LA-UR-15-27783
Approved for public release; distribution is unlimited.

Title: A Critical Examination of Figure of Merit (FOM): Assessing the Goodness-of-Fit in Gamma/X-ray Peak Analysis

Author(s): Croft, S
Favalli, Andrea
Weaver, Brian Phillip
Williams, Brian J.
Burr, Thomas Lee
Henzlova, Daniela
McElroy, R D Jr.

Intended for: Report

Issued: 2015-10-06

A Critical Examination of Figure of Merit (FOM): Assessing the Goodness-of-Fit in γ/X-ray Peak Analysis

S Croft[1], A Favalli[2], B.P.Weaver[2], B.J.Williams[2], T Burr [2,3], D.Henzlova[2] and RD McElroy Jr[1]

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee, US
[2]Los Alamos National Laboratory, Los Alamos, NM, US
[3]International Atomic Energy Agency, Vienna, Austria

croftS@ornl.gov; afavalli@lanl.gov; theguz@lanl.gov; brianw@lanl.gov; T.BURR@iaea.gov; henzlova@lanl.gov; mcelroyrd@ornl.gov

ABSTRACT

In this paper we develop and investigate several criteria for assessing how well a proposed spectral form fits observed spectra. We consider the classical improved figure of merit (FOM) along with several modifications, as well as criteria motivated by Poisson regression from the statistical literature. We also develop a new FOM that is based on the statistical idea of the bootstrap. A spectral simulator has been developed to assess the performance of these different criteria under multiple data configurations.

Keywords: gamma spectroscopy; peak deconvolution; thermoluminescence CGCD; Poisson regression; bootstrap

1. INTRODUCTION

One approach to performing quantitative radionuclide assay involves peak area analysis of gamma-ray spectra. Consider an isolated peak superimposed on an underlying continuum. Assume over some region of interest (ROI), comprising the peak and portions of the continuum on either side, each channel contains a statistically "reasonable" number of counts. The peak and continuum may then be fit using a suitable composite model function by a nonlinear search algorithm (such as nonlinear least squares) that minimizes the normalized chi-squared. After the "best fit" has been obtained in this way one must decide whether the fit can be considered good or not and whether one fit using one model can be considered better that another fit using a different model (line shape, continuum shape, number of peaks, nuclear data library, etc.) or different assumptions (for instance about precision and bias of the data, peak find criteria in cases where a library driven analysis cannot be used).

The situation is not limited to gamma-ray spectroscopy but is a common application. As another example drawn from radiation metrology we mention the method of thermoluminescence computer glow curve deconvolution (TL CGCD) where fitting based on the basic physics of the underlying process, expressed by the so called order of the kinetics employed, is used to describe the shape and symmetry of the overlapping peaks.

The reasons for bad goodness-of-fit (GOF) scores are usually because of either bad model assumptions resulting in a bad fit, or because of bad variance estimates (caused by improper detector set-up, drifts and so forth), or a combination of both. The procedure used to define the continuum is also crucial for

multiplets, especially if the peaks of most interest are small relative to the other peaks in the ROI and/or the continuum.

Balian and Eddy [1] and subsequently Misra and Eddy [2] have considered the problem given that for this kind of comparative decisionmaking the relative magnitude of the normalized chi-squared may not always be a reliable measure. In an attempt to overcome the deficiency of using chi-square as the sole GOF metric, Balian and Eddy [1] proposed a FOM comprising the sum of the absolute deviations over the continuum only channels normalized to the aggregate counts in the continuum and the sum of absolute deviations over the peak only channels normalized to the aggregate counts in the peak. This FOM concept was later revisited and refined by Misra and Eddy [2] to overcome undesirable consequences and this resulted in an improved FOM, IFOM, which is given below:

$$IFOM = \frac{1}{n_b+n_p}\sum_{i=1}^{n_b+n_p}\frac{|\Delta y_i|}{A_p/n_p} \quad (1)$$

where

n_b is the number of channels considered to be "background" continuum

n_p is the number of channels comprising the peak

$n = n_b + n_p$ is the number of channels in the spectral region analyzed

A_p is the peak area, obtained from the model parameters describing the peak and estimated from the fit

$|\Delta y_i|$ is the absolute deviation between the fitted spectrum and the observed spectrum.

2. DISCUSSION

We agree with [1,2] that using normalized chi-square (the FOM) as the sole arbiter of goodness of fit is poor practice.

What is needed is a fairly "scale invariant" GOF measure. That is, we don't want secondary peaks to tend to have better GOF simply because peak misfit is penalized more strongly in dominant peaks. It seems an open-ended question however as to what particular choice for an alternative to standard chi-square (which will tend to have large/bad values in dominant peaks even for the same relative peak misfit as for a secondary peak) will give better performance.

At first look, standard chi-squared is scale invariant (not in a formal mathematical sense, but in a practical sense because each squared error term is divided by a variance or variance estimate). But, peak misfit effects are not scale invariant. FOMs are sensitive to size of peak if just fit by a single component Gaussian (so "peak fitting" errors have bigger effects in bigger peaks, even if data is Poisson).

Despite the significant advances in automated gamma-ray spectral analysis (references [1] and [2]) and our experience is that there is also great value in having a human subject matter expert visually review the quality of the fit.

The FOM and IFOM add to the overall decisionmaking process. However, these tests alone are not sufficient. For instance, analysis of residual plots is extremely informative. Basic statistical tests for bias, sign preference and gross trending should probably also be applied. Systematic behaviors can inform the experienced spectroscopist of differential non-linearity in the electronics, the manifestation of pulse shape tailing caused by incorrectly pole-zero setting, poor charge collection due to for example radiation damage, or by alerting the researcher to the possibility of random and or true coincidence summing. Broad peaks, inconsistent with a prior shape calibration can be flagged because this could be the result of Doppler broadening (a meaningful physical process) or spurious electrical noise in the system (which is unwanted and uninformative). In situations where peak fits are being performed against a library of lines with known energies, systematic deviations may indicate a shift of the energy calibration. The necessity to include additional peaks may also be evidence. Depending on what is known ahead of time about the assay problem, some or all of these tests may be included in the computer algorithms performing the spectral analysis so that the human operator may only need to investigate when alerted to a possible problem. So, although IFOM is a useful check, it is clear from the forgoing discussion that it is not a panacea.

