Model Fitting and Model Selection

G. Jogesh Babu
Center for Astrostatistics
The Pennsylvania State University

1 Introduction

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-square with an iterative procedure. That is

$$\hat{\theta} = \arg\min_{\theta} \chi^2(\theta) = \arg\min_{\theta} \sum_{i=1}^{N} \left(\frac{y_i - M_i(\theta)}{\sigma_i} \right)^2 .$$

Chi-square minimization is a misnomer. It is known as parameter estimation by *weighted least squares*. 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 astronomical model selection and fitting. The procedure depends strongly on Gaussian assumptions. 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 estimators of 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.

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 with other parameters, whether the statistical treatment

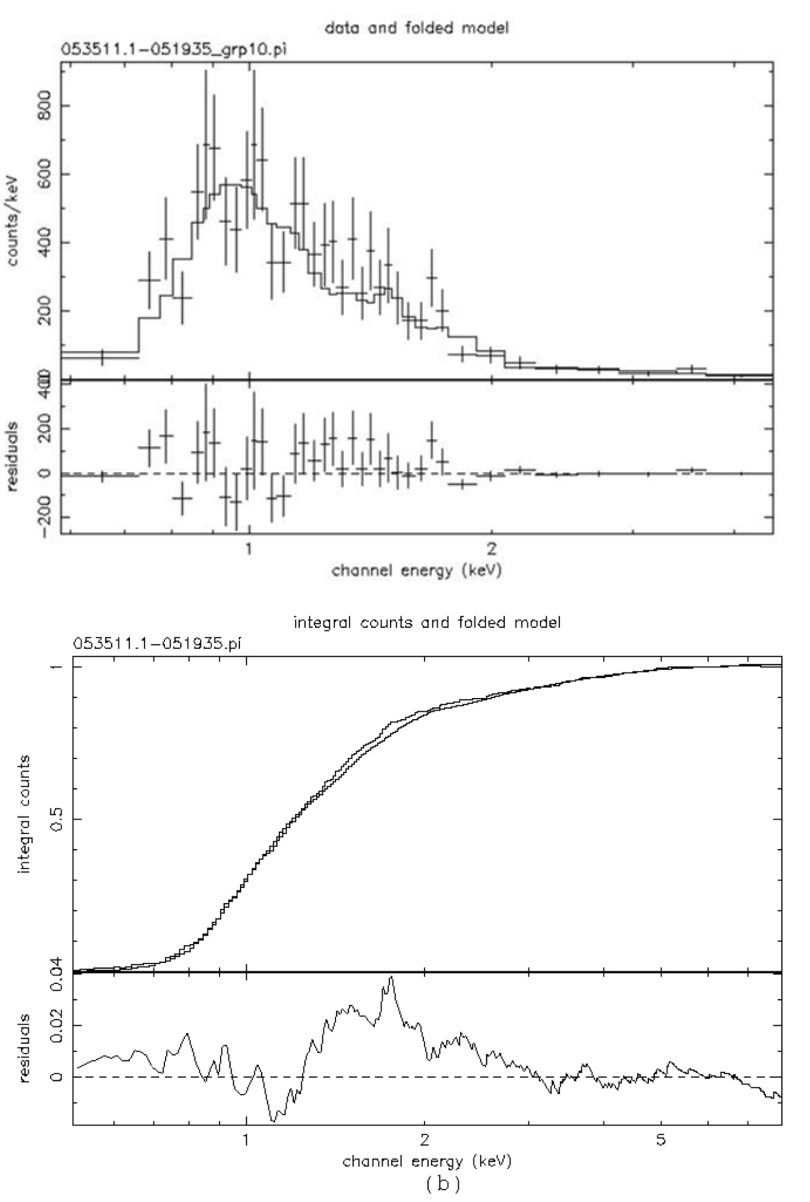

Figure 1: An example of astrophysical model fitting using a spectrum with 264 photons from the *Chandra* X-ray Observatory. (a) Best-fit thermal model (histogram) to differential binned data (separated points with error bars)obtained by minimum-χ^2. Here the absorption parameter has value $A_V \sim 1$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

