泊松计数数据最大似然回归中的系统误差：引入过度离散的 χ^2 分布

Massimiliano Bonamente[1]★1 阿拉巴马大学亨茨维尔分校物理与天文学系，亨茨维尔，阿拉巴马州 *35899*

已接受 XXX。收到日期：YYY；原始版本日期：ZZZ

arXiv:2302.04011v1 [astro-ph.IM] 8 Feb 2023

摘要

本文提出了一种新方法，用于估计计数数据最大似然回归中的系统误差。该方法尤其适用于 X 射线光谱分析：当泊松对数似然或 *Cash* 拟合优度统计量表明拟合不佳，且原因在于数据过度离散时，便可采用此方法。泊松数据中的过度离散被视为内禀模型方差，可利用最大似然 $C_{\rm min}$ 统计量，根据最佳拟合模型进行估计。本文还研究了此类系统误差对 ΔC 似然比统计量的影响；该统计量可用于检验泊松计数数据回归中是否存在嵌套模型分量。本文还引入了一种过度离散 χ^2 分布，它由一个用于描述通常的 ΔC 统计量的 χ^2 分布与一个用于描述数据中过度离散的零均值高斯分布卷积而成。本文建议在存在系统误差时，将该分布作为 ΔC 统计量的首选分布。本文将所述方法应用于类星体 1ES 1553+113的 *XMM–Newton* 数据；这些数据曾用于探测来自前景温热星系际介质（WHIM）的吸收线。本案例研究展示了如何根据数据估计系统误差，以及系统误差对使用 ΔC 统计量探测嵌套分量（例如吸收线）的影响。

关键词：方法：统计；方法：数据分析

1 引言

首批专用于探测太阳系以外X射线的卫星发射升空，具体包括 *Uhuru*卫星 (Giacconi等，1971）以及 *HEAO* 任务（Rothschild等，1979），这标志着X射线天文学领域真正拉开了帷幕。很快，人们便意识到，这些仪器提供的新数据通常以计数或光子数随时间和能量变化的形式呈现，要对其进行恰当的分析和解释，就需要新的统计工具。这些新数据推动了计数数据统计研究的多项进展。

解读早期 X 射线数据所需的工具之一，是将数十年前由 R.A. Fisher 提出（例如 Fisher1922、1934）的最大似然法应用于这些本质上服从泊松分布的整数计数数据。对于正态分布数据，最大似然法会导出 χ^2 统计量，其数学性质相对简单，长期以来一直为统计学家所用（例如 Greenwood & Nikulin 1996）。对于泊松数据，X 射线天文学家 W. Cash 首先证明，可以利用最大似然法导出另一种统计量，该统计量渐近服从 χ^2 分布。如今，这种统计量通常称为 *Cash* 统计量，或 *C*–统计量（Cash1976、1979），之后 Baker 与 Cousins（1984）等人对其作了进一步发展。在统计学的其他领域，等价的基于泊松分布的对数似然函数称为泊松对数似然的偏差（例如 McCullagh & Nelder1989；Cameron & Trivedi2013），或称为 *G* 平方 统计量（Bishopet al. 1975）。

在 X 射线天文学及相关领域中，始终将 χ^2 统计量用于整数计数数据，其依据是：已知当计数数量很大时，泊松分布可由高斯分布很好地近似。[1] 然而，如今已明确，即使在大计数极限下，对于含有自由参数的模型，若将 χ^2 统计量用于泊松数据的参数估计，最小化该统计量仍会导致有偏结果（例如 Humphrey et al.2009；Bonamente 2020）。这种偏差源于 χ^2 统计量需要对数据方差进行估计（即依据 修正的 χ^2 最小化方法，该方法由 Cramer1946，第 424–434 页所述），而以计数数量作为方差估计，会使回归过程中低计数数据点获得较大的权重。

另一方面，*C*统计量不存在这种偏差，因此对于X射线天文学及相关领域中常见的大多数计数数据，它应被视为首选统计量。事实上，主要的X射线拟合软件包（XSPEC和SPEX）现在都提供了使用任一统计量的选项（Arnaud1996；Kaastra 等，1996），而且SPEX 还会在假设数据服从最佳拟合模型的前提下，根据 *C*的近似方法，估算预期的统计量值Kaastra (2017）。尽管如此， *C*统计量在光谱拟合及其他类似应用中的使用仍受到其数学处理相对复杂的阻碍。例如，即使在线性回归这一简单情形下，对泊松数据进行拟合也不像使用 χ^2统计量那样具有简单的解析解，不过最近有研究表明，它存在一种半解析

★ 电子邮件：bonamem@uah.edu（MB）

[1] 例如，Bonamente(2022年）第3章展示了这一近似。

© 2022年 作者

解，可以轻松地通过数值方法实现（Bonamente 与 Spence（2022））。

本文讨论了使用 C–统计量拟合整数计数数据时的一个突出问题，即如何测量超出通常泊松不确定性的可能系统误差。不同于使用 χ^2 统计量拟合高斯数据的情况——在该情况下，可以通过传统的误差平方和方法立即处理系统误差——基于泊松分布的 C–统计量没有这种直接的修改方式。相反，可以采用内在模型方差来处理这一问题，该方差反映最佳拟合模型中存在的不确定性，而不是数据中的额外误差。

本文的结构如下：第 2 节介绍所考虑数据模型的统计性质，第 3 节介绍使用 *Cash* 统计量处理系统误差的新方法，其中包括引入过度离散的 χ^2 分布。第 4 节通过一个案例研究，利用 *XMM–Newton* 对源 1ES 1553+113 的数据来阐释该方法；该源最近由 Nicastro等人 (2018年) 和 Spence 等人 (2023年) 进行了研究；第 5 节给出我们的结论。

2 计数数据的数据模型

本文考虑的数据模型包含 N 次相互独立的测量，其形式如下：

$$(x_i, y_i) \text{ for } i = 1, \ldots, N, \tag{1}$$

其中，x_i 是一个假定已精确获知的自变量（例如光子的波长或能量），而 y_i 是计数的整数（例如给定箱中的计数或光子数）。这是天文学 X 射线光谱的标准模型，但也适用于其他领域的多种数据，通常称为横截面数据。这些数据采用参数模型 $y = f(x)$进行拟合，其中函数 $f(x)$包含 m 个自由参数，需根据最大似然方法确定。

2.1 标准泊松模型

用于估计模型参数的最简单假设是，因变量 y_i 服从泊松变量的分布，

$$y_i \sim \text{Poiss}(\mu_i), \tag{2}$$

其母体均值 $\mu_i = \mu_i(x_i)$未知，是若干可调参数 θ_j, $j = 1, \ldots, m$ 以及自变量的函数。X射线天文学中的常见情况包括幂律模型，或由少数几个参数（如温度和化学丰度）决定的热模型，以及包含吸收线等的更复杂模型。在所有情况下，该数据模型的关键假设都是：给定箱中检测到的计数服从泊松分布；这里的箱即 x 变量中的一个取值范围，由特征值 x_i表示。

这些数据的参数估计采用 R.A. Fisher 开创的常规最大似然方法（例如，参见Fisher（1934）），在本例中利用了数据服从泊松分布以及各测量值相互独立这一性质。在这些假设下，为获得最佳拟合参数及其协方差矩阵而需要最小化的相关统计量是 *Cash* 统计量，或称 C–stat，定义为

$$C = 2\sum_{i=1}^{N}\left(\mu_i - y_i + y_i \ln\left(\frac{y_i}{\mu_i}\right)\right). \tag{3}$$

该统计量由 Cash （1976、1979）和 Baker & Cousins （1984）提出，并且具有一个便利的性质：当每箱计数较多时，它渐近服从 χ^2 分布，

$$C \overset{a}{\sim} \chi^2(N) \tag{4}$$

当模型完全指定时（例如，Kaastra 2017；Bona- mente 2020年，2022年）。对于具有 m 个可调参数的模型，使用泊松统计量进行最大似然回归会得到 Cmin 拟合优度统计量，

$$C_{\min} = 2\sum_{i=1}^{N}\left(\hat{\mu}_i - y_i + y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right)\right), \tag{5}$$

其中 $\hat{}\mu_i$是在最佳拟合参数 $\hat{}\,\theta_j$处计算的。对于每箱计数较多的情况，其渐近分布已由McCullagh(1986)证明近似为

$$C_{\min} \overset{a}{\sim} \chi^2(N - m). \tag{6}$$

这一渐近性质有利于参数估计和假设/模型检验（例如，参见 Kaastra 2017），并类似于适用于 $\chi^2_{\min}$ 统计量的类似结果（Cramer 1946）。

2.2 过度离散泊松模型

泊松分布施加的最大限制是方差等于均值，

$$\text{Var}(y_i) = \mu_i, \tag{7}$$

其中 μ_i 是数据生成过程中 i箱的母体均值。在包括天文学在内的各个科学领域的数据集中，一个常见现象是，测量数据相对于关于（7）的简单假设呈现过度离散，例如参见 Cameron 与 Trivedi （2013）第 3 章中的讨论。数据中这种大于μ_i 方差的原因包括模型不充分——即多变量回归可能需要更多解释变量，或需要对模型进行更好的参数化——或者存在模型内在变异性，使得每个独立测量值 y_i 具有大于泊松分布的方差。

无论起因如何，都可以通过一个函数对过度离散进行建模，

$$\text{Var}(y_i) = \omega_i = \omega(\mu_i, \alpha) \tag{8}$$

该函数以参数 α的形式引入一个额外的自由度。虽然原则上该函数可以采用任何形式，但由 Cameron 与 Trivedi（2013）提出的两种便捷参数化形式为

$$\begin{cases} \omega_i = (1+\alpha)\mu_i & \text{(NB1 model)} \\ \omega_i = \mu_i + \alpha\mu_i^2 & \text{(NB2 model)} \end{cases} \tag{9}$$

这些模型名称源自 Cameron 与 Trivedi （1986）在负二项回归中的应用。本文并未采用负二项分布；负二项分布常见于另一种针对计数数据的回归方法中（例如参见 Hilbe 2011, 2014 $\alpha = 0$），但这种参数化仍然适用。当 时，该模型返回通常的泊松回归。当 $\alpha > 0$时，另一方面，数据不再服从泊松分布。原则上，也可能有 $\alpha < 0$，这对应于 欠离散 数据。不过，这种情况在实践中较少出现，因此通常使用（9）来表示 $\alpha > 0$。广义线性模型（GLM）文献中常见的 NB1 模型等价参数化方式（例如 McCullagh 与 Nelder 1989）是设定

$$\phi = (1+\alpha) \tag{10}$$

其中 $\phi > 1$ 表示过度离散数据。

允许数据存在过度离散，意味着对（3）所描述统计量的最小化不再代表最大似然条件，因为数据不再服从泊松分布。幸运的是，在某些情况下，通常的泊松回归及其通过最小化（$\hat{\theta}_j$）得到的最佳拟合参数，仍适用于过度离散泊松模型。具体而言，Gourieroux 等人(1984b证明，对于包含泊松分布在内的一族指数分布，即使分布设定有误，只要均值设定正确，通过通常的似然方法估计得到的最佳拟合参数仍具有一致性。在这种情况下，这种估计方法称为准最大似然。[2]

