author: Ocean Wong
(Hoi Yeung Wong)
supervisor: Dr Chantal Nobs
Dr Robin Smith
advisor: Prof Alison Bruce
date: May 2020
Organization: Culham Centre for Fusion Energy
Sheffield Hallam University

Abstract

This document attempts to explain why $\frac{\chi^2}{DoF} = 1$ should not be used in case of underdetermined unfolding/fitting; and instead $\chi^2 = 0$ should be used.

The dangers of applying $\chi^2 = 1$ blindly

1 Introduction

In physics we are hard-wired to use $\chi^2 = 1$ as a test of goodness-of-fit. But I warn that this is a dangerous practice to carry into the realm of underdetermined fitting.

This documents will approach in two ways. Section2 will explain it in plain language with the aids of diagram, leveraging on only intuitive understandings of physics; while section 3 requires some familiarity with the χ^2 distribution function.

2 Intuitive Explanation

Let's begin with the first use case that comes to a physicist's mind when χ^2 is mentioned: linear regression.

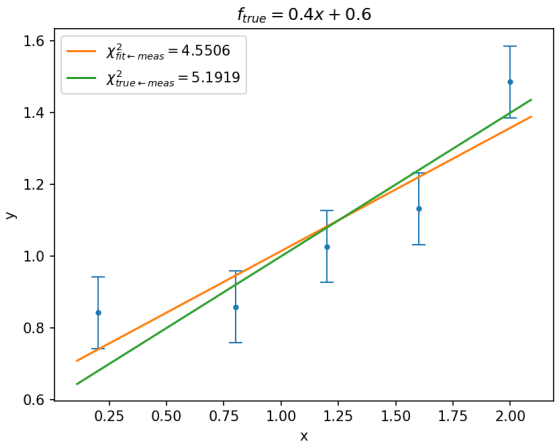

Figure 1: An example of linear regression. The underlying function used is $f_{true} = 0.4x + 0.6$.

Let's say the measurements at x is generated by a true function and some error,

$$y(x) = f_{true}(x) + \sigma_{true}(x) \tag{1}$$

where σ_{true} is a non-deterministic function: its values, after being sampled many times, is Gaussian-ly distributed. f_{true} is deterministic.

By the Law of Large Numbers, when many measurements are taken to obtain a mean,

$$< y(x) >= f_{true}(x) \tag{2}$$

But if only one measurement is taken then one would expect $|y(x) - f_{true}(x)|^2 \approx \sigma^2_{true}(x)$ on average. Sometimes this deviation is overshoots the variance, and sometimes it undershoots.

$$\frac{|y(x) - f_{true}(x)|^2}{\sigma^2_{true}(x)} = 1 \text{ on average} \tag{3}$$

So if we make only one measurement at x_1, one at x_2, one at x_3, ... x_m, then we have

$$\sum_i^m \frac{|y(x_i) - f_{true}(x_i)|^2}{\sigma^2_{true}(x_i)} = m \tag{4}$$

If m measurements were made.

(Don't get too complacent yet, this is NOT the typical chi-squared that we're used to seeing/talking about.) L.H.S. of equation 4 is a measure of how far the measurement deviates from the true distribution. So let's label this quantity, divided by the number of measurements, as $\frac{\chi^2_{true \leftarrow meas}}{DoF}$

$$\frac{\chi^2_{true \leftarrow meas}}{DoF} = \frac{\sum_i^m \frac{|y(x_i) - f_{true}(x_i)|^2}{\sigma^2_{true}(x_i)}}{m} \tag{5}$$

Note that for $\frac{\chi^2_{true \leftarrow meas}}{DoF}$, $DoF = m$ when there are m data-points, since no fitting has been performed.

Following equation 4, on average

$$\frac{\chi^2_{true \leftarrow meas}}{DoF} = 1$$

.

At this stage, as experimentalists, we don't have any knowledge of f_{true}. So unless you have an all-knowing God/deity by your side, who knows both f_{true} and σ_{true}, you won't be able to know the value of $\frac{\chi^2_{true \leftarrow meas}}{DoF}$. (But if such a deity exist and is generous enough to do so, then he can calculate $\frac{\chi^2_{true \leftarrow meas}}{DoF}$ for you using equation 5.)

But unfortunately (to the best of physics' knowledge) such a deity does not exist (or if they did, is a selfish one and won't share f_{true} and σ_{true} with us). Therefore we have to use some other methods than divination to infer f_{true}.

We can construct a function f_{fit} over the same domain as f_{true}'s, to approximate f_{true}. (Orange line in Figure 1)

But how do we know if this f_{fit} is a good approximation of f_{true}?

I mentioned $\frac{\chi^2_{true \leftarrow meas}}{DoF} \approx 1$ on average. If we can, somehow, define an analogous quantity to $\frac{\chi^2_{true \leftarrow meas}}{DoF} \approx 1$ for f_{fit}, and if this quantity = 1, then we can say that f_{fit} *might* be a good fit.

Let's call this analogous quantity $\frac{\chi^2_{fit \leftarrow meas}}{DoF}$. We can't define

$$\chi^2_{fit \leftarrow meas} = \sum_i \frac{|f_{fit}(x_i) - y(x_i)|^2}{\sigma^2_{true}(x_i)}$$

because we still don't know what is $\sigma_{true}(x)$. We have to find a function to replace it.

