统计天体测量

刘牛*

2022 年 12 月 15 日

摘要

本章为《Astrometry for Astrophysics》的第 16 章，由 Anthony G. A. Brown 编写。统计天体测量（Statistical astrometry）指的是从一组恒星或其他源样本中推断天体物理量，通常得到的是该天体物理量的平均值或对其分布的描述，应用广阔，包括：光度定标、星团成员星证认、分离银河系结构成分、探测银晕中低对比度子结构。天体测量星表给出的大量视差测量结果使这些研究变得很有意思。然而，有很多因素使得天体物理量的推断存在偏倚，从而使对天体测量数据的分析变得复杂。本章重点讨论这些复杂因素，以期为在统计研究中最优使用天体测量数据提供指导。

1 使天体测量数据分析复杂化的效应

- 样本不完备性和选择效应（Completeness and selection effects）

  从一本天体测量星表中选择的任何一个样本都存在数据不完备性和选择效应。这两个问题都与天体测量巡天特性和样本的选取方式相关。此外，它们也可能与天区位置或目标源特性相关。因此，所选取的样本几乎不可能代表样本源在天体物理参数空间的真实分布。忽略这一点可能会导致统计推断得到有偏结果。

- 相关误差（Correlated errors）

  一般而言，一个给定天体的天体测量参数误差在统计上是相互不独立的。此外，在不同天体间这些误差也可能相关。忽略这些相关性可能会使所导出的天体物理量分布存在虚假特征。

- 与天区位置相关的系统误差（Systematics as a function of sky position）天体测量观测方式特征会反映在误差和天区相关性的系统变化中，如照相底片巡天中的系统区域误差（systematic zonal error）、HIPPARCOS 和 *Gaia* 项目扫描模式所引入的系统误差。考虑这些系统误差对基于遍及大面积天区的源样本的研究尤为重要。

- 估计天体物理量（Estimating astrophysical quantities）

  在估计某一类天体样本的天体物理量时，首先要从每一个天体的天体测量资料中计算这些天体物理量，再分析这些天体物理量在参数空间中的分布。这些天体物理量包括距离、速度、光度、角动量

*电子邮箱：niu.liu@nju.edu.cn

等。但是，实际测量数据并非是与这些天体物理参数直接相关。如直接的测量量不是天体的距离，而是由于地球绕日运动引起的视差位移。许多天体物理量是天体测量参数的非线性函数。例如，恒星的距离、绝对星等和角动量分别是 $\frac{1}{\varpi}$、$\log \varpi$ 和 $\frac{1}{\varpi^2}$ 的函数。因此，利用天体测量数据简单估计天体物理参数会产生错误结果。唯一可靠的方法是数据的正向建模。

尽管不存在完美解决这些问题的通用方案，不过有许多 “绝佳实践案例” 值得借鉴。

2 了解你的星表（Know your catalog）

为了完全考虑观测误差、完备性和选择效应，理解天体测量星表的内容很重要，至少要对星表的编制过程有基本了解。

以 HIPPARCOS 星表为例，它包含 11 多万颗源（多数为单星）、2500 多个双星和多星系统、60 多颗太阳系内天体和 1 颗类星体。每颗源有 5 参数解和其他附加和补充信息，如天体测量参数之间的相关性、星等、颜色等。

在 HIPPARCOS 星表中，恒星的天球分布明显不同于在光学波段上恒星的真实分布。HIPPARCOS 的观测方式是基于输入星表设置的，因而 HIPPARCOS 恒星天球分布表现出与输入星表相关的特征。最显著的特征是 Gould 带的近邻 OB 星协。因此，在从 HIPPARCOS 星表中选取一组恒星来导出某种平均物理量（如光度函数）时需要慎重处理这些选择效应。*Gaia* 和 HST GSC 星表不会包含这些明确的选择效应；然而，由于不同的完备性，这些星表可能会含有微笑的选择效应，这些效应与天空中源的密度、成像分辨率和探测算法性能的相关。

HIPPARCOS 星表另一个值得关注的特征是赤经坐标、赤经自行和视差的精确度与黄纬强相关，这可以从 HIPPARCOS 的天空扫描方式角度来理解。HIPPARCOS 总是以大角度穿过黄道面，这意味着对于位于黄道面附近的源，黄经方向上的任何位移在测量方向上的投影非常小。这导致黄经坐标和黄经自行存在较大误差，转换到赤道坐标系后就表现为赤经坐标和赤经自行的较大误差。视差误差与黄纬的相关性来源于另一个事实：对于在黄道面附近的源，视差位移主要在黄经方向上。

