arXiv:2603.03168v1 [physics.data-an] 3 Mar 2026

数据展开：从问题表述到结果评估

Nikolay D. Gagunashvili

冰岛大学，*Sæmundargata 2*，*101* 雷克雅未克，冰岛

摘要

粒子物理学、核物理学、粒子天体物理学和辐射防护剂量学中的实验数据，是通过由复杂的传感器、电子设备和软件系统构成的实验设施采集的。测得的谱或截面被视为概率密度函数（PDF），由于分辨率、偏差和效率等因素，它们与真实的概率密度函数存在差异。反卷积被视为估计未知真实概率密度函数的一种过程。要检验理论模型、比较不同实验的结果，以及综合各类研究工作的结果，就必须可靠地估计真实概率密度函数。

为此，可以采用外部和内部质量评估方法。在某些情况下，存在用于评估反卷积质量的外部准则。一个典型例子是模糊图像的反卷积，此时复原图像的清晰度可作为质量指标。然而，定义这类外部准则可能颇具挑战性，尤其是在此前从未进行过相关测量的情况下。

本文讨论了多种不依赖外部信息来评估结果质量的内部准则，以及影响展开分布质量的因素。

关键词： 概率密度函数估计、反卷积问题、系统辨识、质量评估标准

电子邮件地址：nikolay@hi.is (Nikolay D. Gagunashvili)

1. 引言

测量结果由一个实验装置产生，该装置由一套复杂的传感器、电子设备和软件构成。由于实验装置的分辨率、记录效率及其他因素，测得的概率密度函数（PDF） $f(y)$与真实的概率密度函数（PDF） $\phi(x)$有所不同。

作为一个典型例子，粒子穿过探测器材料后，其运动学参数中会加入噪声分量。在某些情况下，粒子可能因能量损失而完全未被记录。

估计未知概率密度函数 $\phi(x)$的问题——该函数可能表示频谱或微分截面——是粒子物理学、核物理学、粒子天体物理学和辐射防护剂量学分析中的主要课题之一。这是数据分析中的常见部分，传统上称为反卷积。计算机建模常用于研究记录过程。

如 [1, 2], 所述，待分析的数据由两组随机变量表示:

1. 测量数据样本是一组 n 独立同分布（IID）随机变量:

$$y_1, y_2, \ldots, y_n \tag{1}$$

其概率密度函数（PDF）为 $f(y)$。

- 真实概率密度函数 $\phi(x)$ 未知。
- 测得的概率密度函数 $f(y)$ 也未知，但可以进行估计。

2. 模拟数据样本是由以下 IID 随机变量对组成的集合： k

$$x_1^s, y_1^s;\ x_2^s, y_2^s;\ \ldots\ x_k^s, y_k^s \tag{2}$$

其概率密度函数分别为 $\phi^s(x)$ 和 $f^s(y)$。

- 生成概率密度函数 $\phi^s(x)$ 是真实概率密度函数 $\phi(x)$ 的一种类比。在常见情况下， $\phi^s(x) \neq \phi(x)$，并且可以估计概率密度函数 $\phi^s(x)$。

- 重建概率密度函数 $f^s(x)$ 是测量概率密度函数$f(x)$ 的模拟量，由此可以估计概率密度函数 $f^s(x)$。

模拟样本用于为测量系统构建数学模型。

针对这种情况，常用的模型是Fredholm积分方程，它描述了记为 $f(y)$ 的测量概率密度函数与真实概率密度函数 $\phi(x)$之间的关系。

$$\int_{-\infty}^{+\infty} R(x,y)A(x)\phi(x)dx = f(y), \tag{3}$$

其中 $A(x)$是记录具有特征 x的事件的概率（接受度）； $R(x,y)$是获得 y 而非 x 的概率（实验分辨率）。在 $R(x,y) = R(y-x)$的特殊情况下，形式解可用傅里叶变换表示如下：

$$\phi(x) = \frac{1}{A(x)} \int_{-\infty}^{+\infty} e^{ipx} \frac{\widetilde{f}(p)}{\widetilde{R}(p)} dp, \tag{4}$$

其中符号 $\sim$ 表示相应函数的傅里叶变换。

如果核函数的傅里叶变换 $\widetilde{R}(p)$在频域的高频区域消失，这意味着测量概率密度函数 $f(y)$不包含关于 $\phi(x)$高频行为的信息。形式上，无法在该频域内由方程(4)确定 $\phi(x)$，正如在接受度 $A(x)$等于零的区域无法定义 $\phi(x)$一样。