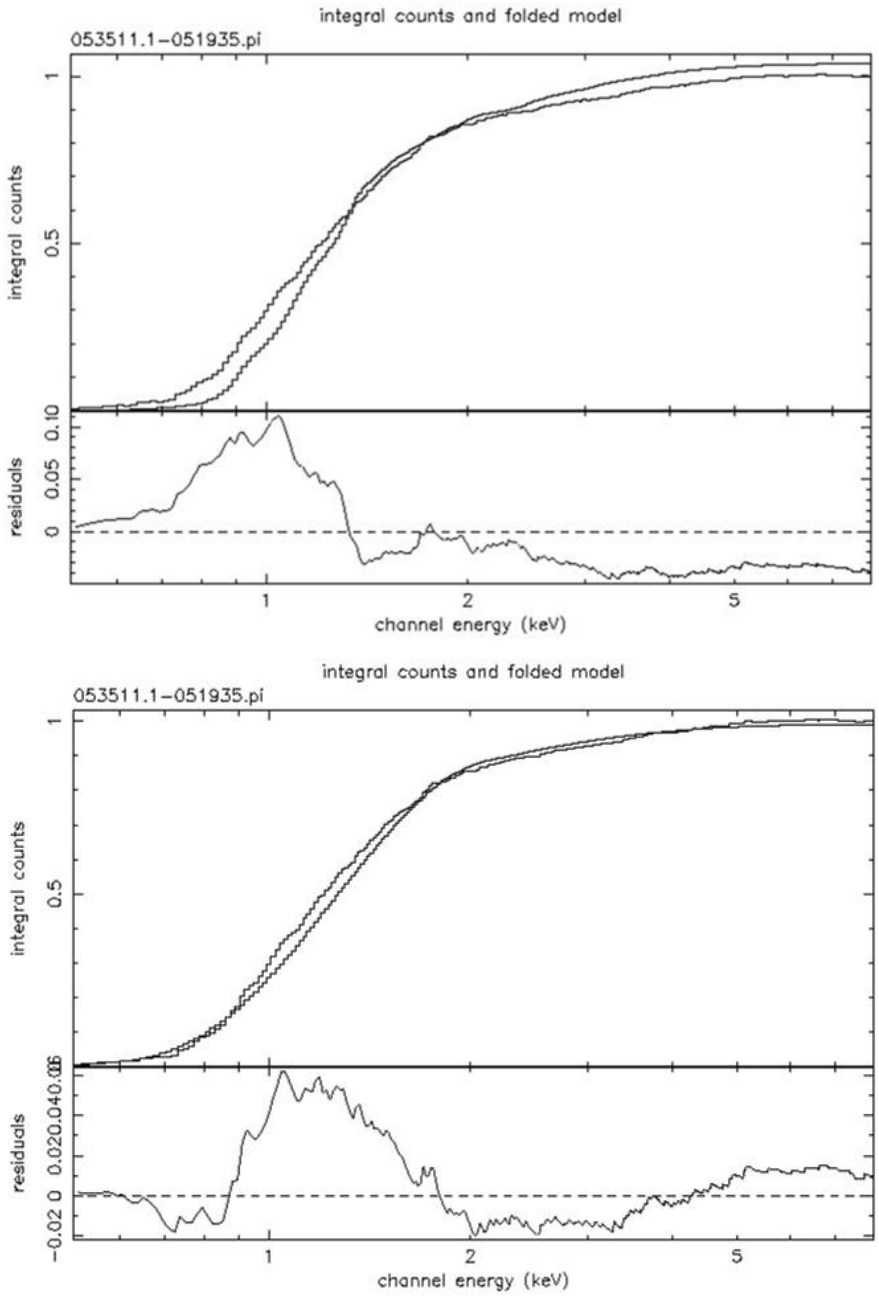

Figure 1: Continued. (c) An example of the correct model family but incorrect parameter value: thermal model with absorption set at $A_V = 10$ mag. (d) An example of an incorrect model family: best-fit powerlaw model with absorption $A_V \sim 1$ mag.

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 Goodness of Fit: Statistics Based on the EDF

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 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.

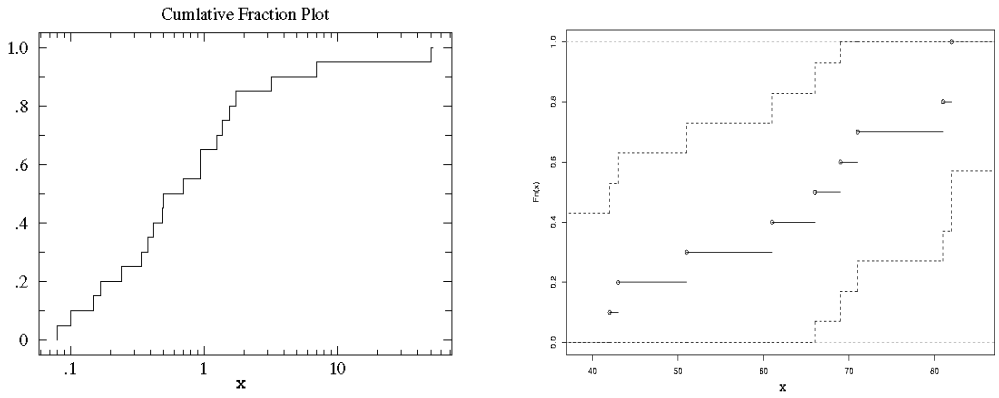

Figure 2: (a) A hypothetical EDF. (b) Confidence bands around the EDF based on the K-S statistic for 90% significance level.

