Different Approaches to Analyzing Multiple Datasets

Version 1.0, January 12, 2005

John Houck

Center for Space Research, Massachusetts Institute of Technology, Cambridge MA 02139

1. Introduction

When multiple spectral datasets are unavoidably coupled, as in the case of marginally resolved point sources, it is usually necessary to analyze the datasets simultaneously. Even when the source of interest is clearly resolved, one may choose to include data from multiple observations of the source in order to derive better constraints on a fitted model.

To simplify the discussion, I will consider a set of point sources for which all count-rates are low enough that photon pileup is insignificant. An application involving high count-rate observations is discussed briefly in §4.

Consider a set of N coupled datasets, each consisting of a spectrum, $D_k(h)$, $k = 1..N$, extracted from some sky region Ω_k. The various regions Ω_k may overlap. To account for the contributions of all the sources to each dataset, compute a set of responses, $R_{kj}(h, E)$ and $A_{kj}(E)$, that describe the response of spectral extraction region k to source j. In terms of these responses, the predicted spectrum for region k may be written in the form

$$C_k(h) = B_k(h) + \sum_j t_j \int \mathrm{d}E \; R_{kj}(h, E) A_{kj}(E) s_j(E) \tag{1}$$

where $B_k(h)$ is the background appropriate to dataset k and where each source is described by a spectrum model $s_j(E)$. Note that (1) assumes that the PSF and QE are such that the effects of the PSF may be folded into the effective area as a fraction (Davis 2001, ApJ 548, 1010).

Multiple observations of a single, steady source may be described by similar system equations. In this case, the sum over j reduces to a single term. Each individual dataset may have a different instrument response, either because different instruments were used or possibly as a result of time-varying instrument characteristics.

The best fit model parameters may be determined by minimizing the χ^2 statistic. For a *joint* fit,

$$\chi^2 = \sum_k \sum_h \frac{1}{\sigma_k(h)^2} \left(C_k(h) - D_k(h)\right)^2 , \tag{2}$$

where $\sigma_k(h)$ is the uncertainty in bin h.

When analyzing multiple observations of a single, steady source, improved model constraints may be obtained from a *combined* fit in which the datasets and models are summed before computing χ^2 so that

$$\chi^2 = \sum_h \frac{1}{\bar{\sigma}_h^2} \left(\sum_k C_k(h) - \sum_k D_k(h) \right)^2 , \tag{3}$$

where $\bar{\sigma}_h$ is the net uncertainty in bin h after combining all the datasets. In §2, I will show that the logic required to implement (3) can also be used to treat the related problem of analyzing spectra from coupled sources.

2. Analysis of Coupled Sources

Most existing X-ray spectral analysis software is designed to handle the case in which each dataset is modeled by a relationship of the form

$$c(h) = b(h) + t \int \mathrm{d}E\ R(h, E)A(E)s(E). \tag{4}$$

With minor modifications, such software should be able to handle systems of the more general form given in equation (1). ISIS already provides support for this kind of analysis.

Write (1) in the form

$$C_k(h) \equiv \sum_j c_{kj}(h) \tag{5}$$

where

$$c_{kj}(h) \equiv b_{kj}(h) + t_j \int \mathrm{d}E\ R_{kj}(h, E)A_{kj}(E)s_j(E). \tag{6}$$

The b_{kj} represent contributions to the instrumental background such that

$$B_k(h) \equiv \sum_j b_{kj}(h). \tag{7}$$

Define an arbitrary set of weights, f_j such that $1 = \sum_j f_j$, so that we can write

$$D_k(h) \equiv \sum_j d_{kj}(h) \equiv \sum_j f_j D_k(h) \tag{8}$$

and $d_{kj}(h) = f_j D_k(h)$. Substituting (5) and (8) into (3), we have

$$\chi^2 = \sum_h \frac{1}{\bar{\sigma}_h^2} \left(\sum_{j,k} c_{kj}(h) - \sum_{j,k} d_{kj}(h) \right)^2 . \tag{9}$$

Since (9) is of the same form as (3), it follows that a simple dataset weighting scheme can be used to analyze spectra from coupled sources using software designed to handle the simpler case defined in (4).

