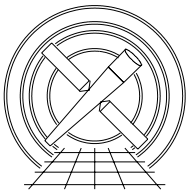
MIT Kavli Institute



Chandra X-Ray Center

How to, & Why (or Why Not) Combine Data

David Huenemoerder, John Davis, John Houck, & Mike Nowak

May 31, 2011

1 All Broken Up

The grating programs carried out by *Chandra* typically have long exposures, and given operational constraints, they are usually divided into two or more distinct exposures. Fewer than half of the targets have only a single observation even after accounting for the calibration monitoring programs.

For programs in which long-term monitoring of spectral variability is not of interest, it is desirable to be able to perform analysis on the data as if it were a single observation. It may seem, at first glance, that to do this one should combine the observational data files. This can certainly be convenient in reducing the number of files to manage. However, it is not necessary to do such since the same effects can be obtained during an analysis session. In some cases, there are advantages to preserving the identity of the individual observations. Here we will explore some of the issues with combining data and some of the methods available for analysis of combined data.

Given that we have multiple datasets which conceptually make a single observation, there are several requirements or preferences for an analyst:

- easy manipulation of multiple datasets, whether two or ten observations, it should be easy to do analysis;
- robust analysis; methods and results should not depend, except statistically, on the number of observations;
- flexible analysis; ability to easily work with different permutations of combined data.

This discussion is relevant to X-ray spectral analysis in general; it is not particular to high-resolution grating spectroscopy, but applies to low-resolution (e.g, CCD non-dispersive) spectroscopy as well.

2 Definition of "Combine"

There are many different data products generated for each *Chandra* observation. By observation, we will refer to the data associated with a single Observation Identifier (ObsID).[1] For a single observation, we can express the spectral counts histogram in terms of the source model and instrumental response as follows:

$$C_m(h) = t_s \int dE\, S(E)\, A_m(E)\, R_m(E,h) + B_{sm}(h) \tag{1}$$

where h refers to our output channel (a wavelength or energy bin), E the model wavelength or energy coordinate bin, and subscript m refers to a diffraction order in the grating's data case, which we will assume implicitly henceforth and only consider a single order. The other terms are

[1] This is usually a sufficient definition. Sometimes there are further subdivisions, such as for "Observation Intervals", or the two frames of "interleaved" mode. We mean the set of files sorted by any such time interval or mode selection.

$C(h)$ is the spectrum, an histogram of counts per coordinate bin ("PHA" (Pulse Height Analyzer) file). This counts spectrum includes whatever background or confusing source counts cannot be explicitly removed (the B-term on the right-hand-side).

t_s is the exposure time in $[\mathrm{s}]$ for the source (plus implicit background) spectrum.

$S(E)$ is the source spectrum, in units of $[\mathrm{photons/cm^2/s/bin}]$;

$A(E)$ is the effective area, or "ARF" (Auxiliary Response File), which has units of $[\mathrm{cm^2 counts/photon}]$;

$R(E,h)$ is the redistribution matrix, or "RMF" (Response Matrix File), which encodes the instrumental energy or wavelength resolution – it is unitless, but not necessarily normalized. For grating data, this encapsulates the Line Spread Function and aperture extraction efficiency;

$B_s(h)$ is the "background" contribution to the source spectrum's extraction region. This term need not be a literal background, but is whatever value is *deemed* "background" – that is, uninteresting for the analysis, but a nuisance term which needs to be included. It could be derived from the same observation as C, a different one, or from models. It is typically obtained from some other region or exposure and then scaled to the source's region and exposure. Whatever it is, it is either not to be convolved with the instrumental response, or already has been. This contains all scaling factors (exposure times and geometric terms).

This will be discussed in more detail below.

Equation 1 is the familiar "forward-folding" expression; since R is typically of a non-diagonal form, this integral cannot be inverted and one must rely on iterative techniques.[2]

2.1 Multiple Datasets

Now we can see that a reasonable high-level definition of a "dataset" is those quantities from Equation 1, $C(h)$, $A(E)$, $R(E,h)$, and $B_s(h)$. In general, these are particular and unique to each observation, and considering different diffraction orders or extraction regions, there could be several sets per observation. To consider multiple datasets appropriate for combined analysis, we introduce the subscript, i, to denote each instance of a like dataset to be considered for combination (the order subscript, m, has been suppressed for clarity):

$$C_i(h) = t_{si} \int dE\, S(E)\, A_i(E)\, R_i(E,h) + B_{si}(h) \tag{2}$$

We can then define our combined data as a summation,

$$\sum_i C_i(h) = \int dE\, S(E) \sum_i [t_{si} A_i(E) R_i(E,h)] + \sum_i B_{si}(h) \tag{3}$$

We here assume that all grids (h, E) are commensurate, otherwise the summations are not valid.[3] (In general, we might also consider a weighting factor, w_i to be applied to each spectrum.) During model optimization, a statistic is defined in the usual way from the summed counts and the summed model counts, using associated uncertainties propagated during the summations of C and B_s.