因此，弗雷德霍姆积分方程3)是不适定的。正则化通过限制容许解空间来应对上述困难，从而将不适定问题转化为适定问题。

弗雷德霍姆方程的例子表明，测得的概率密度函数 $f(y)$对真实概率密度函数 $\phi(x)$具有线性依赖。尽管这种线性依赖是合理的，但在许多实际情况下，仍需仔细研究，以确定这种依赖究竟是线性的还是非线性的。

2. 展开程序的质量评估

外部准则和内部准则均用于评估展开程序的质量。在某些情况下，外部准则是可用的，为评估该程序的可靠性提供了框架。然而，在实验物理学中，定义外部准则可能具有挑战性，

尤其是在不存在先验测量时。在此类情形下，基于模型的假设往往起到指导作用。相反，当外部参考不可用时，用于评估展开结果质量的内部准则就变得至关重要。

下文给出的数据展开过程的质量准则是评估过程不可或缺的一部分。值得注意的是，这些准则不仅有助于确定展开算法参数的最优值，也便于比较不同算法。

展开分布是 $\phi(x)$真实分布的估计量，因此与其存在差异。为了开发一种方法，使其产生尽可能接近 $\phi(x)$的^估计量 $\phi(x)$，必须定义适当的度量来量化这种差异。

2.1. 均方积分误差

^衡量估计量 $\phi(x)$准确性的一种常用度量是均方积分误差（MISE）[3, 4, 5, 6]。为定义MISE，我们首先引入积分平方误差（ISE）：

$$\mathrm{ISE}(\hat{\phi}(x)) = \int \left(\hat{\phi}(x) - \phi(x)\right)^2 dx.$$

于是，MISE可以写为

$$\mathrm{MISE}(\hat{\phi}(x)) = \mathbf{E}(\mathrm{ISE}) = \mathbf{E}\left[\int \left(\hat{\phi}(x) - \phi(x)\right)^2 dx\right].$$

也可以给出MISE的等价公式：

$$\mathrm{MISE}(\hat{\phi}(x)) = \int \mathbf{E}\left[\left(\hat{\phi}(x) - \phi(x)\right)^2\right] dx.$$

将公式代入

$$\mathbf{E}\left[\left(\hat{\phi}(x) - \phi(x)\right)^2\right] = \left(\mathrm{Bias}\left[\hat{\phi}(x)\right]\right)^2 + \mathrm{Var}\left(\hat{\phi}(x)\right)$$

上述表达式中，可得到 MISE 的一个有用表示：

$$\mathrm{MISE}(\hat{\phi}(x)) = \int \left[\left(\mathrm{Bias}\left[\hat{\phi}(x)\right]\right)^2 + \mathrm{Var}\left(\hat{\phi}(x)\right)\right] dx.$$

使均方积分误差达到最小的展开分布，在偏差和方差之间实现了最佳权衡。因此，应优先选择 MISE 值最低的算法。

下面我们将讨论通过阶跃函数近似表示的展开分布的内部质量准则。

将 m 步的位置定义为 $\{a_1, a_2, ..., a_{m+1}\}$，将展开分布的概率密度函数估计定义为

$$\hat{\phi}(x) = \frac{\hat{\phi}_i}{a_{i+1} - a_i} \text{ for } a_i \leq x < a_{i+1}, \tag{5}$$

ˆ其中 ϕ_i 是.....的估计量

$$\phi_i = \int_{a_i}^{a_{i+1}} \phi(x)dx. \tag{6}$$

在这种情况下，MISE 表示为

$$\begin{aligned} \text{MISE} &= \int_{a_1}^{a_{m+1}} \mathbf{E}\left[\hat{\phi}(x)^2\right]dx - 2\int_{a_1}^{a_{m+1}} \phi(x)\mathbf{E}[\hat{\phi}(x)]dx + \int_{a_1}^{a_{m+1}} \phi(x)^2 dx \\ &= \sum_{i=1}^{m} \frac{\mathbf{E}\hat{\phi}_i^{\,2}}{a_{i+1} - a_i} - 2\sum_{i=1}^{m} \frac{\mathbf{E}\hat{\phi}_i}{a_{i+1} - a_i}\phi_i + \int_{a_1}^{a_{m+1}} \phi(x)^2 dx. \end{aligned} \tag{7}$$