实践中，这些结果使我们能够在假定泊松均值正确的前提下，继续使用通常的泊松似然。接下来的问题是估计（ϕ 10）中定义的过度离散参数。泊松准最大似然方法得到的最佳拟合参数渐近服从正态分布，其方差显然取决于 ϕ（例如参见 Cameron 与 Trivedi 2013）。对于数据分析人员而言，最关心的是计数数据中过度离散程度的估计值。针对（ϕ 和α）中两种数据方差模型的标准估计量为

$$\begin{cases} \hat{\phi}_{\mathrm{NB1}} = \dfrac{1}{N-m}\displaystyle\sum_{i=1}^{N}\dfrac{(y_i-\hat{\mu}_i)^2}{\hat{\mu}_i} \\ \hat{\alpha}_{\mathrm{NB2}} = \dfrac{1}{N-m}\displaystyle\sum_{i=1}^{N}\dfrac{(y_i-\hat{\mu}_i)^2-\hat{\mu}_i}{\hat{\mu}_i{}^2}. \end{cases} \tag{11}$$

这些估计量的合理性依据见Cameron 与 Trivedi (2013); Gourieroux 等人 (1984b,a)，不过也有其他估计量可用（例如，参见Cameron 与 Trivedi 1986；Dean 与 Law- less 1989；Dean 1992）。根据式（11）中（9）的两种替代参数化方式，这两个方程给出了计数数据过度离散程度的点估计。这些估计的不确定性可通过标准误差传播法（也称为*delta* 方法）计算。数值 $\hat{\phi}_{\mathrm{NB1}} > 1$ 或 $\hat{\alpha}_{\mathrm{NB2}} > 0$ 表示数据存在过度离散。

2.3 高斯模型

对于 X 射线光谱分析，即使数据服从泊松分布，χ^2 统计量仍然广泛使用。该 χ^2 拟合优度统计量源于采用另一种数据模型，在这种模型中，与（1）相同的数据假定服从一个母体分布

$$y_i \sim \mathrm{Gauss}(\mu_i, \sigma_i^2) \tag{12}$$

其中，除了通常的母体均值 μ_i 外，还需要母体方差 σ_i^2。χ^2 拟合统计量广泛用于整数计数数据，主要是因为它易于使用和解释。其中包括 H. Cramer 提出的自由度缩减定理（Cramer 1946），该定理给出了 χ^2min 统计量的分布形式：

$$\chi^2_{\min} \sim \chi^2(N-m), \tag{13}$$

其中，m 是模型中的自由参数数量，N 是通常的独立高斯分布数据点数量。该结果在相当一般的条件下成立，而这些条件通常都能满足 X 射线天文数据的要求。[3]

式（12）中的母体方差通常通过数据方差近似得到，通常取箱中的计数数量，即$\hat{\sigma}_i^2 = n_i$（参见例如 Bevington& Robinson2003；Bonamente 2022）。这种近似遵循Cramer （1946）修正的最小 χ^2 准则，即渐近分布（13）成立的前提是固定的数据方差。天文学统计学文献已经记载，即使在大计数极限下，此时泊松分布实际上可以很好地用正态分布近似，使用 χ^2 分布仍会导致有偏结果；相比之下，使用 C–stat（例如 Humphrey et al. 2009；Bonamente 2020）。这主要是因为在数据模型（12）中，方差被近似为一个对于计数率较低的箱而言更小的数值，从而在回归中被不恰当地赋予更大的权重。

2.4 使用其他分布进行回归

另一种可能是，假设数据服从其他分布，例如*Conway–Maxwell–Poisson* 分布（例如Conway 与 Maxwell，1962；Shmueli 等，2005；Sellers 与 Shmueli，2010）、广义泊松分布（例如Consul 与 Jain，1973；Famoye，1993、2011、2014），或负二项分布（例如Hilbe）。这些分布引入了附加参数，可方便地对数据方差建模，从而允许数据呈过度离散或不足离散。不过，这些调整会降低使用和解释的便利性，其中包括需要寻找替代的拟合优度统计量，以取代基于泊松分布的偏差或 C–统计量。

鉴于泊松分布相对易于使用，而且人们认为大多数数据集都源自速率固定为 μ_i 的泊松过程，因此在分析天文数据时，通常最好保留泊松分布。由此，本文继续假设数据服从泊松分布，并提出一种能够处理过度离散的新方法。

3 泊松分布计数数据最大似然回归中的系统误差

在X射线天文学及相关领域，通常会将测量 y_i 数据时的不确定性分为两种。第一种称为统计误差，通常指光子计数实验带来的不确定性（例如计数的平方根）。其他所有不确定性来源通常称为系统误差，尽管它们也可能像光子计数过程本身一样，本质上是随机误差。后一类误差的例子包括光子收集过程其他环节所带来的不确定性，例如仪器校准方面的不确定性，或分析其他方面的不确定性。第一类误差与第二类误差的主要区别通常在于：第一类误差是收集过程固有的（即源自 y_i 变量所依据的泊松分布），而第二类误差往往可以通过谨慎地整理和分析数据来减小。尽管这种区分在某种程度上有些武断，本文后文仍将沿用这一划分。

通常，针对高斯数据中存在其他系统误差来源的问题，会修改用于 χ^2 统计量的

[2] 该方法在统计学文献中也称为伪最大似然 。

[3] 该定理也在Bonamente (2022年)的第12节中讨论。

（12）中的方差，直到拟合统计量 χ^2 变得“可接受”。这种方法利用了方差 σ_i^2确实属于数据模型这一事实，见（1）和（12）。例如，Bonamente (2022年）第17章，或 Bevington 与 Robinson (2003）的教材中，都对这些方法进行了介绍。在继续讨论系统误差之前，有必要强调“可接受”一词在假设检验中的含义。假设检验是一种过程，其结果要么是在给定的显著性水平下拒绝零假设，要么只是未能拒绝零假设（例如，参见美国统计学会近期关于 p 值与假设检验的声明，Wasserstein 与 Lazar 2016）。不过，如果理解为不能合理地拒绝某个假设（例如，在3或 5 σ 置信水平下），那么称该假设为“可接受”仍是合理的；但这并不意味着该假设可以作为数据唯一可能的模型而被最终

对于X射线天文学中常见的那类泊松数据，即遵循假设（2）以及由此得出的 C–统计量拟合统计量（3），无法采用这种直接途径。原因在于，泊松数据点的方差等于其均值，因此不能像处理高斯分布及其对应的 χ^2 统计量（根据（12））那样，分别指定二者。统计学文献中已有一种成熟且方便的替代方法，即第2.2节讨论的过度离散泊松回归。这种准最大似然方法既能保留基于泊松分布的最大似然最佳拟合统计量（即偏差或 C–统计量），又能估计过度离散程度。

在这一统计框架下，本文提出了一种新方法，用以处理计数数据中的过度离散及其对*Cash* 统计量的影响。该方法将过度离散解释为内禀模型方差，并以适当的零均值正态分布对其建模。基于这一假设，将这一额外的不确定性来源与通常基于泊松分布的 C–统计量的分布相结合，从而得到一种修正后的分布，可用于对过度离散泊松数据进行回归假设检验。本节余下部分将介绍这一方法。

3.1 使用*Cash*统计量测量内禀模型方差

对于拟合到参数模型的泊松分布数据（如X射线天文学中的常见做法），通过改变对数据和模型在回归中所扮演角色的看法，可以估计系统误差。具体而言，与其要求数据点 y_i 的方差大于根据其泊松分布估计的方差，不如合理地假设模型具有内禀方差。在实践中，这意味着假设给定数据点 y_i 的母模型 μ_i 并非如其希腊字母表示法所暗示的那样是一个固定（且未知）的数值，而是服从一个具有给定方差的分布，或者

$$\mu_i \sim \mathrm{Gauss}(\mu_{i0}, \sigma^2_{\mathrm{int},i}), \tag{14}$$

其中，μ_{i0} 是母体值，σ^2int, i 是内禀方差，i 表示箱。这一假设类似于正态数据存在系统误差时所采用的方法——即数据方差增大时的情形——但额外的方差来源现在与模型相关，而不是与数据相关。当然，内禀方差事先并未知晓，而必须根据数据进行估计。这类似于过度离散泊松模型的情形（见第 2.2 节），该模型需要对过度离散参数进行估计 ϕ 或 α。

要估计模型中的内禀方差，必须假设最佳拟合模型能够合理描述数据。用统计学术语来说，这一零假设可重新表述为：数据与从母模型中抽样得到的结果相容；这种相容性根据所选拟合统计量并在给定置信水平下进行衡量。鉴于泊松数据所选用的拟合统计量是 Cmin 根据（5），模型的可接受性可以表述为：在零假设下，测得的 Cmin 与其母体分布在统计上相符。当每箱计数较大时，无论自由度有多少，McCullagh (1986)表明，该统计量趋近于 χ^2 分布，即

$$C_{\min} \overset{a}{\sim} \chi^2(\nu) \tag{15}$$

其中 $\nu = N - m$ 是自由度的数量。McCullagh（1986年）给出的证明假设模型是线性的（更准确地说，是对数线性的），并基于条件累积量的计算（McCullagh 1984年），因此，要将其外推至非线性模型（例如 Spence 等人（2023年）所使用的模型），还需要进一步的理论论证。据作者所知，迄今尚未获得针对一般非线性模型的（15）的一般证明，但下文将其假定为一般非线性模型的一阶近似。

另一个重要的注意事项是，在低计数区间，（15）通常不适用，因此需要考虑其他因素。关于低计数和少箱区间中 C min 分布的讨论，也见于 McCullagh （1986年），以及 Bonamente （2020年）和 Kaastra （2017）。从定性角度来说，（15）可由大计数极限下泊松分布收敛于正态分布来理解。为清晰起见，本文所给结果的适用范围将在第 3.3 节中总结。

3.1.1 存在系统误差时 Cmin 的渐近分布

现在的问题是，确定内禀模型方差（14）对 Cmin 渐近分布的影响；该渐近分布是在泊松测量的假设下得到的。为此，我们作出以下近似，这些近似对应于常见的实验条件：

(a) 当自由度较大时，$\chi^2(\nu)$ 分布等价于 $N(\nu, 2\nu)$ 正态分布，尤其对于不受 χ^2 分布正定性影响的较大统计量值而言。这与式（15）一致，适用于较大的 ν。较小 ν 的情况将在第3.2节讨论，该节描述内禀方差对 ΔC 统计量的影响。

(b) 我们假设内禀方差

$$\sigma_{\mathrm{int},i} \ll \mu_{i0}/\text{较小},$$

这意味着系统误差的相对幅度较小。因此，该方法采用简单的误差传播法或德尔塔法，估计由假设（Cmin （14）引入的统计量附加方差。

(c) 根据前述两个假设，按照（14），系统误差（或者更准确地说，内禀方差）会导致一个 $Z = C$min 统计量，可将其视为两个独立变量 $Z = X + Y$ 之和，其中

$$\begin{cases} X \sim \chi^2(\nu) \overset{a}{\sim} N(\nu, 2\nu) \\ Y \sim N(0, \hat{\sigma}_C^2), \end{cases} \tag{16}$$