So instead we use

$$\chi^2_{fit \leftarrow meas} = \sum_i^m \frac{|f_{fit}(x_i) - y(x_i)|^2}{\sigma^2_{meas}(x_i)} \tag{6}$$

$$\frac{\chi^2_{fit \leftarrow meas}}{DoF} = \frac{\sum_i^m \frac{|f_{fit}(x_i) - y(x_i)|^2}{\sigma^2_{meas}(x_i)}}{m - n} \tag{7}$$

where $DoF = (m-n)$ for $\chi^2_{fit \leftarrow meas}$ when there are m data-points, fitted to a model with n free parameters. (This condition of $DoF = (m-n)$ is necessary to ensure $\frac{\chi^2_{fit \leftarrow meas}}{DoF} = 1$. Section 3 provides some empirical evidence for this.)

Now *this* (equation 6) is the chi-squared format that everyone is used to seeing/talking about.

We can make multiple measurements at x_i to get multiple values of y as $y(x_i)_1, y(x_i)_2, y(x_1)_3, ...$, and calculate the variance of them $\mathrm{Var}[y(x_i)_1, y(x_i)_2, y(x_1)_3, ...]$, to be used in place of $\sigma^2_{true}(x_i)$. And then we can plug in $y(x_i) = y(x_i)_1$, $\sigma_{meas} = \mathrm{Var}[y(x_i)_1, y(x_i)_2, y(x_1)_3, ...]$ into equation7 to obtain $\frac{\chi^2_{fit \leftarrow meas}}{DoF}$. [1] Or, we can apply certain assumptions, such as assuming $f_{true}(x) + \sigma_{true}(x)$ follows a Poisson distribution, with mean $\lambda = f_{true}$. [2]

Side note: There is a hidden cost for our arrogance. By challenging God and uncovering what He so vehemently hides, we pay the price of inaccuracy: for example, if we measrued $y(x_0) = 2$ counts per minutes, $\sigma_{meas} = \sqrt{2}$ counts per minute. However, the f_{true} could very well have been $f_{true}(x_0) = 3$ counts per minute with an error of $\sigma_{true} = \sqrt{3}$ counts per minute, which is quite far off from the σ_{meas} used to approximate it.

2.1 A graphical representation

The usual procedure of function fitting is as follows:

1. Acquire measurements ($y(x_i)$ for a list of i=1, 2, ...)
2. Propose a function, which usually may contain free parameters. e.g. $f_{fit}(m, c) = mx + c$
   - (If free parameters are present) fit them the best of your (/your computer's) ability by minimizing the root-sum-squared residuals through adjusting these free parameters. (e.g. via adjusting m and c)
3. The optimized f_{fit} is then expected to have $\frac{\chi^2_{fit \leftarrow meas}}{DoF} \approx 1$

Step 2 has the effect of descending down the negative-log-likelihood surface.

[1] This is only a half-truth, written in place of the truth for simplicity. In reality, having made more measurements = obtained more information about the system; it would be foolish to let these new information go to waste. So instead we would plug in $y(x_i) = \overline{y(x_i)} = \frac{\sum_j^M y(x_i)_j}{M}$ if M measurements were made, and $\sigma^2_{meas}(x_i) = \frac{\mathrm{Var}[y(x_i)_1, y(x_i)_2, y(x_1)_3, ...]}{M}$.

[2] Poisson statistics allows us to infer the error σ directly from a single measurement. Let's say we only measure for 1 minute y counts. Then the counts per minute has an error $\sigma_{meas} = \sqrt{y}$.

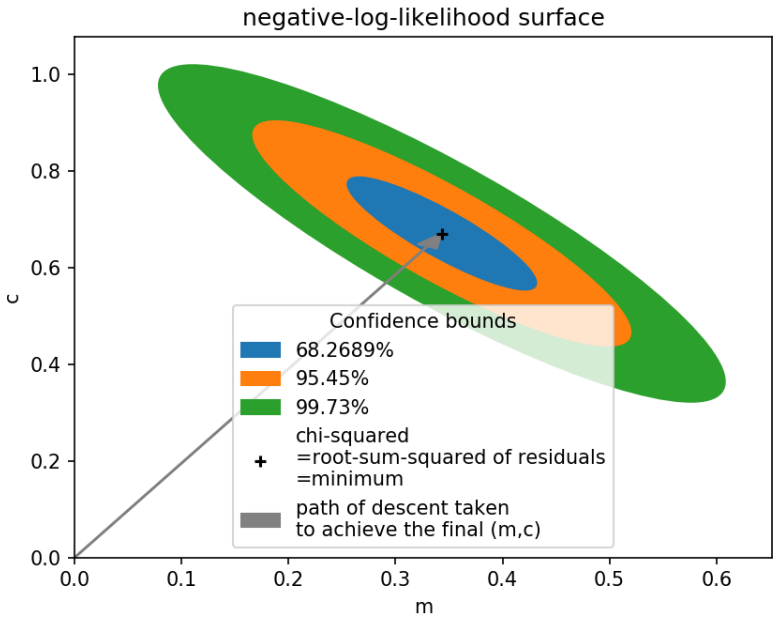

Figure 2: The phase space of the fitting parameters used in figure 1. Each point on this graph corresponds to a combination of (m,c) that gives a $f_{fit} = mx + c$. Thus we can calculate a value of $\frac{\chi^2_{fit \leftarrow meas}}{DoF}$ for each point on this graph. For Gaussian distributions, the chi-square function $= 2\times$ negative log likelihood function. The edges of the concentric ellipses are equivalent to the contour lines encircling the centre of the chi-squared surface bowl.

