《皇家天文学会月报》**000**, 1–16 (2015)　　预印本，2023年4月17日　　使用《皇家天文学会月报》LATEX 样式文件 v3.0 编译

大规模射电干涉测量数据的有损压缩

M Atemkeng,[1]★ S Perkins,[2] E Seck,[3] S Makhathini,[2,4,5] O Smirnov,[2,4] L Bester,[2,4] B Hugo,[2,4]

[1]南非格雷厄姆斯敦 *6139* 罗德斯大学数学系[2]南非射电天文台，黑河公园，菲尔街 *2* 号，天文台区，开普敦，*7925*，南非[3]非洲数学科学研究所（*AIMS*），喀麦隆林贝[4] 南非格雷厄姆斯敦 *6140*，邮政信箱 *94*，罗德斯大学物理与电子学系[5]南非约翰内斯堡扬·斯穆茨大道 *1* 号，*2001*，威特沃特斯兰德大学物理学院

arXiv:2304.07050v1 [astro-ph.IM] 14 Apr 2023

录用：XXX。收到：YYY；原稿形式：ZZZ

摘要

射电望远镜会产生海量数据，随着即将投入运行的望远镜（如平方公里阵列、ngVLA）陆续启用，数据量还将进一步增加。在校准、深度宽场成像、存储和归档以供后续处理的背景下，海量数据是一个亟待解决的重要问题。本文提出一种依赖基线的有损压缩技术，以减少可见度数据量，同时保留视场边缘处的涂抹效应。我们利用矩阵秩的性质，以及低秩近似能够将原始可见度数据表示为若干基本分量之和这一事实，其中每个基本分量对应天空分布的一个特定傅里叶分量。因此，全部可见度数据被表示为来自各基线的数据矩阵集合，而非单个张量。这有利于并行计算，也便于研究依赖基线的秩–r 分解。所提出的方法如下：给定全部可见度数据组成的大型数据集，第一个算法称为简单奇异值分解，将数据投影到由秩为–r 的数据矩阵构成的规则采样空间中。在该空间内，所有基线的数据具有相同的秩，因此所有基线的压缩因子相同。第二个算法称为*BDSVD*，将数据投影到由秩为$^{-}r_{pq}$ 的数据矩阵构成的不规则采样空间中。下标 pq 表示数据矩阵的秩会随基线pq而变化，因此压缩因子依赖于基线。本文以MeerKAT和欧洲甚长基线干涉测量网为参照望远镜，将所提出方法与传统平均和依赖基线平均（BDA）等传统方法进行评估和比较。在相同的空间分辨率阈值下，简单奇异值分解和*BDSVD* 的压缩效果均比传统平均和BDA高出两个数量级。在空间节省率相同的情况下，空间分辨率没有下降，且数据中的噪声方差有所降低，使视场边缘处的信噪比提升至1.5 dB以上。所提出的压缩方法压缩效果更优，但为降低实现复杂度，需要以全数据分辨率处理数据。

关键词：仪器　　：干涉仪，方法：数据分析，方法：数值方法，技术：干涉　　ic

1 引言

射电干涉阵列由成对相关的射电天线组成，可产生称为可见度或空间频率采样值的复数数据。随着天线数量、大规模天空巡测以及光谱和时间分辨率的提高，数据量呈二次方增长。处理和存储如此庞大的可见度数据已成为一项挑战，例如，需要对数据进行校准，并通过重采样和快速傅里叶变换将数据转换到图像空间。在当前射电干涉阵列（如 MeerKAT（Jonas2009）、ASKAP（Johnstonet al. 2008）、NenuFAR（Zarka et al. 2015）、LOFAR（van Haarlemet al. 2013）以及包括平方公里阵列（SKA）（Dewdneyet al. 2009）在内的未来射电干涉仪中，如何处理海量可见度数据，是校准和深度宽视场成像所面临的重要问题。为了解决更紧致源的形态结构，平方公里阵列将通过长基线采样非常高的空间频率。这种高空间频率采样将使平方公里阵列进入前所未有的可见度数据量时代；这需要更多计算资源和改进的数据压缩策略。对校准产品进行长期数据归档是必要的，并将受益于新数据压缩算法的实施。

数据压缩对于降低数据存储、处理和归档所涉及的成本至关重要。虽然有许多数据压缩策略，但选择合适的策略取决于科学研究的具体情况，以及与该研究相关的信息在数据中的编码方式。因此，要实现较高的数据压缩率，就需要了解数据中的信息和科学研究的具体情况。对于给定的数据集，可能只有少数数据点包含了与科学研究相关的大部分信息，而其他数据点则包含噪声或与研究无关的信息（Bobin etal. 2008）。

在可见度数据中，例如，来自系统电子设备的噪声、来自天空非目标区域的信号，以及来自空间频率的.....

★ 电子邮箱：m.atemkeng@gmail.com（MA）

© 2015 作者

可以从数据中移除冗余基线。有一些数学工具能够将这类数据集分解成一种新形式，从而更容易理解数据集中的信息是如何根据重要程度呈现的。用于分解数据集的各类数学工具具有不同的特性，其适用性取决于科学研究的具体情况。选择压缩策略时，还必须考虑其他因素，以便在压缩率、信号损失和科学价值之间取得最佳平衡。

在实际应用中，可见度数据会在有限的时间和频率间隔内进行积分和平均，从而产生去相关效应。为限制去相关，有限的时间和频率间隔会保持得非常小，以保留长基线上的相位项。去相关会使视场（FoV）边缘处源的衰减，并改变其形态。对于给定的视场，如果将时间和频率间隔增大到一定限度，则可以使用传统平均来减少数据量。反卷积例程必须正确考虑涂抹效应，以避免限制图像的动态范围；多位作者对此进行了详细讨论（Cotton 2009；Atemkeng 等人 2016；Tasse et al. 2018；Atemkeng 等人 2020）。

由于只有在长基线上才需要对空间频率进行高分辨率采样，因此，基线依赖平均（BDA）方法近年来逐渐受到关注，可用于压缩可见度数据并抑制视场（FoV）之外的源，同时保持长基线所需的高光谱和时间分辨率。尽管 BDA 有望在短基线上实现数据压缩，并在长基线上保持所需的高光谱和时间分辨率，但仍需进一步研究。测量集 v2.0（MS）（Kemball & Wieringa 2000）格式不支持存储 BDA 所要求的可变频率分箱。原则上，测量集 v2.0 格式可以存储 BDA 所需的可变频率分箱，因此，这种方法需要针对频域的每一种分解方式创建一个频谱窗口。目前的校准算法要求数据具有规则的采样间隔。然而，为了利用 BDA 的压缩能力，校准算法需要进行适配，尤其要适应不同基线上的光谱和时间分辨率差异。在选择 BDA 参数时，选择合适的校准解算间隔也十分重要。成像需要定义明确的插值窗口和加权方案。对于成像算法而言，这并不是问题，因为它们会将非结构化采样点放置在规则网格上。

可见度数据经常会存档以供日后使用；存档前会采用传统平均和/或 BDA。然而，选择压缩参数时必须考虑许多因素，因为某些科学研究（例如 VLBI 科学和谱线研究）以及考虑主波束（PB）变化和电离层影响等因素的校准流程，都要求较高的频谱和时间分辨率。传统平均和/或 BDA 会降低频谱和时间分辨率；对于 BDA，至少会降低较短基线上的分辨率，因此必须谨慎使用。因此，有必要研究数据压缩技术，以便在存档可见度数据时尽可能不降低频谱和时间分辨率。

在本研究中，我们采用低秩近似进行可见度数据压缩。低秩近似通过最小化一个损失函数，来衡量相对于原始数据矩阵具有较低秩的最佳近似数据矩阵，即低秩近似利用一个低秩数据矩阵来表示一个较高秩的数据矩阵，同时几乎不损失或完全不损失信息。一个大小为 $M \times N$ 的数据矩阵 $\boldsymbol{V}$ 具有秩–r 近似，其形式为：

$$\boldsymbol{V} \simeq \boldsymbol{AC}, \tag{1}$$

其中，数据矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{C}$ 的大小分别为 $M \times r$ 和 $r \times N$。秩为 $-r$ 的近似矩阵包含 $r(M+N)$个元素，现在可以用它代替 $\boldsymbol{V}$ 进行存储，以节省内存和/或计算开销，因为如果 r相对较小且 M，N相对较大，那么 $r(M+N)$会显著小于 MN。因此，$1 \leq r < MN/(M+N)$。现在，存储 $\boldsymbol{V}$的秩为 $-r$的近似矩阵只需 $O(r(M+N))$ 的磁盘空间，而非 $O(MN)$。

目前，在大数据领域，许多应用图像处理的研究都通过降维将高秩数据集投影到低秩空间。这是一个特征（维度）提取过程，旨在保留数据中最相关的信息（压缩维度）。降维主要作为预处理步骤，以减少对高性能计算资源的需求，可分为两大类：线性降维和非线性降维。线性降维算法中最著名的低秩近似方法是奇异值分解（SVD）。SVD 将数据分解为多个独立的特征集合，包括相关特征、噪声和冗余分量。SVD 的特性已在其他领域得到充分研究。由于其正交矩阵分解特性，SVD 在信号处理中广受欢迎。多位作者利用 SVD 处理干涉测量数据中的不同任务；例如，Of- fringa 等人（2010）探讨了利用 SVD 将强射频干扰与微弱天文信号分离的方法。为了加快成像和反卷积的速度，Kartik 等人（2017）利用 SVD 对插值和网格化后的可见度数据进行压缩。近期研究提出结合主波束（PB）的全息测量与 SVD，以处理主波束在空间和频谱上的变化；例如，Iheanetu 等人（2019）探讨了利用少量特征确定主波束稀疏表示的可能性，而Ansah-Narh 等人（2020）则探讨了通过选择分解结果中最强的分量来减轻主波束中的噪声。