When data have been combined, we will have more counts per channel, h; this is advantageous if we require Gaussian statistics and such is not acheived for the individual datasets. We will also have a single counts spectrum to visualize, and so can potentially see more features appear above the noise – visualization of the combined data is especially important for guiding analysis and presentation of results.

[2] We are also not considering any time-dependent effects in either the source, response, or background! The assumption is that these quantities pertain to an interval for which time-invariance is valid.

[3] Grids can be defined when data are produced. *Chandra* standard processing and CIAO use standardized spectral grids; analysis tools are available for regridding or re-extraction of products.

2.2 Combined vs. "Joint" Analysis

Note that analysis of combined data as defined above is different from the common practice of fitting and modeling "jointly". Joint analysis means that multiple datasets are modeled simultaneously, but their counts arrays are not summed, nor are the corresponding model counts data summed. This is appropriate and necessary, for example, when fitting data from different instruments (low resolution and high resolution) or data from different observatories (e.g., *Chandra* gratings with Suzaku low-resolution spectra). Joint analysis can also be performed on observations which could be summed if each dataset is in the Gaussian regime or if the appropriate statistic is used, then combined only for visualization.

3 Practical Methods

Given multiple datasets of the same object with the same instrument, how can the data be combined?

3.1 Merging Event Data

First, one should never-ever consider merging the low-level products ("Level 1", event files and associated files). Merging at this level would produce a single set of files for processing to a single set of products, but there are ambiguities and inconsistencies which can make such difficult or error-prone operationally, or just silently produce invalid results. *Just don't do it.*

3.2 Dynamic Method

Use an analysis system which supports dynamic combination of datasets, in other words, one which implements Equation 3 in memory. This means that you do not need to produce any more files than the standard data products for spectral analysis (PHA, ARF, and RMF). Nor do you need to produce files in different combinations (e.g., to manage possible variability, systematics, or gridding issues). The individual datasets can be included or excluded at will in the analysis session, and each unique response, background, and all ancillary data (exposures, backscales – quantities typically read from file headers) are handled robustly and automatically during the optimization.

3.3 File-based Methods

If dynamic data combination is not supported by your favorite analysis package, it is possible to sum PHA and ARF files to produce new such files. For *Chandra* grating spectra, the RMF (which encodes the calibration of the line spread function) depends on grating and order, but rarely on observation (in fact, the grating RMFs are only calibrated and produced for on-axis observations). The only extraction-dependence is on the cross-dispersion region width (the `tg_d` coordinate specified by `tgextract`). Hence, one can typically use one set of RMFs (one per order and grating type) and then sum the PHA and strongly observation-dependent ARF files.

We can put Equation 3 back into the standard form of Equation 1 if we define a mean exposure time and some time-weighted quantities:

$$\mathcal{C}(h) = \mathcal{T}_s \int dE\, S(E)\, \mathcal{R}(E,h) + \mathcal{B}_s(h) \tag{4}$$

where

$\mathcal{C}(h) = \sum_i C_i(h)$ is our summed source plus background counts spectrum,

$\mathcal{T}_s = \frac{1}{N} \sum_{i=1}^{N} t_{si}$ is the mean source observations' exposure time;

$\mathcal{R}(E,h) = \frac{1}{\mathcal{T}_s} \sum_i t_{si} A_i(E) R_i(E,h)$ is an exposure-weighted response, and

$\mathcal{B}_s(h) = \sum_i B_{si}(h)$ is our weighted, summed, scaled background spectrum.

Furthermore, if our redistribution matrix is a constant for the datasets to be summed (as is typically the case for the positive and negative order pairs of each of HEG or MEG gratings, or for LETGS corresponding orders), we can take $R(E,h)$ out of the weighted mean and have

$$\mathcal{C}(h) = \mathcal{T}_s \int dE\, S(E)\, \mathcal{A}(E)\, R_0(E,h) + \mathcal{B}_s(h) \tag{5}$$

with $\mathcal{A}(E) = \frac{1}{\mathcal{T}_s} \sum_i t_{si} A_i(E)$ being an exposure-weighted ARF.

What we ultimately require are thus $\mathcal{C}$ and $\mathcal{B}_s$, our summed source and background spectra, and $\mathcal{A}$, the exposure-weighted ARF. When these are made, care must be taken to include the above definitions of the mean exposure and the exposure-weighted backscale (see below) factors in the files' headers. Given such, analysis is then the same as for any single dataset.

