arXiv:astro-ph/0201547v1 31 Jan 2002

Statistics: Handle with Care
Detecting Multiple Model Components with the Likelihood Ratio Test

Rostislav Protassov and David A. van Dyk *
Department of Statistics, Harvard University

Alanna Connors
Eureka Scientific

Vinay L. Kashyap and Aneta Siemiginowska
Harvard-Smithsonian Center for Astrophysics

February 1, 2008

Abstract

The likelihood ratio test (LRT) and the related F test,[1] popularized in astrophysics by Eadie *et al.* (1971), Bevington (1969), Lampton, Margon, and Bowyer (1976), Cash (1979), and Avni *et al.* (1978) do not (even asymptotically) adhere to their nominal χ^2 and F distributions in many statistical tests common in astrophysics, thereby casting many marginal line or source detections and non-detections into doubt. Although the above references illustrate the many legitimate uses of these statistics, in some important cases it can be impossible to compute the correct false positive rate. For example, it has become common practice to use the LRT or the F test for detecting a line in a spectral model or a source above background despite the lack of certain required regularity conditions. (These applications were not originally suggested by Cash (1979) or by Bevington (1969)). In these and other settings that involve testing a hypothesis that is on the boundary of the parameter space, *contrary to common practice, the nominal χ^2 distribution for the LRT or the F distribution for the F test should not be used*. In this paper, we characterize an important class of problems where the LRT and the F test fail and illustrate this non-standard behavior. We briefly sketch several possible acceptable alternatives, focusing on Bayesian posterior predictive probability-values. We present this method in some detail, as it is a simple, robust, and intuitive approach. This alternative method is illustrated using the gamma-ray burst of May 8, 1997 (GRB 970508) to investigate the presence of an Fe K emission line during the initial phase of the observation.

There are many legitimate uses of the LRT and the F test in astrophysics, and even when these tests are inappropriate, there remain several statistical alternatives (e.g., judicious use of error bars and Bayes factors). Nevertheless, there are numerous cases of the inappropriate use of the LRT and similar tests in the literature, bringing substantive scientific results into question.

1 Introduction

Distinguishing a faint spectral line or a new source from a chance fluctuation in data, especially with low photon counts, is a challenging statistical task. As described in Section 2, these are but

*The authors gratefully acknowledge funding for this project partially provided by NSF grants DMS-97-05157 and DMS-01-04129, and by NASA Contract NAS8-39073 (CXC).

[1]The F test for an additional term in a model, as defined in Bevington (1969) on pp. 208-209, is the ratio

$$F_\chi = \frac{\chi^2(m) - \chi^2(m+1)}{\chi^2(m)/(N-m-1)} = \Delta\chi^2/\chi^2_\nu,$$

where $\chi^2(m)$ and $\chi^2(m+1)$ are the values of χ^2 statistic resulting from fitting m and $m+1$ free parameters respectively and χ^2_ν, in Bevington's (1969) notation, stands for a χ^2 random variable with ν degrees of freedom divided by the number of degrees of freedom ν. In the remainder of this paper we use χ^2_ν to denote a χ^2 random variable with ν degrees of freedom, as this notation is more standard.

two examples in a class of problems that can be characterized in statistical terms as a test for the presence of a component in a finite mixture distribution. It is common practice to address such tests with a likelihood ratio test (LRT) statistic or the related F statistic and to appeal to the nominal asymptotic distributions or *reference distribution*[2] of these statistics (Murakami *et al.*, 1988; Fenimore *et al.*, 1988; Yoshida *et al.*, 1992; Palmer *et al.*, 1994; Band *et al.*, 1995, 1996, 1997; Freeman *et al.*, 1999; Piro *et al.*, 1999, etc.). (See Band *et al.* (1997) for a discussion of the close relationship between the LRT and the F test.) The underlying assumption is that in some asymptotic limit the statistic being used to describe the data is distributed in an understandable way, and hence useful bounds may be placed on the estimated parameters. Unfortunately, the standard asymptotic theory does not always apply to goodness-of-fit tests of this nature even with a large sample size or high counts per bin. Thus, use of these statistics may be misleading. For example, searches for cyclotron scattering and atomic lines in γ-ray bursts based on such uncalibrated statistics may be unreliable.

In nested[3] significance tests such as the LRT or F test, we wish to choose between two models, where one model (the null model) is a simple or more parsimonious version of the other model (the alternative model). We seek evidence that the null model does not suffice to explain the data, by showing the observed data are very unlikely under the assumption that the null model is correct. In particular, we define a test statistic (e.g., the LRT statistic, F statistic, or $\Delta\chi^2$ values) with a distribution that is known at least approximately assuming the null model is correct. We then compute the test statistic for our data and compare the result to the known null distribution. If the test statistic is extreme (e.g., large) we conclude the null model is very unlikely to produce such a data set and choose the alternative model. A test statistic *without a known reference distribution* is without standard justification and is of little direct use; we can neither calibrate an observed value nor compute the false positive rate. Such is the case for the LRT statistic and the F statistic for detecting a spectral emission line, an absorption feature, or *added* model component[4]. Because this use is outside the bounds of the standard mathematical theory, the reference distribution is generally *uncalibrated: unknown and unpredictable*. This problem is fundamental, i.e., intrinsic to the definition of the LRT and F test. It is not due to small sample size, low counts per bin, or faint signal–to–noise ratio. It persists even when Gauss–Normal statistics hold and χ^2 fitting is appropriate.[5] Several authors (Mattox *et al.*, 1996; Denison & Walden, 1999; Freeman *et al.*, 1999) have recognized that the null distribution of the LRT and F statistics may vary from the nominal tabulated values (e.g., the tables given in Bevington, 1969, for the F test). Nonetheless, the inappropriate use of the F test in the astrophysics literature is endemic.

As a rule of thumb, there are two important conditions that must be satisfied for the proper use of the LRT and F statistics. First, *the two models that are being compared must be nested*. Second, *the null values of the additional parameters may not be on the boundary of the set of possible parameter values*. The second condition is violated when testing for an emission line because the line flux must be non-negative and the null value of this parameter is zero, which is the boundary of the non-negative numbers.

Because of the first condition, it is, for example, inappropriate to use the F test to compare a power law with a black body model or to compare a power law model with Raymond Smith

[2] As detailed below, the reference distribution is used to calibrate a test statistic. When choosing between two models, we assume the simpler or more parsimonious model holds and look for evidence that this assumption is faulty. Such evidence is calibrated via the *reference distribution*, the known distribution of the test statistic under the simple model. If the observed test statistic (e.g., LRT or F test) is extreme according to the reference distribution (e.g. $\chi^2_1 > 10.83$), the simple model is rejected in favor of the more complex model.

[3] E.g., the allowed parameter values of one model must be a subset of those of the the other model.

[4] We assume that when testing for the presence of an emission line, model parameters are constrained so the line intensity is greater than zero; similar assumptions are made for absorption features and added model components.

[5] As pointed out by the referee, Wheaton *et al.* (1995) showed that least squares or χ^2 fitting can sometimes be equivalent to maximum likelihood fitting even when Poisson statistics apply. However, their method is not universally applicable since it presumes that the weight matrix is independent of (or only weakly dependent on) the Poisson means (see their Equations 12, 19, and 20), whereas in the Poisson case the weights are the reciprocal of the Poisson means—if the weights are known, there is nothing to estimate.

thermal plasma model. This issue is discussed in Freeman *et al.* (1999) and is not the primary focus of this paper. Instead, we focus on the second condition that disallows, for example, testing for an emission line, an absorption feature, or other *added* spectral components (e.g., a power law, Compton reflection component, black body component, thermal component, etc.) or testing for a quasi-periodic oscillation in timing data, an added Gaussian in light curves, or an *added* image feature (e.g., an added point source). We emphasize that there are many legitimate uses of the LRT and F test, e.g., testing for a broken power law; comparing the variances of two samples; determining if a spectrum is the same or different in two parts of an image, at two difference times, or in two observations; or deciding whether to allow non-solar abundances. Generally, returning to the two rule-of-thumb conditions should guide one as to when the F test and LRT are appropriate. (We note, however, that the reference distributions of these tests are only reliable with a sufficiently large data set, even when the two conditions are met.)

In the remainder of the paper we explain why these standard tests fail and offer alternatives that work. In Section 2, we look at the class of models known as finite mixture models (which allow for multiple model components). We discuss the yet unresolved question of determining the number of components in the mixture and show that testing for a spectral line, a new source, or other *added* model component are special cases of this problem. The LRT and the F test have often been proposed as simple solutions in these cases. The LRT is specifically discussed in Section 3, and is shown to be invalid (i.e., uncalibrated) *in this setting* since as we discussed above its basic criteria are not met. In Section 4, we discuss a number of possible alternatives to the LRT and F statistics including Bayes factors, the Bayesian Information Criterion, and posterior predictive p-values.[6] Complete recipes for all alternatives to the LRT and F test are beyond the scope of this paper. As discussed in Section 4, we focus on posterior predictive p-values because they are conceptually and computationally simple and closely mimic the nested significance tests described above; see the high–redshift quasar example in Sections 3 and 4. With this machinery in place, we investigate a typical example in Section 5; we investigate whether the data support the presence of an Fe K emission line during the initial phase of the GRB 970508. Here neither the LRT nor the F test are appropriate. On the basis of our analysis, the model with a spectral line is clearly preferable. In Section 6, we conclude with several cautions regarding the inappropriate use of statistical methods which are exemplified by the misuse of the LRT.

Throughout the paper we use the LRT for a spectral line as an example, but the discussion and results apply equally to related tests (e.g., the F test) and to other finite mixture models (e.g. how many sources are detected above background; Avni *et al.* (1978)) or even more generally to testing any null models on the boundary of the parameter space.

2 Finite Mixture Models in Astrophysics

We begin with an illustration using a simple spectral model. Consider a source spectrum that is a mixture of two components: a continuum modeled by a power-law $\alpha E^{-\beta}$ and an emission line modeled as a Gaussian profile with a total flux $\tilde{\mathcal{F}}$. The expected observed flux $\mathcal{F}_j$ from the source within an energy bin E_j for a "perfect" instrument is given by

$$\mathcal{F}_j = dE_j \alpha E_j^{-\beta} + \tilde{\mathcal{F}} p_j \qquad \text{for} \quad j = 1, \ldots, J, \tag{1}$$

where dE_j is the energy width of bin j, α is the normalization constant for the power law with an index β, and p_j is the proportion of the Gaussian line that falls in bin j. For simplicity we parameterize the Gaussian line in terms of its location, μ, and assume that its width, σ, is fixed.

To calculate the number of counts observed with a counting detector that is predicted by this model we need to take into account several instrumental effects. The observed spectrum is

[6] A probability-value or p-value is the probability of observing a value of the test statistic (such as χ^2) as extreme or more extreme than the value actually observed given that the null model holds (e.g. $\chi^2_{30} \geq 2.0$) Small p-values are taken as evidence against the null model; i.e., p-values are used to calibrate tests. Posterior predictive p-values are a Bayesian analogue; see Section 4.2.

usually contaminated with background counts, degraded by instrument response, and altered by the effective area of the instrument and interstellar absorption. Thus, we model the observed counts in a detector channel l as independent Poisson[7] random variables with expectation

$$\xi_l = \sum_{j=1}^{J} R_{lj} A_j \mathcal{F}_j e^{-\gamma/E_j} + b_l, \quad l = 1, \ldots, L, \tag{2}$$

where (R_{lj}) is the photon redistribution matrix of size $L \times J$ that summarizes the instrument energy response, A_j is the effective area at energy E_j (normalized for convenience so that $\max_j A_j = 1$), γ is the parameter for simple exponential absorption,[8] and b_l is the expected background counts in channel l. We focus on the task of determining whether the data support the presence of a Gaussian line as in Equation 1.