The formulation of IFOM is arbitrary. The values of IFOM that indicate a good fit (IFOM<0.75%), small flaw (0.75%<IFOM<1.5%), and poor fit (IFOM>1.5%) are empirical, based on a particular experience and subjective. We are aware (see for example [3] and [4]) that the FOM concept has been applied outside the realm of gamma spectroscopy to the domain of TL CGCD. Here the peaks may be strongly asymmetric (non-normal) as for example in glow peaks for first-order kinetics. The traces are of light intensity measured as a current flow in the photomultiplier tube as a function of temperature increasing at a constant rate. The noise structure is therefore quite different from a nuclear counting experiment. The shapes and widths of the contributory peaks are fixed by solid-state effects rather than by a smoothly varying and predictable resolution function. Additionally there is no continuum background term included in the fits. In this case n_b, the number of background bins, would be set to zero and the entire area under the glow peak would appear in the denominator. The suitability of various TL GCD functions when tested against noise free synthetic data can yield FOM (which is equivalent to IFOM in this case) values that are extremely small – close to zero – which is unrepresentative of real data. But the point we wish to emphasize is that IFOM and the numerical values used to judge the quality of a fit were developed in one application space and it is far from clear how the function will perform or be interpreted in another. For now we will focus our discussion and development on the revision of FOM in gamma spectroscopy analysis.

A gamma-ray peak is defined as a Gaussian with various tailing functions and in principle extends many full width half-maxima on either side of the centroid, and so deciding how to partition a region of interest in peak channels and continuum channels seems to us unnatural. The situation is worse for X-ray peaks, which are predominantly Voigtian in shape extending even further (a good reference is [5]).

If chi-square is not universally a good metric, one must question the underlying assumptions of the least squares fitting process. Because, if chi-square cannot be trusted, how is one to reliably estimate

uncertainties on the fitted parameters, in particular in the area of the peaks of interest. IFOM does not help us decide whether the peak area is well defined or not, or whether a weak peak can be detected or not. Other tests are needed for this. Misra and Eddy [2] see a problem when the continuum is weak, for example, since fluctuations can make the goodness-of-fit appear large. Here one must make the distinction between what is statistically significant and what is important from an assay point of view. If the continuum is small in comparison to the strength of the peak we need not be too concerned about how well (either in an absolute or a fractional sense) we estimate it since the impact on the net peak area will be relatively unimportant. In the other extreme, when a weak peak is present on a strong background, fitting the background well becomes critical. In this case having a criterion that emphasizes the peak region may be detrimental. Perhaps a criterion which is either neutral or changes with peak significance would be in order. Further, it is well known that the continuum under a peak is influenced by the counts in the peak at all higher energies, and so it is again unclear how one is to parse a ROI between peak channels and continuum channels.

In cases where a ROI contains multiple overlapping peaks it is not clear how the IFOM formula can be applied since it is cast in the context of a single isolated peak. Indeed for a simple well isolated peak on a smooth continuum, simpler methods than curve fitting might yield satisfactory results and be simpler to give statistical meaning to – for instance a summation across a peak ROI corrected for continuum using an average counts per channel formed from regions below and above the peak. In other words, it is usually when we do not have this simple situation that curve fitting is invoked, the common case being the deconvolution of a complex spectral region. Hence how to evaluate IFOM for the multiple peak case is an important issue although it is not covered in [2].

The concerns over the background continuum being small, we feel, are partially misrepresented in [2] because the statistical foundations of chi-square minimization require that the number of counts per channel be such that each channel can be considered to constitute an independent experiment following Gaussian statistics. Thus, the discussion in [2] is certainly not about the problem of small numbers ("just a few counts"), or the issue that the net peak area may by chance appear negative near or below the detection limit.

In [1] it is stated that narrow peaks might lead to an abnormally large chi-squared value. However we find this statement in need of qualification. In general there are no narrow peaks. Real peaks in a spectrum are expected to have a width that varies in a predictable way with energy (so called energy resolution of a detector). Indeed this provides a means to identify true spectral features from random features. Knowing this the spectroscopic system is normally set up so that the peaks of interest are covered by a reasonable number of channels. On this reasoning we suggest that instead of n_p a more natural definition of number of channels in the peak for use in IFOM would be a multiple of the full width half maximum expressed in channels or a similar metric of peak resolution.

IFOM is a scaled sum of absolute deviations, rather than a sum of standardized deviations squared as is the case for chi-square. There is merit in having a linear absolute metric as opposed to a metric that involves the square since occasional outliers, which can give a large contribution, are deemphasized. However, it seems to us that a deviation only has meaning in comparison to the associated uncertainty. The ratio A_p/n_p may be interpreted as the mean number of counts per channel across the peak. To the extent that peaks tend to be approximately symmetric this quantity can be used to map out the distribution

of peak counts over the channels and has in some sense a meaning from peak to peak. For a Gaussian profile we have

$$A_p = \frac{1}{2}\sqrt{\frac{\pi}{\ln(2)}}HR, \qquad (2)$$

where H is the peak height and R is full width half-maximum. Thus if we replace n_p by a multiple of R as we suggest, we find that the ratio A_p/n_p simply becomes proportional to H, which is a discrete representation of a Gaussian peak placed symmetrically about the centroid and could be said to be the number of counts per channel in the peak channel. A variant on the IFOM that retains the same meaning would then become

$$VFOM = \frac{1}{\mathrm{n}}\sum_{i=1}^{\mathrm{n}}\frac{|\Delta y_i|}{\mathrm{H}}. \qquad (3)$$

In this form several shortcomings are evident. When n becomes large such that the presence of the peak becomes almost irrelevant, how should one interpret VFOM? If the background were flat but subject to Gaussian fluctuations in the number of counts per channel, then a statistical interpretation could be brought to bear but it would have little to say about whether the peak information derived from the fit had meaning (statistical significance) or not. Practical considerations would probably limit the value of n so that the region fitted did not encroach on neighboring peaks and good practice would define it additionally in terms of R. Another problem arises when the supposed peak is weak or absent such that H is small or zero or in the case of an unconstrained fit perhaps even negative. How is one to make sense of VFOM in this case? Of course other tests may be applied in addition to address the question of whether a peak is present or not. We note again that real peaks in a high-resolution spectrometer gamma-spectrum are not purely Gaussian and so the above discussion is only approximate.