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.

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).

6 Model Selection

A good statistical model should be parsimonious (model simplicity), conform fitted model to the data (goodness of fit), and should be easily generalizable. Occam's Razor, a philosophical principle credited to the English Philosopher William of Ockham (1285-1349), that

essentially says that the simplest solution is usually the correct one, is the main guiding principle for statistical modeling. Occam's Razor suggests that we leave off extraneous ideas to better reveal the truth. That is, select a model that adequately accommodates the data. It neither *underfits* that excludes key variables or effects, nor *overfits* that unnecessarily be complex by including extraneous explanatory variables or effects. Underfitting induces bias and overfitting induces high variability. A model selection criterion should balance the competing objectives of conformity to the data and parsimony.

Hypothesis testing is one of the criteria used for comparing two models. Classical hypothesis testing methods are generally used for nested models. However, it does not treat models symmetrically. To set up framework for general model selection, let D denote the observed data and let $M_1, \ldots, M_k$ denote the models for D under consideration. Each model M_j, let $f(D|\theta_j; M_j)$ and $\ell(\theta_j) = \log f(D|\theta_j; M_j)$ denote the likelihood and loglikelihood respectively, θ_i is a p_j dimensional parameter vector. Here $f(D|\theta_j; M_j)$ denotes the probability density function (in the continuous case) or probability mass function (in the discrete case) evaluated at the data D. Most of the methodology can be framed as a comparison between two models M_1 and M_2.

6.1 Special case of Nested Models

The model M_1 is said to be nested in M_2, if some coordinates of θ_1 are fixed, *i.e.* $\theta_2 = (\alpha, \gamma)$ and $\theta_1 = (\alpha, \gamma_0)$, where γ_0 is some known fixed constant vector. In this case, comparison of M_1 and M_2 can be considered as a classical hypothesis testing problem of $H_0 : \gamma = \gamma_0$.

For example, the model M_2 refers to normal with mean μ and variance σ^2, while M_1 refers to normal with mean 0 and variance σ^2. The model selection problem can thus be framed in terms of statistical hypothesis testing $H_0 : \mu = 0$, with free parameter σ. There are some objections to using hypothesis testing to decide between the two models M_1 and M_2, as they are not treated symmetrically by the test in which the null hypothesis is M_1. We cannot *accept* H_0, we can only reject or fail to reject H_0. As larger samples can detect the discrepancies, they tend to make it more likely to reject the null hypothesis.

We now look at three different ideas for testing H_0.

6.2 Three statistical hypotheses tests

The *Wald Test*, the *Likelihood Ratio Test*, and the *Rao's Score Test*, based on maximum likelihood estimators, are collectively referred to in statistical literature as the *Holy Trinity* of statistical hypotheses. These statistical hypotheses tests can be used to test linear and non-linear restrictions among parameters. The three tests are described below in the case of scalar (1-dimensional) parameter θ.

To test the null hypothesis $H_0 : \theta = \theta_0$, the Wald Test uses $W_n = (\hat{\theta}_n - \theta_0)^2 / Var(\hat{\theta}_n)$, the standardized distance between θ_0 and the maximum likelihood estimator $\hat{\theta}_n$ based on a data of size n. The distribution of W_n is approximately the Chi-sqaure distribution with one degree of freedom. In general variance of $\hat{\theta}_n$ is not known, however, a close approximation is $1/I(\hat{\theta}_n)$, where $I(\theta) = E((f'(X;\theta)/f(X;\theta))^2)$ is the Fisher's information, f denotes the probability density function of the random variable X, and f' denotes the derivative of f with respect to θ. Thus $I(\hat{\theta}_n)(\hat{\theta}_n - \theta_0)^2$ has chi-square distribution in the limit, and the Wald test rejects the null hypothesis H_0, when $I(\hat{\theta}_n)(\hat{\theta}_n - \theta_0)^2$ is large.

The Likelihood Ratio Test uses the logarithm of ratio of likelihoods, $\ell(\hat{\theta}_n) - \ell(\theta_0)$, where $\ell(\theta)$ denotes the loglikelihood at θ. While Rao's Score Test (also known as Lagrangian Multiplier Test) uses the statistic $S(\theta_0) = (\ell'(\theta_0))^2/(nI(\theta_0))$, where ℓ' denotes the derivative of ℓ, and as before I denotes the Fisher's Information. That is, if $X, X_1, \ldots, X_n$ denote independent random variables from a common probability density function $f(.;\theta)$, then $\ell'(\theta_0) = \sum_{i=1}^n (f'(X_i;\theta_0)/f(X_i;\theta_0))$. Hence

$$S(\theta_0) = \frac{1}{nI(\theta_0)} \left(\sum_{i=1}^n \frac{f'(X_i;\theta_0)}{f(X_i;\theta_0)} \right)^2 .$$

For example, in the case of data from normal (Gaussian) distribution

$$f(y;(\mu,\sigma^2)) = \frac{1}{\sigma}\,\phi((y-\mu)/\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\,\exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\},$$

where ϕ denotes the standard normal probability density function.