The form of the model specified by Equation 1 is a special case of what is known in the statistics literature as a finite mixture model. The simplest example of a finite mixture model is a population made up of K subpopulations which differ in their distribution of some quantity of interest, x. The distribution of x in subpopulation k may have unknown parameters and is represented by $p_k(x)$. The relative size of subpopulation k is $\omega_k \geq 0$ with $\sum_k \omega_k = 1$. If we sample randomly from the larger population, the distribution of x is

$$f(x) = \sum_{k=1}^{K} \omega_k p_k(x). \tag{3}$$

Finite mixture models are an important class of statistical models with applications in the social, biological, engineering, and physical sciences. A general review of the topic can be found in any of the several useful books describing these models, their applications, and their statistical properties (Everitt & Hand, 1981; Titterington, Smith, & Makov, 1985; McLaughlan & Badford, 1988; Lindsay, 1995).

As an example in high energy spectral analysis, consider again Equation 1. We postulate that there are two "types" or "subpopulations" of photons, those originating from the continuum and those originating from the Gaussian line. The former have distribution in the shape of a power law, the latter as a Gaussian distribution. The relative sizes of the two populations are determined by α and $\tilde{\mathcal{F}}$.

There are many other statistical problems in astrophysics that can be phrased in terms of determining K (i.e., the true number of components or subpopulations) in a finite mixture. For example, consider testing for the presence of a point source in spatial data or a burst in time series data. In both cases, $p_1(x)$ represents the background or steady state model (i.e., model without a point source or without a burst), and we wish to determine if the data support a second component, $p_2(x)$, which represents the point source or burst. Other examples include testing for the presence of a second plasma component in coronal temperature models of late-type stars (Schmitt *et al.*, 1990; Kaastra *et al.*, 1996; Singh *et al.*, 1999). In this case, $p_k(x)$ in Equation 3 represents a plasma component for $k = 1, 2$. Here the key is to determine if a single plasma component (i.e, $K = 1$) suffices to explain the data.

Although the simplified form given in Equation 1 is used for illustration throughout the paper, the difficulty with the LRT (or the F test) that is described in the next section applies equally to the above problems, as does the Bayesian solution suggested in Section 4. It is well known in the statistical literature that determining the number of components, K (e.g., whether to include the Gaussian line in Equation 1) is a challenging problem that defies a standard analytical solution.

[7] Recall, a random variable X is said to follow a Poisson distribution with parameter or intensity $\mathcal{F}$ if $\Pr(X = x) = e^{-\mathcal{F}}\mathcal{F}^x/x!$. In this case $\mathrm{E}(X) = \mathcal{F}$ and we often write $X \overset{\mathrm{d}}{\sim} \mathrm{Poisson}(\mathcal{F})$ (read as X is distributed as Poisson with intensity $\mathcal{F}$). This representation conditions on the intensity parameter, $\mathcal{F}$, which in turn may vary.

[8] For mathematical transparency, we use a simplified exponential absorption throughout the paper except in the analysis of GRB 970508 where standard interstellar absorption is used.

Notice that a formal test of whether we should include component k corresponds to testing whether $\omega_k = 0$, which is the *boundary* of the set of possible values of ω_k. This issue is discussed at length in the several texts referenced above.

3 The Fallible Likelihood Ratio Test

3.1 Mathematical Background

As discussed in Section 1, the LRT and the related F test have often been used to test for the presence of a spectral line. To understand the difficulty with using these tests in this setting, we begin with a formal statement of the asymptotic result that underlies the LRT.

Suppose $\boldsymbol{x} = (x_1, \ldots, x_n)$ is an independent sample (e.g., measured counts per PHA channel or counts per imaging pixel) from the probability distribution $f(x|\boldsymbol{\theta})$, with parameters $\boldsymbol{\theta} = (\theta_1 \ldots, \theta_p)$. We denote the likelihood ratio statistic, $T_{\mathrm{LRT}}(x) = -2 \log R(\boldsymbol{x})$, where

$$R(\boldsymbol{x}) = \frac{\max \prod_{i=1}^{n} f(x_i|\theta_1^T, \ldots, \theta_q^T, \theta_{q+1}, \ldots, \theta_p)}{\max \prod_{i=1}^{n} f(x_i|\theta_1, \ldots, \theta_p)}, \tag{4}$$

where the maxima are found by varying the parameters. In the numerator, the θ^T terms represent parameters which are not varied but held at their 'true' values, i.e., the values assumed under the null model. Under the *regularity conditions*[9] discussed below, if $(\theta_1, \ldots, \theta_q)$ actually equals $(\theta_1^T, \ldots, \theta_q^T)$, the distribution of the likelihood ratio statistic converges to a χ^2 distribution with q degrees of freedom as the sample size, $n \to \infty$. Equation 4 can be written more formally by defining Θ to be the set of possible parameters values of $\boldsymbol{\theta}$. We are interested in testing whether $\boldsymbol{\theta}$ is in some subset of the parameter set, $\Theta_0 \subset \Theta$. We then denote the likelihood ratio statistic, $T_{\mathrm{LRT}}(x) = -2 \log R(\boldsymbol{x})$, where

$$R(\boldsymbol{x}) = \frac{\sup_{\boldsymbol{\theta} \in \Theta_0} L(\boldsymbol{\theta}|\boldsymbol{x})}{\sup_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}|\boldsymbol{x})} \tag{5}$$

with $L(\boldsymbol{\theta}|\boldsymbol{x})$ denoting the likelihood function, $\prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})$. Again, under the regularity conditions, if $\boldsymbol{\theta} \in \Theta_0$, the distribution of the likelihood ratio statistic converges to a χ^2 distribution as the sample size, $n \to \infty$. The degrees of freedom of the χ^2 distribution is the difference between the number of free parameters specified by Θ and by Θ_0. (In Equation 5 we replace the 'max,' i.e., maximum, in Equation 4 with 'sup,' i.e., supremum. This is a technicality which is mathematically necessary when the maximum occurs on the boundary of the parameters space. E.g., $\max_{x>0} 1/(x+1)$ is not defined, but $\sup_{x>0} 1/(x+1) = 1$.)

Examples of the LRT are found widely in scientific applications; see e.g., the many citations in Section 1. As a simple example, suppose one wishes to test whether the spectrum of a source is the same in two spatially separated regions. For simplicity, suppose a simple power law with two parameters (normalization and power law index) is used to model the spectrum. In the denominator of Equation 5, we fit the power law independently in the two regions via maximum likelihood and multiply the two likelihoods evaluated at there respective maximum likelihood estimates. For the numerator, we simply fit the combined data set via maximum likelihood. The resulting test statistic, $-2 \log R(\boldsymbol{x})$ is approximately χ^2_2, i.e., χ^2 in distribution with two degrees of freedom, the difference in the number of free parameters in the two models.

As noted by Cash (1979), this is a remarkably general result with very broad application. It must, however, be used cautiously; it is not a universal solution and does not apply in all settings. (This was also noted by Freeman *et al.* (1999).) In particular, the "regularity conditions" mentioned above are concerned mainly with the existence and behavior of derivatives of $\log L(\boldsymbol{\theta}|\boldsymbol{x})$

[9] The details of these regularity conditions are quite technical in nature and are summarized in Appendix A. A mathematically rigorous statement can be found in Serfling (Sections 4.2 and 4.4, 1980).

with respect to $\boldsymbol{\theta}$ (e.g., $\log L(\boldsymbol{\theta}|\boldsymbol{x})$ must be three times differentiable), the topology of Θ and Θ_0 (i.e., Θ_0 must be in the interior of Θ and both must be open), and the support of the distribution $f(\boldsymbol{x}|\boldsymbol{\theta})$ (i.e., the range of possible values of x cannot depend on the value of $\boldsymbol{\theta}$). Details of these regularity conditions are given in Appendix A. The difficulty with testing for a spectral line or more generally for the existence of a component in a finite mixture model lies in the topology of Θ. The standard theory requires that Θ be an open set—the LRT cannot be used to test whether the parameter lies on the boundary of the parameter space. Consider the source model given in Equation 2. If we set $\boldsymbol{\theta} = (\alpha, \beta, \gamma, \tilde{\mathcal{F}})$ and test $\tilde{\mathcal{F}} = 0$ we are examining the boundary of the parameter space, and the LRT statistic is not (even asymptotically) χ^2 in distribution.

In Appendix B, we show mathematically that the distribution of the LRT is not χ^2 when testing for a spectral line. This is further illustrated via a simulation study in the next section.

3.2 Computing the False Positive Rate for Fully Specified Models

Unfortunately, analytical results for the sampling distribution of the LRT do not exist in many settings, such as testing for a spectral line. Thus, we resort to two simulation studies which illustrate the unpredictable behavior of the distribution of the LRT statistic.

Simulation 1: Testing For an Emission Line. The most common problem in spectral analysis of quasar X-ray data is detection of a fluorescent emission line, Fe-Kα at $\sim 6.4 - 6.8$ keV. The iron line indicates existence of a cold material in the nucleus and its energy and width can constrain the ionization state of the matter and its distance from the central black hole. The line is strong and easily detected in some low redshift, low luminosity sources. However, for high luminosity and high redshift objects its presence in the data is not obvious and the common statistical technique used to support the detection of this line is the F test.

Our first simulation is designed to mimic the ASCA/SIS observation of the high redshift ($z = 3.384$) quasar S5 0014+81(Elvis *et al.*, 1994). We use standard ASCA response matrices in the simulations.[10] The effective area file was created with FTOOLS (ascaarf v.2.67) and the corresponding background file was extracted from the ASCA Background Blank Fields.[11]

We assume that the quasar emission in the observed energy range is described by the model given in Equation 1 with a power law continuum ($\alpha = 1.93 \times 10^{-3}$ counts per cm^2 per keV per second, $\beta = 2.11$), and exponential absorption parameter ($\gamma = 1.69$). These parameter values correspond to fitted values for the quasar observation obtained by van Dyk (2000).

We simulated 200 data sets, $\{\tilde{\boldsymbol{x}}^{(t)}, t = 1, \ldots, 200\}$ and each was fit three times via maximum likelihood using the EM algorithm as described by van Dyk (2000) using three different models:

Model 1: continuum (i.e., power law plus exponential absorption);

Model 2: continuum plus a narrow line with one free parameter, $\tilde{\mathcal{F}}$, and location fixed at 1.55 keV, and width fixed at zero;

Model 3: continuum plus a wide line with two free parameters, $\tilde{\mathcal{F}}$ and location, with the width fixed at $\sigma = 0.28$ keV.