To alleviate the objections just raised in relation to VFOM, we consider a Simple FOM, SFOM, as an indication of whether a fit to a spectrum is reasonable across the whole range. Provided the number of counts per channel exceeds about 20 we may take the Poisson counting uncertainty on the number of counts in a given bin to be well represented by a Gaussian probability distribution with a variance equal to the number of counts in the channel. Thus, we propose the following

$$SFOM = \frac{1}{\mathrm{n}}\sum_{i=1}^{\mathrm{n}}\frac{|\Delta y_i|}{\sqrt{y_i}} \qquad (4)$$

where the target peak is approximately centered in the group of n channels and n is chosen to be a multiple of R.

SFOM can be readily applied to complex multiplets sat on a significant pedestal whereas, as already mentioned, in the case of IFOM it is unclear how to do this for what is both a common and essential case. As an example in the analysis of passive gamma-ray spectra from U and/or Pu for the purpose of determining the relative isotopic composition of the gamma emitting nuclides, the 100 keV region may contain 7-13 overlapping peaks, say, that must be individually intensity quantified. The background continuum function depends on regions above and below the multiplet region and also on the counts per channel across the multiplet zone. The quality of the result depends critically on how well the continuum

is represented. In such cases it is a mistake to only emphasize the peak area as IFOM does and a more global metric such as SFOM would seem to be a better choice.

Another thought emerges if we agree with [1] and [2] that the denominator should be more like a "number of counts" as in VFOM, rather than the root number of counts per channel as in SFOM, and one wants to devise a global GOF metric that covers the entire region of the fit including the continuum and multiple peaks. Namely the quantity $\overline{y_i}$, the mean number of counts per channel across the n channels covered by the fit, would be a reasonable choice retaining the character of IFOM. So the Modified FOM becomes, in this case

$$MFOM = \frac{1}{\mathrm{n}\overline{y_i}} \sum_{i=1}^{\mathrm{n}} |\Delta y_i| = \frac{1}{\mathrm{A}} \sum_{i=1}^{\mathrm{n}} |\Delta y_i| \quad (5)$$

where $A = \sum_{i=1}^{\mathrm{n}} y_i$ is the area of the spectral region fitted, and evaluated from the raw data rather than from the fitted parameters (in contrast to IFOM which uses A_p).

MFOM has the advantage that it treats a region with one or multiple peaks exactly the same. It works for a single peak irrespective of peak size - that is it overcomes the objection about what to do when the peak is weak compared to the continuum, i.e., when what you are really doing is fitting the background. MFOM is a modification of FOM and IFOM that is motivated by our discussion of TL GCD function fitting.

In the next section, we perform a simulation study to understand the behavior of the FOMs proposed above in addition to criteria taken from the Poisson regression literature. In particular, we consider the Neyman chi-square (the FOM), Pearson chi-square, deviance, Akaike Information Criterion (AIC), bias-corrected AIC (AICC), Bayesian Information Criterion (BIC), and our proposed bootstrap FOM (BFOM). Most of these criteria can be found in [6]. Before we continue on with the simulation study, we briefly introduce our proposed BFOM.

Suppose we have observed data and a proposed spectral form we wish to fit. After we fit our model to the data, we calculate the residuals from the fit (i.e., observed spectrum minus estimated spectrum). These are then compared to bootstrap residuals, generated using the following algorithm:

1. Using the fitted model, simulate spectral data.
2. With the new simulated data, refit the proposed model.
3. Calculate the residuals for the simulated data.
4. Repeated steps 1-3 a large number of times.
5. Plot the observed residuals along with the simulated residuals.

The BFOM is simply the fraction of energy bins whose simulated residuals do not contain the observed residuals.

3. NUMERICAL EXAMPLES

We present results of a simulation study conducted to examine the behavior of the criteria introduced in Section 2. We consider three configurations. The first (**Configuration A**) involves a mixture of two Gaussian peaks, where the dominant peak masks the secondary peak. In particular, the dominant peak is centered at 185.72 keV with scale parameter 5 keV and amplitude 2.2 counts/second. The secondary

peak is centered at 190.72 keV (185.72 keV + 5 keV) with scale parameter (4/3)*5 keV and amplitude (2/3)*2.2 counts/second. The Gaussian peaks in this and subsequent configurations have the following form,

$$s_1(E) = A \exp\left[-\frac{1}{2}\left(\frac{E-E_0}{\sigma}\right)\right],$$

where A, E_0 and σ are the amplitude, center and scale of the associated peak. Both peaks have an additive GAMANAL tail of the following form,

$$s_2(E) = P_1 \exp[P_2(E-E_0)]\left(1 - \exp\left[-\frac{P_3}{2}\left(\frac{E-E_0}{\sigma}\right)^2\right]\right),$$

where E_0 and σ are the center and scale of the associated peak. The parameters P_1, P_2, and P_3 take the values 0.5*2.2, 0.1, and 0.5 for the dominant peak and (1/3)*2.2, 0.1, and 0.5 for the secondary peak. The background is a composition of complementary error and linear background functions. The complementary error background has the following form,

$$b_1(E) = \frac{1}{2}P_1 \mathrm{erfc}\left(\frac{E-E_0}{\sqrt{2}\,P_2\,\sigma}\right),$$

where E_0 and σ are the center and scale of the associated peak and $\mathrm{erfc}(\cdot)$ is the complementary error function. This background is associated with the dominant peak in this scenario, with the parameters P_1 and P_2 taking the values 0.25 and 1. The linear background has the following form,

$$b_2(E) = c_0 + c_1 E,$$

with the parameters c_0 and c_1 taking the values -0.05 and 0.001. The source function is defined as the sum of the functions $s_1(E)$ and $s_2(E)$ across both peaks, and the background function is defined as the sum of the functions $b_1(E)$ and $b_2(E)$.

Simulation data were generated from the count rate spectrum defined above under a total of 9 scenarios, specified as a cross between the number of channels (3 levels: 50, 100, and 150 spread evenly in the interval [150 keV, 250 keV]) at which counts are observed, and the amount of time each channel is observed (3 levels: time required to achieve 10%, 5%, and 2.5% relative standard deviation in expected counts at peak maximum). In all, 45 count datasets were generated, representing 5 independent replicates of these 9 scenarios. Figure **1** shows data generated from one of the replicates for all 9 scenarios.