与 $\nu = N - m$。Y 的零均值意味着系统误差

不会为统计量带来净均值或偏差，而 $\hat{\sigma}_C^2$ 表示与系统误差相关的 Cmin 附加方差项，该项将根据数据进行估计。这可确保修改后的 Cmin 统计量分布仍近似服从正态分布，且方差高于标准泊松回归。

3.1.2 内禀方差估计

上一节所述模型引入了两个量：内禀模型方差 $\sigma^2_{\mathrm{int},\,i}$ 以及 Cmin统计量的“设计”方差$\hat{\sigma}_C^2$，后者是零假设可接受所必需的。本节将讨论它们的估计。

对于给定的应用，第一步是估计拟合统计量所需的内禀方差，记为

$$\mathrm{Var}_i(C_{\min}) = \hat{\sigma}_C^2. \tag{17}$$

（17）的含义是，Cmin统计量需要增加一定量$\hat{\sigma}_C^2$的附加方差，才能使测得的统计量与其在零假设下的期望相符。根据（16），Cmin统计量的总方差为

$$\mathrm{Var}(C_{\min}) = 2\nu + \hat{\sigma}_C^2, \tag{18}$$

假设 Cmin统计量仍服从正态分布，并处于每箱计数较大且箱数较多的渐近极限下。

现在的问题转为评估（17）左侧的项，以及它与所求内禀模型方差 $\sigma^2_{\mathrm{int},\,i}$ 的关系见（14）。根据其定义，内禀模型方差可表示为

$$\begin{aligned}\mathrm{Var}_i(C_{\min}) &= \mathrm{E}\left[(C_{\min}(\mu_i) - \mathrm{E}\left[C_{\min}(\mu_i)\right])^2\right] \\ &\simeq \mathrm{E}\left[(C_{\min}(\mu_i) - C_{\min}(\hat{\mu}_i))^2\right],\end{aligned} \tag{19}$$

并采用 $\mathrm{E}\left[C\mathrm{min}(\mu_i)\right] \simeq C\mathrm{min}(\hat{\ }\mu_i)$ 的近似。式（19）的含义是，μ_i 是一个所考虑的随机变量，其分布遵循（14），而 $\hat{\ }\mu_i$ 是通常的最大似然估计。换言之，该方差是相对于分布（14）计算的，而不是相对于根据（2）得到的数据分布计算的——这就是方差（17）和（19）记号中使用 i 下标的原因。事实上，数据 y_i 所遵循的泊松分布决定了 Cmin 的方差，该方差此前已经计算过，即采用（15）中的大计数近似，或依据 Kaastra（2017）针对低计数区间提出的方法计算。

为计算（19），使用 C–统计量，其定义见（3），

$$C_{\min}(\mu_i) - \mathrm{E}\left[C_{\min}\right] = 2\left(\sum_{i=1}^{N}(\hat{\mu}_i - \mu_i) + y_i \ln\left(\frac{\mu_i}{\hat{\mu}_i}\right)\right)$$

$$\simeq 2\left(\sum_{i=1}^{N}(\hat{\mu}_i - \mu_i) + y_i\left(\frac{\mu_i - \hat{\mu}_i}{\hat{\mu}_i}\right)\right) = 2\left(\sum_{i=1}^{N}(\mu_i - \hat{\mu}_i)\left(\frac{y_i}{\hat{\mu}_i} - 1\right)\right)$$

其中假设内禀误差的分数值较小，$\sigma_{\mathrm{int},\,i}/\mu_i$，因此在对数的泰勒级数展开中仅保留了一阶项。采用这一近似，(19) 得出

$$\mathrm{Var}_i(C_{\min}) \simeq \mathrm{E}\left[\left(2\sum_{i=1}^{N}(\mu_i - \hat{\mu}_i)\underbrace{\left(\frac{y_i}{\hat{\mu}_i} - 1\right)}_{(a)}\right)^2\right], \tag{20}$$

其中，方括号标出的项（a）与 μ_i 无关。此外，将和的平方展开后，在假设 μ_i 项彼此不相关的情况下，交叉乘积项的期望值为零。因此可得出这一简单结果：

$$\mathrm{Var}_i(C_{\min}) = \hat{\sigma}_C^2 \simeq 4\left(\sum_{i=1}^{N}\sigma^2_{\mathrm{int},i}\left(\frac{y_i}{\hat{\mu}_i} - 1\right)^2\right), \tag{21}$$

模型的内禀方差为

$$\sigma^2_{\mathrm{int},i} = \mathrm{E}[(\mu_i - \hat{\mu}_i)^2].$$

（21）的右侧项可进一步近似，所依据的假设是 $y_i \sim$ 泊松分布（$\hat{\ }\mu_i$），即平均而言有

$$(\hat{\mu}_i - y_i)^2 \simeq \hat{\mu}_i \simeq y_i.$$

进一步假设所有数据点的内禀相对误差都相同，即 $\sigma\mathrm{int},\,i/\hat{\ }\mu_i$ 为常数，可得到最终结果

$$\frac{\sigma_{\mathrm{int},i}}{\hat{\mu}_i} \simeq \frac{1}{2}\frac{\hat{\sigma}_C}{\sqrt{\sum_{i=1}^{N} y_i}}. \tag{22}$$

方程22 是所需的结果，即该近似将模型的内禀散布相对值（$\sigma\mathrm{int},\,i/\hat{\ }\mu_i$）与所需的设计值 $\hat{\sigma}_C^2$ 联系起来，从而在零假设下确保测得的 Cmin与其母体分布保持一致。[4] 该近似提供了一种简单方法，用于估计模型所需的平均相对内禀方差或百分比内禀方差，以确保拟合可接受。

3.2 使用 ΔC 统计量对嵌套分量进行假设检验时的内禀模型方差

根据（22）从 Cmin统计量估算出的内禀模型方差，使得零假设下的期望统计量与测量值一致。事实上，估算方法最初正是从要求满足（17）这一条件出发，旨在实现这种一致性。第4节将通过详细的案例研究，介绍实际中如何进行这种估算。换言之，内禀模型方差假定拟合是可接受的，因此使用 Cmin进行假设检验已不再有意义。

另一个 Cmin统计量的常见用途（χ^2min统计量也有此用途），是用该统计量

$$\Delta C = C - C_{\min} \tag{23}$$

来检验嵌套模型分量的显著性。该方法最初由Lampton 等人 (1976)针对$\Delta\chi^2$ 统计量提出，其依据是Cramer (1946)中提出的理论。关于 ΔC 统计量的介绍见Bonamente (2022年)第16章，其中主要结果为

$$\Delta C \overset{a}{\sim} \chi^2(l) \tag{24}$$

在测量值数量较大的渐近极限下，且嵌套分量有 l 个感兴趣（或自由）参数时。

ΔC 统计量的用法是计算包含嵌套分量的完整模型的 C最小拟合统计量，以及

[4] 请注意，在这个方差估计中，右侧项 $(y_i/\mu_i - 1)$ 被视为相对于由 (14) 定义的内禀方差而言为常数。根据数据的泊松分布，分别计算 $C_{\min}$ 相对于 y_i 和 $\hat{\ }\mu_i(y_i)$ 的方差，作为 $C_{\min}$ 的“恰当”方差，正如本文前文所述。

计算 C 统计量时，将嵌套模型分量置零（称为约化模型），因此显然 $C \geq C_{\rm min}$ 且 $\Delta C \geq 0$。在这种情况下，最佳拟合模型 $\hat{}\mu_i$ 有 l 个与其嵌套分量相关联的附加参数；当嵌套分量置零时，模型值为 $\mu_i^{'}$。例如，Spence 等人（2023年）表 5 和表 6 中的数据便属于这种情况，其中ΔC 值对应于将一个谱线分量设为零；相较于描述明亮 X 射线类星体连续发射的模型，该分量有 l=1 个或 2 个附加自由参数。这些数据将在第 4 节的案例研究中作更详细的介绍和分析。

作为似然比统计量，ΔC 必须满足若干条件，才能具有渐近分布（24）。Protassov 等人(2002）便于描述了这些条件，其中包括：嵌套分量的零值不得位于允许的参数空间边界上。

3.2.1 ΔC 统计量的内禀方差

从定性角度来说，如果嵌套模型分量的贡献处于根据（22）估算的内禀误差范围内，那么人们会认为无法判断是否需要该嵌套分量。例如，对于连续谱的 10% 以内、类似吸收线的特征，在存在 ±10% 内禀误差时，预计无法对其作出确切检测。

为了定量研究这一问题，考虑约化模型或基线模型 $\mu_i^{'}$ 与最佳拟合的完整模型之间的偏差，

$$\delta_i = \mu_i^{'} - \hat{\mu}_i. \tag{25}$$

该方程仅适用于嵌套模型分量显著的少数几个箱，例如 $i = 1, \ldots\ldots, n$，通常满足 $n \ll N$，而 $\delta_i = 0$ 则适用于所有其他箱。这种情况对应于受限波长范围内、叠加在基线连续谱模型上的嵌套吸收线或发射线型分量所形成的小尺度起伏，在吸收线的探测中经常出现（例如 Nicastro等（2018）；Spence 等人（2023年））。在（25）中，假设偏差的分布

$$\delta_i \sim N(0, \sigma^2_{{\rm int},i})$$

假设基线模型 $\mu_i^{'}$ 是一个随机变量，且其内禀方差与完整模型 μ_i相同，见（14）。其均值为零也代表零假设，即不存在嵌套分量。

因此，ΔC 统计量可近似表示为

$$\Delta C \simeq 2 \sum_{i=1}^{n} \delta_i \left(1 - \frac{y_i}{\hat{\mu}_i}\right),$$

沿用第 3.1.2 节中内禀方差较小的假设。此外，右侧括号中的项与内禀方差无关，这与（$C_{\rm min}$ （20）的方差相同。因此，ΔC 中与内禀模型方差相关的方差可近似为

$${\rm Var}_i(\Delta C) = \hat{\sigma}^2_{\Delta C} \simeq 4 \left(\frac{\sigma_{{\rm int},i}}{\hat{\mu}_i}\right)^2 \sum_{i=1}^{n} \hat{\mu}_i \simeq 4 \left(\frac{\sigma_{{\rm int},i}}{\hat{\mu}_i}\right)^2 \sum_{i=1}^{n} y_i. \tag{26}$$

这个结果与 ${\rm Var}_i(C_{\rm min})$ 所得的方程相同，22)，但求和仅涉及少数几个箱（$n \ll N$），在这些箱中嵌套模型会产生影响。方程26 给出了所求的 ΔC内禀方差，计算时使用了式（22）中的数据内禀方差。

3.2.2 存在内禀方差时 ΔC 统计量的分布

现在，分析转向考察在存在内禀模型方差的情况下，ΔC 在小 ν 极限下的分布问题。本节中，ν 表示嵌套分量中的自由参数数量，也就是相应渐近 χ^2 分布的自由度；而在上一节中，它表示整个回归的自由度。