For each data set, we computed two LRT statistics, $T_2(\tilde{\boldsymbol{x}}^{(t)}) = -2\log(L(\hat{\boldsymbol{\theta}}_1|\tilde{\boldsymbol{x}}^{(t)})/L(\hat{\boldsymbol{\theta}}_2|\tilde{\boldsymbol{x}}^{(t)}))$ and $T_3(\tilde{\boldsymbol{x}}^{(t)}) = -2\log(L(\hat{\boldsymbol{\theta}}_1|\tilde{\boldsymbol{x}}^{(t)})/L(\hat{\boldsymbol{\theta}}_3|\tilde{\boldsymbol{x}}^{(t)}))$, where $\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2$, and $\hat{\boldsymbol{\theta}}_3$ are the maximum likelihood estimate under Models 1, 2, and 3, respectively. Since the data were simulated under the null model (i.e., Model 1), the histograms of the computed LRT statistics in the first two panels of Figure 1 represent the reference distribution of the LRT against these two alternative models. The nominal χ^2 distributions with one and two degrees of freedom are plotted on the histograms and clearly do not suffice. The false positive rates are 2.6% and 1.5% in the nominal 5% tests,

[10] ftp://legacy.gsfc.nasa.gov/caldb/data/asca/sis/cpf/94nov9
[11] ftp://legacy.gsfc.nasa.gov/caldb/data/asca/sis/bcf/bgd/94nov

respectively. In this case, the LRT understates the evidence for an emission line. Correcting the false positive rate should enable us to detect weak lines that would be missed by blind application of the LRT.

Simulation 2: Testing For a Simple Absorption Line. Although the LRT is conservative in both of the tests in Simulation 1, this is not always the case. This can be seen in a second simulation where we consider a simplified absorption line. Although multiplicative model components such as an absorption line do not correspond to testing for a component in a finite mixture, the LRT still does not apply if the null model is on the boundary of the parameter space; such is the case with absorption lines. In this simulation we ignore background contamination, instrument response, and binning. We simulate 1000 data sets each with 100 photons from an exponential continuum, and fit two models:

Model 1: exponential continuum;

Model 2: exponential continuum plus a two parameter absorption line, where the fitted absorption probability is constant across the line which has fixed width, but fitted center.

Again, we computed the LRT statistic for each of the 1000 simulated data set and plotted the results in the final panel of Figure 1. Clearly the LRT does not follow its nominal reference distribution (χ^2 with 2 degrees of freedom) even with this simplified absorption line model; the false positive rate is 31.5% for the nominal 5% test. That is, use of the nominal reference distribution would result in over six times more false line detections than expected.

4 Bayesian Model Checking

Although some theoretical progress on the asymptotic distribution of $T_{\mathrm{LRT}}(\boldsymbol{x})$ when Θ_0 is on the boundary of Θ has been made (e.g., by Chernoff (1954) and specifically for finite mixtures by Lindsay (1995)), extending such results to a realistic highly structured spectral model would require sophisticated mathematical analysis. (See Lindsay (1995) for a simple exception when only ω is fit in Equation 3.) In this section, we pursue a mathematically simpler method based on Bayesian model checking known as posterior predictive p-values (Meng, 1994; Gelman, Meng, & Stern, 1996). As we shall see, this Bayesian solution is simpler and far more generally applicable than the asymptotic arguments required for satisfactory behavior of the LRT.

Posterior predictive p-values are but one of many methods for model checking and selection that may be useful in astrophysics. Our aim here is not to provide a complete catalog of such methods, but rather to provide practical details of one method that we believe is especially promising and little known in astrophysics. In Section 4.3, we provide a brief comparison with several other Bayesian methods.

4.1 The Posterior Predictive P-value

The central difficulty with the LRT and the F test in this setting is that their reference distributions are unknown even asymptotically. Moreover, the distributions likely depend on such things as the particular shape of the continuum, the number of lines, and their profiles and strengths. Thus, it is difficult to obtain any general results regarding such reference distributions even via simulation. The method of posterior predictive p-values uses information about the spectrum being analyzed to calibrate the LRT statistic (or any other test statistic) for each particular measurement. In the simulations described in Section 3.2, we simulated data sets, $\tilde{\boldsymbol{x}}^{(t)}$, using a fixed value of (α, β, γ) and observed the behavior of $T_{\mathrm{LRT}}(\tilde{\boldsymbol{x}}^{(t)})$. Instead of fixing the model parameter at its fitted value under the null model, the method of posterior predictive p-values uses values of the parameter that are relatively likely given the observed counts. That is, *we run a Monte Carlo simulation to access the sampling distribution of the LRT (or other) statistic so that we can calibrate the value*

of the statistic computed on the data and determine a p-value. The Monte Carlo simulation is run using parameter values fit to the data under the null model and *accounts for uncertainty (i.e., error bars)* in these fitted values.

To formalize this, we review Bayesian model fitting which we use to simulate values of the parameter via Monte Carlo. Bayesian model fitting involves the probability of the parameters, given the model: $p(\boldsymbol{\theta}|\boldsymbol{x}, I)$; while "classical" or "frequentist" statistics deals with the converse: $p(\boldsymbol{x}|\boldsymbol{\theta}, I)$. Interested readers are referred to the recent more detailed treatments of Bayesian methods given in Gelman, Carlin, Stern, & Rubin (1995), Carlin & Louis (1996), and specifically for astrophysics, in van Dyk, Connors, Kashyap, & Siemiginowska (2001). One transforms from $p(\boldsymbol{x}|\boldsymbol{\theta}, I)$ to $p(\boldsymbol{\theta}|\boldsymbol{x}, I)$ or visa versa using Bayes theorem, which states

$$p(\boldsymbol{\theta}|\boldsymbol{x}, I) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta}, I)p(\boldsymbol{\theta}|I)}{p(\boldsymbol{x}|I)}, \tag{6}$$

where $p(\boldsymbol{\theta}|\boldsymbol{x}, I)$ is the distribution of the parameter given the data, and any available prior information, I, i.e., *the posterior distribution*, $p(\boldsymbol{x}|\boldsymbol{\theta}, I)$, is the model for the distribution of the data $\boldsymbol{x}$, i.e., *the likelihood*, $p(\boldsymbol{\theta}|I)$ contains information about the parameter known prior to observing $\boldsymbol{x}$, i.e., the *prior distribution*, and $p(\boldsymbol{x}|I)$ is a normalizing constant for $p(\boldsymbol{\theta}|\boldsymbol{x}, I)$. Equation 6 allows us to combine information from previous studies or expert knowledge through the prior distribution with information contained in the data via the likelihood. In the absence of prior information we use a relatively non-informative prior distribution. van Dyk *et al.* (2001) describe how to use MCMC and the Gibbs sampler to simulate from $p(\boldsymbol{\theta}|\boldsymbol{x})$ using spectral models which generalize Equation 1.

Our procedure is given for an *arbitrary* statistic $T(\boldsymbol{x})$ (i.e., an arbitrary function $T(\boldsymbol{x})$ of data $\boldsymbol{x}$) but certainly is valid for $T_{\rm LRT}(\boldsymbol{x})$, i.e., the LRT statistic or the F statistic. We calibrate $T(\boldsymbol{x})$ using the distribution of $T(\tilde{\boldsymbol{x}}^{(t)})$ given the observed data $\boldsymbol{x}$, as the reference distribution, i.e.,

$$p(T(\tilde{\boldsymbol{x}}^{(t)})|\boldsymbol{x}) = \int p(T(\tilde{\boldsymbol{x}}^{(t)}), \boldsymbol{\theta}|\boldsymbol{x}) d\boldsymbol{\theta} = \int p(T(\tilde{\boldsymbol{x}}^{(t)})|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{x}) d\boldsymbol{\theta}, \tag{7}$$

where the second equality follow because $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}^{(t)}$ are independent given $\boldsymbol{\theta}$. (Here and in what follows we suppress explicit conditioning on the prior information, I.) What is important in Equation 7 is that *we do not fix $\boldsymbol{\theta}$ at some estimated value in the reference distribution*; rather we integrate over its uncertainty as calibrated by the posterior distribution.

Although analytical results are typically not available, calibration is easily accomplished via Monte Carlo. Specifically, we

1. Simulate parameter values $\{\boldsymbol{\theta}^{(t)}, t = 1, \ldots, N\}$ from $p(\boldsymbol{\theta}|\boldsymbol{x})$, e.g., using the method of van Dyk *et al.* (2001).

2. For $t = 1, \ldots, N$ simulate $\tilde{\boldsymbol{x}}^{(t)} \overset{\rm d}{\sim} p(\boldsymbol{x}|\boldsymbol{\theta}^{(t)})$, i.e., according to the model simulate N data sets, one for each simulation of the parameters obtained in Step 1. (This is similar to the often used "parametric bootstrapping", but now uncertainties in the parameters are accounted for.)

3. For $t = 1, \ldots, N$ compute the statistic $T(\tilde{\boldsymbol{x}}^{(t)})$. For $T_{\rm LRT}(\tilde{\boldsymbol{x}}^{(t)})$ this involves computing the maximum likelihood estimates for the null and alternative models using each of the N data sets, see Equation 5.

4. Compute the posterior predictive p-value,

$$p = \frac{1}{N} \sum_{t=1}^{N} \mathcal{I}\{T(\tilde{\boldsymbol{x}}^{(t)}) > T(\boldsymbol{x})\}, \tag{8}$$

where $\mathcal{I}\{\text{statement}\}$ is an indicator function which is equal to one if the statement is true and zero otherwise.

The posterior predictive p-value is the proportion of the Monte Carlo simulations that result in a value of $T(\tilde{\boldsymbol{x}}^{(t)})$ more extreme than the value computed with the observed data, $T(\boldsymbol{x})$. If this is a very small number, we conclude that our data is unlikely to have been generated under the posterior predictive distribution. Since this distribution is computed assuming the null model, we reject the null model and investigate alternative models. That is, p is treated as a p-value with small values indicating evidence for the more complex model, e.g., the model with an additional spectral line.

4.2 An Example

Here we illustrate the method of posterior predictive p-values by testing for a spectral line in data obtained from a high redshift quasar used in Simulation 1 of Section 3.2. In particular, we compare Model 3 with Model 1 as described in Section 3.2. Calculations proceed exactly as in Section 3.2, except each $\tilde{\boldsymbol{x}}^{(t)}$ is simulated with a different $\boldsymbol{\theta}^{(t)}$ as simulated from $p(\boldsymbol{\theta}|\boldsymbol{x})$. Five hundred simulations from the resulting posterior predictive distribution of $T_{\mathrm{LRT}}(\boldsymbol{x})$ are displayed in Figure 2 along with a vertical line that correspond to $T_{\mathrm{LRT}}(\boldsymbol{x})$ computed with the observed data. The resulting posterior predictive p-value is 1.6% showing strong evidence for the presence of the spectral line.

4.3 General Advice on Model Checking and Model Selection

Posterior predictive p-values are by no means the only statistical method available for model checking and model selection; in this section we outline some important Bayesian alternatives and discuss our recommendation of using posterior predictive p-values in this case. This material may seem somewhat technical to some readers, but may be skipped since the remainder of the paper is independent.

We begin by emphasizing that detecting model features is a challenging statistical problem indeed; there is no consensus within the statistical community as how best to proceed. Whenever possible, the issue should be avoided by for example focusing on estimating the perhaps very weak strength of a spectral line rather that deciding whether there is a line. Practically speaking, however, we must decide which and how many lines to include in the final analysis and would like statistical tools that offer guidance.

Posterior predictive p-values aim to point out when the null model is inadequate to describe the observed data; more counts than expected under the null model in a narrow energy range is evidence that the null model is missing a spectral line. Since posterior predictive p-values approximate the reference distribution of the test statistic by simulating data using the null model, they tend to favor the null model. That is, posterior predictive p-values tend to be conservative, especially if the test statistic is poorly suited for detecting the model feature in question (Meng, 1994; Bayarri & Berger, 1999). Nonetheless, as illustrated in Section 5 posterior predictive p-values can be used to detect model features and their conservative nature adheres to the scientific standard for burden of proof.