最后，对 HIPPARCOS 星表的统计显示，天体测量参数间的相关性在天球上系统性地变化。一些系统误差是由于将测量实际发生的黄道坐标系转换到最终星表使用的赤道坐标系导致的。大部分系统误差都与扫描模式的细节、观测的时间分布以及观测相对于太阳的位置的分布相关。

类似的与天球坐标相关的系统误差也将会出现在 *Gaia* 和其他大型巡天（如 LSST、Pan-Starrs、GSC 星表等）中。意识到并理解这些因素的起因是很重要的，尤其是在数据中搜寻微弱的大尺度特征时。

3 处理相关误差（Dealing with correlated errors）

一般而言，一颗给定源的天体测量参数是通过用源的运动模型对天体测量数据的数值拟合得到的。这个模型通常是恒星运动的标准模型，但也可能更加复杂，如双星系统的轨道运动模型。这种拟合给出的是对天体测量参数矢量 $\mathbf{a}$ 及其协方差矩阵 $\mathbf{C}_\mathbf{a}$ 的估计。记 $\mathbf{a}^{\text{true}}$ 和 $\mathbf{a}^{\text{est}}$ 分别为 $\mathbf{a}$ 矢量的真值和估计值，E[**x**]

为 **x** 矢量的数学期望，则协方差矩阵可表示为

$$\mathbf{C_a} = \mathrm{E}[(\mathbf{a}^{\mathrm{est}} - \mathbf{a}^{\mathrm{true}})(\mathbf{a}^{\mathrm{est}} - \mathbf{a}^{\mathrm{true}})'] = \mathrm{E}[\mathbf{\Delta a \Delta a}']. \tag{1}$$

其元素 c_{ij} 由下式给出

$$c_{ii} = \sigma_i^2, c_{ij} = \rho_{i,j}\sigma_i\sigma_j. \tag{2}$$

其中，σ_i 为 a_i 分量的标准误差，$\rho_{i,j}$ 为 a_i 分量与 a_j 分量之间的相关系数。

相关系数一般不为零，因为 **a** 的分量估计都基于同一组测量数据得到的。因此，只依靠标准误差来处理天体测量参数的不确定度是不够的。可以通过误差棒的一般形式来考虑协方差矩阵。在真值附件的置信区间可以用下式中的量来描述：

$$z = \Delta\mathbf{a}'\mathbf{C_a^{-1}}\Delta\mathbf{a} \tag{3}$$

z 服从 χ^2_ν 分布，其中 ν 为自由度，一般等于 **a** 的维数。如果矢量 **b** 通过 **a** 的某种转换关系 **f**(**a**) 得到，则 **b** 的协方差矩阵为

$$\mathbf{C_b} = \mathbf{J}\mathbf{C_a}\mathbf{J}' \tag{4}$$

其中，**J** 为转换关系的雅可比矩阵，其元素为

$$J_{ij} = \frac{\partial f_i}{\partial a_j} \tag{5}$$

由此，我们可以计算由观测天体测量参数导出的任意一组变量的协方差矩阵。一个重要的应用就是天体测量参数误差随时间的传递。

在基于天体测量数据的天体物理研究中使用协方差矩阵，一个很好的例子就是 Perryman 等人在 1998 年利用 HIPPARCOS 数据研究毕星团（Hyades cluster）。星团成员星都是基于它们在空间中以共同速度运动（不考虑较小的速度弥散）的假设下选出。单颗恒星的空间速度由天体测量和视向速度计算而得，其协方差矩阵用于判断该恒星的空间运动是否与星团的平均运动一致。随后，对相对于平均速度的速度残差的分析表明，存在具有特定轴方向的速度椭球。此外，速度残差矢量似乎表明速度场中存在剪切或旋转。这些特性将影响星团动力学研究。然而，Perryman 等人 1998 年的工作表明，这些特征都可以被解释为天体测量参数的相关性与毕星团相对于太阳的空间位置和速度两个因素相结合的结果。

天体测量参数的相关性（以及不同源天体测量参数之间的相关性）不能被忽视，需要在数据分析中考虑。

4 通过反演的天体物理参数估算

天体测量巡天计划的主要目的是测量视差，这是对太阳系外天体唯一不需要假定待测源内禀特性的距离测量手段。因此，视差对估计天体物理参数（如恒星光度和在相空间中的恒星分布函数）非常重要。如前所述，最直观的方法就是通过天体测量观测量来反演距离或绝对星等。然而，这种方法会导致错误结果。这里借助利用三角视差数据校准恒星光度（绝对星等）来说明其中的问题。