利用奇异值分解（SVD）压缩可见度数据并非偶然之举；高灵敏度射电干涉阵列主要由短基线构成：奇异值分解利用短基线中的冗余，并能够从数据中分离出噪声。秩–r 对尺寸为 $M \times N$的数据矩阵进行奇异值分解，需要 $O(rMN)$次运算：$O(N^3)$（对于方形数据矩阵的情况）（Gu & Eisenstat 1996；Demmel et al. 2007；Shishkin et al. 2019）。在大数据场景下，这种计算量在实际应用中难以实现。然而，本文所述的奇异值分解方法不应与这样一种方法混淆：该方法将全部可见度数据视为单个张量，并假设不同基线之间具有相似性。我们利用奇异值分解与矩阵秩之间的关系，以及奇异值分解能够将原始可见度数据表示为多个分量之和这一事实，其中每个分量都对应天空分布的一个特定傅里叶分量。因此，全部可见度数据被表示为针对每条基线分别生成的一组矩阵，而非单个张量。这使得可以进行依赖于基线的秩–r 分解，并且能够针对每条基线独立执行，从而实现并行计算。下面给出所提出的方法。给定一个可见度数据集：

- 第一种算法将数据投影到秩为–r 的数据矩阵构成的规则采样空间中。在这个空间里，所有基线的数据具有相同的秩，因此所有基线的压缩因子相同。我们将这种算法称为简单奇异值分解。研究表明，与传统平均相比，简单奇异值分解在压缩可见度数据方面效果显著，压缩倍数高出两个数量级，同时降低了数据中的噪声方差。

• 第二种算法将数据投影到秩为^{-r}pq 的数据矩阵构成的不规则采样空间中。额外的下标 pq 表示数据矩阵的秩会随基线而变化，因此压缩因子取决于基线。我们将这种方法称为基线相关奇异值分解（*BDSVD*）。

需要指出的是，简单奇异值分解 和 *BDSVD* 与传统平均和 BDA 的用途不同。传统平均和 BDA 压缩后的可见度数据可直接作为成像或校准等其他任务的输入，相较于未压缩数据，计算成本更低，但这会降低频谱和时间分辨率。本文中的简单奇异值分解和 *BDSVD* 用于归档可见度数据；在后续使用前，必须先解压数据。这样做的优点是能够保留原始数据的频谱和时间分辨率。各用例总结见表1。本文以串行方式实现简单奇异值分解 和 *BDSVD*，并详细说明并行实现的依据。随后，我们提出一种并行算法，可将计算复杂度从 $O\left(N_b MN^2\right)$降低至 $O\left(N_b MN^2/N_p\right)$；其中 N_b和 N_p分别表示基线数量和计算节点数量。该算法的实现不属于本文工作范围，将留待未来研究。

本文其余部分安排如下：第 2 节概述了射电干涉测量以及奇异值分解背后的理论。第 3 节介绍了这些算法背后的数学形式化。第 4 节分析了所提出算法对图像和噪声的影响。第 5 节讨论仿真结果，第 6 节总结全文并展望未来工作。本文的数学符号汇总于表 2。

2 背景

本节将讨论由射电干涉测量方程（RIME）形式体系（Hamaker 等，1996；Smirnov，2011）所提供的可见度数据测量的数学框架。该框架用于理解、描述和衡量简单奇异值分解和*BDSVD*的性能。

2.1 全天空 RIME 与传统平均

在 RIME 形式体系下，对于单个点源，Smirnov(2011) 建立了时刻 t和频率 ν 下测得的可见度与在某个时间和频率积分区间内平均后的可见度 $[t_0,t_1]\times[\nu_0,\nu_1]$之间的关系。我们将这一关系扩展到全天空 RIME 形式体系：

$$\mathcal{V}(\boldsymbol{u}_{pq}(t,\nu))=\sum_{\boldsymbol{l}}\left(\mathcal{G}_{p\boldsymbol{l}t\nu}\mathcal{I}_{\boldsymbol{l}}\mathcal{G}^{\dagger}_{q\boldsymbol{l}t\nu}\right)\mathcal{K}^{\boldsymbol{l}}_{pqt\nu}. \tag{2}$$

标量项 $\mathrm{K}^{\boldsymbol{l}}_{pqt\nu}=\mathrm{e}^{-2i\pi(\boldsymbol{l}-\boldsymbol{l}_0)\boldsymbol{u}_{pqt\nu}}$描述了天线 p,q 的位置以及基线向量$\boldsymbol{u}_{pq}(t,\nu)$ 的旋转所产生的影响，该向量追踪位于单位向量 $\boldsymbol{l}$ 所指方向上的一个源。相关器引入一个补偿时延 $\boldsymbol{l}_0$，以确保在观测的相位中心处 $\mathrm{K}^{\boldsymbol{l}}_{pqt\nu}\equiv 1$ ：

$$\boldsymbol{u}_{pqt\nu}=\frac{\nu}{c}\begin{bmatrix}u_{pqt}\\ v_{pqt}\\ w_{pqt}\end{bmatrix},\boldsymbol{l}=\begin{bmatrix}l\\ m\\ n\end{bmatrix},\boldsymbol{l}_0=\begin{bmatrix}0\\ 0\\ 1\end{bmatrix}.$$

在上述方程中，† 表示复数转置算子， c 表示光速，I 表示天空亮度，$\mathrm{G}p\boldsymbol{l}t\nu$ 而$\mathrm{G}\dagger q\boldsymbol{l}t\nu$ 分别表示天线 p 和 q的方向相关琼斯矩阵组。

<escape>以下是方程2 在 $[t_0,t_1]\times[\nu_0,\nu_1]$上的平均值的近似:
</escape>

$$V_{pq(t\nu)_i}\simeq\sum_{\boldsymbol{l}}\operatorname{sinc}\frac{\Delta\Phi}{2}\operatorname{sinc}\frac{\Delta\psi}{2}\left(\mathcal{G}_{p\boldsymbol{l}(t\nu)_i}\mathcal{I}_{\boldsymbol{l}}\mathcal{G}^{\dagger}_{q\boldsymbol{l}(t\nu)_i}\right)\mathcal{K}^{\boldsymbol{l}}_{pq(t\nu)_i}, \tag{3}$$

其中 $(t\nu)_i=t_i\nu_i$，其中 $t_i=(t_1+t_0)/2$，且 $\nu_i=(\nu_1+\nu_0)/2$。相位差 $\Delta\Phi$ 和 $\Delta\psi$ 定义如下：

$$\begin{aligned}\Delta\Phi=&\arg\mathcal{G}_{p\boldsymbol{l}t_1\nu_i}+\arg\mathcal{G}^{\dagger}_{q\boldsymbol{l}t_1\nu_i}+\arg\mathcal{K}^{\boldsymbol{l}}_{pqt_1\nu_i}\\&-\left(\arg\mathcal{G}_{p\boldsymbol{l}t_0\nu_i}+\arg\mathcal{G}^{\dagger}_{q\boldsymbol{l}t_0\nu_i}+\arg\mathcal{K}^{\boldsymbol{l}}_{pqt_0\nu_i}\right)\\ \Delta\psi=&\arg\mathcal{G}_{p\boldsymbol{l}t_i\nu_1}+\arg\mathcal{G}^{\dagger}_{q\boldsymbol{l}t_i\nu_1}+\arg\mathcal{K}^{\boldsymbol{l}}_{pqt_i\nu_1}\\&-\left(\arg\mathcal{G}_{p\boldsymbol{l}t_i\nu_0}+\arg\mathcal{G}^{\dagger}_{q\boldsymbol{l}t_i\nu_0}+\arg\mathcal{K}^{\boldsymbol{l}}_{pqt_i\nu_0}\right).\end{aligned}$$

公式 3 中的平均总会导致偏离相位中心的源出现振幅净损失；在实际情况下，由于 $\mathrm{K}^{\boldsymbol{l}}_{pqt\nu}$、$\mathrm{G}p\boldsymbol{l}t\nu$ 和 $\mathrm{G}\dagger q\boldsymbol{l}t\nu$ 中的相位项随时间和频率变化，因此无法满足条件化 $t_1-t_0\to 0$ 和 $\nu_1-\nu_0\to 0$ 。当这种振幅损失仅由标量 $\mathrm{K}^{\boldsymbol{l}}_{pqt\nu}$ 引起时，称为时间和频率去相关（或模糊）。当振幅损失由 $\mathrm{K}^{\boldsymbol{l}}_{pqt\nu}$、$\mathrm{G}p\boldsymbol{l}t\nu$、$\mathrm{G}\dagger q\boldsymbol{l}t\nu$ 的累积相位变化引起时，则使用一般术语相干性损失，因为 $\mathrm{G}p\boldsymbol{l}t\nu$ 和 $\mathrm{G}\dagger q\boldsymbol{l}t\nu$ 也包含随时间和频率变化的复相位。公式 3 表明，对于全天空 RIME，相干性损失可以分别测量为各个单独源贡献之和。在考虑所有复数项引入的相干性损失时，公式 3 中求和的各个单独源的平均值可近似表示为相位差。在上述相干性损失的表述中，我们忽略了由 I 引起的相干性损失，因为本研究关注点源而非扩展源的相干性损失；对于扩展源，I 中的相位项会随时间和频率显著变化。

2.2 BDA 及其实现

为了大幅压缩可见度数据并减轻传统平均导致的振幅损失，已有多位作者探讨了使用 BDA 的潜力（Cotton 2009；Atemkeng 等人 2018；Wijn- holds 等人 2018）。如式3 所示，三个复数项 $\mathrm{K}^{\boldsymbol{l}}_{pqt\nu}$,$\mathrm{G}p\boldsymbol{l}t\nu$ 和 $\mathrm{G}\dagger q\boldsymbol{l}t\nu$ 的乘积受到 $sinc$ 函数的衰减。视场边缘处的衰减程度由这些 $sinc$ 函数的宽度决定，而这些函数取决于每条基线在 $[t_0,t_1]\times[\nu_0,\nu_1]$上采样的空间频率。在长间距基线上，使用较宽的 $[t_0,t_1]\times[\nu_0,\nu_1]$ 进行传统平均会导致较窄的 $sinc$，并使源振幅显著下降。为了使给定视场内所有基线的振幅损失保持一致， $sinc$ 函数的宽度必须保持恒定，因此对数据进行平均的时间-频率间隔会有所变化。该方法对于密集核心干涉阵列至关重要，因为这类阵列会对较短基线的数据进行大幅平均。较长基线需要较小的时间-频率间隔，才能达到 $sinc$ 函数的宽度限制，从而导致压缩程度较低。