To elicit more power from the posterior predictive p-value, Bayarri & Berger (1999) suggest concentrating on the parameters of interest. (More specifically, this method conditions on sufficient statistics for the nuisance parameters when computing the p-value.) Although less conservative than posterior predictive p-values, this method is mathematically and computationally more demanding.

Bayes factors are a popular method for Bayesian model selection; here we briefly compare them with posterior predictive analysis to explain our preference for the latter, at least when well specified prior information is not forthcoming. The primary difficulty with Bayes factors is that relative to posterior predictive p-values, they are much more sensitive to the prior distribution. Consider again testing for a spectral line by selecting between two models

Model 1: a power law with no emission line;

Model 2: a power law with an emission line.

Suppose that Model 1 has two free parameters and Model 2 has four free parameters, the line intensity and location in addition to the power law parameter and normalization. The Bayes factor is defined as

$$B = \frac{\int p(\boldsymbol{Y}|\boldsymbol{\theta}_1, \text{Model 1}) p(\boldsymbol{\theta}_1|\text{Model 1}) d\boldsymbol{\theta}_1}{\int p(\boldsymbol{Y}|\boldsymbol{\theta}_2, \text{Model 2}) p(\boldsymbol{\theta}_2|\text{Model 2}) d\boldsymbol{\theta}_2}, \tag{9}$$

where $\boldsymbol{Y}$ are the counts, $\boldsymbol{\theta}_1$ are the parameters for Model 1 and $\boldsymbol{\theta}_2$ are the parameters for Model 2. The Bayes factor is equivalent to the ratio of the posterior and prior odds for Model 1 versus Model 2. Roughly speaking, large values of B indicate that the data favor Model 1 and small values that the data favor Model 2. Computing a Bayes factors involves marginalizing over (or averaging over) all unknown parameters and thus can be computationally demanding; methods appear elsewhere (Connors, 1997; Freeman *et al.*, 1999; Gregory & Loredo, 1992; Kass & Raftery, 1995).

In a paper describing the natural "Ockham's Razor" property of Bayes factors, Jefferys & Berger (1992) point out that a Bayes factor goes (roughly) like the height of the likelihood at its maximum times its width, divided by the width of the prior. Hence Bayes factors are strongly dependent on the prior distributions for the parameters. This can be seen formally in Equation 9, where the numerator is the *prior predictive distribution* under Model 1; likewise for the denominator. The prior predictive distribution also appears as the denominator in Bayes theorem as a normalizing constant. This distribution quantifies the distribution of the data with uncertainty in the model parameters as described by the prior distribution. Thus, if a highly diffuse prior distribution is used the prior predictive distribution will also be very diffuse. If the prior distribution is improper,[12] neither the prior predictive distribution nor the Bayes factor are defined. (There have been several attempts to define analogous quantities in this situation; see Smith & Spiegelhalter (1980) and Kass & Raftery (1995).) As a result, Bayes factors are *very* sensitive to the choice of prior distribution; the more diffuse the prior distribution for the line parameters the more the Bayes factor will favor Model 1. Thus, when using Bayes factors the prior distributions must be proper and well specified. That is, the prior distribution must not be chosen in an ad hoc fashion, but rather must be a meaningful summary of prior beliefs. We note that difficulties associated with prior specification evaporate when the models compared are discrete with no obvious scientific models between; see Gelman *et al.* (1995, Section 6.5) for examples and discussion. Kass & Raftery (1995) offers a thoughtful review of Bayes factors including numerous examples, computational methods, and methods for investigating sensitivity to the prior distribution.

Another popular method for model selection is the Bayesian Information Criterion (BIC) (Schwartz, 1978). BIC aims to select the model with the highest posterior probability,

$$\text{choose the model that maximizes} \quad \pi_i \int p(\boldsymbol{Y}|\boldsymbol{\theta}_i, \text{Model } i) p(\boldsymbol{\theta}_1|\text{Model } i) d\boldsymbol{\theta}_i, \tag{10}$$

where π_i is the prior probability of model i. Since computations of this posterior probability is generally complicated, Schwartz suggested replacing it with BIC which maximizes the loglikelihood with penalty term $-0.5 q_i \log n$, where q_i is the number of free parameters in model i and n is the sample size. This penalty term favors models with fewer free parameters. Schwartz went on to show that under certain conditions BIC is asymptotically equivalent to using Equation 10. Like the Bayes factor, the posterior probability in Equation 10 is highly sensitive to the choice of prior, including the prior probabilities of the various models. That BIC is not at all sensitive to the choice of prior distribution reflects the fact that it is a poor approximation of Equation 10, at least for small samples.

[12] An improper distribution is a distribution that is *not* integrable and thus is not technically a distribution. One should use improper prior distributions only with great care since in some cases they lead to improper posterior distributions which are uninterpretable.

5 An Fe K Line in GRB 970508?

In this section we examine the X-ray spectrum of the afterglow of GRB 970508, analyzed for Fe K line emission by Piro *et al.* (1999). This is a difficult and extremely important measurement: the detection of X-ray afterglows from γ-ray bursts relies on near-real-time satellite response to unpredictable events and a great deal of luck in catching a burst bright enough for a useful spectral analysis. The ultimate physics of these events is still controversial, but they are among the most distant observable objects in the sky. Detecting a clear atomic (or cyclotron) line in the generally smooth and featureless afterglow (or burst) emission not only gives one of the few very specific keys to the physics local to the emission region, but also provides clues or confirmation of its distance (via redshift).

Piro *et al.* (1999) used the F statistic to determine the significance of the detected Fe K line, as is standard practice. Regularity conditions for the F statistic and the LRT statistic are similar; neither can be used to test whether the true parameter value is on the boundary of the parameter space. Thus, when testing for a spectral line, the F test is equally inappropriate. *We emphasize that we do not highlight any errors particular to Piro* et al. *(1999) but rather illustrate that the standard method for identifying spectral lines is not trustworthy.* In fact, our more rigorous analysis confirms their detection, but with higher significance.

This section is divided into two parts. First we describe the afterglow data, our models with and without an Fe K emission line, and two model fitting strategies, i.e., χ^2 fitting in XSPEC and Bayesian posterior analysis. Second we use posterior predictive p-values to evaluate the strength of the evidence for the Fe K emission line.

5.1 Data and Model Fitting

An X-ray point-like emission associated with GRB 970508 was observed by the BeppoSAX only six hours after the initial γ-ray burst onset. The exposure time of 28ks was long enough to monitor the evolving X-ray spectrum with an average observed flux of order $\sim 10^{-12}$ ergs cm^{-2} s^{-1} in the combined low-energy concentrator spectrometer (LECS) and medium-energy energy concentrator spectrometer (MECS) instruments. The data are plotted in Figure 3. Piro *et al.* (1999) divided the data into two time intervals "1a" versus "1b" and indicated that an emission line, a possible redshifted Fe-K line, can only be present during the initial phase of the observation,[13] i.e., in the data set 1a.

We extracted the data from the BeppoSAX archive[14] in order to investigate the significance of the line. We restrict attention to data sets 1a LECS, 1a MECS, 1b LECS, and 1b MECS and use the default instrument responses and background files provided for this observation. We fix the relative normalization of LECS versus MECS at 0.8 (c.f., Piro *et al.*, 1999), extract the spectra from the original event files, and fit two models:

Model 1 : a simple absorbed power law;

Model 2 : a simple absorbed power law and a Gaussian line at 3.5 keV with a known width, 0.5 keV (cf., Piro *et al.* (1999)).

Both models were fit via both χ^2 fitting and Bayesian posterior analysis. For χ^2 fitting, we use XSPEC v.10 after binning the data so that there are at least 15 counts (source+background) per bin; the results are summarized in Table 1[15] and yield an F statistic for comparing Model 1 and

[13] Piro *et al.* (1999) were specifically interested in this hypothesis which they proposed after preliminary data analysis. See Section 2 in Piro *et al.* (1999).

[14] http://www.sdc.asi.it

[15] The best-fitted values we report for the power law fit (Model 1) are within the intervals defined by error bars in Table 1 (i.e., Observation 1a+1b) in Piro *et al.* (1999). That the two fits are not identical is probably due to differences in binning mechanisms used, to the fact that we fixed the relative normalization LECS/MECS at 0.8 and thus did not fit it, and to possible slight differences in background files used. We also note that the best fitted

Model 2, i.e., $\Delta\chi^2/\chi^2_\nu$, of 4.156673.[16] Because the necessary regularity conditions are not met, it is impossible to calibrate this F statistic using the nominal F reference distribution.

To simulate from the posterior distribution of the parameters of Model 1, we use the Markov chain Monte Carlo method of van Dyk *et al.* (2001); results appear in Figure 4. The Bayesian analysis requires us to specify prior distributions on the model parameters. (See van Dyk *et al.*, 2001, for details of Bayesian spectral analysis.) As illustrated in Figure 5, we have tried to use relatively non-informative prior distributions.[17] Figure 5 compares the marginal posterior distribution of the power law and photon absorption parameters [18] with their marginal likelihood[19] and illustrates the similarity between the two, i.e., the non-informative nature of the prior distribution.

The treatment of background in out Bayesian posterior analysis differs from that of XSPEC. Instead of subtracting the background, we first fit a power law to the background counts and then model the observed counts as Poisson with intensity equal to the sum of the source and fitted background intensities (see van Dyk *et al.*, 2001). To simulate from the posterior predictive distribution, we ran a Markov chain Monte Carlo algorithm, discarding the first 100 simulations and using every fifteenth simulation thereafter to obtain a total of 2000 simulations. Convergence was judged by comparing the simulations to contours of the posterior distribution; see Figure 5. We do not advocate this as a general strategy because of the computational effort required to obtain the contours of the posterior distribution. A more general method based on multiple chains in described in Gelman & Rubin (1992); see also van Dyk *et al.* (2001).

5.2 Calibrating the LRT Accounting for Parameter Uncertainty

To evaluate the reference distribution of a test statistic, posterior predictive analysis accounts for uncertainty in the parameter values in the null model by sampling the parameters from their Bayesian posterior distribution. Thus, following the four-step procedure described in Section 4.1, we first simulate the parameters from their posterior distribution as described in Section 5.1 (Step 1). For each simulation of the parameter, the 1a LECS and MECS and 1b LECS and MECS are then simulated (Step 2) and the LRT statistic is computed (Step 3). The necessary maximum likelihood estimates are computed with the EM algorithm as described by van Dyk (2000). The resulting reference distribution is compared with the observed value of the LRT statistic in Figure 6(a), yielding a posterior predictive p-value of 0.75% (Step 4). Thus, the observed LRT statistic is very unusual in the absence of the Fe K line and we conclude there is evidence for the Fe K line, which is calibrated by the magnitude of the posterior predictive p-values.

line intensity (see Model 2) we obtained agrees with that reported by Piro *et al.* (1999) ($I_{\rm Fe} = (5 +/- 2) \times 10^{-5}$ photons cm^{-2} s^{-1}).

[16] The data sets 1a and 1b LECS and MECS are obtained by splitting two event files (for LECS and MECS) that correspond to data sets 1 LECS and 1 MECS into an initial and a later period. These split versions used by Piro *et al.* (1999) were not available to us, so we did extractions from the event files ourselves based on the information presented in Table 1 in Piro *et al.* (1999). The event files were retrieved from the BeppoSAX archive at http://www.sdc.asi.it. We call our extractions 1a and 1b LECS and MECS, but these files may not be exactly the same as the split data sets analyzed by Piro *et al.* (1999). More specifically, the exposure times for the four data sets we created are as follows: 1a LECS: 7.885 ks, 1b LECS: 9.361 ks, 1a MECS: 10.369 ks, 1b MECS: 17.884 ks.