After that, the goodness-of-fit test performed in step 3 gives a positive result(i.e. $f_{fit} = f_{true}$) if $\frac{\chi^2_{fit \leftarrow meas}}{DoF} = 1$, vice versa.

It so happens that, in this particular fit, $\frac{\chi^2_{fit \leftarrow meas}}{DoF} = \frac{5.1919}{5} \approx 1$, passing the goodness-of-fit test.

Graphically speaking, the black cross in figure 2 (at the bottom of the bowl) is expected to have $\frac{\chi^2_{fit \leftarrow meas}}{DoF} = 1$. Otherwise, we may have a function $f_{fit} \neq f_{true}$.

However, in the underdetermined case, the χ^2 surface is quite different: It has at least one singular direction, i.e. a direction along which the negative-log-likelihood function does not vary, and there is always some points where $\frac{\chi^2_{fit \leftarrow meas}}{DoF}$ can equal 0. (figure 3)

Source of the program	Relies on simultaneously minimizing distance to the *a priori* and $\frac{\chi^2_{fit\leftarrow meas}}{DoF}$ of radionuclide populations (error propagation possible)	Change the spectrum iteratively, starting from the a priori (error propagation not possible)
UMG3.3	MAXED: extremize cross-entropy (in a weird form that I do not agree with) while keeping $\frac{\chi^2_{fit\leftarrow meas}}{DoF}$ at a user-defined value. Can be trapped in sub-optimal points.	GRAVEL
Improved from UMG 3.3 (by me)	IMAXED: maximizes the same quantity as above, but uses a different optimization algorithm so it converges faster, and gives higher quality results. (Gets much closer to the maximum point than MAXED can ever dream of achieving)	
My own invention	Regularization code: extremize the cross-entropy in the correct form	Pseudo-inverse (documentation pending): Apply pseudo-inverse to approach the solution spectrum

Table 1: Categories of unfolding algorithms, and some prominent examples for each categories

Unfolding algorithms can be broadly divided into one of three categories. The first two are listed at the top of column 2 (minimize distance from *a priori*) and 3 (start from *a priori*, take steps until $\frac{\chi^2_{fit\leftarrow meas}}{DoF}$ <threshold) of Table 1. The third kind is parametrisation, i.e. assume the neutron spectrum is a combination of Maxwellian and Watt distribution, reduces the dimensionality of the problem to avoid the problem of underdetermination altogether. However, this dimensionality reduction approach is not mature enough to be applied in Fusion Neutron physics, thus is not discussed here.

Limited by the number of dimension of the medium that I'm presenting through, I can only show an underdetermined system with two neutron bins and one type of radionuclide population ($n = 2, m = 1$). (See appendixA for a refresher on the terms used.)

Let's say we have

a response matrix (shape $= m \times n$) $\underline{\underline{\mathbf{R}}} = \begin{pmatrix} 1 & 2 \end{pmatrix}$,

a true spectrum $\boldsymbol{\phi_{true}} = (1, 2)$;

then the true radionuclide populations is $\boldsymbol{N_{true}} = \underline{\underline{\mathbf{R}}}\boldsymbol{\phi_{true}} = 5$, and $\sigma_{true} = \sqrt{5}$.

And let's say our *a priori* is somewhere in the vicinity of the true spectrum, $\boldsymbol{\phi_{ap}} = (1.25, 2.6)$. (The usual units apply.)

This means that if this experiment is repeated many times, the experimenters will find on average $\boldsymbol{N} = 5$, with a standard deviation $\sigma(\boldsymbol{N}) = \sqrt{5}$.

But unfortunately neutron spectrum measurement experiments are expensive and resource intensive, so it is likely that we will only have one measurement. Let's say in our first experiment, $\boldsymbol{N_{meas}} = 6$. Then we will assume $\sigma = \sqrt{6}$.

The response matrix spans $m = 1$ dimension, while leaving the remaining $n - m = 1$ direction undetermined (singular).

Since the response matrix spans the $\frac{1}{\sqrt{1^2+2^2}}(1,2)$ direction, the singular direction is the direction orthogonal to that, i.e. $\frac{1}{\sqrt{1^2+2^2}}(-2,1)$.

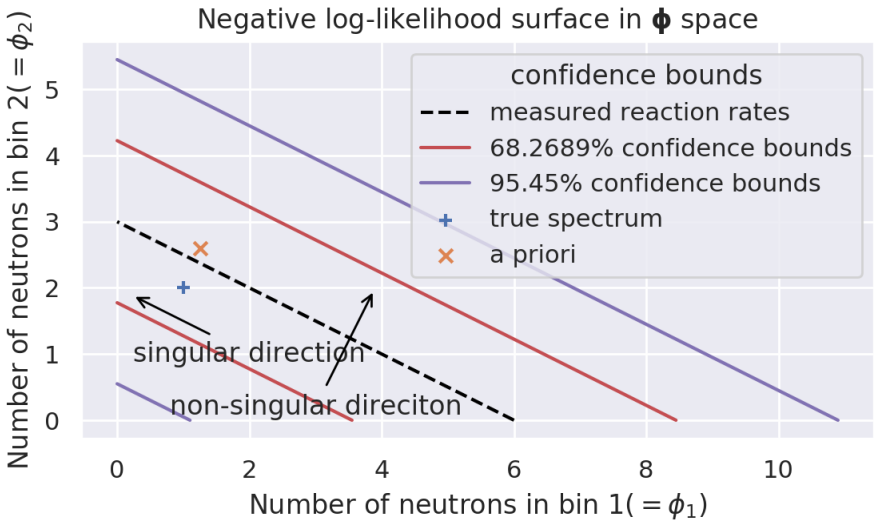

Figure 3: The negative log likelihood function in an underdetermined system must have at least one singular direction, and there must be more than 1 point corresponding to $\frac{\chi^2_{fit\leftarrow meas}}{DoF} = 0$ (Denoted with the dotted line here.)