BDA 实现若能与现有软件、格式和规范集成，将有助于利用压缩优势，并便于使用真实观测数据进行测试。当前无线电天文数据的规范是 MSv2，它基于 CASA Table Data Syste m（CTDS）格式。因此，我们开发了面向该规范和格式的 BDA 实现。

方法	压缩	视场内的频谱和时间分辨率是否得到保留?	能否加快处理速度?
传统平均	yes	no	yes
BDA	yes	no	yes
简单奇异值分解	yes	yes	no
BDSVD	yes	yes	no

表 1. 传统平均和 BDA 与简单奇异值分解 和*BDSVD*的用例。

符号说明	描述
$\boldsymbol{V}$	压缩前 N_b 条基线的数据矩阵
$\boldsymbol{V}_n$	压缩后针对 N_b 基线的数据矩阵
$\mathcal{V}$	未采样的可见度数据
$V_{pqt\nu}$	针对 pq 在 t, ν的采样的可见度数据
$\boldsymbol{u}_{pq}$	基线 pq 向量
$\mathcal{I}_{\boldsymbol{l}}$	方向为 $\boldsymbol{l}$的天空亮度
$\boldsymbol{V}_{pq}$	基线 pq的秩为 $-r$ 的数据矩阵
$\boldsymbol{V}_{pq,\ n}$	$\boldsymbol{V}_{pq}$的秩$-n$ 近似
η_{pqk}	奇异值
CF	压缩因子
N_b	基线数量
M, N	矩阵的大小
ϵ	最大阈值误差
ε	要保留的最小信号百分比
$\boldsymbol{I}^{\mathrm{d}}_{\mathrm{pq}}$	基线 pq 的脏图像
$\widetilde{\boldsymbol{I}}^{\mathrm{d}}_{\mathrm{pq}}$	基线 pq 的压缩脏图像
$\boldsymbol{I}^{\mathrm{d}}$	脏图像矩阵
$\widetilde{\boldsymbol{I}}^{\mathrm{d}}$	压缩脏图像矩阵
$\boldsymbol{I}^{\mathrm{d}}_{\mathrm{loss}}$	损失图像矩阵
$\mathcal{F}$	傅里叶变换算子
S/R	信噪比

表 2. 主要数学符号：小写粗体字母表示向量，大写粗体字母表示矩阵。花体大写字母用于表示函数形式。其他所有符号均为常数。

测量集的行按基线分组、按时间排序，并聚合到各个分箱中，使其 sinc(ΔΦ/2) 不超过 $\mathrm{K}^{\boldsymbol{l}}_{pqt\nu}$ 定义的可接受去相关容差。每个分箱的 sinc(ΔΦ/2) 都会求逆：首先得到 sinc(Δψ/2)，然后得到该分箱的频率积分间隔 $[\nu_0, \nu_1]$。实际应用中，每个分箱的频率间隔可能不同，而这种信道化方式无法方便地适配 CTDS 格式，因为该格式为每个频谱窗口规定了固定数量的信道。

因此，为了将 BDA 数据适配为这种格式，我们通过按信道数的整数因子细分原始频谱窗，将信道化离散化，从而生成表示每个离散化结果的新频谱窗。随后，每个 BDA 分箱输出到输出测量集中的一行，并且每行对应一个不同的频谱窗，从而以牺牲少量压缩率为代价，实现与测量集的兼容。

2.3 低秩近似：奇异值分解

奇异值分解将给定的具有 M 行和 N 列的复数数据矩阵 $\boldsymbol{V}$ 分解为三个相互独立的矩阵：

$$\boldsymbol{V} = \boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{C}^{\dagger}, \tag{4}$$

其中 $\boldsymbol{\Lambda}$ 是一个大小为 $M \times N$ 的对角矩阵，$\boldsymbol{A}$ 是一个大小为 $M \times M$ 的矩阵，而 $\boldsymbol{C}$ 和 $N \times N$ 是酉矩阵。这些矩阵定义如下：

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_r & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}, \quad \boldsymbol{A} = [\boldsymbol{A}_r, \boldsymbol{a}_{r+1}, \cdots, \boldsymbol{a}_M], \quad \boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_r \\ \boldsymbol{c}_{r+1} \\ \vdots \\ \boldsymbol{c}_N \end{bmatrix}. \tag{5}$$

该分解不要求 $\boldsymbol{V}$ 是方阵。大小为 $r \times r$ 的矩阵 $\boldsymbol{\Lambda}_r = \mathrm{diag}(\eta_1, \eta_2, \cdots, \eta_r)$ 是对角矩阵，其对角线上有 $r = \min(M, N)$ 个元素，η_k 是 $\boldsymbol{V}$ 的奇异值。正如 Stewart (1998) 所讨论的，$\eta_k = \sqrt{\lambda_k}$ 其中 λ_k 是 $\boldsymbol{V}\dagger\boldsymbol{V}$ 的特征值。我们注意到，$\boldsymbol{\Lambda}_r$ 对角线上的所有元素均非零，且按降序排列。这里，$\boldsymbol{A}_r = [\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_r]$ 和 $\boldsymbol{C}_r = [\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_r]\mathrm{T}$，其中 T 是转置算子。向量 $\boldsymbol{a}_k$ 的大小为 M，向量 $\boldsymbol{c}_k$ 的大小为 N，它们分别是 $\boldsymbol{V}$的左奇异向量和右奇异向量。$\boldsymbol{a}_k$ 和 $\boldsymbol{c}_k$ 分别是 $\boldsymbol{V}\boldsymbol{V}\dagger$ 和 $\boldsymbol{V}\dagger\boldsymbol{V}$的特征向量。若改写为以下形式，式 4 仍然成立：

$$\boldsymbol{V} = \boldsymbol{A}_r\boldsymbol{\Lambda}_r\boldsymbol{C}_r^{\dagger} \tag{6}$$

$$= \sum_{k=1}^{r} \eta_k \boldsymbol{a}_k \boldsymbol{c}_k^{\dagger}. \tag{7}$$

每个 η_k 都量化了相应分量 $\eta_k\boldsymbol{a}_k\boldsymbol{c}\dagger_k$ 对式 7中相关特征的贡献程度，例如在重建 $\boldsymbol{V}$时。η_k 越大，$\eta_k\boldsymbol{a}_k\boldsymbol{c}\dagger_k$ 对重建 $\boldsymbol{V}$的贡献就越大。注意，$\boldsymbol{a}_k$ 和 $\boldsymbol{c}\dagger_k$ 决定了 $\eta_k\boldsymbol{a}_k\boldsymbol{c}\dagger_k$所含特征的几何形态。

3 拟议的有损压缩方法

本节所述的低秩近似方法应用于原始可见度数据（连续数据），并且逐条基线分别处理。需要指出的是，这些方法也可应用于网格化可见度数据。然而，网格化数据位于规则网格上，所有基线的数据都经过插值并合并在一起，因此难以判断对非零奇异值进行阈值处理时可接受的阈值。此外，由于每条基线观测天空的不同部分，信号失真和衰减取决于基线，而且每份可见度数据的噪声方差也各不相同。这些影响无法在网格化可见度数据中加以考虑。仅存储图像（也就是网格化可见度数据）并不足够，因为校准问题通常占主导地位，而且为了针对特定的连续谱、瞬变和谱线子域生成可用于科学研究的产品，数据大多时候都必须重新校准，这是不可避免的。如果所有基线都卷积到网格间距上，就无法很好地完成这一过程。

根据上述建议，我们采用紧凑且稳健的矩阵形式，对可见度数据（针对给定基线）进行形式化表达。假设天空为非偏振状态，且仅考虑一个通道时间片。

基线 pq 测得的可见度是复数斯托克斯值：

$$V_{pqt_i\nu_j} = \delta_{pqij}(\mathcal{V} + \mathcal{E}), \tag{8}$$

其中 $\delta_{pqij} = \delta(\boldsymbol{u} - \boldsymbol{u}_{pqt_i\nu_j})$ 是一个移至 uv 采样点的狄拉克 delta 函数，$\mathcal{V}$ 在方程 2 中讨论，$\mathcal{E}$ 是均值为零、均方根为 Σ 的干扰随机噪声。假设观测包含 M 个时间片和 N 个信道，则可将单条基线 pq 的可见度数据打包为一个大小为 $\boldsymbol{V}_{pq}$ 的单一数据矩阵，$M \times N$：

$$\boldsymbol{V}_{pq} = \begin{bmatrix} V_{pqt_1\nu_1} & \cdots & V_{pqt_1\nu_N} \\ \vdots & \cdots & \vdots \\ V_{pqt_M\nu_1} & \cdots & V_{pqt_M\nu_N} \end{bmatrix}. \tag{9}$$

该形式体系可扩展至采用正交基、能够完整描述天空的数据结构。我们讨论对 $\boldsymbol{V}_{pq}$ 的压缩，然后在第 4节中详细分析压缩对 Σ 和天空图像的影响。请注意，可以将 $\boldsymbol{V}_{pq}$ 重新划分为子矩阵，并分别求出每个子矩阵的低秩近似。在计算成为瓶颈（即 $MN \to \infty$）且局部低秩矩阵近似具有优势（Lee 等人, 2016）时，应考虑后一种方法。