ˆ表达式 (7) 中的最后一项与估计量 ϕ_i 无关；它是一个常数。因此，使 MISE 最小化的参数和因子的选择不取决于方程的这一部分。

MISE 的另一种广为人知且实用的表示形式为

$$\text{MISE} = \sum_{i=1}^{m} \frac{1}{a_{i+1} - a_i}\mathbf{E}\Big[(\hat{\phi}_i - \phi_i)^2\Big] + \sum_{i=1}^{m} \int_{a_i}^{a_{i+1}} \left(\phi(x) - \frac{\phi_i}{a_{i+1} - a_i}\right)^2 dx \tag{8}$$

或者等价地,

$$\text{MISE} = \sum_{i=1}^{m} \frac{1}{a_{i+1} - a_i}\left[\big(\text{Bias}[\hat{\phi}_i]\big)^2 + \text{Var}(\hat{\phi}_i)\right] + \sum_{i=1}^{m} \int_{a_i}^{a_{i+1}} \left(\phi(x) - \frac{\phi_i}{a_{i+1} - a_i}\right)^2 dx$$

由于

$$\mathbf{E}\Big[(\hat{\phi}_i - \phi_i)^2\Big] = (\mathbf{E}\hat{\phi}_i - \phi_i)^2 + \text{Var}(\hat{\phi}_i) = \big(\text{Bias}[\hat{\phi}_i]\big)^2 + \text{Var}(\hat{\phi}_i).$$

这里，只有第一项受到正则化的影响。第二项定义了与分箱相关的偏差，并表示展开分布的阶跃函数表示所固有的最小偏差。

*2.2. ISE*方差

估计的变异性可定义为

$$\mathrm{Var}(\mathrm{ISE}) = \mathbf{E}\big[\mathrm{ISE}^2\big] - \Big(\mathbf{E}[\mathrm{ISE}]\Big)^2$$

Var(ISE) 值最低的算法更可取，因为它能为展开问题提供更稳定的解。

2.3. 最小条件数

衡量反卷积过程质量的另一项指标是 $\phi_i/(a_{i+1} - a_i)$ 相关矩阵的条件数。ϕ_i 概率的估计量满足以下方程:

$$\sum_{i=1}^{m} \hat{\phi}_i = 1 \tag{9}$$

因此，相关矩阵通常接近奇异。为解决这一问题，建议在排除一个分箱后，确定相关矩阵的最小条件数（MCN）。

$$\mathrm{MCN} = \underset{k}{\mathrm{argmin}}[\mathrm{COND}(C_{-k})], \tag{10}$$

其中，C_{-k} 是移除分箱 k 后的完整相关矩阵。相关矩阵的条件数衡量了展开程序的数值稳定性及其对小扰动的敏感度。特别是，MCN 值最低的算法更可取。

2.4. 均方误差

衡量展开分布 [7] 准确度的一种常用指标是均方误差（MSE）:

$$\mathrm{MSE} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{E}\left(\hat{\phi}_i - \phi_i\right)^2 = \frac{1}{m}\sum_{i=1}^{m}\left[\left(\mathrm{Bias}\big[\hat{\phi}_i\big]\right)^2 + \mathrm{Var}(\hat{\phi}_i)\right] \tag{11}$$

MSE 的主要缺点是无法用于比较采用不同分箱方案得到的两个展开分布的准确性。然而，在为展开程序选择最优分箱方案时，这种比较十分重要。

2.5. 覆盖概率

覆盖概率 [8] 定义为

$$P_{\text{cov}} = \frac{1}{m}\sum_{i=1}^{m} P\left(\hat{\phi}_i - \hat{\sigma}_{\hat{\phi}_i} < \phi_i < \hat{\phi}_i + \hat{\sigma}_{\hat{\phi}_i}\right), \tag{12}$$

ˆ，其中 ˆ$\sigma_{\hat{\phi}_i}$ 是 ϕ_i的估计不确定性（例如标准差）。 P_{cov} 的主要缺点是，它也无法用于比较采用不同分箱方案得到的两个展开分布的准确性。在选择更优的分箱方案用于展开程序时，这种比较十分重要。