4.1 简单光度校准中的统计偏差

假设对某种特定光度型的所有恒星都测量了视差，且这些恒星都在以太阳为中心的空间内均匀分布。待研究的问题是：这一类恒星的平均绝对星等 μ_M 以及相对于平均值的方差 σ_M^2 分别是多少？巡天计划得到的数据是视差测量值 ϖ_0、视星等 m 和它们各自的观测误差（分别记为 σ_ϖ 和 σ_m）。观测误差与视星等相关，恒星越亮，观测误差就越小。回答上述问题最直接的方法就是用下式估计第 i 颗星的绝对星等：

$$\tilde{M}_i = m_i + 5\log\varpi_{0,i} + 5 \tag{6}$$

再从算得的 $\tilde{M}_i$ 的分布中估计 μ_M 和 σ_M^2。在最实际的情形下，存在下述三个问题，将会导致错误的结果。

- 转换偏倚（Transformation bias）

  由天体测量观测估计得到的视差测量值 ϖ_0 总是包含误差且服从某种分布，通常假设其在视差真值 ϖ 附近服从高斯分布，即

  $$P(\varpi_0 \mid \varpi) = \frac{1}{\sigma_\varpi\sqrt{2\pi}} e^{-(\varpi_0-\varpi)^2/2\sigma_\varpi^2} \tag{7}$$

  其中，$P(\varpi_0 \mid \varpi)$ 指的是在给定视差真值 ϖ 的情况下，观测得到 ϖ_0 的概率。在没有系统误差时，ϖ_0 是 ϖ 的无偏估计，即

  $$E[\varpi_0] = \varpi \tag{8}$$

  但对于距离来说却不是：

  $$E\left[\frac{1}{\varpi_0}\right] \neq \frac{1}{\varpi} \tag{9}$$

  对绝对星等来说也是如此：

  $$E[\log\varpi_0] \neq \log\varpi \tag{10}$$

  可以对距离和绝对星等估计中的偏倚建模，不过该模型包含对视差真值的积分，并在较小视差和负视差时失效。有很多工作尝试改正这一偏倚，但它们都依赖于某种消除零和负视差以便计算积分的技巧。这些做法高度可疑，无法解决偏倚改正对视差真值（永远不可能知晓）的依赖问题，在现实情况下具有非常大的方差（即计算得到的偏倚改正值本身就不准确），使得这些“改正”毫无意义。

- 卢茨—凯克尔效应（The Lutz-Kelker effect）

  上述转换偏倚之所以存在，是因为在利用观测视差推断距离或绝对星等时只使用了恒星的观测视差及其误差和视星等。可以通过已掌握的关于样本的其他信息来改进这种方法，如

  - 巡天的星等极限，可以限制 M 的合理取值范围并在给定 M 时为 ϖ 设置下限；
  - 恒星的空间分布，已知是有限的，在太阳近邻可以近似为均匀分布；
  - 恒星的类型，可以为 M 的数值提供先验信息

  Lutz 和 Kelker 在 1973 年提出了这一思路，并写了一篇非常有影响力的论文来展示对基于三角视差光度校准的统计偏倚的改正。该改正通过采用部分贝叶斯方法并引入恒星空间分布的先验信息（即

假设恒星的空间密度恒定）推导出。在给定测量视差 ϖ_0 的情况下，恒星视差真值为 ϖ 的概率为

$$P(\varpi \mid \varpi_0) \propto P(\varpi_0 \mid \varpi)\, \varpi^{-4} \tag{11}$$

上述方程表达了这样一个事实：恒星数量随距离快速增长。这使得真实视差值比观测值小的可能性很高（即更多视差小恒星的观测视差值过大，反之亦然。理解这个说法要把上面的式子反过来看，即看右边的概率代表观测到某一结果的可能性）。他们使用上述视差真值概率分布的表达式来推导出绝对星等的真实值与观测值之差 $\Delta M = M - M_0$ 的数学期望。这就是所谓的" 卢茨—凯克尔效应"。他们给出的结果通常被应用于由视差计算得到的绝对星等。然而，这些改正只在满足下述条件时才适用：