Three candidate statistical models were fit to each dataset by minimizing the Neyman chi-square statistic with respect to the free parameters. The first source model assumes a single Gaussian peak with GAMANAL tailing, the second assumes two Gaussian peaks each with GAMANAL tailing (form of the true model), and the third assumes three Gaussian peaks each with GAMANAL tailing. The background is taken universally to be the sum of complementary error and linear functions as specified in the definition of the true spectrum above. These three candidate statistical models have 10, 16, and 22 free parameters, respectively. Table **1** shows the number of analyses (out of 45 total) that each of the three candidate statistical models produced the best (smallest) value for the criteria considered.

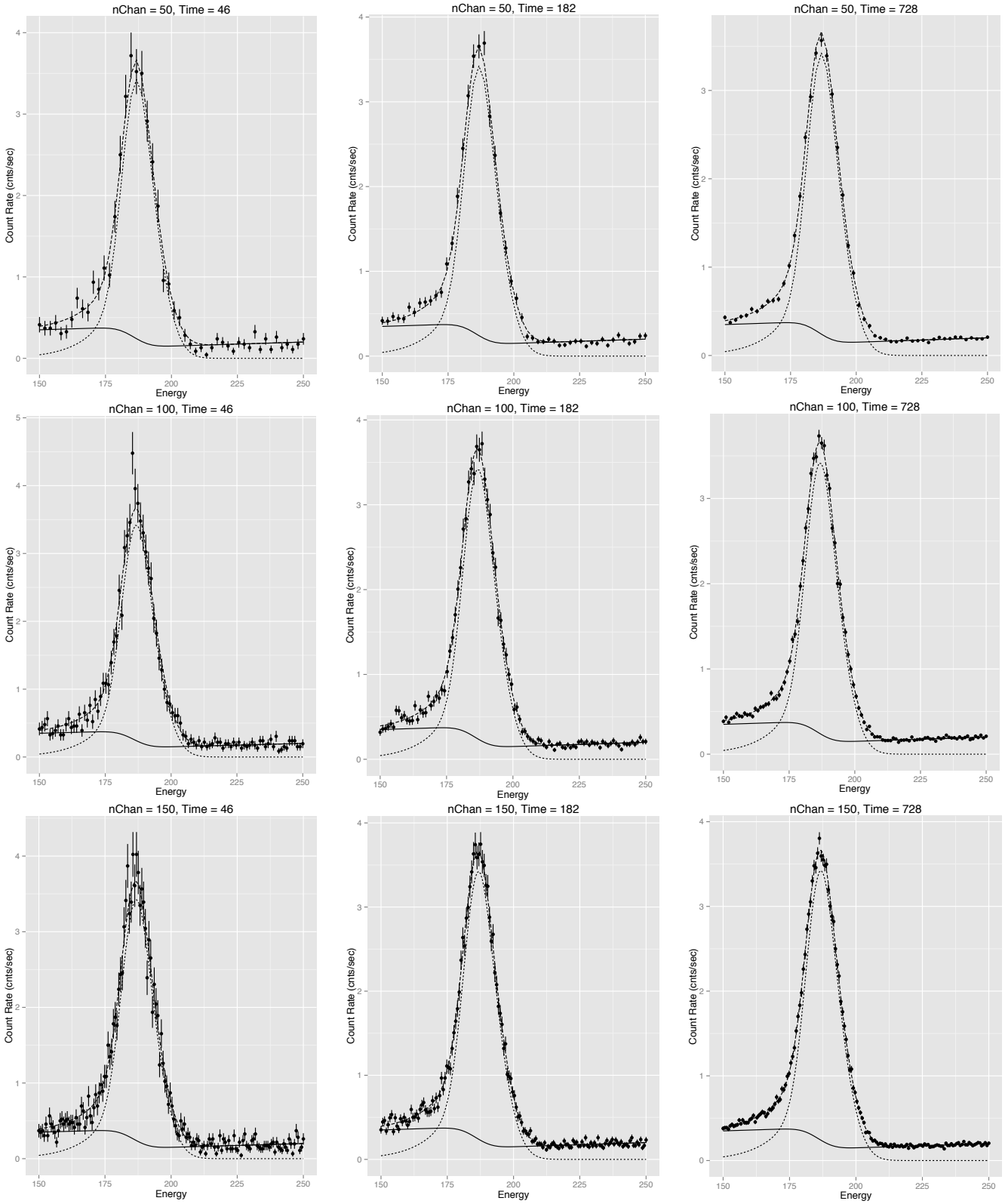

Figure 1. **Configuration A**: Count rate data (solid circles) superimposed on spectrum (long dashed lines), with source (short dashed lines) and background (solid line) components plotted. One standard deviation uncertainty in observed count rate is plotted at each channel.

Non-integer values in this table appear for the BFOM criterion as a result of ties among two or more models for the smallest BFOM error rate.

Table 1. **Configuration A**: Number of analyses each statistical model produced the smallest value of each criterion.

Criterion	1-peak Model	2-peak Model	3-peak Model
Neyman Chi-Square	5	37	3
Pearson Chi-Square	6	36	3
Deviance	5	37	3
IFOM	0	13	32
VFOM	0	11	34
SFOM	0	10	35
MFOM	0	9	36
AIC	11	34	0
AICC	18	27	0
BIC	24	21	0
BFOM	7 2/3	15 1/6	22 1/6

For Configuration A, the chi-square criteria (Neyman, Pearson, and Deviance) do the best job of picking the correct model. These criteria incorporate a model complexity penalty in the sense that more complex models have fewer degrees of freedom for a given dataset. This generally helps these criteria avoid selecting overfit models. Similarly, the information criteria (AIC, AICC, and BIC) either choose the correct model or err on the side of underfitting. The BIC prefers the underfit model with a higher frequency than AIC or AICC, as its model complexity penalty (based on the number of free parameters and sample size) is larger. On the other hand, the proposed figures of merit (IFOM, VFOM, SFOM, and MFOM) strongly prefer the overfit model. This is likely due to the fact that these criteria do not penalize for model complexity, allowing more complex models to potentially achieve a better fit to the data by modeling random noise with their additional flexibility. As we will see with Configuration B, the strength of BFOM is in its ability to detect severely underfit models. In this configuration, Figure **1** visually indicates the presence of one peak, rendering the 1-peak model a potentially viable candidate for data analysis. As a result, BFOM does not detect underfit in almost 20% of the data analyses.