3.1 方法 1：简单奇异值分解

在不造成信息损失的情况下，寻找一个非常小的低秩数据矩阵 $\boldsymbol{V}_{pq}$ 颇具挑战性，因为 $\boldsymbol{V}_{pq}$ 中的各个元素具有很强的相关性。在实际观测中，由于系统电子设备的影响，以及每个可见度数据点都来源于一对天线电压之间的相关性等因素，$\boldsymbol{V}_{pq}$ 几乎不可能实现不相关。这意味着，不存在一个秩为$-n$ 的低秩近似 $\boldsymbol{V}_{pq,\ n}$，能够完美重建秩为$-r$的 $\boldsymbol{V}_{pq}$：

$$\boldsymbol{V}_{pq,n} \simeq \boldsymbol{V}_{pq},\ 1 \le n < r. \tag{10}$$

$$\boldsymbol{V}_{pq,n} = \mathbf{A}_{pq,n}\boldsymbol{\Lambda}_{pq,n}\mathbf{C}^{\dagger}_{pq,n} \tag{11}$$

$$= \sum_{k=1}^{n} \eta_{pqk}\boldsymbol{a}_{pqk}\boldsymbol{c}^{\dagger}_{pqk}. \tag{12}$$

为简便起见，式 11 也可以写成克罗内克积的形式，$\otimes$在向量化之后：

$$\mathrm{vec}(\boldsymbol{V}_{pq,n}) = (\mathbf{C}^{\dagger}_{pq,n} \otimes \mathbf{A}_{pq,n})\boldsymbol{\eta}_{pq}, \tag{13}$$

其中 $\boldsymbol{\eta}_{pq} = [\eta_{pq1}, \eta_{pq2}, \cdots,\ \eta_{pqn}]^{\mathrm{T}}$。

这种简单奇异值分解只允许所有基线采用相同的压缩因子，因为 n 是固定的。压缩因子是未压缩数据的数据点数量与压缩数据的数据点数量之比。$\mathbf{A}_{pq,n}$ 和 $\mathbf{C}^{\dagger}_{pq,n}$ 的元素是复数，而 $\boldsymbol{\Lambda}_{pq,n}$ 的元素是实数。就计算机存储需求而言，如果一个复数元素计作一个元素，那么一个实数元素应计作 0.5。因此，如果 $\boldsymbol{V}_{pq}$ 的大小为 $M \times N$，那么可以证明，计算 $\boldsymbol{V}_{pq,n}$ 所需子矩阵中的元素数量为 $n(M + N + 0.5)$，由此得到总体压缩因子 CF为：

$$CF = \frac{MN}{n(M + N + 0.5)}. \tag{14}$$

在这种情况下，必须谨慎选择 CF，以确定所有基线上要保留的奇异值数量：

$$n = \left\lceil \frac{MN}{CF(M + N + 0.5)} \right\rceil, \tag{15}$$

其中，$\lceil 。 \rceil$ 是向上取整算子。压缩损失根据张量范数计算：

$$||\boldsymbol{V} - \boldsymbol{V}_n|| \le \epsilon||\boldsymbol{V}||, \tag{16}$$

其中 ϵ 是用于控制可见度数据张量 $\boldsymbol{V}$ 和 $\boldsymbol{V}_n$ 之间差异的最大阈值误差；这两个张量分别表示压缩前后的 N_b 基线数据。在本研究中，我们将张量范数定义为每条基线的 Frobenius 范数之和，$||。||_{\mathrm{F}}$：

$$||\boldsymbol{V} - \boldsymbol{V}_n|| = \sum_{pq} ||\boldsymbol{V}_{pq} - \boldsymbol{V}_{pq,n}||_{\mathrm{F}}. \tag{17}$$

公式 17可通过对每条基线中未保留的奇异值的欧几里得范数求和来计算：

$$||\boldsymbol{V} - \boldsymbol{V}_n|| = \sum_{pq} \sqrt{\sum_{k=n+1}^{r} \eta^2_{pqk}}. \tag{18}$$

请注意，在式 18中，k 从 $n+1$ 到 r，即 $(r-n)$ 奇异值不会保留在压缩中。$||\boldsymbol{V}||$ 定义如下：

$$||\boldsymbol{V}|| = \sum_{pq} ||\boldsymbol{V}_{pq}||_{\mathrm{F}} \tag{19}$$

$$= \sum_{pq} \sqrt{\sum_{k=1}^{r} \eta^2_{pqk}}. \tag{20}$$

可以指定要保留的信号最低百分比 ε，而不是使用误差阈值：

$$||\boldsymbol{V}_n|| \ge \varepsilon||\boldsymbol{V}||, \tag{21}$$

$$||\boldsymbol{V}_n|| = \sum_{pq} ||\boldsymbol{V}_{pq,n}||_{\mathrm{F}} \tag{22}$$

$$= \sum_{pq} \sqrt{\sum_{k=1}^{n} \eta^2_{pqk}}. \tag{23}$$

如果给定约束 ϵ 或 ε，则可以按照算法 1 所示，以较低的计算开销获得满足该约束的对应 n 值。请注意，这些范数可以直接通过计算 $\mathrm{vec}(\boldsymbol{V}_{pq,n})$、$\mathrm{vec}(\boldsymbol{V})$ 和 $\mathrm{vec}(\boldsymbol{V}_n)$ 的欧几里得范数获得；它们分别是 $\boldsymbol{V}_{pq,n}, \boldsymbol{V}$ 和 $\boldsymbol{V}_n$ 的向量化形式。

3.2 天空中的源位置和基线对奇异值分布的影响

如2.1 节所述，由于 $\mathrm{K}^{\boldsymbol{l}}_{pqt\nu}$,$\mathrm{G}p\boldsymbol{l}t\nu$ 和 $\mathrm{G}\dagger\, q\boldsymbol{l}t\nu$ 中相位项的变化，以及基线几何结构和分布导致的奈奎斯特采样差异，各基线观测到的天空区域各不相同。根据这些信息，我们注意到去相关程度取决于基线。基于这一发现，提出了以下问题：所有基线上的奇异值分布是否相同？如果不同，各基线上的奇异值强度和衰减有何差异？奇异值强度和衰减的变化能否用于大幅压缩数据？压缩又受哪些因素限制？

为回答上述问题，图1 展示了三个基线中前20 个奇异值：1 央斯基源位于相位中心（顶行面板）、偏离相位中心 5 度（中行面板）和 10 度（底行面板）时，最长基线（左列面板）、中等长度基线（中列面板）和最短基线（右列面板）的情况

数据： V；ϵ或ε
结果： 要保留的奇异值数量，n
$n \leftarrow 0$；$V_{pq,\ n} \leftarrow \mathbf{0}$
如果给定 ϵ 则
  do
    对于所有基线 pq **执行**
      $V_{pq,\ n} \leftarrow V_{pq,\ n-1} + \eta_{pqn} \boldsymbol{a}_{pqn} \boldsymbol{c}^{\dagger}_{pqn}$
    end
    $n \leftarrow n+1$
  当 $\frac{\sum_{pq} ||V_{pq} - V_{pq,\ n}||_F}{\sum_{pq} ||V_{pq}||_F} > \epsilon$ 和 $n \leq r$；
否则
  do
    对于 所有基线 pq **进行**
      $V_{pq,n} \leftarrow V_{pq,n-1} + \eta_{pqn} \boldsymbol{a}_{pqn} \boldsymbol{c}^{\dagger}_{pqn}$
    end
    $n \leftarrow n+1$
  当 $\frac{\sum_{pq} ||V_{pq,n}||_F}{\sum_{pq} ||V_{pq}||_F} \times 100 < \varepsilon$ 和 $n \leq r$；
结束

算法 1： 为简单奇异值分解查找n。此处，所有V_{pq}均取自V

时的奇异值。这些奇异值通过模拟 1.4 GHz 下的 MeerKAT望远镜获得。模拟对应每秒 10 000 个时间片，总带宽为 0.8 MHz，分为 10 个信道，每个信道宽 80 千赫。该图展示了奇异值在不同基线上的分布的一些重要特征，因此值得详细说明。无论源的位置如何，每条基线上的第一个分量都能捕获原始数据中的大部分特征。此外，在不同基线和源位置上，奇异值的强度和衰减情况各不相同：与较长基线相比，较短基线上的奇异值衰减更快，而较长基线上的衰减较慢。在观测的相位中心处，奇异值衰减更快；随着源远离相位中心，衰减程度逐渐减缓。换言之：

- 对于位于相位中心的源，所有基线中的第一分量捕获了原始数据中的全部特征，换句话说，η_{pq1} 非常大，$\forall k > 1$，$\eta_{pqk} \sim 0$。这一结果与已知情况一致；对于所有基线，在观测的相位中心处，去相关可以忽略不计。
- 奇异值的分布取决于基线；在较长的基线上，奇异值分散到多个分量中，而对于较短的基线，只有前几个分量能够捕捉特征；例如，设 pq 和 $\alpha\beta$ 是长度为$||\mathbf{u}_{pq}||_2 < ||\mathbf{u}_{\alpha\beta}||_2$的两条基线；$\exists i$使得 $\forall k > i$：

$$\eta_{pqi} \gg \eta_{\alpha\beta i} \text{ and } \eta_{\alpha\beta k} \gg \eta_{pqk}, \tag{24}$$

其中，$||_{\circ}||_2$ 为欧几里得范数。这一结果与较长基线观测到的振幅损失比短基线严重这一事实一致。

- 奇异值的分布范围取决于位置；与近场源相比，远场源的奇异值分布在更多分量中。这很容易解释——与近场源相比，远场源会经历严重的振幅损失。

上述结果表明，如果我们按照第 n 3.1 节所述，对所有基线都保留相同数量的分量（即 $r-n$ 丢弃相同数量的分量），这将直接影响基线压缩误差，$||V_{pq} - V_{pq,\ n}||_F$；与长基线相比，短基线上的误差会较小。图 2展示了图 1中模拟所产生的这些压缩误差，其中每条基线均保留了前八个分量。图中，压缩误差以基线长度为横轴绘制；左、中、右图分别对应源位于相位中心、5 度和 10 度时的压缩误差。观察结果如下：