- 恒星在空间中的分布时均匀的；
- 样本在一定的距离范围内是完备的；
- 相对视差误差是恒定的。

另外，当恒星的光度函数在某一绝对星等范围内分布均匀时，只能通过假设 ϖ/ϖ_0 存在下限且要将样本限制在 $\sigma_\varpi/\varpi_0 < 0.175$ 来计算改正值。这些需求使得 Lutz-Kelker 改正在很大程度上是毫无意义的。这一问题在该论文发表后很快就被发现，其他学者也试图通过引入对光度函数的附加假设或放宽对恒星空间分布的假设并考虑自行来推导出更加可靠的偏倚改正。

- 马尔姆奎斯特偏倚（Malmquist bias）这种偏倚是由以下事实造成：随着样本距离增加，任何一个在星等上完备的样本将会系统性地包含更多亮星，这是由于对视亮度的选择和源绝对星等的内禀扩散结合起来引起的。在前述的光度校准问题中，样本的视星等极限会使得优先选择更近的或者本质上更亮的恒星。这将使得样本中的绝对星等和视差分布不具备代表性，从而引起 μ_M 估计中的系统误差，即"Malmquist bias"。正如卢茨—凯克尔效应，在对样本采取某种假设后，可以推导出对马尔姆奎斯特偏倚的 "改正"。但是，这些改正也存在如同卢茨—凯克尔改正一样的问题。

4.2 用其他方法最小化偏倚

由于前文讨论的直接方法存在诸多问题，可以尝试其他替代方法。其中一种思路就是将研究的问题重新表述，来避免视差的非线性变换。

例如，在研究造父变星周光关系的零点问题时，Feast 和 Catchpole 在 1997 年改写了周光关系，使得零点与视差存在线性关系，从而在不用担心计算绝对星等值时引入的转换偏倚对推导造父变星周光关系零点的影响。类似地，Arenou 和 Luri 在 1990 年提议使用 "基于天体测量的光度"（astrometry based luminosity）

$$a_V = \varpi_0 10^{0.2V-1} \tag{12}$$

（指数里面似乎是 0.2V+1）。这一恒星光度物理参数可用于建立不受卢茨—凯克尔偏倚影响的赫罗图，并且适用于负视差和相对视差误差较大的情形。然而，需要注意的是，诸如 a_V 之类的物理参数，尽管随视差线性变化，但是与 V 是非线性关系。这意味着使用新的非线性转换时偏倚依旧存在，即使是在 V 的相对误差远小于视差相对误差的情况下。此外，这种形式上的改写不能解决由样本选择效应引起的偏倚。

4.3 数据正向建模

前文提到的处理天体测量数据的思路主要是将天体测量观测量转换成天体物理参数，然后将得到的天体物理参数固定在转换值上。相应的" 改正项" 只能在极不现实的假设下才能计算，且只适应于移除负视差和小视差值的样本。这样截断数据的处理方式是非常浪费的，且会使结果的偏倚增大。此外，数据本身也存在一种内含在测量过程中的系统误差，与对测量结果的处理无关。而前文讨论的偏倚与数据本身无关，完全是由对数据处理不当引起的。即使是对没有系统误差的数据，偏倚问题依然存在。

保留数据原本的样子，通过数据建模来估计目标天体物理参数，可以完全避免偏倚问题。即由模型预言天体测量观测量，再将它们与实际的观测结果比对，从而确定模型参数的数值。这种方法的缺点是参数估计依赖于模型，不过正如前文所讨论的那样，直接方法也是依赖于模型的。这体现在偏倚改正值的推导需要引入大量对样本特性的假设上。在数据建模中也可以考虑样本特性和选择效应。作者以简单的贝叶斯光度校准为例来说明数据建模方法。

4.3.1 贝叶斯光度校准

观测数据形式同第 16.4 节描述的一样，不过所研究的问题变了：在给定观测量 $\mathbf{o} = \left\{\varpi_{\mathrm{o},i}, m_i\right\}$ $(i = 0, \ldots, N-1)$ 的情况下，μ_M 和 σ_M 最有可能的取值为多少？可以依据贝叶斯定理将该事件发生的概率表述出来：

$$P\left(\mu_M, \sigma_M, \mathbf{t} \mid \mathbf{o}\right) = \frac{P\left(\mathbf{o} \mid \mu_M, \sigma_M, \mathbf{t}\right) P\left(\mu_M, \sigma_M, \mathbf{t}\right)}{P(\mathbf{o})} \tag{13}$$

上式各项参数的含义为：

- $P\left(\mu_M, \sigma_M, \mathbf{t} \mid \mathbf{o}\right)$: 在给定观测数据 $\mathbf{o}$ 下，μ_M、σ_M 和每颗恒星的视差和绝对星等真值 $\mathbf{t} = \{\varpi_i, M_i\}$ 的联合分布;
- $P\left(\mathbf{o} \mid \mu_M, \sigma_M, \mathbf{t}\right)$：在给定观测量真值的情况下得到观测数据的概率，即似然函数;
- $P\left(\mu_M, \sigma_M, \mathbf{t}\right)$：观测量真值和光度分布模型参数的联合分布，代表的是对恒星空间分布和光度分布函数的先验信息;
- $P(\mathbf{o})$：获得观测数据的概率，在本问题中可被视为归一化常数.