The second configuration (**Configuration B**) also consists of a mixture of two Gaussian peaks. Both the dominant peak and its GAMANAL tail are specified as in Configuration A, while the secondary peak is centered at 200.72 keV (185.72 keV + 3*5 keV) with scale parameter (5/3)*5 keV and amplitude (1/3)*2.2 counts/second. In this configuration, the secondary peak does not have a GAMANAL tail. The source function is defined as the sum of the functions $s_1(E)$ and $s_2(E)$ for the dominant peak and $s_1(E)$ for the secondary peak. The background function for this configuration is identical to that defined for Configuration A.

Five independent replicates of the same 9 scenarios as in Configuration A resulted again in 45 total count datasets. Figure **2** shows data generated from one of the replicates for all 9 scenarios. In this configuration, the presence of a second peak is visually more obvious than it was for Configuration A.

Three candidate statistical models were fit to each dataset by minimizing the Neyman chi-square statistic with respect to the free parameters. These models are identical to those utilized in Configuration A,

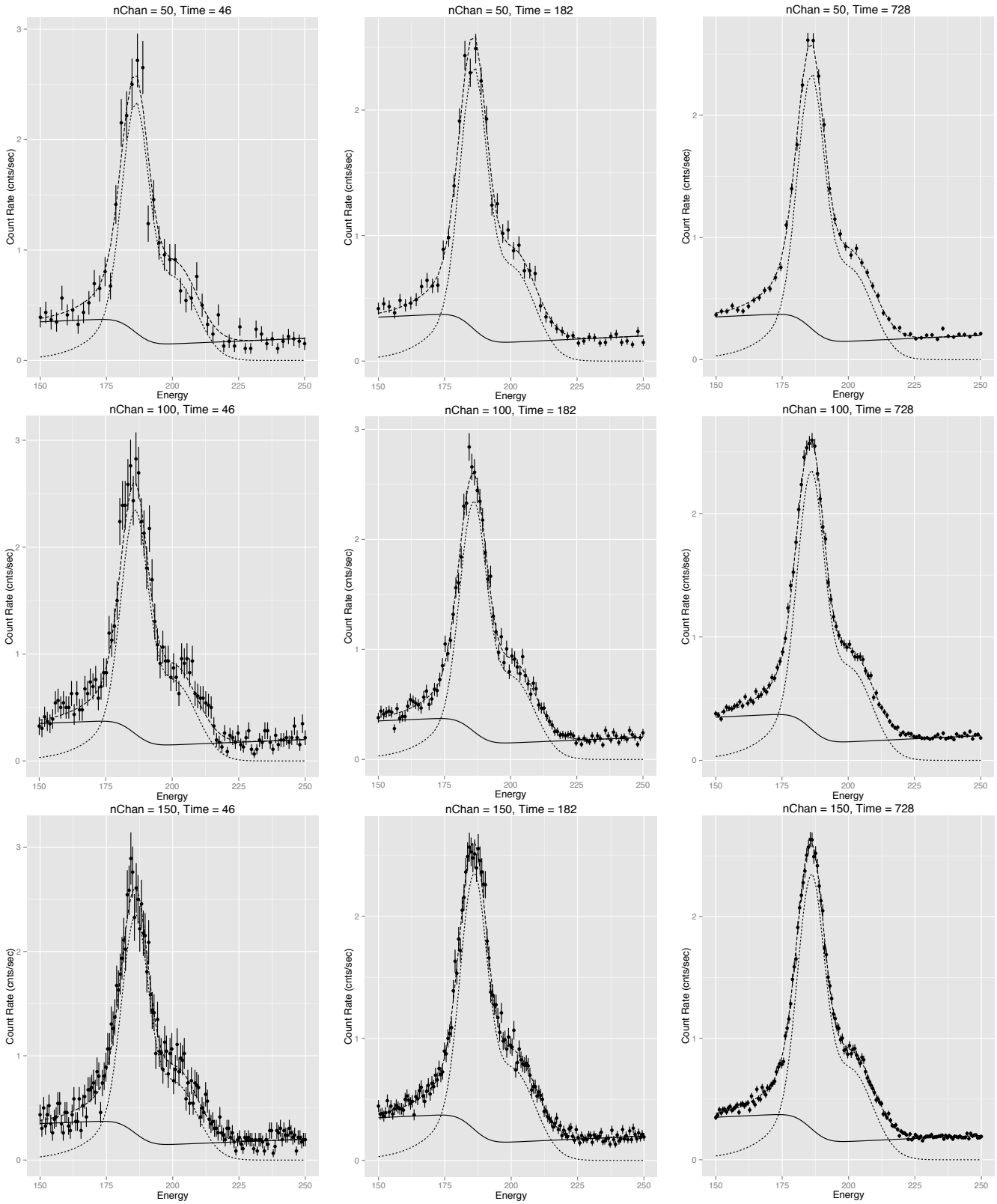

Figure 2. **Configuration B**: Count rate data (solid circles) superimposed on spectrum (long dashed lines), with source (short dashed lines) and background (solid line) components plotted. One standard deviation uncertainty in observed count rate is plotted at each channel.

except a GAMANAL tail is only associated with a single peak. These three candidate statistical models have 10, 13, and 16 free parameters, respectively. Table **2** shows the number of analyses (out of 45 total) that each of the three candidate statistical models produced the best (smallest) value for the criteria considered.

Table 2. **Configuration B**: Number of analyses each statistical model produced the smallest value of each criterion.

Criterion	1-peak Model	2-peak Model	3-peak Model
Neyman Chi-Square	1	38	6
Pearson Chi-Square	1	40	4
Deviance	1	39	5
IFOM	0	20	25
VFOM	5	21	19
SFOM	1	19	25
MFOM	1	23	21
AIC	2	41	2
AICC	4	41	0
BIC	9	36	0
BFOM	4 2/3	20 1/6	20 1/6

These results are similar to those observed from Configuration A, although the correct model is now selected at a higher frequency for every criterion. This is due to the more obvious presence of a second peak. As before, the larger model complexity penalty causes BIC to select an underfit model more often than for any other criterion, but this only occurs for datasets generated with the largest (10%) relative standard deviation at peak maximum when underfitting is more plausible. The proposed figures of merit (IFOM, VFOM, SFOM, and MFOM) again prefer the overfit model more often than do the other criteria.