- At the phase centre, $||V_{pq} - V_{pq,8}||_F \sim 0$ for all baselines pq.
- 对于更长的基线，$||V_{pq} - V_{pq,\ 8}||_F$ 对远场源迅速增大。

- 对于任意两条长度不匹配的基线，例如 pq 和 $\alpha\beta$，其中 $||\mathbf{u}_{pq}||_2 < ||\mathbf{u}_{\alpha\beta}||_2$，我们有：

$$||V_{pq} - V_{pq,8}||_F < ||V_{\alpha\beta} - V_{\alpha\beta,8}||_F. \tag{25}$$

我们注意到，对于偏离相位中心的源以及长基线，奇异值衰减缓慢。由于奇异值的大小η_{pqk} 表示分量$\eta_{pqk}\boldsymbol{a}_{pqk}\boldsymbol{c}_{\dagger}{}_{pqk}$对基线$pq$的原始数据所贡献特征的程度，因此，视给定基线下奇异值的衰减情况而定，可以舍弃奇异值较小的分量。这样做后，我们观察到：

- 在较短的基线上，可以舍弃大量分量，而源振幅几乎不会损失，甚至完全没有损失。这能实现大幅数据压缩，同时对图像保真度几乎没有影响。
- 若要尽可能避免源振幅和图像保真度的损失，则在较长的基线上只能舍弃少量分量。

这使压缩取决于基线。然而，对于给定的视场，一种复杂的压缩方法是在每条基线上仔细选择 n，使得公式 25 中的误差在所有基线上都相等。该方法将在第 3.3节中介绍。

3.3 方法 2：基线相关奇异值分解（*BDSVD*）

BDSVD 方法会为每条基线确定相应的保留分量数，从而使不同基线之间的压缩误差保持一致。例如，对于任意两条满足 $||\mathbf{u}_{pq}||_2 < ||\mathbf{u}_{\alpha\beta}||_2$ 的基线 pq 和 $\alpha\beta$，分别求出分量数 n_{pq} 和 $n_{\alpha\beta}$，使得：

$$||V_{pq} - V_{pq,n_{pq}}||_F \equiv ||V_{\alpha\beta} - V_{\alpha\beta,n_{\alpha\beta}}||_F, \tag{26}$$

其中

$$V_{pq,n_{pq}} = \sum_{k=1}^{n_{pq}} \eta_{pqk} \boldsymbol{a}_{pqk} \boldsymbol{c}^{\dagger}_{pqk}. \tag{27}$$

需要注意的是，现在会针对每条独立基线验证最大阈值误差 ϵ（或要保留的最小信号百分比 ε），而不是针对式 16（或式 21）进行验证：

$$||V_{pq} - V_{pq,n_{pq}}||_F \leq \epsilon ||V_{pq}||_F \tag{28}$$

$$||V_{pq,n_{pq}}||_F \geq \varepsilon ||V_{pq}||_F. \tag{29}$$

直观而言，由于 $||\mathbf{u}_{pq}||_2 < ||\mathbf{u}_{\alpha\beta}||_2$这意味着 $n_{pq} < n_{\alpha\beta}$紧致源的奇异值累积强度在所有基线上都相等，

$$\sum_{k=1}^{n_{pq}} \eta^2_{pqk} = \sum_{k=1}^{n_{\alpha\beta}} \eta^2_{\alpha\beta k}. \tag{30}$$

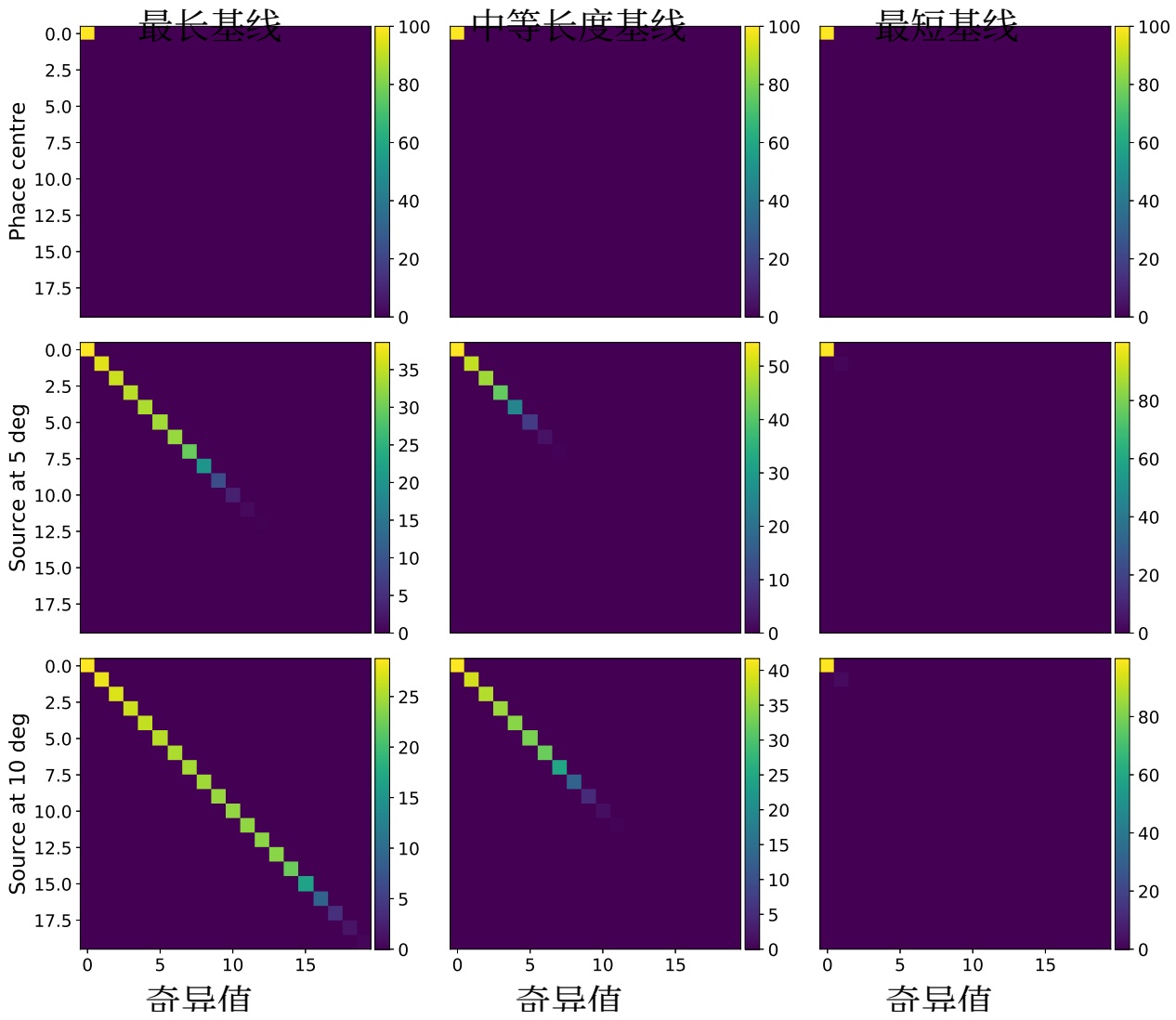


图 1. 三个基线的前 20 个奇异值；相位中心处（顶行面板）、距相位中心 5 度（中间行面板）和 10 度（底行面板）处，1 央斯基源对应的最长基线（左列面板）、中等长度基线（中列面板）和最短基线（右列面板）。这些奇异值通过模拟 1.4 GHz 下的 MeerKAT望远镜得到。数据以 1 秒和 80 千赫的采样间隔采样，观测时长为 166 分 40 秒，带宽为 0.8 兆赫。

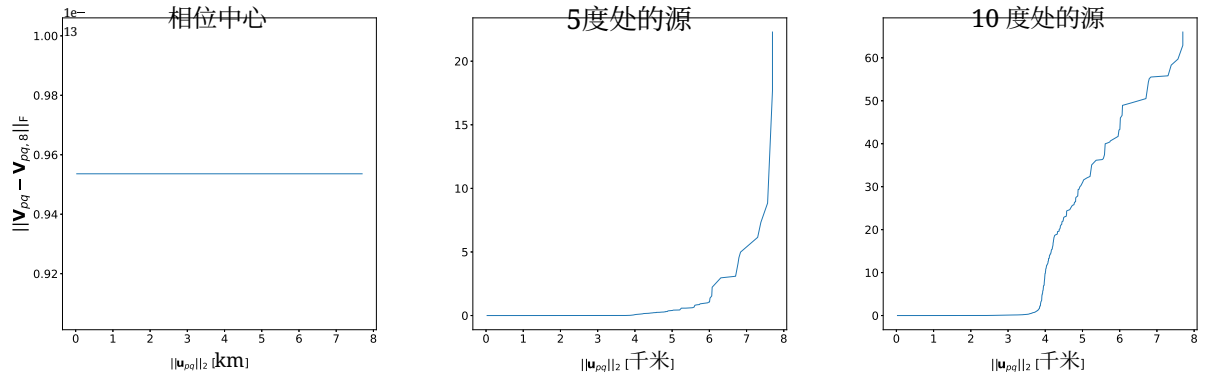


图 2. 图 1中的模拟压缩误差，其中保留每条基线的前八个分量。在相位中心，所有基线的压缩误差都趋于零。对于远场源，长基线上的压缩误差迅速增大。与长基线相比，短基线上的压缩误差较小。