估计 μ_M 和 σ_M 需要最大化后验概率 $P\left(\mu_M, \sigma_M, \mathbf{t} \mid \mathbf{o}\right)$。可以将上述方程改写为

$$\begin{aligned} &P\left(\mu_M, \sigma_M, \mathbf{t} \mid \mathbf{0}\right) \propto \\ &\prod_i \exp\left[-\frac{1}{2}\left(\frac{\varpi_{o,i} - w_i}{\sigma_{w,i}}\right)^2\right] \times \exp\left[-\frac{1}{2}\left(\frac{m_i - M_i + 5\log\varpi_i + 5}{\sigma_{m,i}}\right)^2\right] \\ &\times \varpi_i^{-4} \times \frac{1}{\sigma_M\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{M_i - \mu_M}{\sigma_M}\right)^2\right] P\left(\mu_M\right) P\left(\sigma_M\right) \end{aligned} \tag{14}$$

上式中的各项对应着不同的假设和先验信息

- 前两项：假设视差和视星等的测量误差是高斯型;
- ϖ_i^{-4}：假设恒星在以太阳为中心的空间内均匀分布;

- 第三个高斯函数：假设样本中恒星的光度分布函数为高斯函数;
- $P(\mu_M)$ 和 $P(\sigma_M)$：对 μ_M 和 σ_M 取值范围的先验信息。

接下来的工作就是根据上述方程算出后验似然函数，之后通过对 “nuisance parameters” $\mathbf{t}$ 进行积分，得到 (μ_M, σ_M) 的边缘分布函数，再依据这个分布确定估计值 $\widetilde{\mu_M}$ 和 $\widetilde{\sigma_M}$，如取平均值。问题在于 $P(\mu_M, \sigma_M, \mathbf{t} \mid \mathbf{o})$ 是个高维空间的函数，无法利用解析方法计算。此时需要借助数值方法来对后验分布进行采样，建立后验分布。常用的方法就是 Markov Chain Monte Carlo (MCMC) 方法。基本思路是通过可控的随机行走来遍历参数空间 $(\mu_M, \sigma_M, \mathbf{t})$ 进行采样。采样将生成参数点在参数空间中的分布，其密度正比于后验似然函数。从参数点的分布可以画出 μ_M 和 σ_M 的直方图，用来代表在对其他模型参数积分后，μ_M 和 σ_M 的后验概率分布函数。这反映了在考虑所有由观测误差和先验信息含糊导致的不确定性后得到的对光度函数参数的后验认知。

之后作者使用模拟数据作为算例，来演示该算法。

- 假设该巡天数据包含 $N = 400$ 颗同一类型的恒星;
- 恒星的绝对星等服从 $N(9, 0.49)$ 的分布，即 $\mu_M = 9$，$\sigma_M = 0.7$;
- 恒星在 1 pc 和 100 pc 距离范围内均匀分布，即 10 mas $\leq \varpi_i \leq$ 1000 mas ;
- 该巡天计划在一定距离上是完备的;
- 已知观测误差 $\sigma_{\varpi,i}$ 和 $\sigma_{m,i}$ 随视星等变化（根据光子噪声）且对于亮星包含校准本底;
- 假设我们已知恒星的分布是均匀的且已知视差真值分布的极值;
- 待估参数为 μ_M 和 σ_M;
- μ_M 的先验分布假设为 μ_M 在 −5 和 +16 之间均匀分布;
- σ_M 的先验分布假设为 σ_M^2 在 0.01 和 1.0 之间的概率密度正比于 $\dfrac{1}{\sigma_M^2}$;

MCMC 方法使用了 25 万个”burn-in” 步长生成一百万个迭代样本，并将每第 150 个样本存储下来作为观测数据。生成的观测视差包含小视差（小于 10 mas）和负视差，这些值也用于参数估计。得到的结果如表1。可见，截断数据会使估计值更加偏离真值。同时，施加 Lutz-Kelker 改正也会使估计值变小，远离真值。