Notice that all contour lines are aligned in the singular directions. This is because, in an underdetermined system, any changes in $\boldsymbol{\phi}$ along the singular direction(s) will not affect the outputted radionuclide populations. In other words these contour lines are also "iso-radionuclide-population" lines. Any $\boldsymbol{\phi_{test}}$ on the same line will produce the same set of radionuclide population(s).

Thus the χ^2 surface forms a parabolic "trench" instead of the familiar bowl shape in figure 2.

By examining figure 5 and 4, it becomes intuitive why setting $\frac{\chi^2_{fit\leftarrow meas}}{DoF}$ is an incorrect practice.

Bearing in mind that our goal is to find a solution $\boldsymbol{\phi_{sol}}$ as close to $\boldsymbol{\phi_{true}}$ as possible, setting $\frac{\chi^2_{fit\leftarrow meas}}{DoF} = 1$ gives a solution worse than setting $\frac{\chi^2_{fit\leftarrow meas}}{DoF} = 0$.

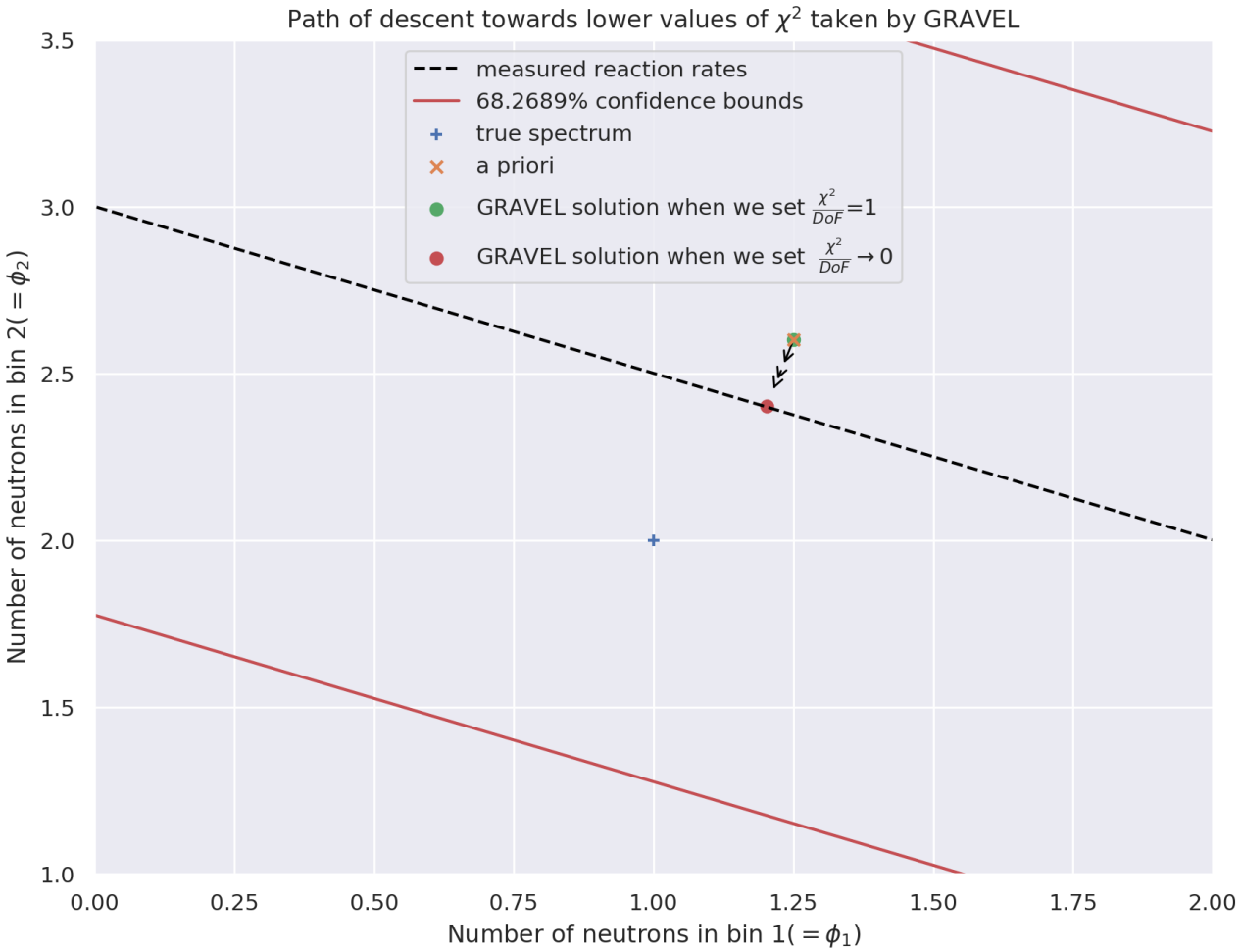

Figure 4: The solutions that GRAVEL will output when different termination values of $\frac{\chi^2_{fit \leftarrow meas}}{DoF}$ is inputted. Note that in the first case, the *a priori* was outputted directly, as the solution started in a place where $\frac{\chi^2_{fit \leftarrow meas}}{DoF}$ <termination value=1. (In figure 3, $\frac{\chi^2_{fit \leftarrow meas}}{DoF} < 1$ within the red lines, $\frac{\chi^2_{fit \leftarrow meas}}{DoF} < 4$ within the purple lines.)