2.6. 后分辨率

如果能够估算后分辨率，就可以了解有效分辨率函数相较于实验装置固有分辨率的改善情况。 [7, 9]中给出了估算后分辨率函数的示例。

3. 影响内部和外部质量标准的因素和参数。

展开算法中的不同因素和参数会影响展开分布的质量，具体定义见上一节介绍的质量评估方法。这些因素和参数列举如下：

1. 测量系统是线性还是非线性的。2. 用于计算响应矩阵 R的概率分布函数 $\phi^s(x)$，以及它是否接近真实分布$\phi(x)$。3. 用于计算响应矩阵 R 的方法（系统辨识），可以是：

- 传统方法 [10], 如果模拟中使用的生成分布 $\phi^s(x)$ 与真实分布 $\phi(x)$ 不一致，可能会导致展开结果出现偏差。
- 文献 [1]中提出的一种系统辨识方法 [11] 。

4. 用于创建测量系统数学模型的模拟事件数量 k 。

5. 实验事例的数量 n。6. 实验（测量）分布中的区间数量。7. 展开分布中的区间数量 m 。8. 分箱类型：

- 等距分箱。
- 非等距分箱，可使用 k 均值聚类方法 [12] 定义；此外，在多维情况下，还可使用 Voronoi 方法 [13] 定义。9. 展开算法的正则化参数。在 Richardson-Lucy 方法中， [14, 15] 指迭代次数。10. 展开分布的初始猜测。当实验统计量较小时，这一点对 Richardson-Lucy 方法尤其重要。

MISE、Var(ISE) 和 MCN 可用作内部质量准则，用于比较在其所依赖的因素和参数发生变化时，不同的展开算法。该方法首次应用于[1, 2]。

将 MSN 和 P_{cov} 用作内部质量准则的适用范围相当有限。如 [10] 所示，当采用相同的等距分箱时，可以比较不同的算法。然而，若两个不同的算法以两种不同的分箱方案给出展开分布，则无法对它们进行比较，正如 [2]中的做法。

4. 结论

反卷积可视为一种估计未知概率密度函数的过程。影响展开结果质量的主要因素和参数已得到识别并加以表述。估计质量通过均方积分误差（MISE）、积分平方误差的方差（Var(ISE)），以及估计分布相关矩阵的最小条件数（MCN）来表征。将展开分布与相应的展开结果质量评估一同呈现，能够显著提升对实验数据的物理解释。

参考文献

参考文献

[1] N. D. Gagunashvili。《数据展开的均方积分误差优化》。《计算物理通讯》，308:109473，2025。[2] N. D. Gagunashvili。《采用均方积分误差优化的Richardson–Lucy反卷积与数据展开的比较分析》。《计算物理通讯》，318:109894，2026。[3] B. W. Silverman。《统计学与数据分析中的密度估计》。Chapman and Hall，1986。[4] A. B. Tsybakov。《非参数估计导论》。施普林格，2009。[5] D. W. Scott。《多元密度估计：理论、实践与可视化》。约翰·威立父子公司，2015。[6] M. P. Wand和M. C. Jones。《核平滑》。Chapman & Hall，1995。[7] V. P. Zhigunov。作为逆问题的分辨函数改进。《核仪器与方法*A*辑》，322(1-2):183，1983。[8] M. Kuusela和V. M. Panaretos。基本粒子谱的统计展开：经验贝叶斯估计与偏差校正的不确定性量化。《应用统计年刊》，9:1671，2015。[9] M. Schmelling。采用离散值惩罚函数的正则化展开。《核仪器与方法》，1056:168566，2023。[10] L. Brenner、R. Balasubramanian、C. Burgard、W. Verkerke、G. Cowan、P. Verschuuren和V. Croft。使用RooFitUnfold比较反卷积方法。《国际现代物理学杂志*A*辑》，35(24):2050145，2020。[11] D. Graupe。《系统识别》。Krieger and Huntington„ 1976。[12] P. N. Tan、M. Steinbach、A. Karpatne和V. Kumar。《数据挖掘导论》。培生，2018。[13] G. Voronoi。连续参数在二次型理论中的新应用。《纯粹与应用数学杂志》，133:97，1907。

[14] W. H. Richardson. 图像复原的基于贝叶斯的迭代方法。《美国光学学会杂志》，62(1):55–59，1972。

[15] L. B. Lucy. 用于校正观测分布的迭代技术。《天文学杂志》，79:745，1974。