当然上述方法也存在若干问题：

- 在实际测量中，视差的真实分布是不可能准确预知的。不过，可以通过引入更多参数来调整 $P(\varpi)$，如 $P(\varpi) = \varpi^{a+4}$ 并将视差的上下限设为未知量。
- 更严格地说，任何实际测量得到的样本总是包含选择效应的，如星等极限。这些选择效应也可以作为已知或含待估参数的选择函数，被纳入在正向模型中。

这里想传达的主要信息是基于概率推断的正向建模可以考虑天体测量数据中的复杂因素，包括负视差和视差相对误差较大的结果、（相对）观测误差的变化和相关误差。

Condition	μ_M	σ_M^2
True	9	0.49
All sample	8.96 ± 0.05	0.50 ± 0.06
Only positive parallaxes	8.8	3.2
$0 < \sigma_\varpi/\varpi_o \leq 0.175$	8.7	0.5

表 1: 贝叶斯光度估计结果。

4.4 天体测量数据建模的其他例子

Luri 等人 1996 年提出了一种基于最大似然函数数据建模的光度校准方法。该方法不仅使用了视星等和视差数据，也考虑了自行和视向速度。在模型方面，考虑了光度函数、运动学、恒星的空间分布、恒星类型、观测误差和选择效应。所有视差数据都被使用了。最重要的一点是使用所有可用数据。尽管这会使数据分析变复杂，却极大地增加了数据包含的信息量。例如，自行可以作为距离的替代，从而为恒星光度提供了额外的约束。对自行、视向速度和视差的运动学分析可以区分不同恒星族。

另一个例子是星团天体测量数据的运动学建模。假设星团成员星都以相同的空间速度运动，数据就可以用这种运动模型描述。此时考虑星团内部速度场。星团模型包含自身视差作为待估参数。三角视差测量结果与运动学视差（由恒星自行与横向速度算得）的结果可以提升视差的测量精度。其基本原因在于使用了成员星自行中包含的距离信息。运动学模型还提供了一种不同于分光方法的视向速度测量方法。

建模是一种强大的统计天体测量工具，尤其在使用所有可用数据以及其他巡天计划的补充数据时。如对 *Gaia*，LSST 等，建模将是一种不可或缺的数据分析方法。

5 建议

作者根据前文的描述，总结出来了一些指导性建议，可作为使用统计天体测量来研究天文问题的参考。

- 要意识到几乎所有天体物理量都是天体测量参数的非线性函数，这主要是因为计算这些物理量时要使用的是距离而非视差，如从自行到横向速度、恒星轨道的角动量和能量、作用角变量等
- 仔细考虑要研究的问题。要测定哪些天体物理参数？直接方法是否可取？是否需要改写（重新表述）问题以避免非线性变换？
- 理解所用星表，选择样本时要有明确定义的标准。只有详细了解样本的特性，才能消除最终结果中的潜在偏倚。
- 在使用直接方法时永远不要试图改正 Malmquist 或 Lutz-Kelker 效应。在大部分情况下，这两种效应同时发生，使得标准改正失效。此外，标准改正所隐含的假设几乎不可能成立。

- 通过预测观测量并判断模型在数据空间的拟合优度来进行数据建模。这可以完全避免由非线性转换引入的偏倚问题，同时为考虑选择效应提供了一种直观方法。在这方面，贝叶斯数据分析是一个非常强大的工具。
- 对于特定样本或单个恒星，可以只考虑视差相对误差较小（如小于 10%）的恒星。但是，这隐含了对真实视差的截断（真实视差较大往往对应着较小的视差相对误差）且偏向于亮星（视差绝对视差较小）。因此，结果中仍可能出现严重偏倚。
- 使用所有可用数据，不仅是自行和视向速度，也包括测光和恒星天体物理特性（光谱型、光度型、金属丰度等）。这些补充信息会对距离和光度提供额外的约束，并可用区分恒星类型。此外，测光或分光距离可以用于替代相对不准确的视差，这些距离将会在 *Gaia* 项目之后得到很好的校准。
- 在使用补充信息时，需要注意结果可能存在循环相关。恒星距离和天体物理参数的校准本质上与一小部分标准星的视差测量值紧密相连，因此需要确保这些校准与标准星的视差数据相符。
- 不要忽略天体测量误差中的相关性以及它们随天空位置的变化
- 在没有充分理由时不要丢弃负视差或者低精度的视差，这些数据也包含信息。