GRAVEL takes steps towards the centre-line of the trench, monitoring the $\frac{\chi^2_{fit \leftarrow meas}}{DoF}$ after each step, and terminates when $\frac{\chi^2_{fit \leftarrow meas}}{DoF}$ drops below a user-defined value.

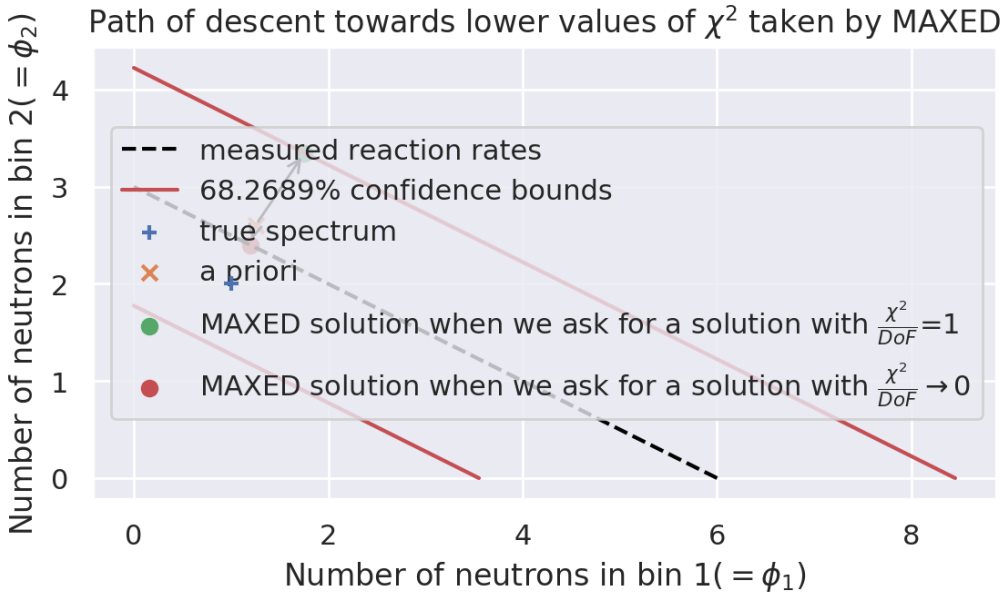

Figure 5: The solutions that MAXED will output when using different input of "desired $\frac{\chi^2_{fit \leftarrow meas}}{DoF}$"

MAXED uses the value of $\frac{\chi^2_{fit \leftarrow meas}}{DoF}$ inputted by the user as part of its constraint, and extremize the cross-entropy between the solution spectrum and the *a priori* spectrum.

To make this argument more convincing, I can repeat the process above in a Monte Carlo approach: while keeping both the *a priori* and the ϕ_{true} fixed, simulated measurements of the radionuclide populations $\boldsymbol{N}$ can be generated via the non-deterministic function $\boldsymbol{N_{true}} + \sigma_{true}(\boldsymbol{N})$, and the unfolded solution can be plotted as a dot. We will then generate two collections of dots for each algorithm, one for the $\frac{\chi^2_{fit \leftarrow meas}}{DoF} = 1$ solutions for each algorithm, and one for the $\frac{\chi^2_{fit \leftarrow meas}}{DoF} \to 0$ solutions for each algorithm. We can then see that the $\frac{\chi^2_{fit \leftarrow meas}}{DoF} = 1$ cloud of dots is spread further away from $\boldsymbol{\phi_{true}}$.

However, this exercise is rather time-consuming and does not convincingly prove that the same holds in higher dimensions. Therefore I have opted to not make such a graph.

3 The relationship between chi-squared (the quantity), and chi-squared (the distribution)

This whole time we have been discussing the reduced-chi-squared $= \frac{\chi^2_{fit \leftarrow meas}}{DoF} \approx 1$ if we're not unlucky. However, I have not yet quantified how close to unity is '≈ 1'. This ambiguity will be resolved in the following section; along with it we will discover another piece of evidence, alluding to the fact that using $\frac{\chi^2_{fit \leftarrow meas}}{DoF} = 1$ is wrong.

Each time we perform an experiment and perform a fitting procedure, we will obtain a slightly different set of measurements, and a slightly different value of $\frac{\chi^2_{fit \leftarrow meas}}{DoF}$. And although we can't know it, this set of measurements will also have a value of $\frac{\chi^2_{true \leftarrow meas}}{DoF}$.

Let's say we repeat the "experiment" 16 times. Then we have a list of 16 $\chi^2_{fit \leftarrow meas}$ and 16 $\chi^2_{true \leftarrow meas}$.