In this configuration, the BFOM is better able to distinguish the underfit model, selecting it in only approximately 10% of the analyses. Figure **3** shows the results of BFOM when the 1-peak model (left

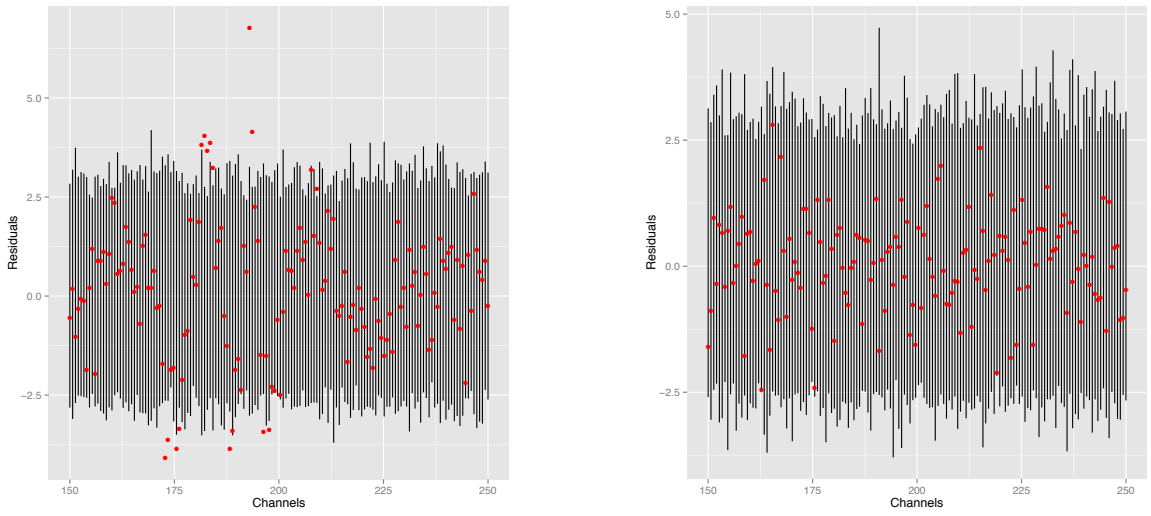

Figure 3. Range of BFOM residuals (vertical lines) from fitting 1-peak (left) and correct 2-peak (right) models, with observed count data residuals (red circles) superimposed.

panel) and (correct) 2-peak model (right panel) are fit to the count data from one of the replications of the 150 channel, 2.5% relative standard deviation at peak maximum scenario. The observed count data residuals fall within the range of the BFOM residuals at most channels when the correct 2-peak model is fit to these count data, while this is clearly not the case for the fitted 1-peak model. Not only do the observed count data residuals fall outside the range of the BFOM residuals at a higher number of channels, but the former also demonstrate a periodic type of behavior often observed when "non-negligible" peak(s) are missing from the fitted statistical model.

The final configuration (**Configuration C**) is an approximate representation of a portion of the 235U spectrum. This spectrum contains a mixture of 5 Gaussian peaks having centers, scales, and amplitudes given in Table **3**.

Table 3. Centers, scales, and amplitudes for 5 Gaussian peak representation of 235U spectrum.

	Center (keV)	Scale (keV)	Amplitude (counts/sec)
Peak 1	143.76	6	0.24
Peak 2	163.33	6	0.15
Peak 3	185.72	5	2.20
Peak 4	202.12	7.5	0.05
Peak 5	205.32	7.5	0.23

The source function is the sum of the contributions to count rate due to each of the five peaks. Peak 3 is the dominant peak, while we expect Peak 4 to be difficult to detect due to its relatively large scale and small amplitude compared against the background. The background is a composition of sigmoid and constant background functions. The sigmoid background has the following form,

$$b_1(E) = \frac{P_1}{1+\exp\left(\frac{E-E_0}{0.75\, w\, P_2}\right)},$$

where E_0 and $w = \sigma\sqrt{8\log 2}$ are the center and FWHM (corresponding to scale σ) of the associated Gaussian peak. This background is associated with the dominant peak in this scenario, with the parameters P_1 and P_2 taking the values 0.25 and 1. The constant background has the following form,

$$b_2(E) = c_0,$$

with the parameter c_0 taking the value 0.15. The background function is defined as the sum of the functions $b_1(E)$ and $b_2(E)$.

Five independent replicates of 9 scenarios similar to those covered in the previous two configurations resulted again in 45 total count datasets. The only difference for this configuration is the number of channels at which counts are observed (3 levels: 200, 250, and 300 spread evenly in the interval [110 keV, 250 keV]). Figure **4** shows data generated from one of the replicates for all 9 scenarios.

Six candidate statistical models were fit to each dataset by minimizing the Neyman chi-square statistic with respect to the free parameters. The first source model assumes three Gaussian peaks, the second assumes four Gaussian peaks, the third assumes five Gaussian peaks (form of the true model), and the