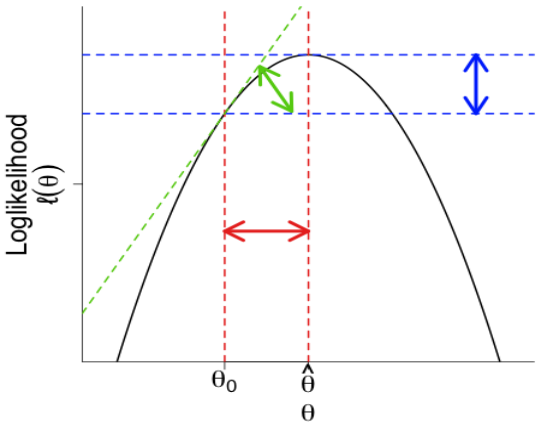

Figure 3: Wald Test is based on the distance between $\hat{\theta}_n$ and θ_0; the Likelihood Ratio Test is based on the distance from $\ell(\theta_0)$ to $\ell(\hat{\theta}_n)$, the loglikelihoods; the Rao's Score Test is based on the gradient of the loglikelihood at θ_0.

The three tests are equivalent to the first order of asymptotics, but differ to some extent in the second order properties. No single test among these three is uniformly better than the others.

In the regression context with data $y_1, \ldots, y_n$ and Gaussian residuals, the loglikelihood ℓ is given by

$$\ell(\beta) = \log \prod_{i=1}^n \frac{1}{\sigma}\,\phi((y_i - \mathbf{x}_i'\beta)/\sigma).$$

6.3 Information Criteria based model selection

If the model M_1 happens to be nested in the model M_2, the largest likelihood achievable by M_2 will *always* be larger than that achievable by M_1. It suggests adding a penalty on "larger" models would achive a balance between overfitting and underfitting. This leads to the so called *Penalized Likelihood approach.*

The traditional maximum likelihood paradigm, as applied to statistical modeling, provides a mechanism for estimating the unknown parameters of a model having a specified dimension and structure. Hirotugu Akaike extended this paradigm in [1] by considering a framework in which the model dimension is also unknown. He proposed a framework where both model estimation and selection could be simultaneously accomplished. Grounding in the concept of entropy, Akaike proposed *an information criterion* (AIC), which is now popularly known as *Akaike's Information Criterion*, and is defined for model M_j, as $-2\ell(\hat{\theta}_j)+2p_j$. The term $2\ell(\hat{\theta}_j)$ is known as the *goodness of fit* term, and $2p_j$ is known as the *penalty* term. This penalty term increase as the complexity of the model grows. AIC is generally regarded as the first model selection criterion, and it continues to be the most widely known and used model selection tool among practitioners.

One advantage of AIC is that it does not require the assumption that one of the candidate models is the "true" or "correct" model. It treats all the models symmetrically, unlike hypothesis testing. AIC can be used to compare nested as well as non-nested models. AIC can also be used to compare models based on different families of probability distributions. One of the disadvantages of AIC is the requirement of large samples especially in complex modeling frameworks. In addition, it is not *consistent*, in the sense that if p_0 is the correct number of parameters, and $\hat{p} = p_i$ $(i = \arg\min_j\ (-2\ell(\hat{\theta}_j)+2p_j))$, then $\lim_{n\to\infty} P(\hat{p} > p_0) > 0$. That is even if we have very large number of observations, $\hat{p}$ does not approach the true value.

Bayesian Information Criterion (BIC), sometimes called the *Schwarz Bayesian Criterion* is another popular model selection criteria. Unlike AIC, BIC defined as

$$-2\ell(\hat{\theta}_j) + p_j \log n$$

is consistent. Like AIC, the models need not be nested to be compared using BIC.

Conditions under which these two criteria are mathematically justified are often ignored in practice. Some practitioners apply them even in situations where they **should not be** applied. AIC penalizes free parameters less strongly than does the Schwartz's BIC. A note of caution: sometimes, these criteria are given a minus sign so the goal changes to finding the maximizer.

References

[1] Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In *Second International Symposium on Information Theory*, (B. N. Petrov and F. Csaki, Eds). Akademia Kiado, Budapest, 267-281.

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

[3] 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.

[4] 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.

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

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

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

[8] Lehmann, E. L. (1998). *Elements of Large-Sample Theory*. Springer.

[9] 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 .

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

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