1. Introduction: The ΔC statistic and the significance of nested model components

Similar to the the $\Delta\chi^2$ statistic for Gaussian data, the

$$\Delta C = C_{true} - C_{min} \qquad (1)$$

statistic is asymptotically distributed as $\chi^2(m)$, thanks to the Wilks theorem, where m is the number of free parameters. This is expected to be true for any number of counts-per-bin, therefore even in the limit of sparse data when neither C_{true} or C_{min} are individually χ^2-distributed (see Sect. 16.4 in Bonamente 2022).

It is useful to emphasize certain mathematical requirements to use such likelihood-ratio statistics as ΔC and $\Delta\chi^2$ to determine the significance of a model component(see Protassov et al 2002):

(a) the additional model component must be nested

(b) the null value of the model parameter(s) may not be at the boundaries of the allowed parameter space. (For example, you may not use ΔC for an absorption line component, if the normalization is not allowed to be positive too.)

Protassov R., van Dyk D. A., Connors A., Kashyap V. L., Siemiginowska, A.(2002). Statistics, Handle with Care: Detecting Multiple Model Components with the Likelihood Ratio Test. ApJ, 571, 545

Bonamente, M. Statistics and Analysis of Scientific Data, Springer, 3rd Ed. (2022) https://link.springer.com/book/10.1007/978-981-19-0365-6

2. Systematic errors and what to do when the goodness-of-fit statistic is not acceptable: the overdispersed chi-square distribution.

It is a common situation that the goodness of fit, in this case C_{min}, is not acceptable, yet the model generally follows the data without systematic trends. In such cases, it is possible to consider whether there are other sources of variance in the data that have not been considered.

For Gaussian data, $y_i \sim Gauss(\mu_i, \sigma^2_i)$, the traditional route is to identify additional sources of variance, σ^2_{sys} , and typically add these variances prior to the ML regression,

$$\sigma^2_{new,i} = \sigma^2_i + \sigma^2_{sys} \qquad (2)$$

For Poisson data, $y_i \sim Poiss(\mu_i)$, there is no direct way to provide additional variance, unlike in the case of Gaussian data. This is an intrinsic limitation of the Poisson regression. Motivated by this limitation, I have developed a new method to account for systematic errors in the Poisson ML regression (Bonamente 2023).

Bonamente, M. (2023). **Hypothesis testing with the Cash statistic for overdispersed Poisson count data.** MNRAS in press, https://arxiv.org/abs/2302.04011

The method is based on introducing an *intrinsic model variance* associated with the model itself, while retaining the Poisson distribution for the data. This means that the parent model mean

$$\mu_i \sim \text{Gauss}(\mu_{io}, \boldsymbol{\sigma^2_{int,i}}) \qquad (3)$$

is no longer a fixed (yet unknown) number, but it is a random variable. It is necessary to estimate this intrinsic variance from the data. This can be accomplished by setting a goal for the total variance of the C_{min} statistic as

$$\text{Var}(C_{min,sys}) = \text{Var}(C_{min}) + \boldsymbol{\tilde{\sigma}^2_c} \qquad (4)$$

where in the large-count limit $\text{Var}(C_{min}) = 2\,\nu$, and the additional term $\tilde{\sigma}^2_C$ is a design variance that ensures consistency with the data (for example, require consistency of C_{min} at a preset p-value). In practice, this means treating the fit statistic as the sum of two random variables,

$$C_{min,sys} = X + Y \quad \text{with}$$

$$X = C_{min} \sim \chi^2(\nu) \sim N(\nu, 2\nu)$$
$$Y \sim N(0, \tilde{\sigma}^2_C)$$