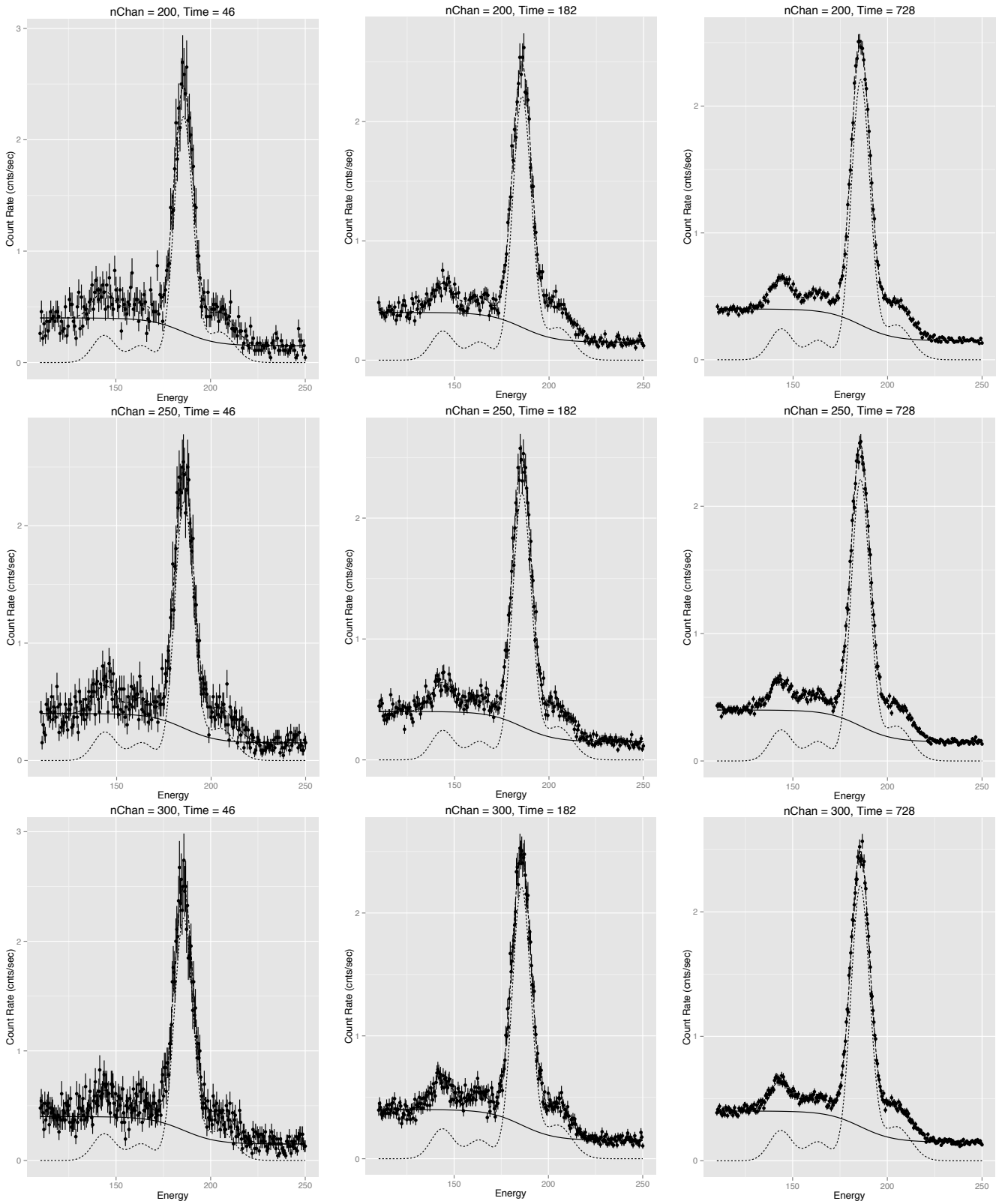

Figure 4. **Configuration C**: Count rate data (solid circles) superimposed on spectrum (long dashed lines) with source (short dashed lines) and background (solid line) components plotted. One standard deviation uncertainty in observed count rate is plotted at each channel.

fourth assumes six Gaussian peaks. The background for these four source models is taken to be the sum of sigmoid and constant functions as specified in the definition of the true spectrum above. The fifth source model assumes four Gaussian peaks, and the sixth source model assumes five Gaussian peaks. The background for these last two source models is taken to be the sum of complementary error and constant functions, where the complementary error background is defined previously in the discussion of Configuration A. These six candidate statistical models have 12, 15, 18, 21, 15, and 18 free parameters, respectively. Table **4** shows the number of analyses (out of 45 total) that each of the six candidate statistical models produced the best (smallest) value for the criteria considered.

Table 4. **Configuration C**: Number of analyses each statistical model produced the smallest value of each criterion. S (sigmoid) and CE (complementary error) indicate background model components.

Criterion	3-peak Model S	4-peak Model S	5-peak Model S	6-peak Model S	4-peak Model CE	5-peak Model CE
Neyman Chi-Square	0	19	1	1	24	0
Pearson Chi-Square	1	22	2	0	20	0
Deviance	0	23	2	1	19	0
IFOM	0	11	9	16	2	7
VFOM	0	3	6	22	2	12
SFOM	0	7	8	12	2	16
MFOM	0	6	8	13	2	16
AIC	1	22	1	0	21	0
AICC	1	22	1	0	21	0
BIC	8	20	0	0	17	0
BFOM	3 7/12	7.7	8 1/3	8 29/30	8 13/60	8.2

Qualitatively these results are very similar to those observed for Configuration B. As before, BIC selects the underfit 3-peak model far more often than any other criterion, but again this only occurs for datasets generated with the largest (10%) relative standard deviation at peak maximum when an underfit model is more plausible. The proposed figures of merit IFOM and VFOM again prefer the overfit (6-peak) model. Interestingly, SFOM and MFOM prefer the 5-peak model with the wrong background (complementary error vs. sigmoid) but with the overfit model a close second. The chi-square and information criteria clearly favor a 4-peak model, but have difficulty distinguishing the (correct) sigmoid and (incorrect) complementary error backgrounds. As previously mentioned, Peak 4 is extremely difficult to detect above the background in this configuration, and these results confirm the difficulty of doing so given the assumed precision in the generated datasets. BFOM successfully detects underfitting in the 3-peak model, but otherwise has difficulty distinguishing among the remaining models.

Two additional scenarios in configuration C are considered in which highly precise data are available. Specifically, suppose five independent replicates of data at 1% and 0.1% relative standard deviation at peak maximum are collected at 700 channels evenly spaced between 110 keV and 250 keV. Table **5** shows the number of analyses (out of 5 total in each additional scenario) that each of the six candidate statistical models produced the best (smallest) value for the criteria considered.

Table 5. **Configuration C**: Number of analyses each statistical model produced the smallest value of each criterion (1% and 0.1% relative standard deviation at peak maximum). S (sigmoid) and CE (complementary error) indicate background model components.