假设测量值数量较大，ΔC 统计量的抽样分布近似服从 $\chi^2(\nu)$分布，其均值为 ν，方差为 2ν。对于嵌套模型分量的 ΔC 统计量，通常采用的自由参数数量较少，例如 $\nu = l = 1$ 或 2（如嵌套分量中增加一个或两个参数），因此用正态分布近似 $\chi^2(\nu)$分布的效果并不好。图1展示了 $\chi^2(\nu)$分布在 $\nu = 1$ 和 2 时较重的右尾，与相应的高斯分布 $N(\nu,\ 2\nu)$相比，这会使 $\chi^2(\nu)$ 分布的临界值随着p 增大而渐近增大。当 $p = 0.9$，即单侧置信区间为 90% 时，这两种分布的临界值相近；而在更大的p值下，高斯近似会系统性地低估临界值。显然，一般而言，不建议对 $\chi^2(\nu)$分布采用正态近似；在 ν 较大时研究 Cmin 分布的做法也曾如此。

对于 ΔC 统计量，式（16中的分布经过修改，假设其母 B 分布是两个变量之和，$B = X + Y$ 这两个变量分别具有

$$\begin{cases} X \sim \chi^2(\nu) \\ Y \sim N(0, \hat{\sigma}^2_{\Delta C}), \end{cases} \tag{27}$$

其中 $\nu = l$ 是附加嵌套分量的自由参数个数。此处记号表示，B 是一个过度离散的χ^2 分布；当根据式（14）因系统误差而产生的附加方差^为零时，该分布便成为通常的()分布。

当 ΔC 的自由度较少时，有两种方法可根据式（27）估计过度离散 ν 统计量的临界值及其他参数。第一种方法是用正态分布粗略近似 $\chi^2(\nu)$ 分布；正如前文已提醒的那样，使用时应谨慎。下文第3.2.3节主要出于说明目的介绍这种方法。第二种方法是对式（27）中的两个分布进行卷积——即 χ^2 分布与均值为零的正态分布——并在第3.2.4 节将其作为过度离散 χ^2 分布加以介绍。实际上，对于几乎所有应用，这都应是首选方法。

3.2.3 对 $\chi^2(\nu)$ 分布的正态近似

该方法与上文第 3.1 节的方法完全等价，并利用母体 $\chi^2(\nu)$ 分布的正态近似，并使用式（18）中的方差线性相加来确定能够纳入系统误差的 ΔC 分布的临界值。鉴于 ν 通常是一个较小的数，该方法的结果是近似的，因此通常更适合使用下文所述的过度离散 χ^2 分布。不过，考虑到该方法简单，而且有助于说明问题，仍值得进一步讨论。

如图1所示，$\chi^2(\nu)$ 分布的临界值大于相应的 $N(\nu, 2\nu)$ 分布，尤其是

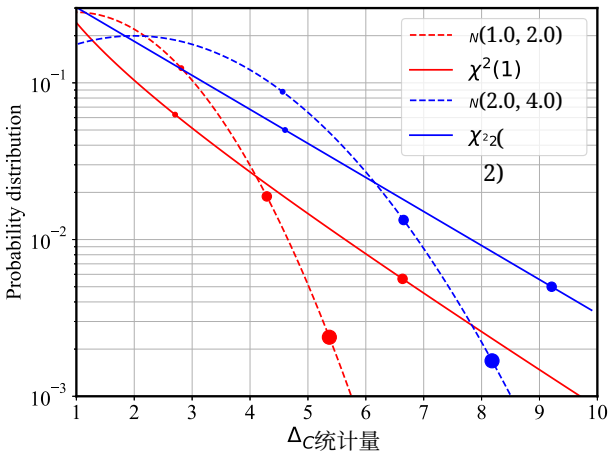


图1。选定自由度下 χ^2 与高斯分布的比较。按标记大小递增顺序，还报告了这些分布的 $p = 0。9、0。99、0。999$单侧临界值。

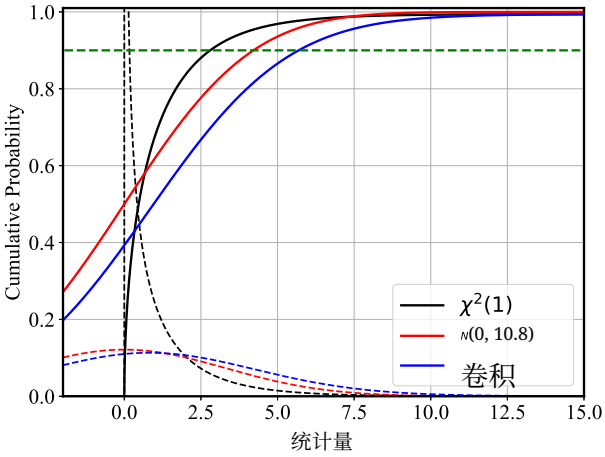


图 2. 过度离散 χ^2 分布，由方程 28 的卷积得到，其中 $\nu = 1$ 表示 χ^2 分布的自由度，且 $\sigma = 3。7$ 为零均值高斯分布的标准差。

对于较小的残差概率，由于 $\chi^2(\nu)$ 分布的尾部较重，这一方法是近似且保守的：如果根据渐近 χ^2 分布，零假设确实成立，那么在 p 置信水平下，零假设绝不会被错误地判定为可接受（即有必要加入附加分量）。另一方面，在 p 置信水平下拒绝零假设（即没有必要加入附加分量）则可能被错误判定。

使用这种近似时，应说明原假设是在置信水平 $\leq p$下被拒绝的，其中 p 是检验所选用的置信水平。该近似的优点是，可以直接利用（18）中的方差加和以及高斯分布函数来计算临界值。此外，图 1 表明，对于 $\chi^2(\nu)$ 和 $N(\nu, 2\nu)$，两种分布的 90% 临界值几乎相同 $\nu = 1$，2，这表明选择 $p = 0.9$ 很可能使这种近似带来的误差最小。鉴于下文将介绍一种更精确的方法，这里不再进一步研究该近似；通常也不建议在大多数应用中使用它。

3.2.4 过度离散的 χ^2 分布

一种更精确的研究方法是，按照（ ΔC 27）模型研究 ΔC 统计量的分布，即保留泊松贡献对应的 $\chi^2(\nu)$ 分布。这意味着，尽管由于假设参与的随机变量相互独立，方差仍会按照（18）线性相加，但根据（27）， ΔC 的分布并非高斯分布。在这种情况下， ΔC 统计量的母体分布变为两个分布的卷积，

$$f_B(z; \nu, \sigma^2) = \int_{-\infty}^{+\infty} f_{\chi^2}(z - y; \nu) f_N(y; 0, \sigma^2)\, dy \tag{28}$$

其中， $f_{\chi^2}(x;\ \nu)$ 是 $\chi^2(\nu)$ 随机变量的概率分布函数，其中 $\nu = l$ 是嵌套分量的自由参数个数； $f_N(x; 0,\ \sigma^2)$ 是 $N(0,\ \sigma^2)$ 随机变量的概率分布函数，其中 $\sigma^2 = \hat{\sigma}^2_{\Delta C}$ 表示设计方差。该母体分布将被称为

称为 B（ν、 σ^2）分布或过度离散的 χ^2 分布

其参数为 ν 和 σ^2。这种卷积通常不会使分布函数 f_B得到解析表达式，因此需要进行数值计算。 [5]

与测量值 $z = \Delta C$ 相关联的 p 值为

$$F_B(\Delta C; \nu, \sigma^2) = 1 - p, \tag{29}$$

其中 F_B 是过度离散 χ^2 分布的累积分布。概率分布函数 $B(\nu,\ \sigma^2)$ 如图 2 所示，图中给出了一个具有代表性的参数案例。表 1列出了该过度离散 χ^2 分布在选定的自由度数 ν 和内禀模型方差 σ^2下的临界值，也可用于计算 p值。通过对式（ χ^2 28）进行数值积分，可以轻松获得更全面的过度离散 χ^2 分布表。

建议在存在内禀模型方差的情况下，将此 $B(\nu, \sigma^2)$ 分布作为具有 ν 自由度的 ΔC 统计量的首选分布，其中内禀模型方差按照式（27）建模为均值为零、方差为 σ^2 的高斯分布。参数$\sigma = \hat{\sigma}_{\Delta C}$ 表示内禀标准差，这意味着范围 $C\mathrm{min} \pm \hat{\sigma}_{\Delta C}$ 约包含该 ΔC 统计量预期内禀变异的 68%，而这种变异由内禀模型方差导致。此内禀“设计”标准差根据式（26）计算，并将嵌套模型分量的具体细节纳入考量。例如，对于自由度为 $\nu = 1$ 的 ΔC 统计量——例如，检验含有 1 个额外自由参数的嵌套分量——当$\hat{\sigma}_{\Delta C} = 1$ 取某一数值时，通常的 90% 临界值会从 2.7 增至 3.1。若$\hat{\sigma}_{\Delta C} = 5$ 取更大的值，临界值则升至 7.7；这意味着在存在此类系统误差时， ΔC 统计量的任何测量值只要介于 2.7 和 7.7 之间，在 90% 置信水平下仍具有统计可接受性。因此显而易见，低估或忽视系统误差会导致对额外嵌套分量进行假设检验时得出错误结论。

[5] 例如，可以使用 `python` 中的 `scipy.signal.fftconvolve`例程轻松估算卷积。该例程使用快速傅里叶变换对离散化分布进行卷积。

过度离散的 χ^2 分布也适合作为 Cmin 统计量本身的母分布。事实上，如第3.1.1节所述，存在系统误差时的 $Z = C$ min 统计量是两个独立变量之和，其中第一个变量为 $X \sim \chi^2(N-m)$。因此，当存在内禀模型方差 $\hat{\sigma}_C^2$ 时，参数为 $\nu = M - n$ 和 $\sigma^2 = \hat{\sigma}_C^2$ 的过度离散 χ^2 分布是 Cmin 统计量的首选分布。当 $\nu \gg 1$ 时（大多数回归都是如此），假设数据处于大计数范围，该分布渐近收敛于期望为ν、方差为 $2\nu + \hat{\sigma}_C^2$的正态分布。因此，对于该范围内的许多应用，如第3.1.1节所述，在存在系统误差时，使用 Cmin 统计量的正态近似可能更为方便。

显然，式（28）中的卷积会产生包含负值的分布，这与 χ^2 分布不同，后者是正定的。当 ν 数值较小，而 σ^2 内禀方差较大时，尤其如此；较大的内禀模型涨落可能导致 C 统计量实际上低于 Cmin。在实践中，关注的是针对这种过度离散 χ^2分布，使用单侧置信区间进行假设检验。因此，即使负尾显著（如图2所示），也不会引起数据分析人员的直接关注。