For example, consider two marginally resolved point sources. To simplify the notation, let the same redistribution function, $r(h, E)$, and exposure time, t, apply to both extracted spectra. The extracted spectra may be described by a system of equations

$$C_a(h) = B_a(h) + t \int \mathrm{d}E\ r(h, E)a_{aa}(E)s_a(E) + t \int \mathrm{d}E\ r(h, E)a_{ab}(E)s_b(E) \tag{10}$$

$$C_b(h) = B_b(h) + t \int \mathrm{d}E\ r(h, E)a_{ba}(E)s_a(E) + t \int \mathrm{d}E\ r(h, E)a_{bb}(E)s_b(E) \tag{11}$$

where the energy-dependent effect of the point-spread function is included in the effective area functions, $a_{kj}(E)$. I have labeled the two extraction regions a and b to reduce confusion in the discussion below.

To make this example more concrete, I will assume the analysis is being done using ISIS. Other packages may handle this problem differently.

In ISIS, the `combine_datasets` function is used to indicate which datasets are to be combined in the sense of (3). For example,

```
id = combine_datasets ([1,2])
```

specifies that the contribution of datasets `1` and `2` to the total χ^2 is to be computed according to (3). The return value, `id`, is an integer which acts as a label for the specified dataset combination. This function also supports defining the weights, f_j, discussed above:

```
id = combine_datasets ([1,2], [0.5, 0.5])
```

specifies that the contribution of datasets `1` and `2` to the total χ^2 is

$$\chi^2 = \sum_h \frac{1}{\bar{\sigma}_h^2} \left(\sum_k C_k(h) - \sum_k f_k D_k(h) \right)^2 \tag{12}$$

where $k = 1, 2$ and $f_1 = f_2 = 0.5$. Unless otherwise specifed, the default weights are $f_k = 1$, $k = 1..N$.

This functionality can be used to construct a system of equations such as (1). To set up the equation for $C_a(h)$, load the data from extraction region a, *twice*, creating datasets 1 and 2. Specify that these two datasets are to be combined, with weights $f_1 = f_2 = 0.5$:

```
a_id = combine_datasets ([1,2], [0.5, 0.5]);
```

These weights ensure that the data from extraction region a are not counted twice when χ^2 is computed. The model computed for dataset 1 will represent the contribution from $s_a(E)$ to extraction region a. The model computed for dataset 2 will represent the contribution from $s_b(E)$ to extraction region a. Assign responses to datasets 1 and 2 accordingly. To set up the equation for $C_b(h)$, repeat the same procedure, creating datasets 3 and 4, which are to be combined in a similar fashion to represent the contribution from each source to extraction region b.

Having loaded all the data and responses, we must construct a fit-function so that, when evaluated for datasets 1 and 3, it returns $s_a(E)$ and, when evaluated for datasets 2 and 4, it returns $s_b(E)$. Such a fit-function may be constructed by making use of the intrinsic variable `Isis_Active_Dataset`, which takes on the value of the dataset index for which the fit-function is currently being evaluted. In this case, the necessary fit-function would have the form

```
define coupled_sources ()
{
   switch (Isis_Active_Dataset)
   {
      case 1 or case 3:
      return s_a();
   }
   {
      case 2 or case 4:
      return s_b();
   }
}

fit_fun ("coupled_sources()");
```

where `s_a()` and `s_b()` are suitably defined S-Lang functions that compute the respective source models.