根据这些信息，式 32是正确且有效的。

$$||\boldsymbol{V}_n|| = \sum_{pq} \sqrt{\sum_{k=1}^{n_{pq}} \eta_{pqk}^2} \tag{31}$$

$$= N_b \sqrt{\sum_{k=1}^{n_{pq}} \eta_{pqk}^2}. \tag{32}$$

在上述内容中，要保留的分量数量取决于基线，因此压缩因子也取决于基线：

$$CF_{pq} = \frac{MN}{n_{pq}(M+N+0.5)}, \tag{33}$$

例如 $CF_{pq} > CF_{\alpha\beta}$，因为 $n_{pq} < n_{\alpha\beta}$，较短基线对应的数值大于较长基线。与式 14不同，*BDSVD* 总体上 CF 遵循以下关系：

$$CF = \frac{N_b MN}{(M+N+0.5)\sum_{pq} n_{pq}}. \tag{34}$$

算法 2 描述了一个迭代过程，用于求出与基线相关的分量数量 n_{pq}，给定 ϵ 或 ε。

对于压缩因子 CF，空间节省 SS 的测量方式如下：

$$SS = (1 - CF^{-1}) \times 100\%. \tag{35}$$

3.4 计算复杂度与并行化

低秩近似存在计算方面的缺点。计算一个超大矩阵的奇异值分解（SVD）的成本很高。对于本文讨论的可见度数据压缩，在某些情况下（例如，为了归档数据），我们可以忽略这一计算缺点，因为我们每次只需对数据进行一次压缩。在其他情况下，例如平方公里阵列（SKA）这类处理大数据的射电干涉仪，每条基线上的可见度数据量都非常庞大，即使是一次性压缩也可能成为一项挑战，因为求解精确的 CF_{pq}需要依次计算每条基线上的奇异值，这是一项计算成本高昂的任务。例如，如果每条基线的可见度数据大小为 $M \times N$，其中 $M \geq N$，那么对于 N_b 条基线，顺序执行的全秩–n 简单奇异值分解的计算复杂度为：

$$s_{\text{cost}} \sim O(N_b MN^2). \tag{36}$$

BDSVD 与 简单奇异值分解 在相同 CF的情况下具有相近的计算规模。然而，好消息是，可见度数据具有由行数据构成的某种天然基。这些信息可用于通过并行算法加速压缩。对于给定的基线，其行可见度数据可以共享给多个计算节点；例如，对于每个基线，我们可以将行数据细分为大小为 $\boldsymbol{V}^{[\mathrm{i}}_{pq}{}^{]}$ 的多个数据块，并并行计算所有数据块的奇异值分解。下标 $^{[i]}$ 表示 $i^{\prime}$ 第 h 个行可见度数据块。我们在算法 3中简要讨论了这一分布式压缩过程，但仍需进一步研究，以评估其在实践中的计算效率，并确定在可见度数据达到何种规模时应采用该算法。

算法 3 表明，我们可以并行计算每条基线的 SVD（第 1 行）；假设有 N_p 个并行节点。此外（如第 5 行所示），对于每条基线，要找到 $\mathbf{A}pq$，$\boldsymbol{\Lambda}_{pq}$ 和 $\mathbf{C}pq$，需要并行计算所有 $\boldsymbol{V}^{[\mathrm{i}}_{pq}{}^{]}\boldsymbol{V}^{[i]\dagger}_{pq}$。$\boldsymbol{V}^{\dagger}_{pq}\boldsymbol{V}_{pq}$则通过对所有 $\boldsymbol{V}^{[\mathrm{i}}_{pq}{}^{]}\boldsymbol{V}^{[i]\dagger}_{pq}$求和得到。后者表明，算法3的复杂度按以下方式扩展：

$$p_{\text{cost}} \sim O\left(N_b MN^2/N_p\right). \tag{37}$$

理想情况下，我们希望证明式 36具有强扩展性，但这超出了本文的讨论范围。

4 图像与噪声的影响：解析量化

实际应用中，奇异值分解并非完全正交，分解后的每个分量中都同时包含噪声和信号。因此，需要在去除信号和噪声的量之间进行权衡。简单奇异值分解 和*BDSVD* 也会造成少量信号损失，并保留少量噪声。现在，我们的目标是提供数学模型，以量化所有基线中信号损失和噪声去除的贡献。

4.1 对图像的影响

由于傅里叶变换 F 是线性的，单个基线的压缩脏图像 $\widetilde{\boldsymbol{I}}_{\mathrm{d}\,\mathrm{pq}}$是以下各项之和：

$$\widetilde{\boldsymbol{I}}^{\mathrm{d}}_{\mathrm{pq}} = \mathcal{F}\boldsymbol{V}_{pq,n_{pq}} \tag{38}$$

$$= \sum_{k=1}^{n_{pq}} \eta_{pqk}\boldsymbol{I}^{\mathrm{d}}_{pqk}, \tag{39}$$

其中，$\eta_{pqk}\boldsymbol{I}_{\mathrm{d}pqk}$ 是基线 pq 观测到的天空图像的 $k^{t}h$ 线性无关分量，而 F 是一个酉线性算子。根据帕塞瓦尔定理，我们有：

$$\boldsymbol{I}^{\mathrm{d}}_{pqk} = \mathbf{a}^{\prime}_{pqk}\mathbf{c}^{\prime\dagger}_{pqk}, \tag{40}$$

其中，$\mathbf{a}^{\prime}_{pqk}$ 和 $\mathbf{c}^{\prime}_{pqk}$ 是向量。奇异值，η_{pqk} 表示 $\eta_{pqk}\boldsymbol{I}_{\mathrm{d}pqk}$ 对 $\widetilde{\boldsymbol{I}}_{\mathrm{d}\,\mathrm{pq}}$的质量和保真度的贡献程度。仔细查看式 39 可知，由于傅里叶变换的线性和酉性质，可见度域和图像域中的奇异值相等。这意味着，选择在哪个域压缩数据并不重要；最终，在可见度域中保留的分量数量与图像域中相同。

或者，脏图像可由对所有基线上的式子39求和得到：

$$\widetilde{\boldsymbol{I}}^{\mathrm{d}} = \sum_{pq} \widetilde{\boldsymbol{I}}^{\mathrm{d}}_{\mathrm{pq}} \tag{41}$$

$$= \sum_{pq} \left(\sum_{k=1}^{n_{pq}} \eta_{pqk}\boldsymbol{I}^{\mathrm{d}}_{pqk} \right). \tag{42}$$

若 $\boldsymbol{I}_{\mathrm{d}} = \sum_{pq}\mathrm{F}\boldsymbol{V}_{pq}$ 是未压缩数据的傅里叶变换（即天空的未压缩图像），则为量化每个像素的信号净损失，$\boldsymbol{I}_{\mathrm{dloss}}$，采用以下差值作为标准保真度指标：

$$\boldsymbol{I}^{\mathrm{d}}_{\mathrm{loss}} = \boldsymbol{I}^{\mathrm{d}} - \widetilde{\boldsymbol{I}}^{\mathrm{d}} \tag{43}$$

$$= \sum_{pq} \left(\sum_{k=1}^{r} \eta_{pqk}\boldsymbol{I}^{\mathrm{d}}_{pqk} - \sum_{k=1}^{n_{pq}} \eta_{pqk}\boldsymbol{I}^{\mathrm{d}}_{pqk} \right). \tag{44}$$

对于 $n_{pq} < r$，有

$$\boldsymbol{I}^{\mathrm{d}}_{\mathrm{loss}} = \sum_{pq} \sum_{k=n_{pq}+1}^{r} \eta_{pqk}\boldsymbol{I}^{\mathrm{d}}_{pqk}, \tag{45}$$

其中 $\boldsymbol{I}_{\mathrm{dloss}}$ 的各项均不为 0。

数据： V；ϵ或 ε
结果： 保留的奇异值数量列表，适用于每条基线 $\{n_{pq}\}$，$\forall pq$

$n_{pq} \leftarrow 0$；$V_{pq,\ n_{pq}} \leftarrow \mathbf{0}$
如果给出了 ϵ，则
  对于所有基线 pq **执行**
    do
      $V_{pq,n_{pq}} \leftarrow V_{pq,n_{pq}-1} + \eta_{pqn_{pq}} a_{pqn_{pq}} c^{\dagger}_{pqn_{pq}}$
    当 $\frac{||V_{pq}-V_{pq,n_{pq}}||_{\mathrm{F}}}{||V_{pq}||_{\mathrm{F}}} > \epsilon$ 和 $n_{pq} \leq r$；
    $n_{pq} \leftarrow n_{pq}+1$
  end
否则
  对于所有基线 pq **执行**
    do
      $V_{pq,n_{pq}} \leftarrow V_{pq,n_{pq}-1} + \eta_{pqn_{pq}} a_{pqn_{pq}} c^{\dagger}_{pqn_{pq}}$
    当 $\frac{||V_{pq,n_{pq}}||_{\mathrm{F}}}{||V_{pq}||_{\mathrm{F}}} \times 100 < \varepsilon$ 和 $n_{pq} \leq r$；
    $n_{pq} \leftarrow n_{pq}+1$
  end
结束

算法 2： 使用*BDSVD*查找 $\{n_{pq}\}$，$\forall pq$。此处，所有 V_{pq} 均取自 V。

数据： V；N_p
结果： V 的奇异值分解按基线分解：$\{\mathbf{A}pq, \mathbf{\Lambda}_{pq}, \mathbf{C}_{pq}\}$，$\forall pq$

对 所有基线 pq，并行**执行**
  对 所有 i，并行**执行**
    $V^{\dagger}_{pq} V_{pq} + = V^{[i]\dagger}_{pq} V^{[i]}_{pq}$
  end
  $\mathbf{C}_{pq}\mathbf{\Lambda}^2_{pq}\mathbf{C}\dagger_{pq} = V\dagger_{pq} V_{pq}$
  $\mathbf{\Lambda}_{pq} = \sqrt{\mathbf{\Lambda}^2_{pq}}$
  对所有 i，并行**执行**
    $\mathbf{A}^{[\,]}_{pq} = V^{[\,]}_{\mathrm{ii}pq} \mathbf{C}_{pq} \mathbf{\Lambda}^{-1}_{pq}$
  end
  $\mathbf{A}_{pq} = \{\mathbf{A}^{[\,]}_{pq}\}_{\mathrm{i}}$，$\forall i$
end