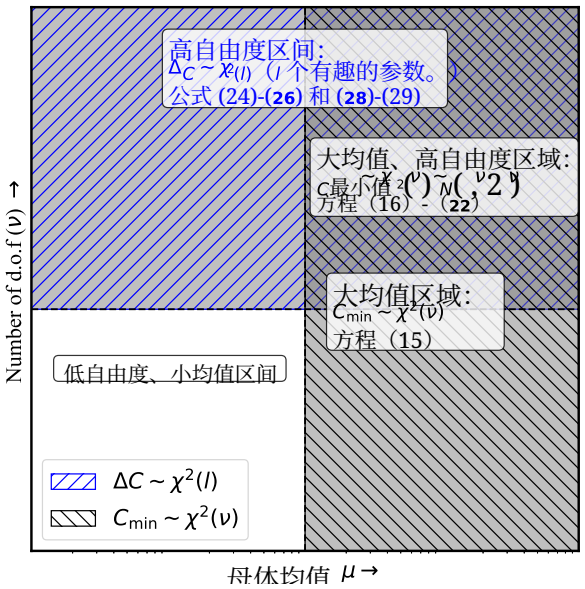


图 3.本文主要结果的适用范围。本文主要结果对应的公式以粗体显示。对于数据点较少且母体均值较小的数据集（左下角），其参数空间范围是本文结果通常不适用的部分。

3.3 渐近分布的适用范围

选择在式（Cmin （15）和（16）中，以及在 ΔC 式（24）和（27）中采用渐近分布，旨在获得相关过度离散统计量的解析表示，并将系统误差考虑在内。出于同样的原因，系统误差也采用简单的正态分布进行建模。显然，实现本文所讨论的方法并不需要这些假设。具体而言，内禀方差估计（见第3.1.2节）也可以使用不同的分布来估计 X 和 Y 变量，它们分别表示 C min 统计量和系统误差。然而，如果分布没有简单的解析形式，就需要进行数值卷积，因此无法保证式（22）中的简洁结果仍然适用。

类似地，定义过度离散 χ^2 分布的卷积（28）可以在任意两个分布之间进行。例如，如果分析者决定放弃系统误差服从正态分布的假设，就可以采用这种做法；Lee et al. (2011) 和 Xuet al. (2014) 就曾在处理系统误差的另一种框架中采用过这一做法。

本节结果的适用范围总结于图 3。具体而言，公式（22）适用于大均值区间，并且对于任意自由度，均适用于 Cmin；而（ χ^2 ）中给出的过度离散分布28适用于母体均值的所有取值，因此即使在低均值区间也适用，但要求自由度较大。

各适用范围之间的界限并不明确，可以近似估计如下。

(a) 在高自由度范围内，当测量值数量较大时，对数似然 ΔC统计量近似服从 χ^2分布。该结果可由Wilks (1938) 定理推出，该定理表明，对数似然在相差一个 χ^2阶$1/\sqrt{n}$ 的项之内服从Bonamente 2022的第 16 章也有论述）。因此，一个近似的经验法则是， $\nu \geq 20$左右可能足以达到百分比级别的精度，而且大多数天文数据集通常都处于这一范围。

(b) 大均值区间中，Cmin 服从 χ^2 分布，这主要是由以下近似所驱动：用具有相同均值和方差的正态分布近似泊松分布，这种近似在左右时通常相当准确。例如，Kaastra（2017）和Bonamente（2020年）借助数值计算和模拟，给出了低均值区间中 C 统计量的均值和方差近似值。

4 应用于 1ES 1553+113 的 X 射线数据

我们将第3 节所述的方法应用于*XMM–Newton*对 1ES 1553+113 的光栅光谱；相关内容详见另一篇论文（Spence 等人（2023））。Nicastro等人(2018年) 使用这些数据探测来自视线方向上温热星系际介质的 O VII 吸收。这些数据也代表了本文所讨论的数据模型。光谱见图4，其中数据点表示每个箱中的总计数（整数）y_i，水平标记表示最佳拟合模型μ_i。由于校准较差，若干波长范围被排除；此外，排除范围相邻的若干箱的大小小于数据处理过程中采用的默认分箱大小 20 mÅ，因此计数较少。

就本文而言，连续谱建模的细节并不重要，当前分析所关注的主要统计量总结于表2中。假设检验的方法和系统误差的确定仅依赖于该表中列出的数值。作为参考，下面对分析过程作简要总结。用于拟合数据的模型由13–33 Å波长范围内包含若干自由参数的样条连续谱，以及用于建模局部背景的附加分量组成。具体而言，样条连续谱以1 Å为间隔进行参数化，每个区间对应一个自由参数。对于这些数据，在所关注的波长范围内，背景通量平均为源通量的 ~ 10%，且其波动远小于源的波动，因为背景是在比点源更大的区域内收集的

表1. 过度离散的 $\chi^2(\nu)$ 分布的临界值，随内禀方差（$\hat{\sigma}_C$）的若干代表值以及选定的 p–值而变化。括号内为标准 $\chi^2(\nu)$ 分布对应的临界值。

$\hat{\sigma}_C$	$\nu = 1$						$\nu = 2$						$\nu = 3$					
	$p = 0.9$		$p = 0.99$		$p = 0.999$		$p = 0.9$		$p = 0.99$		$p = 0.999$		$p = 0.9$		$p = 0.99$		$p = 0.999$	
1.0	3.1	(2.7)	7.0	(6.6)	11.8	(10.8)	4.9	(4.6)	9.5	(9.2)	14.1	(13.8)	6.5	(6.3)	11.6	(11.3)	16.5	(16.3)
2.0	4.1	(2.7)	7.9	(6.6)	12.6	(10.8)	5.6	(4.6)	10.2	(9.2)	14.9	(13.8)	7.1	(6.3)	12.3	(11.3)	17.2	(16.3)
5.0	7.7	(2.7)	13.4	(6.6)	18.5	(10.8)	8.9	(4.6)	15.0	(9.2)	19.9	(13.8)	10.1	(6.3)	16.6	(11.3)	22.0	(16.3)
10.0	14.0	(2.7)	24.7	(6.6)	33.4	(10.8)	15.1	(4.6)	25.8	(9.2)	33.7	(13.8)	16.2	(6.3)	27.1	(11.3)	35.3	(16.3)
15.0	20.3	(2.7)	36.3	(6.6)	49.2	(10.8)	21.4	(4.6)	37.2	(9.2)	48.8	(13.8)	22.5	(6.3)	38.4	(11.3)	50.2	(16.3)
20.0	26.7	(2.7)	47.9	(6.6)	65.1	(10.8)	27.8	(4.6)	48.8	(9.2)	64.0	(13.8)	28.8	(6.3)	49.9	(11.3)	65.4	(16.3)

。光谱分析采用SPEX 3.0拟合软件包（Kaastra 等，1996），该软件使用计数总数（源加背景）来计算 C 统计量。有关数据和数据分析的完整说明见配套论文（Spence 等人（2023）），其中包括对背景及其对拟合统计量计算影响的分析。

4.1 假设检验与系统误差的必要性 rors

使用 C–stat进行假设检验的方法在Kaastra (2017)和Bonamente (2022)中有详细描述。对于此处的应用，由于这些数据处于大–N 和大–均值范围内，该方法得到了显著简化；在这种情况下，预计 Cmin统计量服从 $\chi^2(N-m)$分布，其中 m 是拟合中的自由参数个数。已知一个具有ν 自由度的 χ^2 随机变量，其均值为 ν ，方差为 2ν，因此需要将测得的统计量值与该分布的置信区间进行比较。

为方便起见，下文用范围$\nu \pm \sqrt{2\nu}$ 表示自由度为 ν 的χ^2 分布的均值上下一个标准差范围。值得指出的是， Cmin的这一范围与SPEX 软件包报告的结果略有不同；后者假定拟合统计量的期望值等于式(3)中各组成项的 N 期望值之和，在均值较大时渐近为 $N \pm \sqrt{2N}$ ，且不考虑可调参数的个数(Kaastra 2017)。尽管表2中的两个范围仅有微小差异，仍建议读者参阅Bonamente (2022)第16.3节及Bonamente (2020)，其中详细描述了使用 C–stat进行假设检验的方法，包括基于Wilks (1938)的定理对 C–stat自由度缩减问题的已知结论。如本文前文所述（见第3.1节），McCullagh(1986)也讨论了这一问题；他证明， C min均值的一阶近似为 N，而二阶近似给出了 m 修正，从而得到期望值 $\nu = N - m$。因此，在本分析中，我们假设 C–stat存在自由度缩减，其大计数极限下 Cmin统计量的母体均值为 ν 。

假设检验基于 p 拟合的值，其定义为：在数据取自母模型这一假设下（即所考虑的模型能够准确表示数据），测得的统计量（在本例中为 $C_{\text{最小}}$值）大于或等于测量值的概率。对于总体$C_{\text{最小}}$=1862.7 统计量及其基于$\chi^2(N-m)$渐近分布的期望值（见表2），测量值落在距零假设下的母体均值 $z = +7.1$ 个标准差处，对应的 1478 ± 54.4 零假设所对应的 p–值小到可忽略不计。对于单独考虑的 RGS1 和 RGS2 数据的两个拟合统计量，也可得出类似结论。因此，显然测得的拟合统计量与零假设不相容。

拟合统计量高于预期的情况——无论是C统计量还是 χ^2 ——在X射线天文学中相当常见。如果在较宽的波长区域内，最佳拟合模型显著高于或低于数据（文献中有不少此类例子，例如针对另一数据集的Bonamente 等人（2003）图4），最佳做法是认定当前模型不可接受，从而弃用该模型并尝试其他模型。然而，在本例中，残差没有系统性趋势，也就是说，最佳拟合模型看起来总体而言能够合理地近似数据。因此，分析者面临两种选择：（a）弃用该模型；或（b）调查是否存在其他不确定性来源，可能导致拟合统计量的实测值与预期值不匹配。本节采用后一种方法，并使用第3 节的方法，估算模型中的内禀方差大小，使模型与数据达到可接受的一致程度。该方法假设残差中不存在结构，也就是说，数据中的附加方差大致均匀地分布在所有数据点之间。因此，采用这种方法时，分析者必须确保对残差的分析支持这一假设。对于这里的数据，Spence 等人(2023年)提供的分析确实表明，残差中不存在显著结构。

此处还值得指出，在天体物理学文献中，通常会报告拟合统计量的约化值（通常为 $\chi^2\text{red} = \chi^2\text{min}/\nu$ 的值），并认为只要该约化值“接近”1，拟合就是可接受的。更具统计依据且定量的方法是报告拟合的 p 值，正如本节前文所述，因为 χ^2 或 C 统计量的约化值与 1 的接近程度，很大程度上取决于数据点数量或自由度。对于总体 Cmin 统计量及其基于表 2 中所报告的 $\chi^2(N-m)$ 渐近分布所得的期望值，发现其 p 值几乎为零，尽管约化后的 Cmin 统计量值为 1.29，因此可能会被错误地认为“接近”1。

在根据第 3 节的方法估计系统误差之前，我们使用方程 9 和11 来估计数据中的过度离散参数。对于 NB1 参数化，我们估计

$$\hat{\phi} = \begin{cases} 1.58 \pm 0.05 & \text{for RGS1 data} \\ 1.35 \pm 0.05 & \text{for RGS2 data} \end{cases} \tag{30}$$