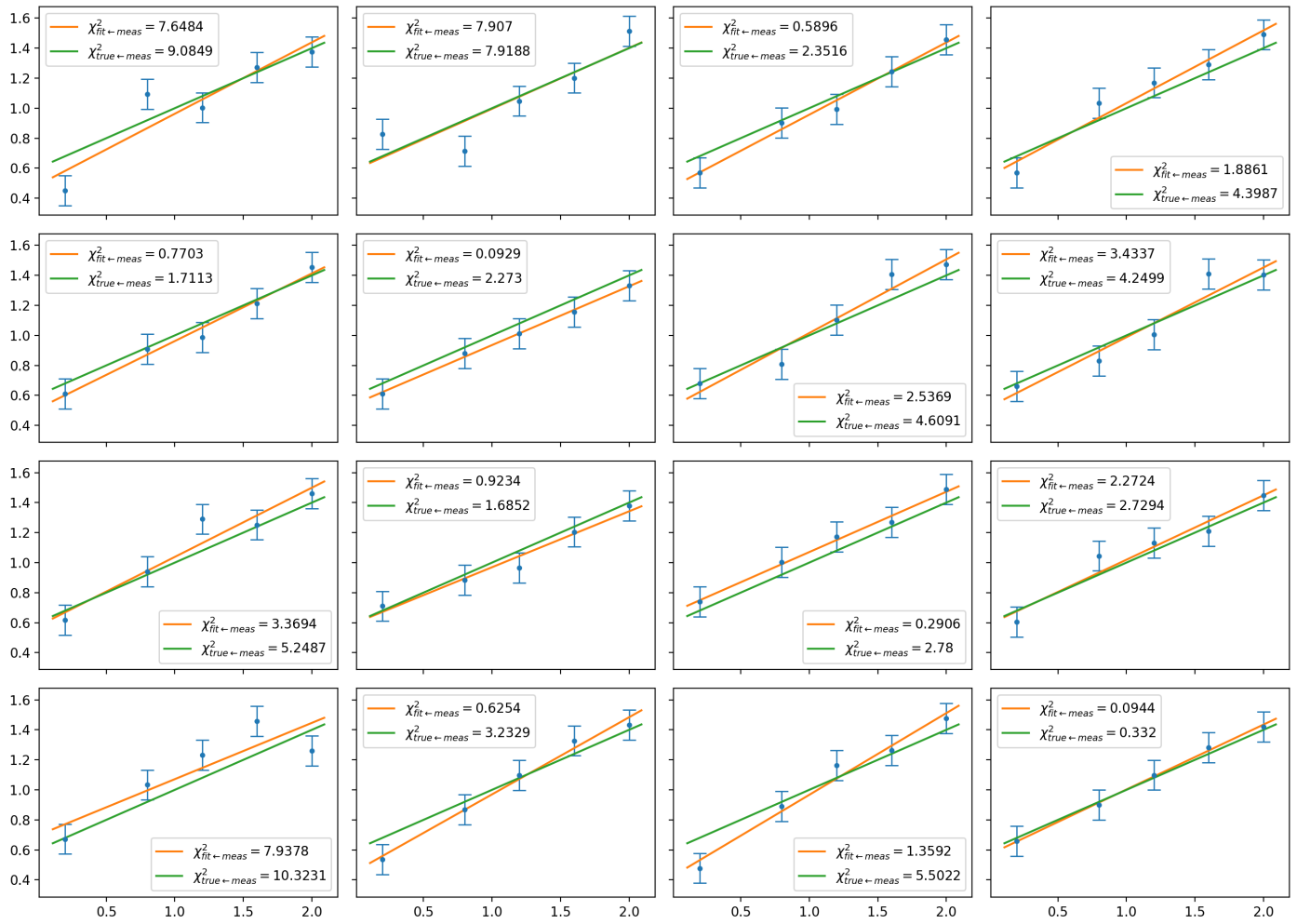

Figure 6: Two distributions of χ^2 values are obtained after 16 simulated experiments, each "experiment" uses 5 data-points.

Note that $\frac{\chi^2_{fit\leftarrow meas}}{DoF}$ is necessarily smaller than or equal to $\frac{\chi^2_{true\leftarrow meas}}{DoF}$ in each simulated experiment. This is because there may exist an f_{fit} which fits the data better than f_{true} (i.e. with a smaller value of root-sum-squared residuals). This is the first hint that $\frac{\chi^2_{true\leftarrow meas}}{DoF}$ and $\frac{\chi^2_{fit\leftarrow meas}}{DoF}$ should follow different distributions.

We can go even further and create 1000 of the graphs above, and plot the distribution of these χ^2's on a histogram.

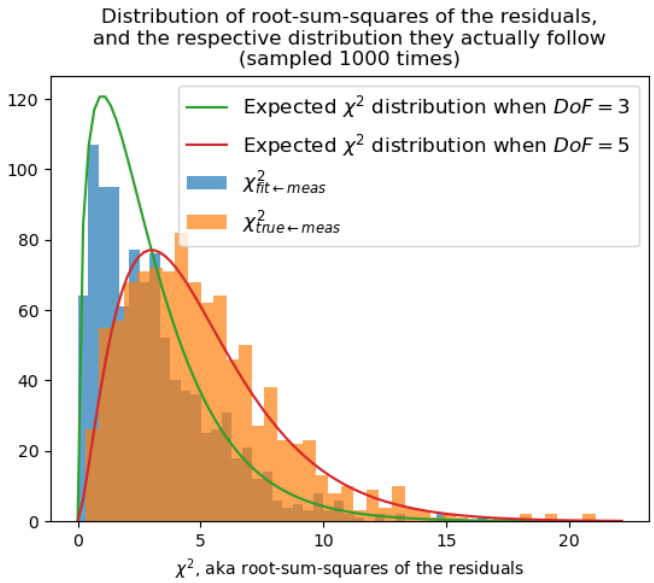

Figure 7: Distribution of χ^2 values in 1000 simulated experiments. 5 data points are generated per simulated experiment. Note that $\frac{\chi^2_{fit \leftarrow meas}}{DoF}$ and $\frac{\chi^2_{true \leftarrow meas}}{DoF}$ follows different distribution.