[17] We imposed a flat prior on the normalization parameter, independent Gaussian relatively diffuse priors on photon index and nH ($N(2, \sigma = .61)$ and $N(6, \sigma = 2.1197)$ respectively), and for Model 2 a gamma prior on $\tilde{\mathcal{F}}$, $p(\tilde{\mathcal{F}}) \propto \exp(-0.05\tilde{\mathcal{F}})$.

[18] The marginal posterior distribution of the photon index, β, and photon absorption, γ_{ph}, is obtained from Model 1 by integrating out the normalization parameter, α. I.e,. $p(\beta, \gamma_{ph}|x, \text{no line is present}) = \int p(\boldsymbol{\theta}|x, \text{no line is present})\mathrm{d}\alpha$. Computing the marginal posterior distribution requires numerical integration of the joint posterior distribution and is a computationally expensive task. On a dual-processor 200Mhz Ultra-2 Sun workstation with 256M RAM it took the program 183994.27 CPU time (in seconds) to run. We do not advocate such computation in general, but rather use the marginal posterior distribution to verify the method in this specific case.

[19] The marginal likelihood of the photon index, β, and the photon absorption ,γ_{ph}, is obtained from the likelihood function $p(x|\alpha, \beta, \gamma_{ph}, \text{no line is present})$ by integrating out the normalization parameter, α, over a large, but finite interval containing the range of physically meaningful values of α.

We can repeat this procedure using the posterior distribution under Model 2 to simulate the parameters in Step 1. That is, we can treat the model with the emission line as the null model. The resulting p-value (65.71%) indicates no evidence against Model 2. In this simulation, the data are generated *with* a spectral line that is exactly the best fit line for the observed data. Thus, since the LRT statistic is designed to detect this very spectral line as discussed in Section 4.3, it is no surprise that the model is found to agree with the data. As we shall see, this situation persists; test statistics that are not well designed to detect discrepancies between the model and the data will never result in p-values that make us question the null model.

The LRT statistic and F statistic are not the only quantities that can be used in posterior predictive checking. In fact, posterior predictive p-values can be used to access how well a particular important feature of the data is explained by a model. For example, another measure of the strength the Fe K line is the maximum likelihood estimate of its intensity, $\tilde{\mathcal{F}}$.[20] Having already simulated from the posterior distribution of the model parameters and the corresponding simulated data sets, we need only compute the maximum likelihood estimate of $\tilde{\mathcal{F}}$ for each data set to construct the reference distribution for $\tilde{\mathcal{F}}$. The results for both the model with and without the Fe K line appear in Figure 6(c)-(d). Qualitatively, the conclusion drawn from Figure 6(c) is the same as with the LRT statistic; the data is explained better by the model with the Fe K line.

If the continuum parameters (or more generally the parameters of the null model) are *very well* constrained, we can perform an approximate calibration of a test statistic by fixing the parameters of the null model at their fitted values rather than simulating them from their posterior distribution. That is, we can mimic the simulations described in Section 3.2, using the "parametric bootstrap". We illustrate this procedure by calibrating the F statistic calculated in Section 5.1. Specifically, we use a simulated reference distribution that we obtain by simulating N data sets under a fully specified null model, (i.e., Model 1) with parameters fixed at the best-fitted values in Table 1. The simulations have the following steps:

1. Simulate N data sets (`fakeit` in XSPEC v.10) according to the null model with parameters fixed at their best-fit values, with the same effective area, instrument response, and background files as well as the exposure time as the initial analysis (i.e. simulate $\tilde{\boldsymbol{x}}^{(t)} \overset{\text{d}}{\sim} p(\boldsymbol{x}|\boldsymbol{\theta})$, for $t = 1, \ldots, N$). Each simulated data set is appropriately binned, using the same binning algorithm as that used to bin the real data.

2. Fit the null and alternative model (i.e., Models 1 and 2) to each of the N data sets and compute the F statistic, $\Delta\chi^2/\chi^2_\nu$ for $t = 1, \ldots, N$.

3. Compute the approximate p-value,

$$p = \frac{1}{N}\sum_{t=1}^{N} \mathcal{I}\{T_F(\tilde{\boldsymbol{x}}^{(t)}) > T_F(\boldsymbol{x})\},$$

   where $T_F(\boldsymbol{x})$ represents the F statistic.

We emphasize that the simulation results rely on the assumption that the data were actually generated under the null model with parameters fixed at their best fit values. A probability histogram of the simulated F statistics can be used to calibrate the F statistic and compute a p-value; see Figure 6(e). The value of the F statistic is the 93.77th percentile of $F(1, 13)$ distribution (i.e., F distribution with degrees of freedom 1 and 13: $F(1, 13) = \chi^2_1/(\chi^2_{13}/13)$). Thus, the nominal p-value is 6.23%. The simulation, however, gives somewhat stronger evidence against the null model reporting a p-value of 3.8%. Unfortunately, this calibration is contingent on the accuracy of the model used to simulate data in Step 1.

[20] We define $\tilde{\mathcal{F}}$ as the rate per second of counts due to the Fe K line for the MECS instrument before absorption with the maximum effective area.

6 Statistics: Handle with Care

Although the LRT is a valuable statistical tool with many astrophysical applications, it is not a universal solution for model selection. In particular, when testing a model on the boundary of the parameter space (e.g., testing for a spectral line), the (asymptotic) distribution of the LRT statistic is unknown. Using this LRT and its nominal χ^2 distribution can lead to unpredictable results (e.g., false positive rates varying from 1.5% to 31.5% in the nominal 5% false positive rate test in Monte Carlo studies). Thus, the LRT should not be used for such model selection tasks.

The lesson to be learned from misapplication of the LRT is that there is no replacement for an appreciation of the subtleties involved in any statistical method. Practitioners of statistical methods are forever searching for statistical "black boxes": put the data in and out pops a p-value or a fitted model. When working with the sophisticated models that are common in spectral, spatial, or temporal analysis as well as other applications in astrophysics, such black boxes simply do not exist. The highly hierarchical structures inherent in the data must be, at some level, reflected in the statistical model. Stochastic aspects of instrument response, background counts, absorption, pile-up, and the relationship between spectral, temporal, and spatial data must be accounted for. With such structured models, over-simplified, off the-shelf methods such as assuming Gaussian errors (e.g., χ^2 fitting) lead to unpredictable results. Standard tests, such as the LRT, Cash, or F statistics, sometimes are appropriate (e.g., testing whether a mean parameter is equal to a specified value) and sometimes are not (e.g., testing for the presence of a spectral line) and are never appropriate with small data sets.

Even more sophisticated methods can have pitfalls. Frequentist methods (e.g., maximum likelihood with asymptotic frequentist error bars), for example, typically rely on large sample sizes which may not be justifiable in practice (e.g., for mixture models). Non-parametric methods require fewer parametric assumptions but often grossly simplify the structure of the underlying model, discarding scientific information. Sometimes even these methods require strong assumptions such as knowing the underlying model completely (e.g., Kolmogorov-Smirnov goodness-of fit tests). Bayesian methods easily accommodate the hierarchical structure in the data and do not require asymptotic (large data set or many measurements) approximations. The computational tools required for highly structural models, however, require careful implementation and monitoring; determining convergence of Markov chain Monte Carlo methods is rarely automatic. Moreover, although Bayesian statistical summaries (e.g., error bars) are mathematically consistent summaries of information, they may not exhibit the frequentist properties (e.g., coverage rates) that might be expected. Nevertheless, the computational difficulties that sometimes exist with Bayesian analysis are much easier to overcome than the conceptual difficulties that may arise in other frameworks (e.g., unknown sampling distributions of test statistics). Thus, Bayesian methods are best equipped to handle highly structured models, but we emphasize that like any statistical method, they must be used with knowledge, sophistication, and care.

Appendices

A Regularity Conditions for the LRT

Here we state the regularity conditions required for the standard asymptotic behavior of the LRT. (Our presentation follows Chapter 4 of Serfling (1980) which should be consulted for details.) Let $X_1, ..., X_n$ be independent identically distributed random variables with distribution $F(x; \boldsymbol{\theta})$ belonging to a family $\mathcal{F} = \{F(x; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$, where $\Theta \subset \mathcal{R}^k$ is open and $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)$. $F(x; \boldsymbol{\theta})$ are assumed to possess densities or mass functions $f(x; \boldsymbol{\theta})$, that satisfy

1. For each $\boldsymbol{\theta} \in \Theta$, each $i = 1, ..., k$, each $j = 1, ..., k$, and each $l = 1, ..., k$, the derivatives

$$g\frac{\partial \log f(x;\boldsymbol{\theta})}{\partial \theta_i}, \quad \frac{\partial^2 \log f(x;\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}, \quad \frac{\partial^3 \log f(x;\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j \partial \theta_l} \tag{11}$$

exist, all x;[21]

2. For each $\boldsymbol{\theta}_* \in \Theta$, there exist functions $g(x)$, $h(x)$, and $H(x)$ (possibly depending on $\boldsymbol{\theta}_*$) such that for $\boldsymbol{\theta}$ in a neighborhood $N(\boldsymbol{\theta}_*) \subset \Theta$ the relations

$$\left|\frac{\partial f(x;\boldsymbol{\theta})}{\partial \theta_i}\right| \leq g(x), \quad \left|\frac{\partial^2 f(x;\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right| \leq h(x), \quad \left|\frac{\partial^3 \log f(x;\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j \partial \theta_l}\right| \leq H(x) \tag{12}$$

hold, all x and all $1 \leq i, j, l \leq k$, with

$$\int g(x)dx < \infty, \int h(x)dx < \infty, \text{ and } \int H(x)f(x;\boldsymbol{\theta})dx < \infty \text{ for } \boldsymbol{\theta} \in N(\boldsymbol{\theta}_*) \tag{13}$$

3. For each $\boldsymbol{\theta} \in \Theta$ the information matrix

$$\mathbf{I}(\boldsymbol{\theta}) = \left[E\left\{\frac{\partial \log f(x;\boldsymbol{\theta})}{\partial \theta_i}\frac{\partial \log f(x;\boldsymbol{\theta})}{\partial \theta_j}\Big|\boldsymbol{\theta}\right\}\right]_{k\times k} \tag{14}$$

exists and is positive definite.