表 2. *XMM–Newton* 对 1ES 1553+113 光谱的拟合统计量

统计量	值	注释
	RGS1 和 RGS2 数据的合并结果	
$C_{\rm min}$	1862.7	(1861.8 来自 Spence 等人（2023）)⋆
χ^2	1730.6	（未用于最小化）
N	1526	数据点数量
m	48	自由参数数量
$f = N - m$	1478	自由度数量
期望的 C–统计量	1526。4± 55.3 1478 ± 54.4	如 SPEX 所报告 根据 $\chi^2(N-m)$ 近似值。
	RGS1 数据	
$C_{\rm min}$	1023.5	
N_1	801	
χ^2	1023.5	
m_1	48	得到 $N_1 - m_1 = 753$ 自由度
预期的 C–统计量	753。0 ± 38。8	根据 $\chi^2(N-m)$ 近似值。
	RGS2 数据	
$C_{\rm min}$	788.2	
N_2	725	
χ^2	788.2	
m_2	44	20–24 Å 范围内的数据被忽略， 从而得到 $N_2 - m_2 = 681$ 个自由度。
期望的 C–统计量	681.0 ± 36.9	根据 $\chi^2(N-m)$ 近似。

⋆细微差异是由于 SPEX 软件所采用的近似造成的，与我们的分析无关。

其中，不确定性基于简单的误差传播法或德尔塔法，并假设估计的均值已知且准确。鉴于 $\hat{\phi} > 1$，该估计证实数据中存在额外的方差来源。对于 NB2 参数化，我们估计

$$\hat{\alpha} = \begin{cases} 1.5 \pm 0.3 \times 10^{-3} & \text{for RGS1 data} \\ 0.6 \pm 0.2 \times 10^{-3} & \text{for RGS2 data} \end{cases} \tag{31}$$

其中 α 参数乘以每箱平均计数的平方，因此其绝对值低于 NB1 模型。与前一种参数化一样，该值 $\hat{\alpha} > 0$ 表明存在过度离散。$\hat{\phi}$ 和 $\hat{\alpha}$ 参数为数据相对于采用通常泊松对数似然得到的最佳拟合模型存在过度离散提供了定量证据。然而，这些参数并未涉及过度离散或内禀模型方差对拟合统计量的影响；以下将使用第 3 节介绍的方法对此进行研究。

4.2 使用 C–统计量估计系统性模型不确定性

4.2.1 初步考虑

鉴于表 2 中报告的统计量数值，分别独立分析这两台 *XMM–Newton* 仪器是合理的。这样做也展示了如何根据多个相互独立的数据集确定系统误差。首先需要指出，所用模型不允许应用于 RGS1 和 RGS2 的模型之间存在自由归一化。因此，我们首先仅针对 RGS1 和 RGS2 数据重复进行相同的拟合，以探究拟合统计量不佳的原因。通过这一操作，我们得到的拟合在统计上与表 2底部报告的拟合相当；该处给出的是 Spence 对 RGS1 与 RGS2 联合回归所得统计量的贡献等人 (2023年）。因此，我们得出结论：两台仪器之间的交叉归一化误差并非拟合不佳的原因 $C_{最小}$ 值。因此，我们以表 2 中的数值为依据，估算系统误差。

4.2.2 确定 C*min* 方差的设计值

首先要确定 Cmin 统计量的内禀方差的一个“设计”值，使其与测量值相符。从定性角度来说，这个值必须使 Cmin 统计量的分布函数与当前不相符的测量值达成一致（例如，RGS1 的测量值为 1023.5，而期望值和标准差分别为 753.0 ± 38.8）。设 F_C Cmin 统计量的母体分布的累积分布函数为，则可要求该统计量的测量值满足

$$F_C(C_{\rm min}) = 1 - p, \tag{32}$$

其中，p 是一个较高的概率，例如 $p = 0.9$ 或 0.99。方程 32 表示，要求测量值 Cmin 成为 p 分位数；根据拟合统计量的抽样分布，超过该测量值的剩余概率仅为 $1 - p$。方程 32采用单侧拒绝域 Cmin $> C$crit，这符合使用 χ^2 分布进行假设检验的惯例。在（32）中，Cmin 是根据现有数据对该统计量进行的实际测量值。内禀方差问题会在统计量足够大、超过其期望值时出现，因此，（32）中的统计量测量值可理解为位于拒绝域边界，即 Cmin $= C$crit。

当 $C_{最小}$统计量服从正态分布时（对于大量测量值以及在大计数极限下均是如此，本应用也属于这种情况），则（32）等价于这样一个熟知的要求：测量值超过均值若干个预先确定的标准差，

$$C_{\rm min} = \mathrm{E}[C_{\rm min}] + \beta \times \sqrt{\mathrm{Var}(C_{\rm min})} \tag{33}$$

其中 $\beta = 1$。28 或对于 $p = 0$。9 或 0.99。必须求解方程（32 或33），以得到内禀方差 $\hat{\sigma}^2_C$，其中

$$\mathrm{Var}(C_{\rm min}) = \sigma^2_C + \hat{\sigma}^2_C \tag{34}$$

并且 σ^2_C 表示未计入内禀模型不确定性时该统计量的母体方差。内禀方差的设计值可由（32）（或在拟合统计量服从高斯分布的假设下由（33））以及（34）立即计算得到：

$$\hat{\sigma}^2_C = \frac{(C_{\rm min} - \mathrm{E}[C_{\rm min}])^2}{\beta^2} - \sigma^2_C. \tag{35}$$

对于表 2中的统计量，并选择β = 2。33 （对应于 99% 的单侧置信水平），方程35 分别对 RGS1 和 RGS2 数据给出了以下内禀方差值：

$$\begin{cases} \hat{\sigma}_{C,1} = & 109.4, \text{ for a prediction of } 753 \pm 116; \\ \hat{\sigma}_{C,2} = & 57.0, \text{ for a prediction of } 681 \pm 67.9. \end{cases} \tag{36}$$

将这些内禀方差合并后，对于整个 RGS1+RGS2光谱，得到

$$\hat{\sigma}_C = 123.4, \text{ for an overall prediction of } 1478 \pm 134.8. \tag{37}$$

方程 36和 37 是所求的“设计方差”估计值。根据最大似然泊松分布，它们是使最佳拟合模型与数据在统计上相符所需的方差

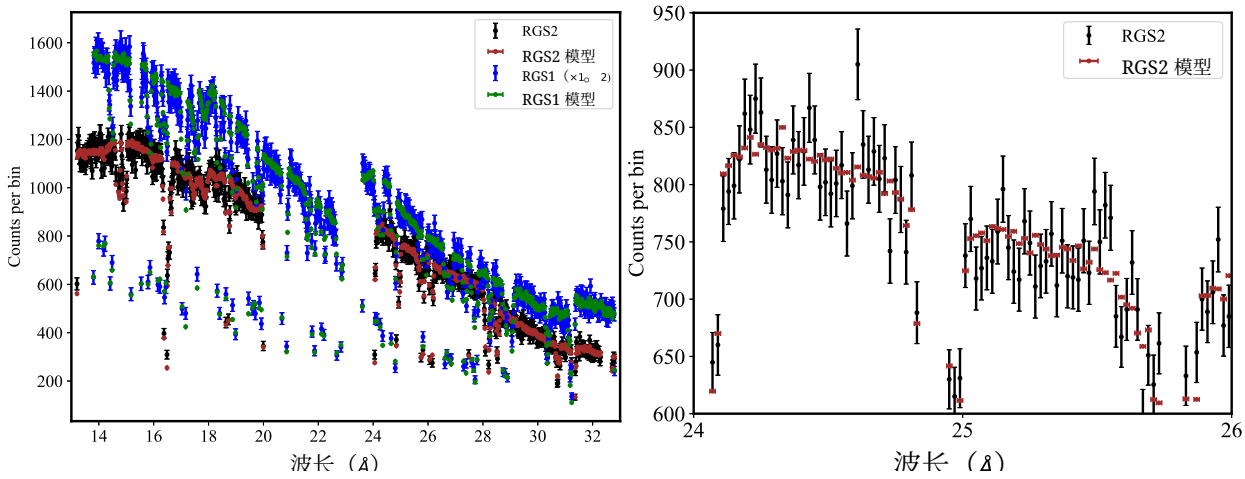


图 4. 左：类星体 1ES 1553+113 的 *XMM–Newton* 光谱，其中包含来自 RGS1 和 RGS2 相机的两条独立光谱。为便于观察，将 RGS1 数据平移了 1.2 倍。右：窄波长区间内光谱的放大图，用于展示数据和模型。对于每个数据点，黑点表示计数的整数值（源加背景），误差棒表示泊松误差，而在 C–统计量中并未显式使用该误差。

$C_{\text{最小}}$ 拟合优度统计量。这些设计方差根据式（$C_{\text{最小}}$ 34），加到原始的“统计”方差上（即基于数据泊松计数统计的方差），从而得到各统计量的总体方差。请注意，给定通常测得的 min.7 和修正后的期望值 .8，总体 Cmin 统计量的 z 分数现在为 $z = 2.85$，对应于测量值超过 $p \simeq 0.002$ 的单侧零假设概率。

4.2.3 内禀方差的估计

对于拟合统计量内禀方差的这些设计值，使用 (22) 可得出内禀模型误差。RGS1、RGS2 和合并数据集的平均计数分别为 728.6、756.2 和 741.7。因此，该方法得出

$$\frac{\sigma_{\text{int},i}}{\hat{\mu}_i} = \begin{cases} 0.072 \text{ for RGS1} \\ 0.039 \text{ for RGS2} \\ 0.058 \text{ for combined RGS1 and RGS2.} \end{cases} \tag{38}$$

这些数值表明，最佳拟合模型存在一定程度的系统不确定性：RGS1、RGS2 和合并数据分别为 7.2%、3.9% 和 5.8%。

因此，这种方法利用简单的解析方法，估算了单独的 RGS1 和 RGS2 光谱以及合并光谱的内禀模型方差。根据（38），这些方差是以每个箱中最佳拟合模型的分数误差来估算的，假设相对系统误差在整个光谱中均匀分布。必须将这些数值与已知的 *XMM–Newton* 仪器系统误差进行比较。 Spence 等人 (2023年)指出，对于本分析所用大小的箱，典型系统误差约为每箱计数的百分之几（另见 Marshall 等人 2021）。因此，本文分析结果中相对于均值为 4%–7% 的波动，与仪器已知的系统误差水平一致。

最后，将估计的过度离散参数 $\hat{\phi}$（来自（30））的值与（38）中的内禀模型方差进行比较也很有用。过度离散参数 $\hat{\phi} = 1。3 - 1。6$ 刻画了每个测量值中的额外方差，其取值为 $\phi = 1。0$ 表示方差等于均值。方差的平方根 $\omega_i = \hat{\phi}\,\mu_i$（或 $\omega_i = \mu_i + \alpha\mu_i^2$ 对于另一种参数化）需要除以每箱的平均计数（μ_i），才能评估额外方差所隐含的数据中额外分数系统误差的水平。对于这些数据中每箱数百个计数的典型计数率，估计值 $\hat{\phi}$ 和 $\hat{\alpha}$ 所对应的分数波动比 $\phi = 1$的情形大几个百分点，这总体上与（38）中估计的百分比级内禀模型方差一致。