3.4 The Background Term

Our generalized background term can be treated in a similar fashion as the source spectrum, though ignoring the instrumental responses, and considering only "instrumental" space – that is, the detector channel h, and not energy or wavelength. Here we will only consider the "noisy" kind of background – those uniformly spatially distributed, random, time-invariant counts from internal (electronic) sources or from the sky. Such are typically extracted in regions adjacent to the source spectrum from the same exposure, or from separate exposures. (They may, however, have a characteristic spectral shape.)

Other kinds of "background" are left as exercises for the reader (but they may be able to be expressed in the simplified forms described here).

Consider extracting a background spectrum, $B_i(h)$ from the $i^{\rm th}$ observation with exposure time t_{bi}, from the region $\Omega_{bi}(h)$, with an underlying (and *constant*) background rate of r_b $[\mathrm{counts\,area^{-1}\,s^{-1}\,bin^{-1}}]$. The "area", $\Omega(h)$, is a generalized area; for a dispersed spectrum, it may be the angular width of the extraction region, but for an imaging spectrum, it could be a number of pixels or a sky area. We will call it "area", and it is related to the FITS `BACKSCAL` keyword quantity. We will also consider that Ω may be a function of h, which is possible for a non-uniform dispersed spectrum background region. (Since it eventually occurs in a ratio with another such term, the units cancel, and *possibly* dependence on h.)

Our background spectrum can thus be expressed as

$$B_i(h) = t_{bi}\, r_b\, \Omega_{bi}(h) \tag{6}$$

As before, we can now define some mean quantities: a mean exposure time, and an exposure-weighted area to write:

$$\mathcal{B}(h) = \sum_i B_i(h) = r_b\, \mathbf{\Omega}_b(h)\, \mathcal{T}_b \tag{7}$$

where

$\mathcal{T}_b = \frac{1}{N} \sum_{i=1}^{N} t_{bi}$ is the mean background observations' exposure time;

$\mathbf{\Omega}_b(h) = \frac{1}{\mathcal{T}_b} \sum_i t_{bi} \Omega_{bi}(h)$ is an exposure-weighted area.

4 Together Again

We can now finish our assembly of multiple datasets from files by defining the background term in Equations 4 and 5. The source regions also have associated extraction areas, $\Omega_{si}(h)$ and mean, $\mathbf{\Omega}_s(h)$. We can therefore express our "background" contribution to the source region as

$$\mathcal{B}_s(h) = \mathcal{B}(h)\, \frac{\mathcal{T}_s}{\mathcal{T}_b}\, \frac{\mathbf{\Omega}_s(h)}{\mathbf{\Omega}_b(h)} \tag{8}$$

We equate our familiar FITS keyword, BACKSCAL with Ω (and it is a *column* instead of a keyword if it – or more strictly, the ratio – depends on h!). This means that we can treat summing of the source spectra and background spectra identically:

- sum the source counts spectra ($\mathcal{C}(h)$);
- sum the background counts spectra ($\mathcal{B}(h)$);
- compute the mean source exposure ($\mathcal{T}_s$);
- compute the mean background exposure ($\mathcal{T}_b$);
- compute the exposure-weighted mean backscale for the source region ($\mathbf{\Omega}_s(h)$);
- compute the exposure-weighted mean backscale for the background region ($\mathbf{\Omega}_b(h)$);
- compute the exposure-weighted response (ARF, or ARF*RMF) ($\mathcal{A}(E)$ or $\mathcal{R}(E,h)$);

Given these files, which are of standard types with standard keywords, subsequent analysis is the same as for a single dataset.

5 Implementations

5.1 Merging Event Data

While there are CIAO programs for merging Level-1 datasets (e.g., `dmmerge`), don't use them for the purpose of "simplifying" analysis by producing fewer files to process. Data will be ambiguous, inconsistent, or incomplete. We re-repeat, *just don't do it.*

5.2 Dynamic Merging

The only current analysis system we know of which explicitly supports dynamic merging (including vector BACKSCAL values) is the CXC's ISIS[4] program. (The primary function is `combine_datasets`; see its documentation for use and associated functions). High-level visualization of combined data is not yet built-in to ISIS, but an ISIS package is available in `fancy_plots.sl`, which is included in the *TGCat* analysis package distribution.[5] (This is used for the *TGCat* on-line interactive plotting of combined data.)

The CIAO interactive package, Sherpa, is scriptable and extensible, so in principle, a Python package could probably be written to implement dynamic data combination; any custom package would need to provide a user-defined statistic which sums the datasets flagged for combination, and would need to provide the combined model evaluation.