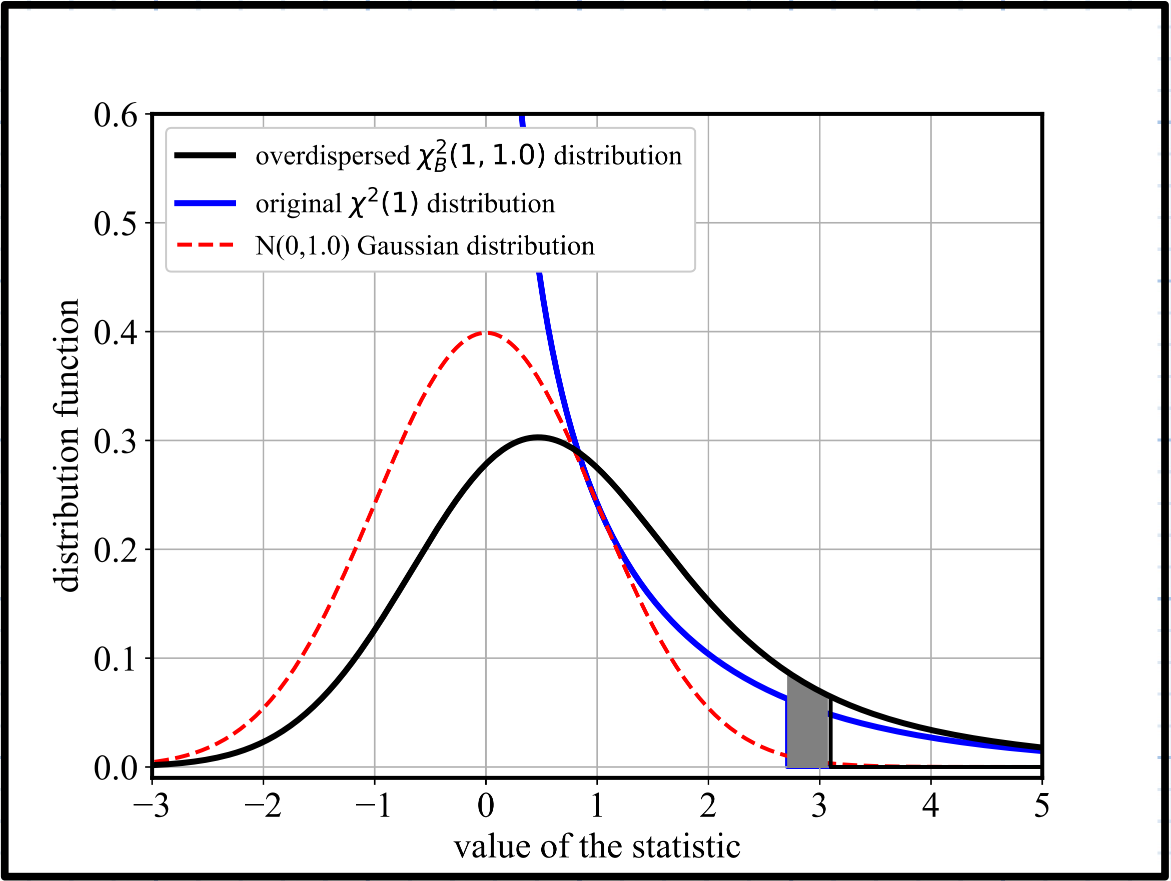
The $C_{min,sys}$ fit statistic is now distributed as an overdispersed $\chi_B^2(\nu, \boldsymbol{\tilde{\sigma}_c^2})$ distribution, which is the convolution of a $\chi^2(\nu)$ distribution with an $N(0, \boldsymbol{\tilde{\sigma}_c^2})$.

To estimate the intrinsic variance in the model, it can be shown that is possible to connect the two variances in (3) and (4) via

$$\boldsymbol{\sigma}_{\mathbf{int,i}}/\tilde{\mu}_i = (½)\ \tilde{\boldsymbol{\sigma}}_{\mathbf{C}}\ /\ (\Sigma\ y_i)^{1/2} \qquad (5)$$

This results in a simple estimate of the intrinsic model variance from the data, based on a *design variance* $\tilde{\sigma}^2_C$.

The ΔC fit statistic is also distributed as an overdispersed $\chi_B^{\,2}(\nu,\ \tilde{\sigma}_C^{\,2})$ distribution, where m is the number of additional parameters in the nested component


overdispersed $\chi^2_B(1, 1.0)$ distribution
original $\chi^2(1)$ distribution
N(0,1.0) Gaussian distribution
distribution function
value of the statistic