算法 3： 求解 $\{\mathbf{A}pq, \mathbf{\Lambda}_{pq}, \mathbf{C}_{pq}\}, \forall pq$，使用并行计算。这里，所有的 V_{pq} 均取自 V，而 N_p 是计算节点的数量。

4.2 噪声滤波和信噪比

关于噪声滤波，有三个重要问题值得探讨：简单奇异值分解 和 *BDSVD* 的使用：i）对于在所有基线上保留相同数量分量的简单奇异值分解，噪声表现如何？ii）BDSVD及其变化的压缩因子如何影响各基线上的噪声？滤波后的噪声与简单奇异值分解和传统平均相比有何不同？iii）使用简单奇异值分解和BDSVD时，哪种基线上的噪声会被大量滤除？本节将通过适当的条件化和讨论来解答这些问题。

紧凑型可见度数据矩阵的条目 V_{pq} 来自对$\mathcal{V}$ 与噪声$\mathcal{E}$ 之和进行采样，如公式 8所示。因此，每个 η_{pqk} 反映了从 V +E 中采样的分量所贡献的信号强度。较小的 η_{pqk} 对应于 V_{pq} 中受到噪声严重污染的分量，这意味着保留 η_{pqk} 较大的分量，应该等同于去除噪声并保留有用信号。这意味着，使用简单奇异值分解和*BDSVD*，压缩数据中的噪声分别在 $n < r$ 和 $n_{pq} < r$的情况下有所降低。下面，我们将详细介绍简单奇异值分解和*BDSVD* 相较于传统平均，在相同 CF下的噪声滤波能力。

对于相同的压缩因子，简单奇异值分解和*BDSVD* 的可见度噪声惩罚解析估计值，是相对于传统平均的噪声降低量：

$$\Xi_X = \frac{\Sigma_X}{\Sigma_{avg}}, \tag{46}$$

其中，Σ_X 和 Σ_{avg} 分别是使用简单奇异值分解（或*BDSVD*）和传统平均压缩后的噪声。目标是表明 $\Xi_X < 0$，这意味着 $\Sigma_X < \Sigma_{avg}$。

假设所有基线和样本的噪声项均具有恒定的均方根值 Σ，对信号进行平均去噪时n_{avg} 个样本，噪声的降低幅度被充分证明为$\Sigma_{avg} = \Sigma/\sqrt{n_{avg}}$前提是样本之间的噪声不相关；这证实了 $\Sigma_{avg} < \Sigma$。如前所述，奇异值分解去除了数据中的部分噪声，因此，$\Sigma_X < \Sigma$。然而，在相同的压缩因子下，无法轻易通过解析方法看出 $\Sigma_X < \Sigma_{avg}$（详细的数学讨论请参阅附录A）。第5 节将通过实测数据表明，在相同的压缩因子下，$\Sigma_X < \Sigma_{avg}$ 因此 $\Xi_X < 0$。

在图像域中，中心像素的噪声惩罚估计值由下式给出。

$$\Xi^W_X = \frac{\sum_{pqt_iv_j} W^2_{pqt_iv_j} \Xi^2_X}{(\sum_{pqt_iv_j} W_{pqt_iv_j})^2}, \tag{47}$$

其中，$W_{pqt_iv_j}$ 为每个可见度的成像权重。

如上所述，较小的分量η_{pqk} 噪声较大，因此在需要考虑噪声时，可以将其作为移除对象。此外，较大的分量η_{pqk} 包含了需要保留的大部分信号。简单奇异值分解和*BDSVD* 仅保留在给定阈值处理条件下具有较大η_{pqk} 的分量，不满足该阈值的分量则会被舍弃。*BDSVD* 与简单奇异值分解的不同之处在于，所保留的分量数n_{pq} 取决于基线。如第3.2节所述，前η_{pqk} 个分量在较短基线上较大，并且与较长基线相比，下降得更快；而在较长基线上，η_{pqk} 则缓慢下降（图1 清楚地展示了η_{pqk}的这一特征）。这意味着，如果像简单奇异值分解那样，在所有基线上保留相同数量的分量，那么较长基线上一些包含信号的分量就会被舍弃，从而使压缩后的数据失真。另一方面，在较短基线上，除了所有包含强信号的分量外，还会保留若干噪声分量。*BDSVD* 利用了这一缺点，构建出基线相关的保留分量数，从而确保每条基线上仅保留信号强度较大的分量。与简单奇异值分解相比，在相同的 CF下，这种方法能够大幅滤除噪声，同时保持信号保真度。基于这一点，在相同压缩因子的情况下，我们可以写出：

$$\Sigma_{\mathrm{svd}} > \Sigma_{\mathrm{bdsvd}} \implies \Xi^W_{\mathrm{svd}} > \Xi^W_{\mathrm{bdsvd}}, \tag{48}$$

<escape>其中，Σ奇异值分解和 ΣBDSVD（分别表示 Ξ^W奇异值分解和 Ξ^W BDSVD）采用简单奇异值分解和*BDSVD*时产生的噪声（分别为噪声惩罚）。从解析角度来看，式48 清楚地表明，与简单奇异值分解相比，*BDSVD* 能够降低噪声。第5 节将通过实测结果验证式48中的解析结论。</escape>

我们用于测量压缩图像中每个像素的信噪比（ S/R ，以分贝为单位）的指标为：

$$S/R = 10\log_{10}\frac{\widetilde{\mathcal{I}_I}}{\sigma_{\text{pix}} + c_{\text{noise}}}, \tag{49}$$

其中 $\widetilde{\mathcal{I}_I}$ 是 $\mathcal{I}_I$ 的压缩版本；天空中不存在任何电子设备造成的损坏，也不存在会在信号传向仪器的路径上产生干扰的效应（见式2）。在此表述中，$c_{\text{噪声}}$ 表示来自视场之外各源的信号，而$\sigma_{\text{像素}}$ 噪声是脏图像中的逐像素噪声：

$$\sigma^2_{\text{pix}} = \frac{\sum_{pqt_i\nu_j} W^2_{pqt_i\nu_j}\Sigma^2_X}{(\sum_{pqt_i\nu_j} W_{pqt_i\nu_j})^2}. \tag{50}$$

5 仿真

本节采用 MeerKAT 和欧洲甚长基线干涉测量网（EVN）作为参考望远镜，评估并比较各方法的性能：传统平均、BDA、简单奇异值分解 和 *BDSVD*。评估三项不同的指标：i) 测量源振幅损失与基线长度的关系，以及基线依赖压缩因子与基线长度和东西向基线长度的关系；ii) 测量振幅损失与源在天空中位置的关系；iii) 为测量频谱和时间分辨率，测量点扩散函数（PSF）形状与源在天空中位置的关系；iv) 测量信噪比（S/R）与源在天空中位置的关系。

请注意，对于二维可见度数据的秩–n 奇异值分解，$\boldsymbol{V}_{pq,\,n}$ 如公式 11 所定义； M和 N分别是时间步数和信道数。在定义了$\boldsymbol{V}_{pq,\,n}$的空间中，行数据和列数据分别是时间观测和频率观测。当$\boldsymbol{V}_{pq,\,n}$ 被投影到秩–n奇异值分解空间时，时间观测和频率观测会混合在一起，因此难以完全确定对秩–n 进行压缩的方向。

为了将所提出的方法与传统平均和 BDA 进行比较，我们需要沿明确的压缩方向（例如时间或频率方向）评估上述指标。举例来说，为了评估时间方向上的指标，我们采用以下策略：$\boldsymbol{V}_{pq,\,n}$ 从大小为 $M \times N$ 的数据矩阵缩放为大小为 $M_1 \times M_2 \times N$的数据矩阵，其中秩–n 奇异值分解仅沿时间方向应用，分块大小为 $M = M_1 \times M_2$。本节采用的记号 $CF = CF_t \times CF_\nu$ 表示数据在时间方向压缩 CF_t 倍，在频率方向压缩CF_ν 倍。本节不讨论频率方向上的压缩（$CF_\nu = 1$）；不过，如果将$\boldsymbol{V}_{pq,\,n}$ 缩放为 $M \times N_1 \times N_2$，并且仅对$\boldsymbol{V}_{pq,\,n}$的$N = N_1 \times N_2$ 个条目应用秩–n 奇异值分解，则在频率方向上也能观察到相当的性能。

5.1 MeerKAT望远镜

对于工作在 1.4 GHz 的 MeerKAT望远镜，在 166 分 40 秒内模拟一个点源，共得到 10 000 个时间步，每步 1 秒。模拟的总带宽为 0.8 MHz，分为 10 个信道，每个信道宽 80 千赫。工作在 1.4 GHz 的 MeerKAT望远镜观测完整视场，其半径为 ~ 2。25 度；此处为未受时间和频率模糊影响时，主波束的第二个零点，时间步长为 1 秒，信道宽度为 80 千赫。该高分辨率数据集被缩放至 $100\times100\times10$，并在大小为 100×100 的每个频率块上分别应用简单奇异值分解 和*BDSVD*。压缩因子

我们采用 $CF = 25 \times 1$、 $CF = 35 \times 1$ 和 $CF = 50 \times 1$，这分别对应时间方向上 96。0%、97。14% 和 98。0% 的空间节省率。对于该高分辨率数据集， $CF = 50\times1$是所提出方法所能达到的最大时间压缩因子。这是因为秩-1奇异值分解的 $CF_t = 50$ ；换言之，所有基线中只保留一个分量，且如公式 33 100×100 $100+100+0.5$ ~ 50 所述，压缩因子变为 /(。)。