When the model is evaluated, the model contributions from datasets 1 and 2 will be summed, yielding $C_a(h)$, and the corresponding data values will be summed, yielding $D_a(h)$ (the spectrum from extraction region a). Similarly, the combination of datasets 3 and 4 will yield $C_b(h)$ and $D_b(h)$. Finally, the total χ^2 will be computed using (2) as for a joint fit, yielding:

$$\chi^2 = \sum_h \frac{(C_a(h) - D_a(h))^2}{\sigma_a(h)^2} + \sum_h \frac{(C_b(h) - D_b(h))^2}{\sigma_b(h)^2}. \tag{13}$$

Admittedly, the current setup process is somewhat involved. But this process can be simplified by providing a function which takes as input a list of datasets, responses and spectral model specifications, and then automates the process of loading the data and responses and defining an appropriate fit-function.

3. Advantages and Disadvantages

The method of fitting combined datasets described above has several advantages when compared with the traditional alternative in which all spectra are summed and responses combined in a separate step before model fitting begins.

Each individual dataset is treated with the appropriate instrument response and all are combined self-consistently. There is no need to provide additional software to combine responses. By keeping the components separate, it is easy to interactively examine subsets of the data by omitting certain datasets.

The implementation is simplified by requiring that all spectra to be combined share a common grid and that the same bins are noticed or ignored in all spectra. One can often optimize the process of combining N datsets by computing the spectral model once on a sufficiently high resolution wavelength grid and then rebinning that function onto the N separate grids. When the spectral model computation is the most expensive part of the computation, this optimization speeds the process considerably because $N-1$ separate model computations are eliminated in each iteration of the minimization algorithm.

On the other hand, even with this optimization, treating each individual spectrum with the appropriate response can be computationally expensive. When the analysis includes a large number of datasets, each with a large number of spectral bins, the cost of the large number of forward-folding computations may become prohibitive. If the variation in instrument responses is relatively small compared to the noise level in the data, this additional computational expense may not be justified.

The user should judge whether or not this approach is worth the computational expense, given the quality of the available spectral data and the corresponding range of variation in instrument response. Generally speaking, the technique of combining datasets is most useful when there are significant variations in instrument response, either because different instruments were used or because the data are of sufficient quality that even small response variations may matter.

4. Applications

4.1. Extended Sources

In the analysis of extended sources which span multiple CCD chips, it is often desirable to extract a spectrum from a region which spans a chip gap. In cases where only a few such spectra are to analyzed, one can choose to either combine the responses separately and analyze the summed data or to combine the datasets as in (1) and (3). But when automating the analysis of such observations, as is necessary in spectral mapping (cluster temperature maps for example), it may be simpler to use the approach of combining datasets because, otherwise, it is necessary to compute a large number of different response combinations in advance. In analyzing multiple observations of an extended source taken at different roll angles, a similar problem arises. In such cases it is most natural to extract spectra from each observation for a specific sky region and then to combine these spectra using the appropriate responses as in (1) and (3).

4.2. High Resolution Spectroscopy

In analyzing dispersed spectra from HETG for which photon pile-up is not important, one can usefully combine datasets from plus and minus orders, from HEG and MEG and from higher orders. Although the overlapping orders in LETG spectra can be treated using this approach, in that case it is probably simpler to assign multiple responses to the same dataset. ISIS supports this approach via the `assign_rsp` function.

Although most of the discussion has centered on low count-rate observations, broader applications are possible. For example, consider HETG observations of bright sources in which photon pile-up in the dispersed spectrum becomes significant (Davis 2005, private communication). Because the dispersed spectrum spans multiple CCDs, the dither pattern complicates the response near the chip gaps, making the effective response a combination of the responses on each side of the chip gap. When photon pile-up is not important, the linear dependence (4) of the count-rate on the instrument response allows one the option of either combining the relevant responses and analyzing the summed spectra or combining datasets from neighboring chips as in (1) and (3). But when photon pile-up is important, the dependence of the count-rate on the instrument response becomes nonlinear and one no longer has the option of separately combining the relevant responses. In this case, the spectra from each chip must be combined as in (1) and (3).