Consider $\Theta_0 \subset \Theta$ such that the specification of Θ_0 may be equivalently given as a transformation

$$\theta_1 = g_1(\nu_1, ..., \nu_{k-r}), \ ... \ , \ \theta_k = g_k(\nu_1, ..., \nu_{k-r}), \tag{15}$$

where $\nu = (\nu_1, ..., \nu_{k-r})$ ranges through an open set $N \subset \mathcal{R}^{k-r}$. E.g., if $k = 3$ and $\Theta_0 = \{\boldsymbol{\theta} : \theta_1 = \theta_1^*\}$, we then may take $N = \{(\nu_1, \nu_2) : (\theta_1^*, \nu_1, \nu_2) \in \Theta_0\}$ and the functions g_1, g_2, g_3 to be

$$g_1(\nu_1, \nu_2) = \theta_1^*, \ g_2(\nu_1, \nu_2) = \nu_1, \ g_3(\nu_1, \nu_2) = \nu_2.$$

Assume further that g_i possess continuous first order partial derivatives and that

$$\mathbf{D}_\nu = \left[\frac{\partial g_i}{\partial \nu_j}\right]_{k\times(k-r)}$$

is of rank $k - r$. Alternatively, if Θ_0 is defined by a set of r ($r \leq k$) restrictions given by equations

$$R_i(\boldsymbol{\theta}) = 0, \quad 1 \leq i \leq r$$

(e.g., in the case of a *simple* hypothesis $\Theta_0 = \{(\theta_1^0, \theta_2^0, \theta_3^0)\}$, we have $R_i(\boldsymbol{\theta}) = \theta_i - \theta_i^0$, $i = 1, 2, 3$), we require that $R_i(\boldsymbol{\theta})$ possess continuous first order derivatives and that

$$\mathbf{C}_{\boldsymbol{\theta}} = \left[\frac{\partial R_i}{\partial \theta_j}\right]_{r\times k}$$

is of rank r. Let $\boldsymbol{\theta}^* \in \Theta$ denote the true unknown value of the parameter $\boldsymbol{\theta}$. Define the null hypothesis to be $H_0 : \boldsymbol{\theta}^* \in \Theta_0$. Then if H_0 is true, the LRT statistic (see Equation 4) is asymptotically distributed as χ^2 with r degrees of freedom.

[21] Implicit here is the requirement that the support of the distribution be independent of θ, otherwise there would be a θ and an x for which the derivatives in (11) would not exist.

B Testing for Lines: A Misapplication of the LRT

Under the required regularity conditions, the asymptotic χ^2 distribution of the LRT is based on the asymptotic normality of the maximum likelihood estimate[22] with mean equal to the true parameter value. If the true value is on the boundary of the parameter space, however, the mean of the maximum likelihood estimate cannot possibly be the true value of the parameter since maximum likelihood estimates are always in the parameter space. (Clearly $\mathrm{E}(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta})$ can not be $\boldsymbol{\theta}$ if $\hat{\boldsymbol{\theta}}$ is, for example, always greater than $\boldsymbol{\theta}$.) Thus, one of the regularity conditions required for the LRT is that Θ, the parameter space, be an open set. This is not the case when we test for the presence of a component in a finite mixture model. Consider the source model given in Equation 1. If we set $\boldsymbol{\theta} = (\alpha, \beta, \gamma, \tilde{\mathcal{F}})$ and test $\tilde{\mathcal{F}} = 0$ we are examining the boundary of the parameter space, and the LRT statistic is not (even asymptotically) χ^2 in distribution.

We illustrate this point again using the model specified by Equation 1. For simplicity we assume in this example that there is no background or instrument response, effective area and absorption are constant, and (α, β, γ) are fixed. Thus, the only unknown parameter in Equation 1 is the intensity $\tilde{\mathcal{F}}$. Our goal is to test whether the data are consistent with the null model, i.e., $\tilde{\mathcal{F}} \in \Theta_0 = \{\tilde{\mathcal{F}} = 0\}$ or if there is evidence for the more general alternative model, i.e., $\tilde{\mathcal{F}} \in \Theta = \{\tilde{\mathcal{F}} \geq 0\}$. Let Y_j denote the counts in channel j and let τ be the exposure time, then $P(Y_j|\mathcal{F}_j\tau) = (\exp(-\mathcal{F}\tau))(\mathcal{F}_j\tau)^{Y_j}/(Y_j!)$, or $Y_j \sim \mathrm{Poisson}(\mathcal{F}_j\tau)$ and the log-likelihood is given by

$$l(\tilde{\mathcal{F}}|\boldsymbol{Y}) = \sum_{j=1}^{J} [Y_j \log(\mathcal{F}_j\tau) - \mathcal{F}_j\tau] \tag{16}$$

where $\mathcal{F}_j$ is given in Equation 1. The first derivative with respect to $\tilde{\mathcal{F}}$ is given by

$$l'(\tilde{\mathcal{F}}|\boldsymbol{Y}) = \sum_{j=1}^{J} \left[Y_j \frac{p_j}{\mathcal{F}_j} - p_j\tau \right], \tag{17}$$

where $\boldsymbol{Y} = (Y_1, ..., Y_J)$ and p_j is as in Equation 1. Since

$$l''(\tilde{\mathcal{F}}|\boldsymbol{Y}) = -\sum_{j=1}^{J} Y_j \frac{p_j^2}{\mathcal{F}_j^2}, \tag{18}$$

$l''(\tilde{\mathcal{F}}|\boldsymbol{Y}) \leq 0$, and $l'(\tilde{\mathcal{F}}|\boldsymbol{Y})$ is a decreasing function of $\tilde{\mathcal{F}}$. If

$$l'(\tilde{\mathcal{F}} = 0|Y) \leq 0, \tag{19}$$

(i.e. the data happen to fluctuate so that the first derivative ≤ 0) the maximum likelihood estimate for $\tilde{\mathcal{F}}$ must be 0 (see Figure 7). In this case the LRT statistic, $-2\log R(\boldsymbol{Y})$ must be 0 since the maximum likelihood estimate under the null and the alternative models are both 0 and thus $R(\boldsymbol{Y}) = 1$. To compute the false positive rate of the LRT we assume there is no spectral line (i.e., $\tilde{\mathcal{F}} = 0$ as in the null model) and compute the probability that we reject the null model in favor of the alternative model. According to the central limit theorem, the distribution of Y_j converges to Gaussian as τ increases. Since the expectation of the first term of the summand in Equation 17 is equal to the second term, $l'(\tilde{\mathcal{F}}|\boldsymbol{Y})$ converges to a Gaussian with mean 0 as τ increases. Therefore asymptotically the LRT statistic equals zero 50% of the time, whence the reference distribution of the LRT statistic cannot be a χ^2 distribution with any number of degrees of freedom. Examples of this kind are well known to statisticians and in this case it can be shown that the distribution

[22] The maximum likelihood estimate is a statistic with a sampling distribution. A theorem in mathematical statistics establishes that under the same regularity conditions required for the LRT to be asymptotically χ^2 the distribution of the maximum likelihood estimate becomes Gaussian (i.e., Normal) as the sample size increases; see Serfling (1980).

of the LRT statistic is itself a *mixture* distribution taking on the value zero with probability 50% and follows a χ^2_1 with probability 50% (e.g., see Section 5.4 in Titterington *et al.* (1985)). Mattox *et al.* (1996) notice this same behavior in a Monte Carlo evaluation of the null distribution of the LRT statistics for testing for point source in Energetic Gamma Ray Experiment Telescope data. This mixture distribution should be used to calibrate the LRT statistic when testing for a single line when all other parameters are fixed.

In practice, this may mean that in cases where the continuum is extremely well constrained by the data, and the width and position of the possible line are known, the LRT or F test could underestimate the true significance by about a factor of two. But there is no guarantee that this will occur in real data; particularly when the continuum is not well constrained the true significance can be underestimated or overestimated.

References

Avni, Y., 1978, A&A, 66, 307.

Band, D., Ford, L., Matteson, J., Briggs, M., Paciesas, W., Pendleton, G. & Preece, R. 1997, *ApJ* 485, 747

Band, D., et al., 1995, *ApJ* 447, 289

Band, D., et al., 1996, *ApJ* 458, 746

Bayarri, M.J. and Berger, J.O., 1998, Quantifying Surprise in the Data and Model Verification, *Bayesian Statistics 6*, pp. 53-82, (Oxford University Press)

Bevington, P. R., 1969, *Data Reduction and Error Analysis for the Physical Sciences*. (New York: McGraw-Hill)

Carlin, B. P. & Louis, T. A. 1996, Bayes and Empirical Bayes Methods for Data Analysis (London: Chapman & Hall)

Cash, W. 1979, *ApJ* 228, 939

Chernoff, H. 1954, *The Ann. of Math. Statistics* 25, 573

Connors, A. 1997, in "Proceedings of the Fifth Workshop in Data Analysis in Astronomy (DAA V)", p 251, 1997, ed. Di Gesu, V., Duff, M.J.B., Heck, A, Maccarone, M.C., Scarsi, L. and Zimmeran, H.U., World Scientific Publishing Co., London.

Denison, D. G. T. & Walden, A. T., 1999, *ApJ* 514, 972

Eadie, W.T., Drijard, D., James, F.E., Roos, M., and Sadoulet, B., 1971, Statistical Methods in Experimental Physics (New York: North-Holland)

Elvis, M., Matsuoka, M. Siemiginowska, A., Fiore, F., Mihara, T. & Brinkmann, W. 1994, *ApJ* 436, L55

Everitt, B. & Hand, D. 1981, Finite Mixture Distributions (London: Chapman & Hall)

Fenimore, E., et al., 1988, *ApJ* 335, L71

Freeman, P. E., et al., 1999, *ApJ* 524:753-771

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. 1995, Bayesian Data Analysis (London: Chapman & Hall)

Gelman, A. & Rubin, D. B. 1992, Statistical Science, 7, 457

Gelman, A., Meng, X.-L. & Stern, H. 1996, *Statistica Sinica* 6, 733

Gregory, P. C. & Loredo, T. J. 1992, ApJ, 398, 146

Jefferys, W.H. and Berger, J.O., 1992, Ockham's Razor and Bayesian Analysis, *American Scientist*, 80, pp. 64-72, (Oxford University Press)

Kass, R.E. and Raftery, A.E. 1995, *JASA*, Vol.90, No. 430, 773-795.

Kaastra, J. S., Mewe, R., Liedahl, D. A., Singh, K. P., White, N. E. & Drake, S. A. 1996, *A&A* 314, 547

Lampton, M., Margon, B. and Bowyer, S., 1976, *ApJ* 208, 177.

Lindsay, B. G. 1995, Mixture Models: Theory, Geometry, and Applications (Hayward, CA and Alexandria, VA: The Institute of Mathematical Statistics and the American Statistical Association)

Mattox, J. R., et al., 1996, *ApJ* 461, 396

McLaughlan, G. J. & Badford, K. E. 1988, Mixture Models: Inference and Applications to Clustering (New York: Dekker)

Meng, X.-L. 1994, *The Ann. of Statistics* 22, 1142

Murakami, T., et al., 1988, *Nature* 335, 234

Palmer, D. M., et al., 1994, *ApJ* 433, L77

Piro, L., et al., 1999, *ApJ* 514, L73

Schmitt, J., Collura, A., Sciortino, S., Vaiana, G., Harnden, Jr., F. & Rosner, R. 1990, *ApJ* 365, 704

Schwartz 1978, "Estimating the dimension of a Model", *Annals of Statistics*, 6, 461-464

Serfling, Robert, J. 1980, Approximation Theorems of Mathematical Statistics (New York: Wiley & Sons)

Smith, A.M.F., and Spiegelhalter, D.J. 1980, "Bayes Factors and Choice Criteria for Linear Models", *J. of the Royal Statistical Society, Ser B,* 42, 213-220.

