-S) statistic is popular, although other EDF based statistics such as the Cramer-von Mises (C-vM) and Anderson-Darling (A-D) statistics have better sensitivity for some data-model differences. However, as we review in §3 below, *the goodness-of-fit probabilities derived from the K-S or other EDF statistics are usually not correct when applied in model fitting situations with estimated parameters.* Astronomers are thus often making errors in EDF model fitting.

Figure 1c illustrates another major astrostatistical question: When a "good" model is found with parameters θ_0, what is an acceptable range of parameter values around θ_0 consistent with the data? In the example shown, we might ask: "What is the confidence interval of absorption consistent with the data at 99% significance?" This question is not simple to answer. The scientist must specify in advance whether the parameter of interest is considered in isolation or in consort

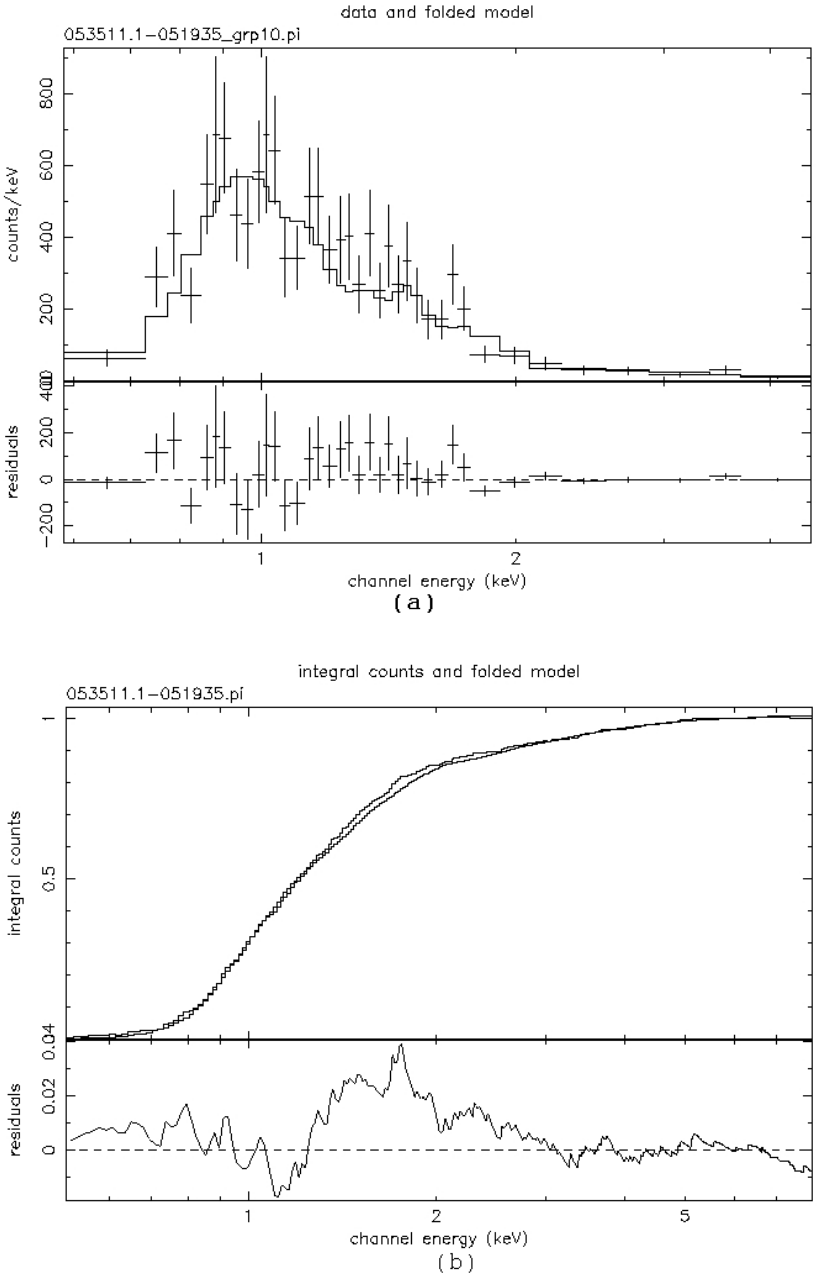

Figure 1: An example of astrophysical model fitting using a spectrum with 264 photons from the *Chandra* X-ray Observatory. (a) Best-fit thermal model (histogram) to differential binned data (separated points with error bars)obtained by minimum-χ^2. Here the absorption parameter has value $A_V \sim 1$ mag. Data-minus-residuals appear in the bottom plot. (b) Thermal model (smooth curve) obtained by minimizing the K-S statistic to the integral EDF (step function). The resulting parameters very similar to the χ^2 fit.