At this point it becomes clear how the probability distribution known as "χ^2 distribution" get its name. As we repeat simulated experiments and accumulate χ^2's values, they starts to follow respective χ^2 distributions.

Since 5 measurements are taken per simulated experiment, $\frac{\chi^2_{true \leftarrow meas}}{DoF}$ follows the distribution of χ^2 for 5 degrees of freedom. $\frac{\chi^2_{fit \leftarrow meas}}{DoF}$ is generated after fitting the five data-points to a function with two free parameters, reducing its DoF from five to three. Therefore it follows the χ^2 distribution for 3 degrees of freedom.

The situation above is generated with over-determined system, i.e. more data-points than there are free parameters. But in case of underdetermined unfolding, we have more free parameters in the model than there are data-points.

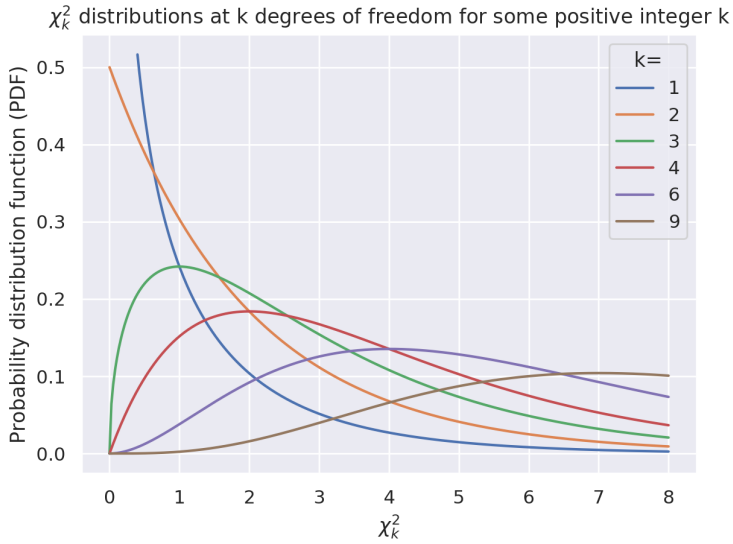

Figure 8: The χ^2 distributions at various positive degrees of freedoms.

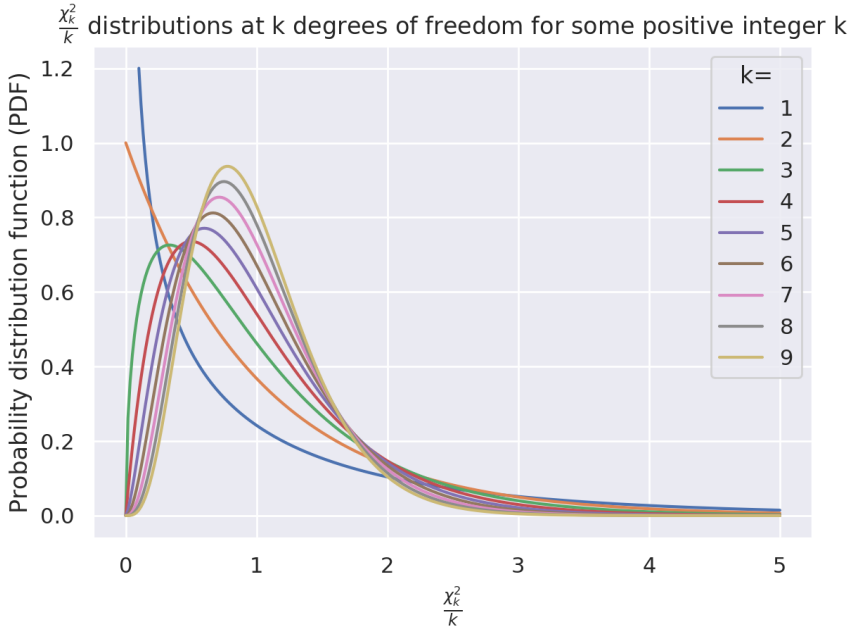

Figure 9: The PDF of $\frac{\chi^2_k}{k}$ for various k. k is used in place of DoF here. Note that both the denominator and the numerator are dependent on k.

For all positive integer values of k, the mean (first moment) of the χ^2_k distribution is k. This is the reason why we are usually told that $\frac{\chi^2_{fit\leftarrow meas}}{DoF}$ should equal 1.

(For your information, the median of the χ^2 distribution is $\approx k(1-\frac{2}{9k})^3$ and the mode is $max(k-2,0)$. The variance $\sigma^2 = 2k$, therefore as k increase, the standard deviation from the mean of $\frac{\chi^2_k}{k}$ decreases as $\sigma(\chi^2) = \frac{2}{\sqrt{k}}$.)

And as the number of degrees of freedoms increases, the $\frac{\chi^2_k}{k}$ distribution concentrates around 1. In other words, the more redundant measurement the experimenters makes, the more likely they will find their $\frac{\chi^2_{fit\leftarrow meas}}{DoF} \to 1$, with decreasing likelihood to deviate from 1.