Singh, K., Drake, S., Gotthelf, E. & White, N. 1999, *ApJ* 512, 874

Titterington, D. M., Smith, A. F. M. & Makov, U. E. 1985, Statistical Analysis of Finite Mixture Distributions (New York: Wiley & Sons)

van Dyk, D. A. 2000, *Preprint*

van Dyk, D. A., Connors, A., Kashyap, V. L., & Siemiginowska, A. 2001, *ApJ* 548, 224

Yoshida, A., Murakami, T., Nishimura, J., Kondo, I. & Fenimore, E. E. 1992, Gamma-Ray Bursts: Observations, Analyses and Theories, Ho, C., Epstein, R. I. & Fenimore, E. E., eds. (Cambridge: Cambridge University Press), 399

Wheaton, W. A., et al., 1995, *ApJ* 438, 322

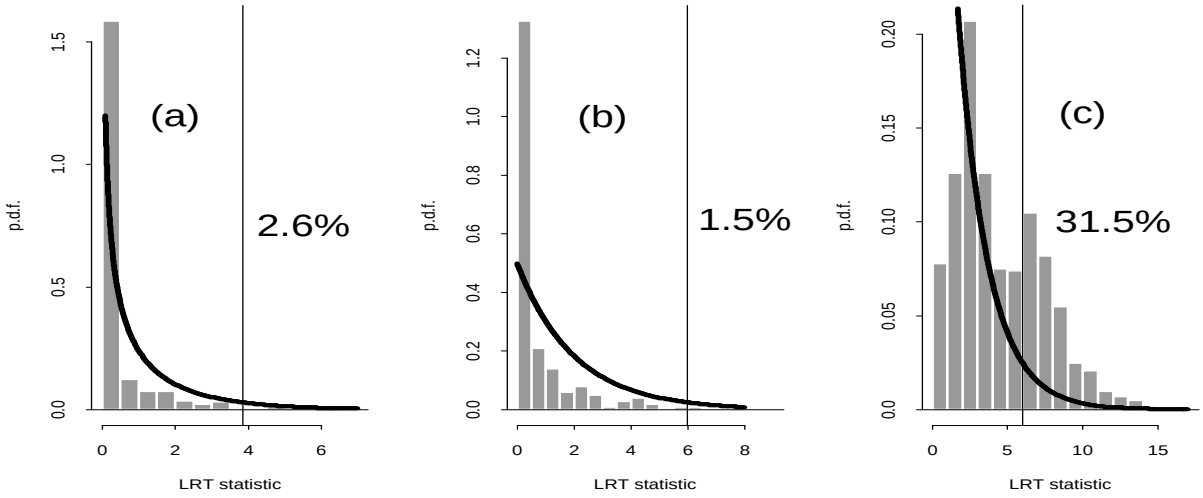


Figure 1: The null distribution of the LRT test statistic. The histograms illustrate the simulated null distribution of the LRT statistic in three scenarios and should be compared with nominal χ^2 distributions which are also plotted. As detailed in Section 3.2, histogram (a) corresponds to testing for a narrow emission line with fixed location, histogram (b) corresponds to testing for a wide emission line with fitted location, and histogram (c) corresponds to testing for an absorption line. The vertical lines show the nominal cut off for a test with a 5% false positive rate; note the actual false positive rates vary greatly at 2.6%, 1.5%, and 31.5%. The label on the y-axis stands for the probability density function (p.d.f.).

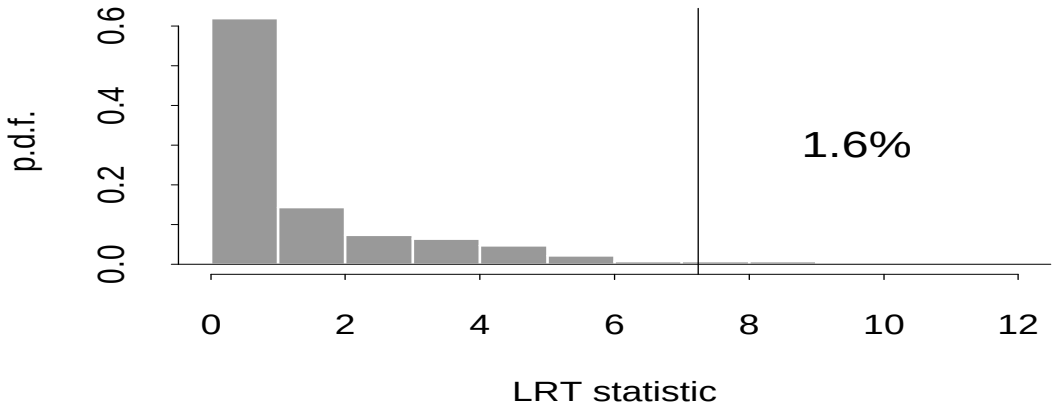


Figure 2: The posterior predictive distribution of the LRT test statistic. This histogram gives the expected variation under $p(T(\tilde{x})^{(t)}|x)$ for the quasar image. Note the observed value, $T(x) = 7.24$, giving evidence against the null (no spectral line) model.

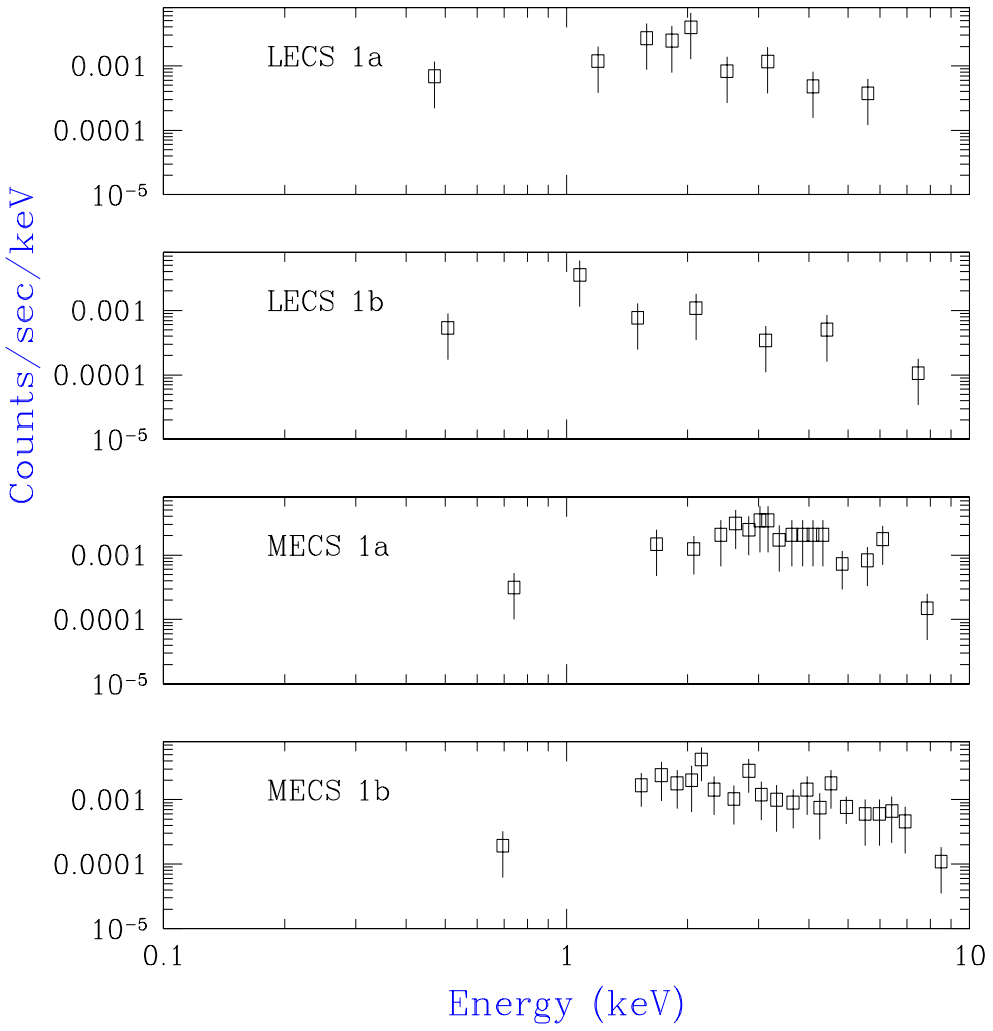


Figure 3: The GRB 970508 observation. The four panels corresponding to the four data sets: LECS 1a & 1b and MECS 1a & 1b. The posited spectral line is at 3.5 keV with width of .5keV in LECS 1a and MECS 1a; it is not present in LECS 1b and MECS 1b.

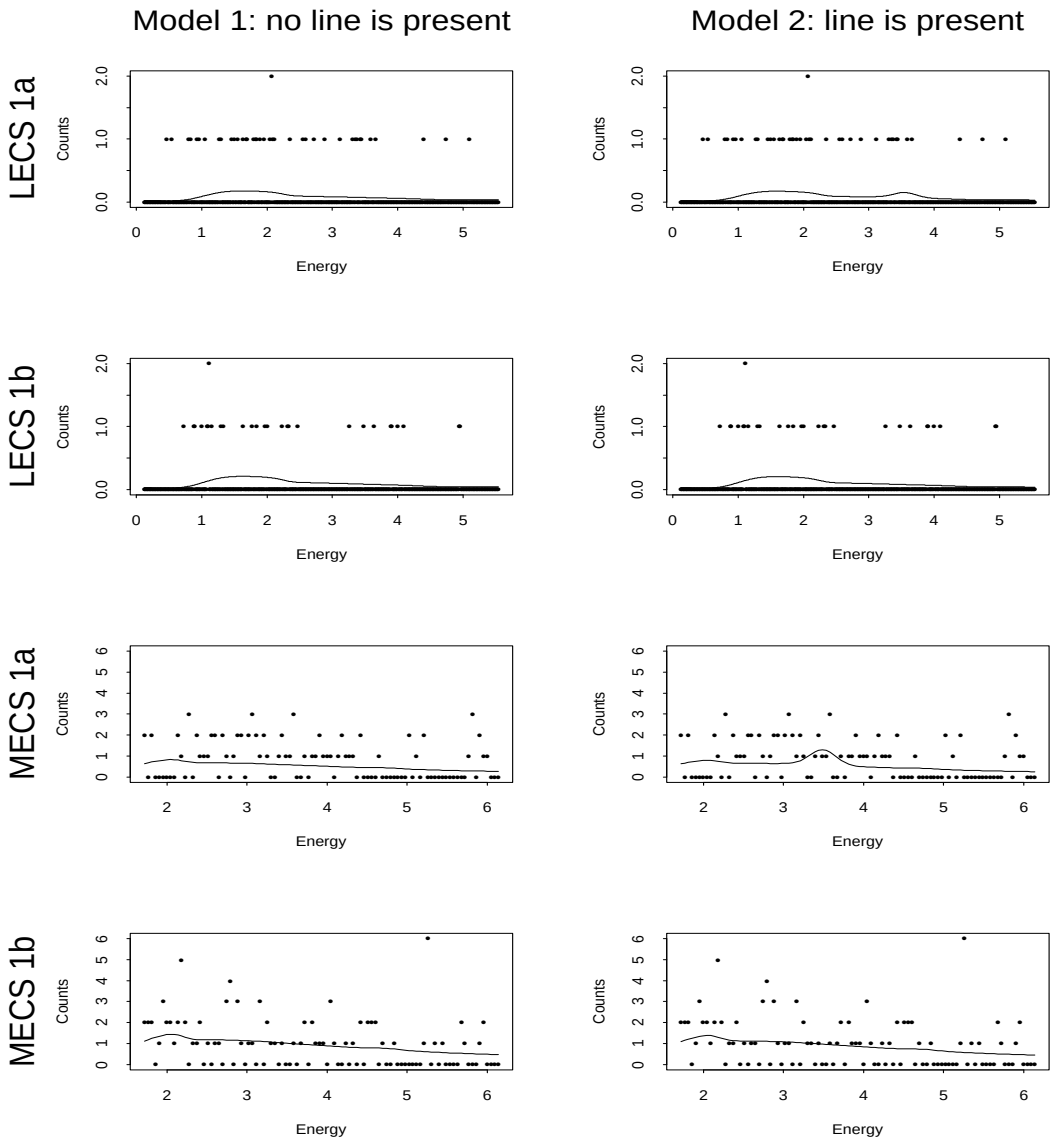


Figure 4: Posterior Analysis of GRB 970508. The black dots correspond to the counts registered in the pha channels. There are four data sets: LECS 1a & 1b and MECS 1a & 1b. Model 1 stands for a simple power law with photon absorption and Model 2 for a simple power law with photon absorption and a Gaussian line at 3.5 keV with fixed width .5keV. Model 2 assumes further that the line is only present in LECS 1a and MECS 1a, but not in LECS 1b or MECS 1b. The posterior distribution of the parameters for each model was studied separately using the method of van Dyk *et al.*(2001). The curves in the plots represent the fitted expected counts and were computed by fixing parameters in each model at their posterior means.