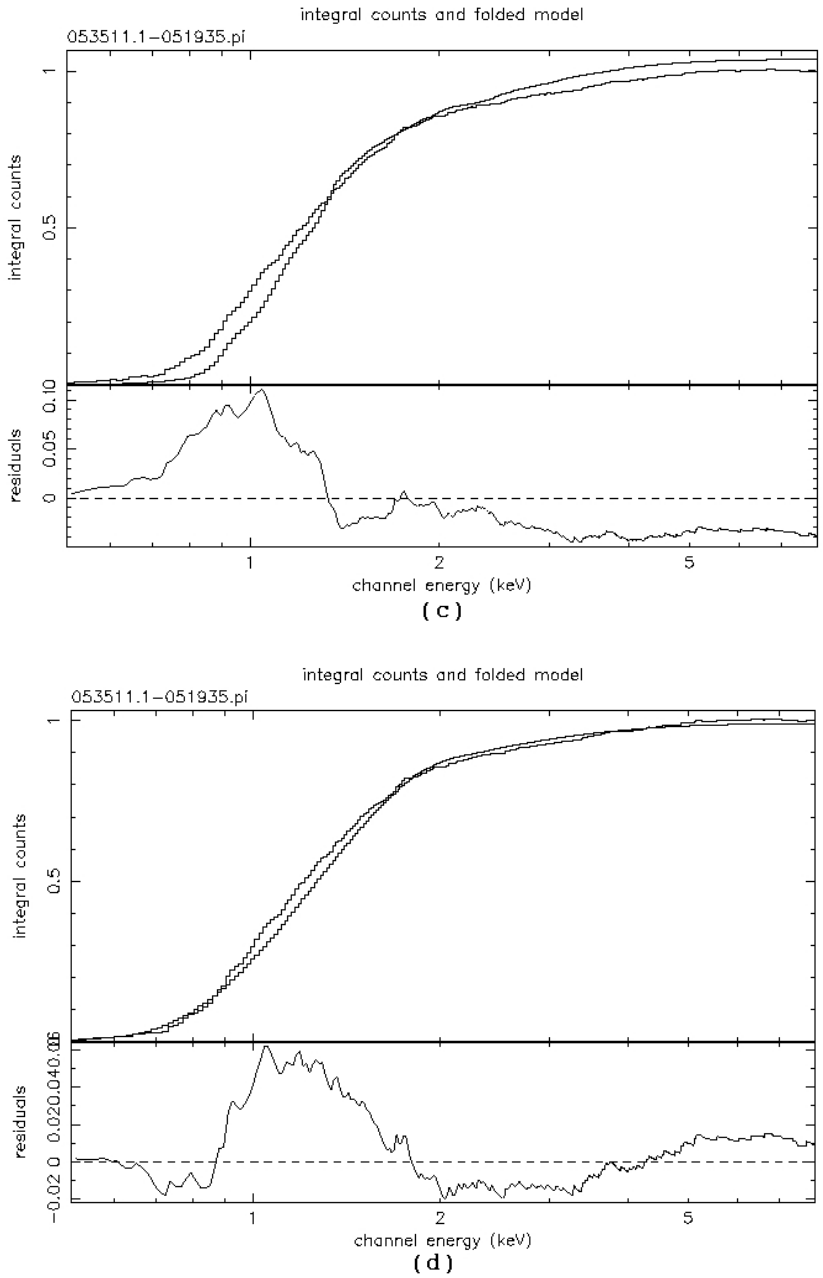

Figure 1: Continued. (c) An example of the correct model family but incorrect parameter value: thermal model with absorption set at $A_V = 10$ mag. (d) An example of an incorrect model family: best-fit powerlaw model with absorption $A_V \sim 1$ mag.

with other parameters, whether the statistical treatment involves binned histograms or EDFs, and whether 67% (1σ equivalent), 90% or 99.7% (3σ equivalent) values should be reported. The statistician must decide which statistic to use, whether normal approximations are valid, and how extraneous model parameters should be treated.