But knowing the behaviour of $\frac{\chi^2_{fit\leftarrow meas}}{DoF} \to 1$ as $DoF \to \infty$ is not important to us. We want to know what happens when DoF decreases, and eventually becomes negative as the number of free parameters exceeds the number of measurements.

$$\chi^2_k(x) = \frac{1}{2^{k/2}\Gamma\left(k/2\right)} x^{k/2-1} \exp\left(-x/2\right) \tag{8}$$

We can use the equation above to infer what happens when $k \leq 0$. As $k \to k'$ for any negative integer k', $\frac{1}{\Gamma(k)} \to 0$, thus $\chi^2_k(x) = 0 \forall x \neq 0$. Since we have defined χ^2_k to be a probability distribution, integrateing the area under its curve must give unity, i.e.

$$\int_{x=0}^{x\to\infty} \chi^2_k(x) = 1$$

. Therefore,

$$\chi^2_k(x) = \delta(x) \qquad\qquad \forall k \in \mathbb{Z}^{\leq 0} \tag{9}$$

when k is a nonpositive integer, with mean, mode, median and variance all $= 0$.

In other words, the expected $\frac{\chi^2_{fit\leftarrow meas}}{DoF} = 1$ if $DoF > 0$, else $\frac{\chi^2_{fit\leftarrow meas}}{DoF} = 0$.

In light of this, I propose that we set $\chi^2_{fit\leftarrow meas} = max(m-n,0)$ for a dataset with m data-points fitted to a model with n degrees of freedoms, assuming that the model

is correct. Therefore we should set the value of $\chi^2_{fit\leftarrow meas}$ to a very small value for GRAVEL (since GRAVEL will never reach $\chi^2_{fit\leftarrow meas} = 0$ with finite computing time) and set $\chi^2_{fit\leftarrow meas} = 0$ for MAXED.

4 Conclusion

Previous users and creators of unfolding program MAXED and GRAVEL has the misconception that $\frac{\chi^2_{fit\leftarrow meas}}{DoF} = \frac{\chi^2_{true\leftarrow meas}}{DoF} = 1$, which is incorrect. The degrees of freedom in $\frac{\chi^2_{true\leftarrow meas}}{DoF} = m =$ number of data-points, while $\frac{\chi^2_{fit\leftarrow meas}}{DoF} = m - n =$ number of data-points - number of free parameters adjustable by the fitting algorithm.

The expectation value of $\chi^2_{fit\leftarrow meas} = max(0, m-n)$, as shown in section 3. Therefore in case of underdetermined unfolding when $m < n$, we should use $\chi^2_{fit\leftarrow meas} = 0$.

Appendices

A Neutron spectrum unfolding

Neutron spectrum unfolding using activation foils involves the following procedure:

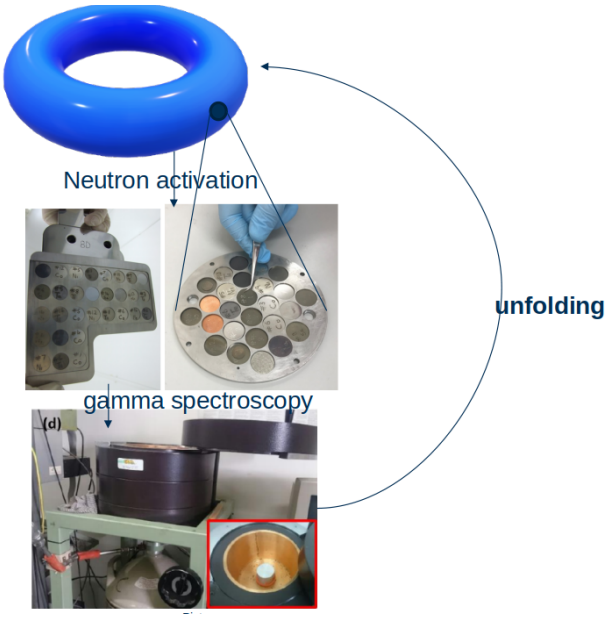

Figure 10: Unfolding procedure

1. Bathe a selection of foils in the unknown neutron spectrum, thus activating them;
2. Extract them from the test environment;
3. Measure the radionuclide populations inside each foil, by measuring the intensities of known gamma peaks;
4. Infer the neutron spectrum using an unfolding algorithm.

The algorithm in 4 requires the following inputs:

1. List of radionuclide populations $\boldsymbol{N}_{measured}$ measured at step 3, consisting of m types of radionuclides;
2. An *a priori* spectrum $\boldsymbol{\phi_{ap}}$, similar to the expected output, discretized into n bins;

3. The response matrix $\underline{\underline{\mathbf{R}}}$, an $m \times n$ matrix, which theoretically converts the spectrum into the radionuclide populations,

$$\boldsymbol{N} = \underline{\underline{\mathbf{R}}}\boldsymbol{\phi}_{true}$$

Note that the unfolding algorithm does not know about the true spectrum $\boldsymbol{\phi}_{true}$; the only information it has about the true-spectrum is $\boldsymbol{N}_{measured}$. Whenever χ^2 is mentioned below, it refers to a scalar value that measures the difference between two list of radionuclide populations $\boldsymbol{N}$'s : $\boldsymbol{N}_{test\ solution} = \underline{\underline{\mathbf{R}}}\boldsymbol{\phi}_{test\ solution}$ and $\boldsymbol{N}_{measured} = \underline{\underline{\mathbf{R}}}\boldsymbol{\phi}_{true}$.