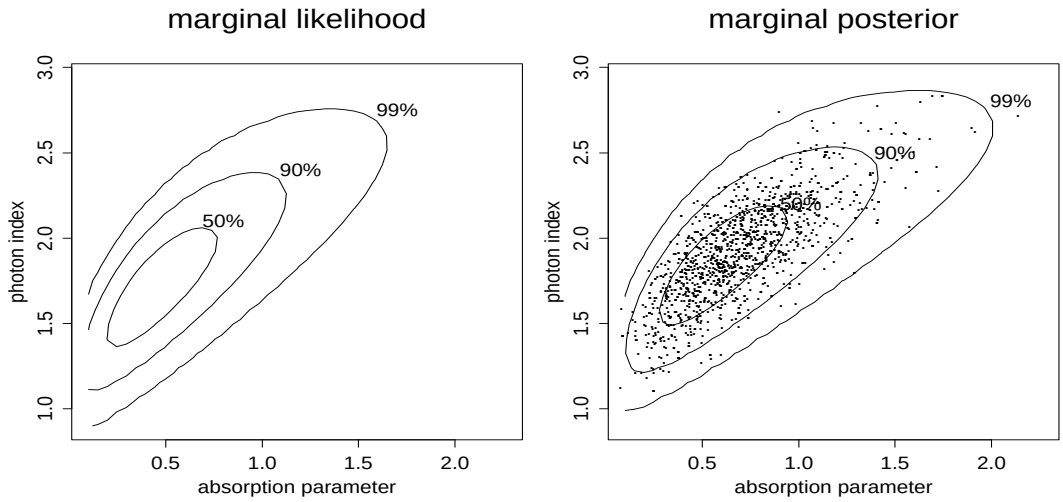


Figure 5: Contours of the marginal likelihood and marginal posterior distribution. The contours corresponds to 50%, 90%, and 99% of the area under each surface; in the case of the marginal posterior distribution, these corresponds to posterior probabilities. Comparing the two plots illustrates that the prior distribution is relatively uninformative. In the plot of the marginal posterior distribution we compare the numerically computed contours with 1500 Monte Carlo simulations generated with the Gibbs sampler; the Monte Carlo simulations are displayed as points and follow the contours well.

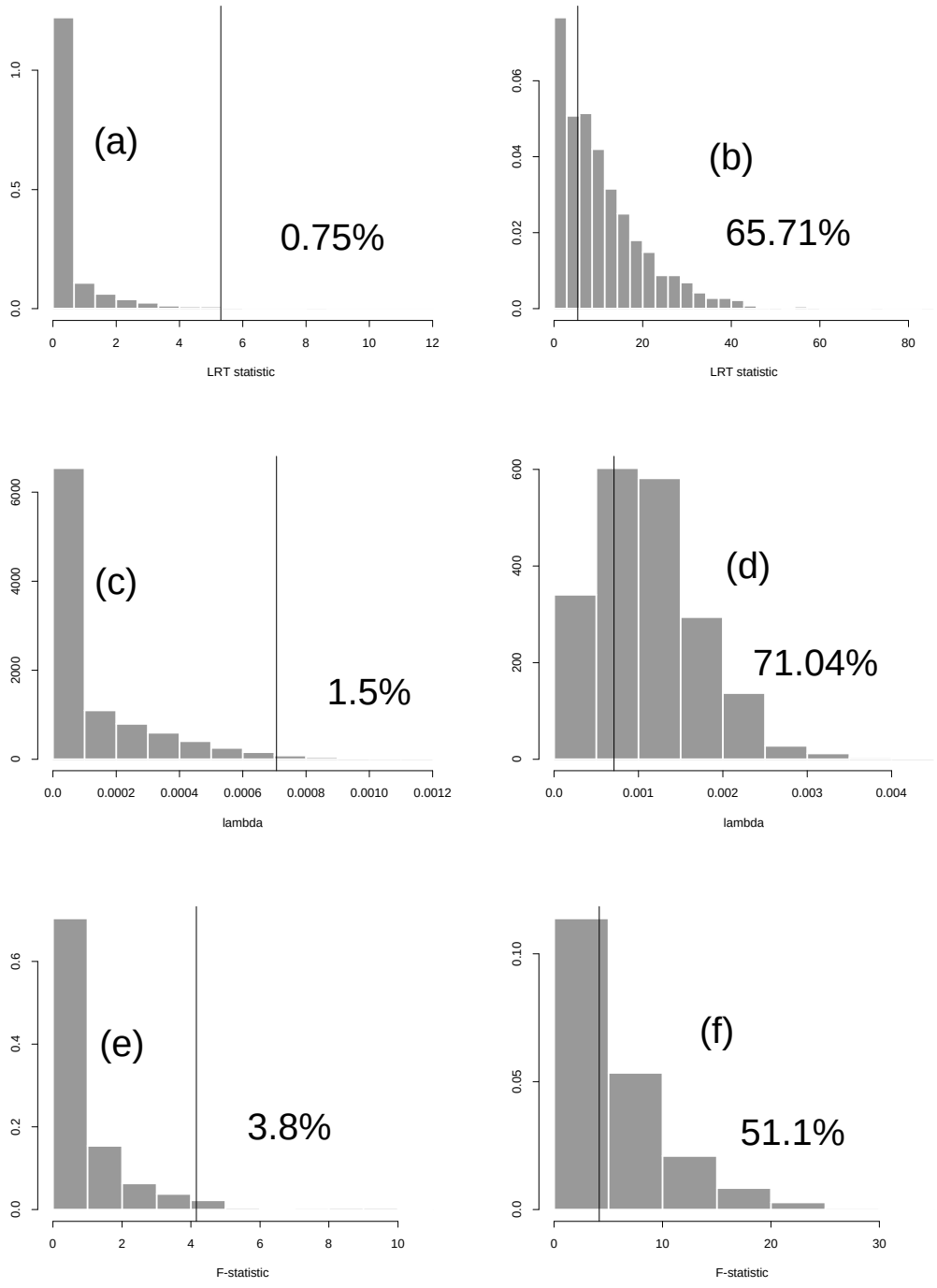


Figure 6: Analysis of GRB 970508. **(a)-(b)**: 2000 simulations from (a) $p(T_{\mathrm{LRT}}(\tilde{x})|x, \text{no line is present})$ and (b) $p(T_{\mathrm{LRT}}(\tilde{x})|x, \text{line is present})$ were used to produce these probability histograms of the posterior predictive distributions for the LRT statistic. Percentages indicate the mass of the corresponding distribution to the right of the vertical line at 5.3, the observed value of the LRT statistic. **(c)-(d):** Observed value of the maximum likelihood estimate for $\tilde{\mathcal{F}}$ at 0.0007 against probability histograms of 2000 simulations from the posterior predictive distribution of this quantity under 2 models: (c) without a spectral line and (d) with a spectral line. We give percentages of the mass of the distribution to the right of the vertical line at 0.0007. Histogram (c) illustrates that the model with a spectral line is clearly preferable. **(e)-(f):** Observed value of the F statistic (vertical line at 4.2) against probability histograms of 1000 simulations obtained from null distributions under two fully specified models: (e) model 1 with parameters as they appear in Table 1 and (f) model 2 with parameters as they appear in Table 1. Percentages indicate the mass of the corresponding distribution to the right of 4.2.

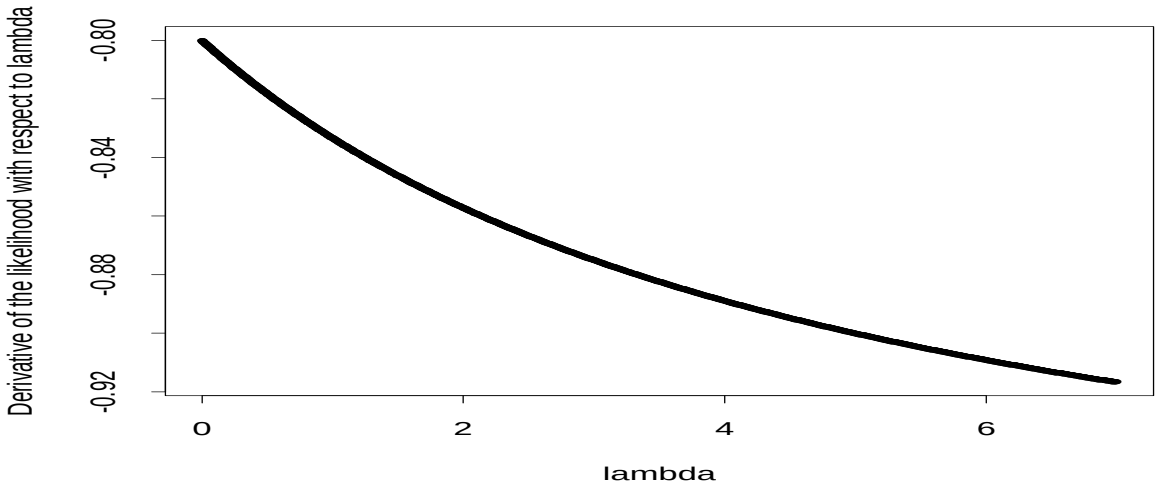


Figure 7: If $l'(\tilde{\mathcal{F}} = 0|Y) \leq 0$, then $l'(\tilde{\mathcal{F}}|Y) \leq 0$ for any $\tilde{\mathcal{F}} \geq 0$ since $l'(\tilde{\mathcal{F}}|Y)$ is a monotonically decreasing function of $\tilde{\mathcal{F}}$, recall: $l''(\tilde{\mathcal{F}}|Y) \leq 0$ for all $\tilde{\mathcal{F}}$ and Y. But this implies that $l(\tilde{\mathcal{F}}|Y)$ is maximized at $\tilde{\mathcal{F}} = 0$ whenever $l'(\tilde{\mathcal{F}} = 0|Y) \leq 0$. This figure is a qualitative illustration.

Model	norm	photon index	n_H (10^{22})	Gaussian norm
Model 1: No emission line $\chi^2 = 25.74467$, $\nu = 14$	3.6424E-04	1.928	0.6397	**n/a**
Model 2: Emission line is present at 3.5keV (width 0.5 keV) $\chi^2 = 19.50732$, $\nu = 13$	4.3427E-04	2.186	0.70393	4.7633E-05

Table 1: Minimal χ^2 fits from XSPEC for data sets 1a+1b.