Finally, Figure 1d treats a broader scientific question: Are the data consistent with *different families* of astrophysical models, irrespective of the best-fit parameter values within a family? We illustrate this here by obtaining the best-fit model using a nonthermal power law X-ray spectrum rather than a thermal plasma X-ray spectrum. Among statisticians, these are called 'non-nested' models. Even decisions between nested models can be tricky; for example, should the dataset in Figure 1 be modeled with thermal models with arbitrary elemental abundances, or is the assumption of solar abundances adequate?

3 Inference for Statistics Based on the EDF

Figure 2a shows a hypothetical EDF, the cumulative frequency distribution function of the data. The three commonly used statistics, for inference on F, based on EDF mentioned above are:

Kolmogorov-Smirnov (K-S): $\sup |F_n(x) - F(x)|$

Cramér-von Mises (C-vM): $\int^{x} (F_n(x) - F(x))^2 \, dF(x)$,

and Anderson - Darling (A-D): $\int \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} \, dF(x)$.

Here F_n is the EDF, F is the model distribution function, and "sup" means the supremum. The K-S statistic is most sensitive to large-scale differences in location (*i.e.* median value) and shape between the model and data. The C-vM statistic is effective for both large-scale and small-scale differences in distribution shape. Both of these measures are relatively insensitive to differences near the ends of the distribution. This deficiency is addressed by the A-D statistic, a weighted version of the C-vM statistic to emphasize differences near the ends.

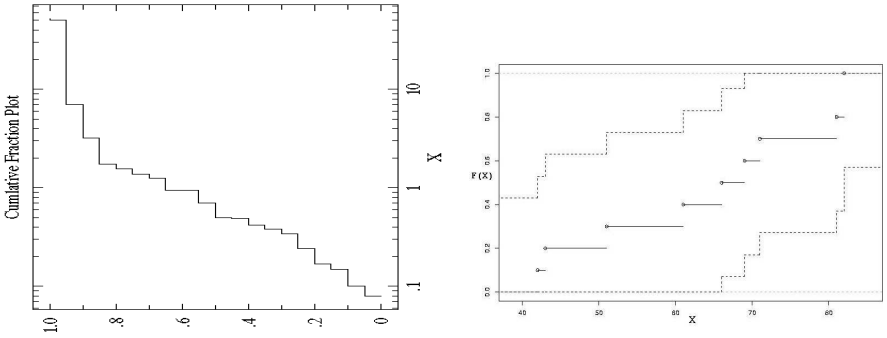

Figure 2: (a) A hypothetical EDF. (b) Confidence bands around the EDF based on the K-S statistic for 90% significance level.

The power of these statistics is that they are distribution-free as long as F is continuous. That is, the probability distribution of these statistics is free from F. Consequently, the confidence bands for the 'unknown' distribution F can be obtained from standard tables of K-S, C-vM or A-D probabilities which depend only on the number of data points and the chosen significance level. A typical confidence band based on Kolmogorov-Smirnov test resembles Figure 2b.

But all these statistics are no longer distribution-free under two important and common situations: when the data are multivariate, or when the model parameters are estimated using the data. We address these situations here.

3.1 Failure of the multivariate case

Let (X_1, Y_1) be a data point from a bivariate distribution F on the unit square. Simpson (1951) shows that if F_1 denotes the EDF of (X_1, Y_1), then

$$P\big(|F_1(x,y) - F(x,y)| < .72, \text{ for all } x, y\big) \begin{cases} > 0.065 & \text{if} \quad F(x,y) = xy^2 \\ < 0.058 & \text{if} \quad F(x,y) = xy(x+y)/2. \end{cases}$$

Thus, the distribution of the K-S statistic varies with the unknown F and hence is not distribution-free when two or more dimensions are present. The K-S statistic still is a measure of "distance" between the data and model, but probabilities can not be assigned to a given value of the statistic without detailed calculation for each case under consideration. Several methodological studies in the astronomical literature discuss two-dimensional K-S tests. The results may be unreliable to degrees that can not readily be calculated.