4.3 使用 χ^2 统计量处理数据中的系统误差及其他考量

鉴于第 4.2 节的分析表明，模型存在量级为 $\sim 5\%$ 的内禀不确定性，自然会想知道，如果将这种不确定性改用于数据，会对 χ^2 统计量产生怎样的影响。进行这一比较，是为了在统计量无法接受时，与一种较为传统但不太准确的数据系统误差评估方法进行一致性检验。

为此，我们将式（38）中的平均内禀误差应用于每个数据点，并将其与常规泊松误差按平方和开方的方式合并。更改数据误差后，χ^2 统计量变为

$$\chi^2 = \begin{cases} 874.8 \text{ for RGS1 } (\Delta\chi^2 = -68.5) \\ 626.1 \text{ for RGS2 } (\Delta\chi^2 = -101.82) \\ 1500.9 \text{ for RGS1+RGS2 } (\Delta\chi^2 = -170.4). \end{cases}$$

相对于表2中的数值，主要结果是，在数据中加入 6.5% 的系统误差后，$\Delta\chi^2 = -170.4$ 会减小，使该统计量更接近其在给定自由度 ($\nu = 1478$) 下的期望值。因此，这一练习表明，使用第 4.2 节所述的新方法，根据泊松数据和C–统计量估算出的内禀模型方差是合理的：对数据和 χ^2 统计量施加幅度相同的系统误差，会使该统计量更接近其期望值，这与使用内禀模型方差时 Cmin 的变化类似。

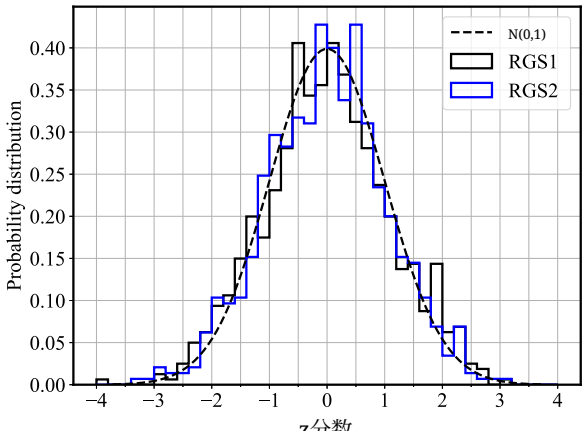


图5. RGS1 和 RGS2 数据的 z–分数分布。两个样本的均值和标准差分别为：RGS1 z分数为 0.00 和 1.08，RGS2 z分数为 -0.01 和 1.04。

此外，绘制最佳拟合残差对应的 z– 分数的概率分布也很有参考价值；此处采用通常的泊松计数作为数据方差。图 5所示的这些分布与标准正态分布没有显著偏离，而根据零假设，预期分布正是标准正态分布。柯尔莫哥洛夫–斯米尔诺夫检验（Kolmogorov 1933）也证实了这一观察。该检验是一种使用未分箱数据的经验分布函数（EDF）检验。单样本柯尔莫哥洛夫–斯米尔诺夫检验统计量分别为 $D_N = 0.033$ 和 0.018，对应 RGS1 和 RGS2 数据；相应的 p–值分别为 0.33 和 0.97，表明数据与母模型高度吻合。 [6] 因此，柯尔莫哥洛夫–斯米尔诺夫检验对检测几个百分点量级的系统误差并不敏感，也就是说，对本案例研究数据中存在的这一量级误差并不敏感。对这些分布进行正态性检验所得的安德森–达林经验分布函数统计量（Anderson & Darling1952；Stephens 1974）的值分别为 $A^2 = 0.32$ 和 0.21，对应 RGS1 和 RGS2 数据；两组数据的 99% 临界值均为 1.09，这表明无法拒绝数据服从正态分布的零假设。因此，与单样本柯尔莫哥洛夫–斯米尔诺夫检验类似，安德森–达令检验无法发现 z 分数偏离正态性的情况。 [7]

4.4 使用ΔC 统计量时内禀误差对嵌套分量的影响

方程 26 提供了一种方法，可用于估计内禀模型方差导致的 ΔC 统计量的不确定性。对于采用 1ES 1553+113*XMM–Newton* 数据的案例研究，相关统计量列于表 3，摘自 Spence 等人(2023年)。用于检测嵌套线分量的回归现在是在预期谱线特征附近的窄波段 ±1 Å 内进行，而非采用宽波段拟合（如表 2结果所示）。对于表 3中的数据，基线模型由一个简单的幂律模型构成，并辅以一个高斯线模型（谱线，位于 SPEX 中）。这些谱线模型描述了平滑连续谱在以下位置的偏离

[6] 该检验的介绍见Bonamente （2022年）第19章，例如；该检验由 `kstest` Python 脚本实现，可通过 `scipy.stats` 软件包免费获取。

[7] 安德森–达令检验在 `scipy.stats`软件包中以 `anderson` 的形式实现。

几个百分点的水平，因此有必要确定类似水平的内禀误差如何影响对这些特征进行检测的显著性。表3结果所依据的数据示例见Spence 等人(2023年)。

使用（26）时，最大的不确定性在于求和中独立 n 数据点的数量。对于吸收线类模型，例如`line`的高斯轮廓，确定独立数据点数量的一种简单方法是计算吸收线特征的等效宽度。在所有情况下，等效宽度都小于仪器的分辨率，因此可以简单估算为 $n = 1$，这意味着模型只影响一个数据点，也由此得出本征方差的保守估计（即可能的最小估计）。对于三条谱线，每箱的典型计数分别为800、500和400，且本征模型误差为$\sigma_{\rm int,i}/\hat{}$ μ_i = 0。058，如第4.2.3节估算，式（26）得出的估计为

$$\hat{\sigma}^2_{\Delta C} = \begin{cases} 10.8 & \text{for line 1} \\ 6.7 & \text{for line 2} \\ 5.4 & \text{for line 3.} \end{cases} \tag{39}$$

根据第 χ^2 节所述，在存在内禀误差时 ΔC 统计量的修正3.2分布，现在可以在考虑系统误差影响后，判断表3 中的三条谱线是否仍具有统计显著性。结果列于表p的“–值”栏3。

第一条谱线的统计显著性从 p值为 1% 降至 ~ 7%。因此可以得出结论，如果如本文所述的方法所假定的那样，将数据相对于表2 中最佳拟合模型的附加方差解释为系统误差的额外来源，那么位于 $z = 0$。1876 的吸收线很可能是由统计涨落造成的，不太可能是真正的天体信号。本文不宜进一步评论这条假定吸收线的天体物理学意义，因为本报告的主要目的是建立一种分析系统误差的方法。因此，结论是，像Spence 等人(2023年)分析的光谱这类泊松数据中的系统误差来源很容易处理，而且其影响可能很显著。

第二条吸收线最初由 Nicastro 等人（2018年）报告，其显著性也降低了近 ~ 100 倍。对于这条吸收线，必须考虑到其红移是偶然识别出来的，因此分析中还增加了一个自由度（即吸收线的中心波长）。对于这类偶然发现的源，Kaastra等人（2006年）和Bonamente（2019年）介绍了一种方法，用以处理多次独立发现涨落的机会，这种情况称为红移试验。鉴于有大量独立机会可以发现随机涨落，Spence 等人（2023年）估计其数量级为 $N \simeq 100$，Bonamente （2020年）所述的方法得出了经校正的 p–值，列于表2 [8]最右侧的几列中。采用基于过度离散 χ^2 分布修订后的 p–值，经过红移试验校正的 P 值现在约为 0.003，这大致对应于正态分布的 $\pm 3\sigma$ 双侧置信水平。因此显而易见，即使嵌套模型分量的检测看似“显著”（2 个自由度下为 $\Delta C = 29.9$），也可能因百分比量级系统误差的影响而变得远不那么显著。

对于第三条谱线，不宜进行相同的分析，因为它

[8] 当 p 较小时，该方法近似等于 $P \simeq p\,N$。

表3。 ΔC 幂律模型加高斯线特征拟合的统计量，针对 *XMM–Newton* 数据（1ES 1553+113），转载自Spence 等人（2023年）。“*p*–值”列对应回归的常规 p 值，而“*P* 值”列对应于纳入 *p*–值，其中包含了对未知红移谱线进行盲搜相关的不确定性。

目标谱线	ΔC	ν	*p*–值		*P* 值（含红移试验）	
			（无系统误差）	过度离散 χ^2 分布。	（无系统误差）	过度离散 χ^2 分布。
O VII $z = 0。1876$	6.6	1	0.010	0.066	–	–
O VII $z = 0。4339$	29.9	2	$3。2 \times 10^{-7}$	$2。6 \times 10^{-5}$	$4。1 \times 10^{-5}$	3.3×10^{-3}
O VII $z = 0。3551$	8.2	2	0.017	0.880		–

即使不考虑内禀方差，该结果在统计上也不显著（根据过度离散的 χ^2 分布，概率为 88%）。红移试验校正不适用于第一条谱线，因为其中心波长固定为已知值，因此没有进行红移试验。

4.5 典型实现的步骤总结

本节简要总结本文所述方法的典型实现步骤，旨在通过参数模型的最大回归及 *C*min统计量估计泊松计数数据中的系统误差，并使用 ΔC 统计量评估嵌套模型分量的显著性。

1. 确定设计方差 $\hat{\sigma}^2_C$，根据（35），其中 $C_{\rm min}$是参数回归中测得的统计量，且 ν 为自由度。此步骤需要选择一个 β 值，以衡量测得的统计量与其母体分布的一致性。在大计数区间，$C_{\rm min}$ 统计量的期望值为 ν；关于低计数区间的讨论见第 4.1节。

2. 分数系统误差 $\sigma_{\rm int,i}/\hat{}\mu_i$ 根据（22）估算，其中 N 是独立泊松数据点的数量，$\hat{}\mu_i$ 是第 i 个数据点的最佳拟合模型。该数值表示与每个数据点的最佳拟合模型相关的分数不确定性。

3. 需要评估嵌套分量的显著性时，首先要确定或估算n 嵌套模型存在的箱数，然后根据（$\hat{\sigma}^2_{\Delta C}$）估算方差26。

4. 与 ΔC 具有附加 $\nu = l$参数的额外嵌套分量相关联的统计量，随后与母体过度离散χ^2 分布 $B(l,\ \hat{}\sigma^2_{\Delta C})$ 进行比较，该分布定义于（28）中。通常，这通过对方程（29）求数值解来实现，该方程给出与 *p*统计量相关联的 ΔC 值。

所有这些步骤所需的计算工作量都很小，因为大多数结果都是初等解析公式，而且对过度离散的 χ^2 分布进行数值积分也是一项简单的工作，因为该分布具有简单的单峰特性。

5 讨论与结论

本文提出了一种简单的定量方法，利用 *Cash* 统计量，估计将泊松计数数据拟合到参数模型时，最大似然拟合中的系统不确定性。这种情况在分析天文源的 X 射线光谱时很常见，例如本案例研究中的 1ES 1553+113 (Spence 等人（2023）所介绍的类型。该源是一个明亮的类星体，其跨较宽波长范围的拟合优度统计量在形式上不可接受，但数据并未显著偏离最佳拟合模型，也不存在系统性偏差。