5.3 File-based Merging

The analysis with packages such as XSPEC and Sherpa have no built-in support for dynamic combination of data. Hence there are several options for combining multiple data files externally.

`dmarfadd` is a CIAO program to sum ARFs. This is used to produce the standard grating ARFs by summing the ARF made for each detector element, and so does the exposure weighting, and averaging of the exposure time. It can also be used for non-dispersive ARFs.

`addresp` is a CIAO program for summing ARFs or RMFs.

[4] ISIS, The Interactive Spectral Interpretation System, `http://space.mit.edu/cxc/isis/`

[5] TGCat utilty ISIS scripts are available from `http://space.mit.edu/cxc/analysis/tgcat/index.html`

`combine_spectra` is a CIAO program for summing PHA files, ARFs, and RMFs. (Note in the caveats that only constant `BACKSCAL` is supported (even if a scalar quantity for each observation.)

`add_pha` is an ISIS script for summing PHA files[6]; this is intended to be used in conjunction with `dmarfadd`. (Note that scalar `BACKSCAL` values are properly averaged, but vector `BACKSCAL` is not supported).

`ftools`[7] include several programs for manipulation of spectral files:

`addspec` for adding two or more PHA files; `addarf` for adding two or more ARFs; `addrmf` for adding two or more RMFs; `marfrmf` for multiplication of an ARF by an RMF. (But beware the `ftools` expression parser, which can make it very difficult to handle file names including "/" or "–", for example.)

6 Pros & Cons of the Methods

6.1 Dynamic Combination

Pros

- Unique, well defined responses for each dataset; models for individual datasets are treated consistently for any fit kernel (e.g., pileup).
- Flexible dataset management – can include/exclude any dataset.
- Other applications are supported by the infrastructure, such as fitting spectra of coupled sources, with overlapping regions (see the ISIS help for `combine_datasets` for an explicit example).

Cons

- Multiple files to manage during analysis;
- Greater computer memory requirements for storage of individual datasets;
- Additional software required for visualization of combined data.

6.2 File Combination

Pros

- Fewer files to manage during analysis;
- More memory efficient;
- No additional visualization software is required.

Cons

- Extra preparation is required to produce merged data files, beyond the standard set;
- For each permutation of combined data, a separate set of files must be prepared;
- Need to merge counts and responses consistently, without ambiguity.

[6] For `add_pha` source and documentation, see `http://space.mit.edu/cxc/analysis/add_pha/`.

[7] "A General Package of Software to Manipulate FITS Files"
`http://heasarc.gsfc.nasa.gov/docs/software/ftools/ftools_menu.html`

7 Caveats

7.1 Multiple Sources vs. Multiple Observations

Note that with *Chandra* grating data, there is a subtle distinction between adding orders in one dataset (analogous to multiple sources in one field in imaging observations; e.g., to combine -1 and $+1$), and combining the same order in different observations (like the same source in multiple observations). In the first case, you increase the effective area (ARF) for constant exposure. In the latter, you increase the exposure for constant area. In the definitions given in Equations 4-5, we made no such distinction in computing $\mathcal{T}$ or $\mathcal{A}$. This is sufficient when computing counts since they involve the product, $\mathcal{T}\mathcal{A}$. However, care should be taken if computing count-rates, since the mean exposure and weighted area might be strictly incorrect.

7.2 ARF & RMF, vs. RSP

While use of ARF and RMF matched pairs (whether individual or combined) is sufficient, it is probably "safest" to use the product, the "RSP" (the response), which includes the appropriate weighted sum of $ARF * RMF$, since in general any combination necessarily requires their product.

7.3 Special Cases

In general, the integrand of Equation 1 need not be a linear function of the source model and responses. In fact, with CCD-photon pileup, the response is a non-linear function of the source model. In that case, it is easy to see that we cannot effect a physical sum of response files (especially the RMF) since we do not know the responses until we have a source model. Usually, there is no need to sum counts for analysis of piled spectra. For grating spectra, however, it is possible to have pileup in some region of a spectrum but still wish to combine counts in other spectral regions (given the very high dynamic range of the instrument); here dynamic merging is the *only* valid option.

For imaging data, the RMF is much more observation-dependent than for grating spectra, so an exposure-weighted response is required.

It is possible to have an variable sized extraction region for grating spectra, but have scalar `BACKSCAL` values. In fact, this is the default for Chandra LETG/HRC-S spectra: the extraction regions are not of constant width, but they are of constant width ratio. Hence, we only store the relative values in the headers as scalars. In this case, $\mathbf{\Omega}_s(h)/\mathbf{\Omega}_b(h) = \mathrm{a\ constant}$.