3.2 Failure when parameters are estimated from the data

The K-S statistic is also no longer distribution-free if some parameters are estimated from the dataset under consideration. For example, consider the question whether the illustrated X-ray spectrum supports a powerlaw in addition to a thermal model (Figure 1d). It may seem natural to find the best-fit powerlaw and best-fit thermal models by a procedure such as maximum likelihood, compute the K-S statistic for each case, and evaluate which model is acceptable using the probabilities in standard tables. But it has long been established that the K-S probabilities are incorrect in this circumstance (Lilliefors 1969). The K-S probabilities are only valid if the model being tested is derived independently of the dataset at hand; *e.g.* from some previous datasets or from prior astrophysical considerations.

4 Bootstrap resampling: A good solution

Fortunately, there is an alternative to the erroneous use of K-S procedure, although it requires a numerically intensive calculation for each dataset and model addressed. It is based on bootstrap resampling, a data-based Monte Carlo method that has been mathematically shown to give valid estimates of goodness-of-fit probabilities under a very wide range of situations (Babu and Rao 1993).

We now outline the mathematics underlying bootstrap calculations. Let $\{F(.;\theta) : \theta \in \Theta\}$ be a family of continuous distributions parametrized by θ. We want to test whether the univariate dataset $X_1, \ldots, X_n$ comes from $F = F(.;\theta)$ for some $\theta = \theta_0$. The K-S, C-vM and A-D statistics (and a few other goodness-of-fit tests) are continuous functionals of the process, $Y_n(x;\hat{\theta}_n) = \sqrt{n}\big(F_n(x) - F(x;\hat{\theta}_n)\big)$. Here F_n denotes the EDF of $X_1, \ldots, X_n$, $\hat{\theta}_n = \theta_n(X_1, \ldots, X_n)$ is an estimator of θ derived from the dataset, and $F(x;\hat{\theta}_n)$ is the model being tested. For a simple example, if $\{F(.;\theta) : \theta \in \Theta\}$ denotes the Gaussian family with $\theta = (\mu, \sigma^2)$, then $\hat{\theta}_n$ can be taken as $(\bar{X}_n, s_n^2)$ where $\bar{X}_n$ is the sample mean and s_n^2 is the sample variance based on the data $X_1, \ldots, X_n$. In the astrophysical example considered in §2, F is the family of thermal models with three parameters.

In the case of evaluating goodness-of-fit for a model where the parameters have been estimated from the data, the bootstrap can be computed in two different ways: the *parametric bootstrap* and the *nonparametric bootstrap*. The parametric bootstrap may be familiar to the astronomer as a well-established technique of creating fake datasets realizing the parametric model by Monte Carlo methods (*e.g.* Press et al. 1997). The actual values in the dataset under consideration are not used. The nonparametric bootstrap, in contrast, is a particular Monte Carlo realizations of the observed EDF using a "random selection with replacement" procedure.

We now outline the mathematics underlying these techniques. Let $\hat{F}_n$ be an estimator of F, based on $X_1, \ldots, X_n$. In order to bootstrap, we generate data $X_1^*, \ldots, X_n^*$ from the estimated population $\hat{F}_n$ and then construct $\hat{\theta}_n^* = \theta_n(X_1^*, \ldots, X_n^*)$ using the same functional form. For example, if $F(.;\theta)$ is Gaussian with $\theta = (\mu, \sigma^2)$ and if $\hat{\theta}_n = (\bar{X}_n, s_n^2)$, then $\hat{\theta}_n^* = (\bar{X}_n^*, s_n^{*2})$.

4.1 Parametric Bootstrap

The bootstrapping procedure is called parametric if $\hat{F}_n = F(.;\hat{\theta}_n)$; that is, we generate data $X_1^*, \ldots, X_n^*$ from the model assuming the estimated parameter values $\hat{\theta}_n$. The process $Y_n^P(x) = \sqrt{n}\big(F_n^*(x) - F(x;\hat{\theta}_n^*)\big)$ and the sample process $Y_n(x) = \sqrt{n}\big(F_n(x) - F(x;\hat{\theta}_n)\big)$ converge to the same Gaussian process Y. Consequently, $L_n = \sqrt{n}\,\sup_x |F_n(x) - F(x;\hat{\theta}_n)|$ and $L_n^* = \sqrt{n}\,\sup_x |F_n^*(x) - F(x;\hat{\theta}_n^*)|$ have the same limiting distribution. For the K-S statistic, the critical values of L_n can be derived as follows: construct B resamples based on the parametric model ($B \sim 1000$ should suffice), arrange the resulting L_n^* values in increasing order to obtain 90 or 99 percentile points for getting 90% or 99% critical values. This procedure replaces the incorrect use of the standard probability tabulation.