在这种情况下，数据分析人员面临着一个选择：要么完全拒绝该模型，要么确定是否存在泊松分布数据中未包含的其他误差来源。当分析者认为采取后一种做法是合理的时，第 3 节所述的方法可导出式 22 中的主要结果，提供一种简单的定量方法，用于估计模型中存在的内禀方差，使数据与最佳拟合模型在预先设定的 *p*值下相符。对于本文及 1553+113 Spence 等人 Spence 等人 (2023年) 所提供的 1ES 数据，我们发现，复杂参数模型拟合中高于预期的 *C*统计量，可以自然地用最佳拟合模型中存在 ∼ 5% 的系统不确定性来解释。该方法完全不再使用更常用的 χ^2统计量，因为它不适用于对整数计数且服从泊松分布的数据进行回归，而大多数天文 X 射线光谱正属于此类数据（例如，参见 Humphrey 等，2009；Bonamente，2020）。

内禀模型方差的一个关键应用在于它对 ΔC 统计量的影响。该统计量是一个似然比统计量，通常用于检验附加嵌套分量是否显著（例如，Protassov等，2002）。根据第3.2节所述的方法，可以证明，内禀模型方差会为 ΔC 统计量带来一个额外项，该项可建模为均值为零、设计方差为$\hat{\sigma}^2_{\Delta C}$的正态分布。此附加分布应与通常的$\chi^2(\nu)$分布相加；后者表示嵌套模型分量具有 $\nu = l$ 个自由参数时，ΔC统计量中的泊松变异。由此引入过度离散的 χ^2 分布，以取代检验附加嵌套分量时通常使用的 $\chi^2(\nu)$ 分布。这个新引入的分布是两个贡献概率分布 $\chi^2(\nu)$和 $N(0, \hat{}\sigma^2_{\Delta C})$的卷积；在存在系统误差且零假设为嵌套分量不存在的情况下，它是通常 ΔC统计量的母体分布，因此为嵌套分量的假设检验提供了一种简单方法。选择将系统效应导致的额外变异归因于模型而非数据后，ΔC 统计量本身的计算保持不变，因此对嵌套分量进行假设检验只需计算这种新的过度离散 χ^2 分布的临界值或 *p*值，例如表1所列的此类数值。

在对 1ES 1553+113 源进行的案例研究中，使用了 ΔC统计量来检验是否存在额外的吸收线分量。最佳拟合模型中的 ∼ 5% 系统误差会导致 ΔC 分布产生附加方差，并降低检测到这些谱线的显著性水平，正如表3所报告的那样。对于

例如，所考虑的第一条谱线由 Spence 等（2023）认定为可能的 O VII 共振线，位于 $z = 0.1876$ 处；其显著性将从 99% 水平降至 93% 水平（即，p–值从 0.01 增加到 0.07）。Spence 等（2023）报告的另一条吸收线此前由 Nicastro 等（2018）发现，其检测显著性也因内禀模型方差的影响而大幅降低（见表3中的第二条谱线），但仍高于 99% 置信水平（p–值为 0.003）。因此，在使用似然比统计量（例如 ΔC 或 $\Delta\chi^2$）检测额外嵌套分量时纳入系统误差，可能会影响基于计数数据的天体物理推断，例如本案例研究所考察类型的高能谱。这一点在诸如从温热星系际介质（WHIM）中检测 X 射线吸收线等领域尤为重要；在这些领域，微弱的天体物理信号与仪器有限的分辨率共同导致了多项检测（例如 Ahoranta 等 2021年、2020年；Nevalainen 等 2019年；Kovács 等 2019年；Bonamente 等 2016年；Ren 等 2014年；Fang 等 2010年；Nicastro 等 2005年；Fang 等 2007年、2002年），而这些检测可能会受到此类系统误差的影响。

Cash 统计量的开发，是为了应对新型整数计数天文 X 射线数据的出现，并将其作为泊松数据的首选最大似然拟合统计量（例如 Cash 1976、1979）。在统计学的其他领域中，该统计量通常被称为泊松对数似然偏差（例如 Cameron 与 Trivedi 2013）。最近，作者还开发了一种新的半解析方法，用于获得计数数据线性回归的最佳拟合参数；此前，这一问题只能通过数值方法求解（Bonamente 与 Spence（2022））。本文第 4.5 节概述了这些方法的典型实现，预计对大多数应用而言都很容易实施。鉴于 C–stat 通常是对这些数据进行最大似然分析时适用的拟合统计量，因此，本文介绍的方法为 X 射线天文学家（或任何其他计数数据分析人员）提供了另一种工具，使其能够使用 *Cash*统计量分析数据。

数据可用性

与 *XMM–Newton* 1ES 源相关的 X 射线数据可通过 heasarc.gsfc.nasa.gov/cgi- bin/W3Browse获取。光谱等处理后的数据可向作者索取。用于实现和复现本文主要结果（即过度离散 χ^2 分布（28）及表 1 中该分布的临界值）的 Python数值程序可在以下网址获取：

`https://github.com/bonamem/overdispersedChi2.`

参考文献

Ahoranta J., Finoguenov A., Bonamente M., Tilton E., Wijers N., Muzahid S., Schaye J., 2021,《天文学与天体物理学》, 656, A107 Ahoranta J. et al., 2020年,《天文学与天体物理学》, 634, A106 Anderson T. W., Darling D. A., 1952,《数学统计年刊》, 23, 193 Arnaud K. A., 1996, 收录于第五届天文数据分析软件与系统会议, Jacoby G. H., Barnes J., 编，第 101 卷，第 17 页 Baker S., Cousins R. D., 1984,《核仪器与物理研究方法》, 221, 437 Bevington P. R., Robinson D. K., 2003, 物理科学中的数据约简与误差分析。麦格劳-希尔，第三版 Bishop Y., Fienberg S., Holland P., 1975, 离散多元分析：理论与实践：Yvonne M.M. Bishop、Stephen E. Fienberg 和 Paul W. Holland。麻省理工学院出版社 Bonamente M., 2019年,《应用统计学杂志》，46，1129 Bonamente M., 2020年，《应用统计学杂志》，47，2044 Bonamente M., 2022年，《科学数据的统计与分析》。施普林格，物理学研究生教材，第三版 Bonamente M., Joy M. K., Lieu R., 2003,《天体物理学杂志》，585，722 Bonamente M., Nevalainen J., Tilton E., Liivamägi J., Tempel E., Heinämäki P., Fang T., 2016, 皇家天文学会月报，457，4236 Bonamente M., Spence D., 2022年，《应用统计学杂志》，49，522 Cameron A. C., Trivedi P. K., 1986,《应用计量经济学杂志》，1，29 Cameron C., Trivedi P., 2013, 计数数据回归分析（第二版）。剑桥大学出版社 Cash W., 1976,《天文学与天体物理学》，52，307 Cash W., 1979,《天体物理学杂志》，228，939 Consul P. C., Jain G. C., 1973,《技术计量学》，15，791 Conway T. R., Maxwell W. L., 1962,《工业工程学杂志》，12，132 Cramer H., 1946, 数理统计方法。普林斯顿大学出版社，普林斯顿 Dean C., Lawless J. F., 1989,《美国统计学会杂志》，84，467 Dean C. B., 1992,《美国统计学会杂志》，87，451 Famoye F., 1993,《统计学通讯：理论与方法》，22，1335 Fang T., Buote D. A., Humphrey P. J., Canizares C. R., Zappacosta L., Maiolino R., Tagliaferri G., Gastaldello F., 2010,《天体物理学杂志》，714，1715 Fang T., Canizares C. R., Yao Y., 2007,《天体物理学杂志》，670，992 Fang T., Marshall H. L., Lee J. C., Davis D. S., Canizares C. R., 2002,《天体物理学杂志》，572，L127 Fisher R., 1922,《皇家统计学会杂志》，85，87 Fisher R., 1934, 研究工作者的统计方法，第五版。Oliver and Boyd, 爱丁堡 Giacconi R., Kellogg E., Gorenstein P., Gursky H., Tananbaum H., 1971,《天体物理学杂志》，165，L27 Gourieroux C., Monfort A., Trognon A., 1984a,《计量经济学》，52，701 Gourieroux C., Monfort A., Trognon A., 1984b,《计量经济学》，52，681 Greenwood P., Nikulin M., 1996, 卡方检验指南。威利 Hilbe J. M., 2011, 负二项回归，第二版。剑桥大学出版社 Hilbe J. M., 2014, 计数数据建模。剑桥大学出版社 Humphrey P. J., Liu W., Buote D. A., 2009,《天体物理学杂志》，693，822 Kaastra J. S., 2017,《天文学与天体物理学》，605，A51 Kaastra J. S., Mewe R., Nieuwenhuijzen H., 1996, 收录于《天体与实验室等离子体的紫外和X射线光谱学》，Yamashita K., Watan- abe T., 编，第411–414页 Kaastra J. S., Werner N., Herder J. W. A. d., Paerels F. B. S., de Plaa J., Rasmussen A. P., de Vries C. P., 2006,《天体物理学杂志》，652，189 Kolmogorov A., 1933,《意大利精算师学会杂志》，4，1 Kovács O. E., Bogdán Á., Smith R. K., Kraft R. P., Forman W. R., 2019,《天体物理学杂志》，872，83 Lampton M., Margon B., Bowyer S., 1976,《天体物理学杂志》，208，177 Lee H. et al., 2011,《天体物理学杂志》，731，126 Marshall H. L. et al., 2021,《天文学杂志》，162，254 McCullagh P., 1984,《生物计量学》，71，461 McCullagh P., 1986,《美国统计学会杂志》，81，104 McCullagh P., Nelder J., 1989, 广义线性模型。Chapman & Hall/CRC，第二版 Nevalainen J. et al., 2019,《天文学与天体物理学》，621，A88 Nicastro F. et al., 2018,《自然》，558，406 Nicastro F. et al., 2005,《天体物理学杂志》，629，700 Protassov R., van Dyk D. A., Connors A., Kashyap V. L., Siemiginowska A., 2002,《天体物理学杂志》，571，545 Ren B., Fang T., Buote D. A., 2014,《天体物理学杂志》，782，L6 Rothschild R. et al., 1979,《空间科学仪器》，4，269 Sellers K. F., Shmueli G., 2010,《应用统计年刊》，4，943 Shmueli G., Minka T. P., Kadane J. B., Borle S., Boatwright P., 2005,《皇家统计学会杂志：C辑（应用统计）》，54，127

Spence D., Bonamente M.,Nevalainen J., Tuominen T., Ahoranta J., Liu W., Wijers N., 2023年，皇家天文学会月报，已投稿,
0 Stephens M. A., 1974,《美国统计学会杂志》，69，730
Wasserstein R. L., Lazar N. A., 2016, 美国统计学家，70，129
Wilks S. S., 1938, 数学统计年刊，9，60 Xu J. et al., 2014,《天体物理学杂志》，794，97

本文根据作者准备的 TEX/LATEX 文件进行排版。