Criterion	3-peak Model S	4-peak Model S	5-peak Model S	6-peak Model S	4-peak Model CE	5-peak Model CE
Neyman Chi-Square	0 0	3 0	0 5	0 0	2 0	0 0
Pearson Chi-Square	0 0	3 0	0 5	0 0	2 0	0 0
Deviance	0 0	3 0	0 5	0 0	2 0	0 0
IFOM	0 0	1 1	3 3	1 1	0 0	0 0
VFOM	0 0	0 1	4 3	1 1	0 0	0 0
SFOM	0 0	0 0	4 4	1 1	0 0	0 0
MFOM	0 0	0 1	4 3	1 1	0 0	0 0
AIC	0 0	3 0	0 5	0 0	2 0	0 0
AICC	0 0	3 0	0 5	0 0	2 0	0 0
BIC	0 0	3 4	0 1	0 0	2 0	0 0
BFOM	0 0	1.45 1	1.2 2	0.45 2	0.45 0	1.45 0

The proposed figures of merit (IFOM, VFOM, SFOM, and MFOM) select the correct model for a plurality of replications in both highly precise data scenarios. The chi-square and information criteria still prefer the 4-peak models in the 1% scenario, but with the exception of BIC they shift to identifying the correct model in the 0.1% scenario. BIC prefers a 4-peak model in both scenarios, although it is able to better identify the correct functional form of the background in the 0.1% scenario. The data is sufficiently precise in both scenarios to rule out the underfit 3-peak model, and in the 0.1% scenario to rule out both models possessing complementary error background components. Figure **5** plots BFOM residuals in the 0.1% scenario for the (incorrect) 5-peak model with complementary error background component (left panel) and for the correct 5-peak model (right panel). Note the identifiable regular patterns in the observed residuals when fitting the incorrect model.

The above analyses highlight the fact that extreme precision in the observed count data may be necessary to reliably identify the correct model, particularly in configurations where one or more peaks are masked by more dominant peaks as in configurations A and C. It appears that the proposed figures of merit (IFOM, VFOM, SFOM, and MFOM) shift from preferring an overfit model to identifying the correct model once sufficient precision in the observed count data has been attained. On the other hand, due to the masking of Peak 4 in this configuration, the chi-square and information criteria shift from preferring

an underfit model to identifying the correct model as observed count data precision increases. This trend appears to be substantially slower for BIC due to its larger model complexity penalty relative to the other information criteria.

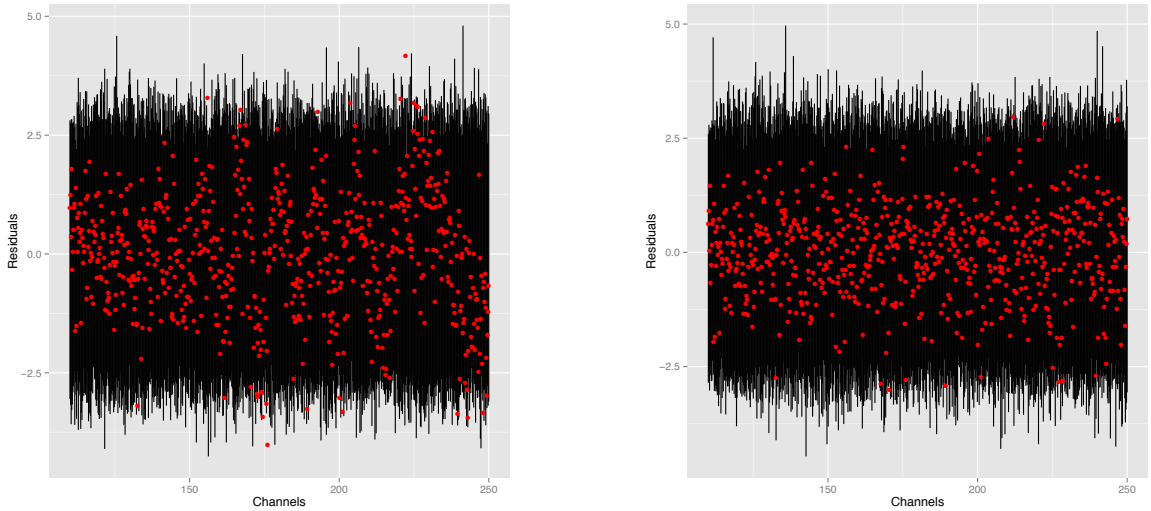

Figure 5. Range of BFOM residuals (vertical lines) from fitting (incorrect) 5-peak model with complementary error background component (left) and correct 5-peak model (right), with observed count data residuals (red circles) superimposed.

4. CONCLUSION

The simulation study of the previous section emphasizes the point that selection of criteria for establishing goodness-of-fit of a model spectrum to count data is a challenging endeavor. None of the criteria considered perform universally better than all others across the quality range in the count data generated.

For count data having relative standard deviation at peak maximum in the range 2.5% - 10%, the proposed FOMs (IFOM, VFOM, SFOM, and MFOM) tend to prefer models that are overfit (too many parameters) relative to the correct model. The extra degrees of freedom are apparently used to model random noise in order to reduce the FOM value, as none of these FOMs possess a model complexity penalty that would provide a counterbalance to this behavior. On the other hand, the information criteria (particularly BIC) may prefer models that are underfit (too few parameters) when individual peak(s) are difficult to detect relative to background. This behavior results from their model complexity penalties overriding goodness-of-fit when the data are too noisy to infer real peak(s) relative to background.

All the criteria investigated converge to identifying the correct model in the scenarios for which the count data have relative standard deviation at peak maximum in the range 0.1% - 1%, although in the case of BIC the larger model complexity penalty still results in preferring an underfit model as best.

The BFOM can be useful for detecting incorrect models by producing a high error rate (fraction of observed residuals outside the range of simulated residuals) and by examination of observed and BFOM residual plots. In the latter, incorrect models often result in observed residuals that exhibit noticeable patterns rather than the expected random noise under the correct model. BFOM is not helpful for

detecting overfit models that have as a special case the correct model, as it does not possess a model complexity penalty as is the case with the proposed FOMs.

It remains an open question to identify criteria that perform universally well across a broad range in quality of the observed count data.

REFERENCES

[1] HG Balian and NW Eddy, Figure-of-merit (FOM), an improved criterion over the normalized chi-squared test for assessing goodness-of-fit of gamma-ray spectral peaks, *Nuclear Instruments and Methods*, 145(1977)389-395.

[2] SK Misra and NW Eddy, IFOM, a formula for universal assessment of goodness-of-fit of gamma ray spectra, *Nuclear Instruments and Methods*. 166(1979)537-540.

[3] R Chen, SWS McKeever, *Theory of Thermoluminescence and Related Phenomena*, World Scientific, 1997.

[4] YS Horowitz and D Yossian, Computerised glow curve deconvolution: Application to thermoluminescence dosimetry, *Radiat Prot Dosimetry*, 60(1)(1995)3.

[5] K Debertin and RG Helmer, *Gamma-And-X-Ray Spectroscopy with Semiconductor Detectors*, North-Holland, 1988.

[6] J. Hilbe, *Negative Binomial Regression*, Cambridge University Press, 2011.