4.2 Nonparametric Bootstrap

The nonparametric bootstrap involving resamples from the EDF;

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

is operationally easy to perform but requires an additional step of bias correction

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

The sample process Y_n and the bias corrected nonparametric process Y_n^N converge to the same Gaussian process Y. That is, $L_n = \sqrt{n}\sup_x |F_n(x) - F(x;\hat{\theta}_n)|$ and $J_n^* = \sup_x |\sqrt{n}\left(F_n^*(x) - F(x;\hat{\theta}_n^*)\right) - B_n(x)|$ have the same limiting distribution. The critical values of the distribution of L_n can then be derived as in the case of parametric bootstrap. For detailed understanding of the regularity conditions under which these results hold see Babu and Rao (2004).

5 Confidence Limits Under Misspecification of Model Family

We now address the more advanced problem of comparing best-fit models derived for non-nested model families; *e.g.* the powerlaw vs. thermal model fits in Figure 1. Essentially, we are asking 'How far away' is the unknown distribution underlying the observed dataset from the hypothesized family of models?

Let the original dataset $X_1, \ldots, X_n$ come from an unknown distribution H. H may or may not belong to the family $\{F(.;\theta) : \theta \in \Theta\}$. Let $F(.,\theta_0)$ be the specific model in the family that is 'closest' to H where proximity is based on the Kullback-Leibler information, $\int \log\big(h(x)/f(x;\theta)\big)dH(x) \geq 0$, which arises naturally due to maximum likelihood arguments and has advantageous properties. Here h and f are the densities (*i.e.* derivatives) of H and F.

If the maximum likelihood estimator $\hat{\theta}_n \to \theta_0$, then $U_n(x;\hat{\theta}_n) = \sqrt{n}\big(F_n(x) - F(x;\hat{\theta}_n)\big) - \sqrt{n}\big(H(x) - F(x;\theta_0)\big)$ converges weakly to a Gaussian process U (Babu and Rao 2003). In this (nonparametric bootstrap) case, $Y_n^N(x) = \sqrt{n}\big(F_n^*(x) - F(x;\hat{\theta}_n^*)\big) - \sqrt{n}\big(F_n(x) - F(x;\hat{\theta}_n)\big)$, and U_n converge to the same Gaussian process. For the K-S statistic, for any $0 < \alpha < 1$,

$$P\big(\sqrt{n}\sup_x |F_n(x) - F(x;\hat{\theta}_n) - (H(x) - F(x;\theta_0))| \leq C_\alpha^*\big) \; - \alpha \to 0,$$

where C_α^* is the α-th quantile of $\sup_x |\sqrt{n}\left(F_n^*(x) - F(x;\hat{\theta}_n^*)\right) - \sqrt{n}\big(F_n(x) - F(x;\hat{\theta}_n)\big)|$. This provides an estimate of the distance between the true distribution and the family of distributions under consideration (Babu and Bose 1988).

References

[1] Babu, G. J., and Bose, A. (1988). Bootstrap confidence intervals. *Statistics & Probability Letters*, **7**, 151-160.

[2] Babu, G. J., and Rao, C. R. (1993). Bootstrap methodology. In *Computational statistics*, Handbook of Statistics **9**, C. R. Rao (Ed.), North-Holland, Amsterdam, 627-659.

[3] Babu, G. J., and Rao, C. R. (2003). Confidence limits to the distance of the true distribution from a misspecified family by bootstrap. *J. Statistical Planning and Inference*, **115**, no. 2, 471-478.

[4] Babu, G. J., and Rao, C. R. (2004). Goodness-of-fit tests when parameters are estimated. *Sankhyā*, **66**, no. 1, 63-74.

[5] Bevington, P. R. (1969). Data reduction and error analysis for the physical sciences. McGraw-Hill.

[6] Getman, K. V., and 23 others (2005). Chandra Orion Ultradeep Project: Observations and source lists. *Astrophys. J. Suppl.*, **160**, 319-352.

[7] Lilliefors, H. W. (1969). On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown. *Journal of the American Statistical Association*, **64**, No. 325, 387-389

[8] Press, W. H. et al. (1997). Numerical recipes in C: The art of scientific computing. Cambridge Univ. Press.

[9] Simpson, P. B. (1951). Note on the estimation of a bivariate distribution function. *Ann. Math. Stat.*, **22**, 476–478.