Th传统平均和 BD 所采用的策略 A 是：

- 创建了两个低分辨率测量集，分箱大小分别为 $25s \times 5kHz$ 和 $50s \times 5kHz$，用于接收分别采用 $CF = 25 \times 1$ 和 $CF = 50 \times 1$进行传统平均后的重采样可见度。
- 使用第2.2节所述的方法，创建第三个测量集，用于接收经过 BDA 降采样的可见度，其中 $CF = 25 \times 1$ 和 $CF = 50 \times 1$。
- 创建第四个测量集，其副本与高分辨率测量集相同，用于接收经重采样的可见度数据，以进行简单奇异值分解 和*BDSVD*。无需备份策略来保存简单奇异值分解 和*BDSVD*的解压缩后的可见度数据。采用的压缩因子为$CF = 25 \times 1$ 和 $CF = 50 \times 1$ ，用于简单奇异值分解；对于*BDSVD* ，则采用$CF = 25 \times 1$ 和 $CF = 35 \times 1$ 。由于 $CF = 50 \times 1$ 是最大压缩因子，采用 $CF = 50 \times 1$ 的*BDSVD* 将等同于采用 $CF = 50 \times 1$的简单奇异值分解 。因此，我们不使用 $CF = 50 \times 1$ 运行*BDSVD* 。

5.1.1 振幅与东西向基线长度

这项初步测试旨在量化一个振幅为 1 央斯基、位于 2.25 度处的单个点源所产生的去相关；该位置对应于 MeerKAT 在 1.4 GHz 下主波束的第二个零点。结果见图3。

对于 $\varepsilon = 99\%$，图3（上图）以对数刻度展示了基线依赖压缩因子 CF_{pq} 随东西向基线长度增加的变化。秩–n 奇异值分解相关方法 CF_{pq} 在根据算法33 计算所有^npq 后，由方程2得出；而对于 BDA， CF_{pq} 则按照Atemkeng 等人(2018)中的讨论计算。这是本研究的一项重要结果，因此值得详细解释。

尽管我们理解 CF_{pq} BDA 的Atemkeng 等人(2018)中讨论的 CF_{pq} ，但*BDSVD* 的东西向投影基线呈现出不同的模式和强度。可以观察到，对于 BDA， CF_{pq} 会随东西向投影基线长度增加而严格递减；而对于*BDSVD*，长度相同的东西向投影基线可能具有不同的 CF_{pq} ，其中一些CF_{pq} 显著大于另一些。不过，这一现象仍表明，与较长的东西向基线相比，较短的东西向基线上压缩了更多数据；但其模式和强度与 BDA 并不相同。图3 （中图）展示了 CF_{pq} 随基线长度增加的变化。我们观察到，当采用 BDA 或BDSVD时，长度相同的基线可能具有不同的 CF_{pq}。而 CF_{pq} 在*BDSVD* 下呈现出更明显的递减趋势，在 BDA 下则呈随机分布；这证实去相关并非基线长度的函数，而是东西向基线长度的函数。在相同的总体压缩因子 CF下，BDA 和*BDSVD* 表现出截然不同的 CF_{pq}。

图 3 （下图）显示了源振幅（1 央斯基）随东西向投影长度增加而变化的情况。从结果可以清楚看出，*BDSVD* 的表现优于 BDA。与 *BDSVD* 相比，BDA 会使源振幅衰减，而 $\varepsilon = 99\%$ 以上的源振幅得以保留。此外，我们观察到，使用*BDSVD* 时，对于某些东西向基线长度相同的情况，去相关源的振幅并非恒定。事实上，如

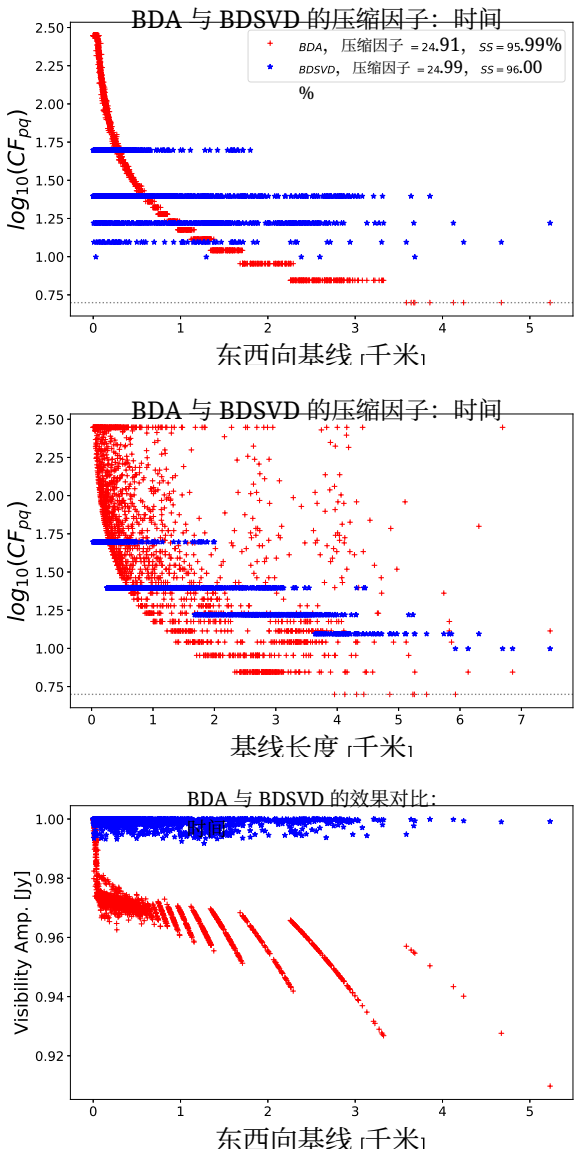


图3. 上图显示了随东西向基线长度增加而变化的基线依赖压缩因子 CF_{pq}（对数刻度），中图则以对数刻度显示了 CF_{pq} 随基线长度的变化。下图显示了 1 央斯基点源的振幅随东西向基线长度增加的变化。与*BDSVD* 相比，BDA 的差异明显不同 CF_{pq}；而*BDSVD* 在保留 1 央斯基源振幅方面优于 BDA。

图 3（上图）所示；CF_{pq} 会因某些相同的东西向基线长度而异，从而导致不同程度的去相关。BDA 也能观察到类似的现象：属于同一平均范围的东西向基线长度，其去相关会更快显现。例如，所有长度介于 2.2 km 和 3.5 km 之间的东西向基线都属于同一平均范围，其中 ~ 2 个分箱会被一起平均。由于同一平均范围内的东西向基线长度各不相同，因此会产生不同程度的去相关。

5.1.2 振幅与源位置

本节考察振幅去相关随天空中源位置变化的结果，并测量和比较各压缩方法所产生的噪声惩罚。为测量热噪声，使用上述高分辨率数据集，在空天空中填入 1 央斯基热噪声，并应用不同的压缩方法。

图 4 展示了时间压缩的结果。关于这些结果，有几点值得注意。为了保留至少 $\varepsilon = 99\%$ 的源振幅，需要采用 $CF = 25\times1$ 的压缩因子。在这一压缩范围内，简单奇异值分解和 BDSVD 能够使源振幅几乎保持平坦，仅出现可忽略的 1% 衰减，该衰减从半径约为 ~ 4 度的位置开始向外延伸；而对于 BDA 和传统平均，1% 的衰减分别从半径约为 ~ 1.32 度和 ~ 0.8 度的位置开始向外延伸。采用 $CF = 50\times1$ 的压缩因子时，对于 $\varepsilon = 95\%$ 的源振幅，可提供半径达 ~ 2.9 度的视场；而在这一压缩因子下，传统平均和 BDA 所能提供的视场半径分别最多仅为 ~ 1.32 度和 ~ 2 度。

还可以注意到，传统平均仅能为 $CF = 25\times1$ 提供半径为 ~ 1.32 度的视场；而使用简单奇异值分解获得相同的半径为 ~ 1.32 度的视场时，数据压缩率可以比传统平均高两个数量级。传统平均或 BDA 无法将数据压缩 $CF = 25\times1$ 倍，同时在 1% 衰减率下实现半径为 ~ 2.25 度的视场。然而，在这一 1% 衰减率下，*BDSVD* 可为 $CF = 35\times1$ 实现半径为 ~ 2.25 度的视场。*BDSVD* 兼具 BDA 和简单奇异值分解的优势，充分发挥了自身潜力。

其值为 $\Xi \sim 1$ 对于 BDA，这一结果也已在 Atemkeng 等人 (2018) 中观察到；在相同的压缩因子下 CF，BDA 和传统平均的噪声降低程度均随 $\sqrt{CF}$变化。 $\Xi < 1$ 对于 简单奇异值分解的值，说明了奇异值分解常用于信号去噪。 Ξ 对于 *BDSVD*的值低于 简单奇异值分解，这证实了第 4.2 节讨论的理论结果；*BDSVD* 的噪声性能优于简单奇异值分解，且压缩因子相同。

5.1.3 PSF 畸变与源位置的关系

如第 1 节所述，对可见度数据进行平均处理的一个负面影响是导致点扩散函数（PSF）发生畸变，其中较长的基线是造成这种畸变的主要因素。天空中某一给定位置处 PSF 的振幅可用于衡量信号损失，而其宽度（例如半高全宽 FWHM 处的宽度）则描述了源在天空该位置处的扩展程度。本节通过测量应用各类时间压缩方法后 PSF 在半高全宽（FWHM）处的宽度，来评估 PSF 畸变。我们重新使用第 5.1 节中讨论的包含 $1s\times80kHz$ 个时间区间的模拟高分辨率数据集，比较所有不同方法下的三个压缩因子 $CF = 25\times1$、$CF = 35\times1$ 和 $CF = 50\times1$。

结果如图 5 和 6所示。图 5 展示了当高分辨率数据集采用以下方法进行压缩时，位于 $CF = 25\times1$ 的 2.25 度处一个源的点扩散函数（PSF）：BDSVD （左上）、简单奇异值分解 （右上）、BDA（左中）和传统平均（右中）。图 5 还展示了归一化点扩散函数的垂直方向（左下）和水平方向（右下）。通过目视检查可以清楚地看出，所有压缩方法得到的点扩散函数宽度均有所不同，其中传统平均得到的点扩散函数比其他方法更宽。尽管 BDA 是一种能够保留振幅损失的潜在压缩方法，但其产生的点扩散函数

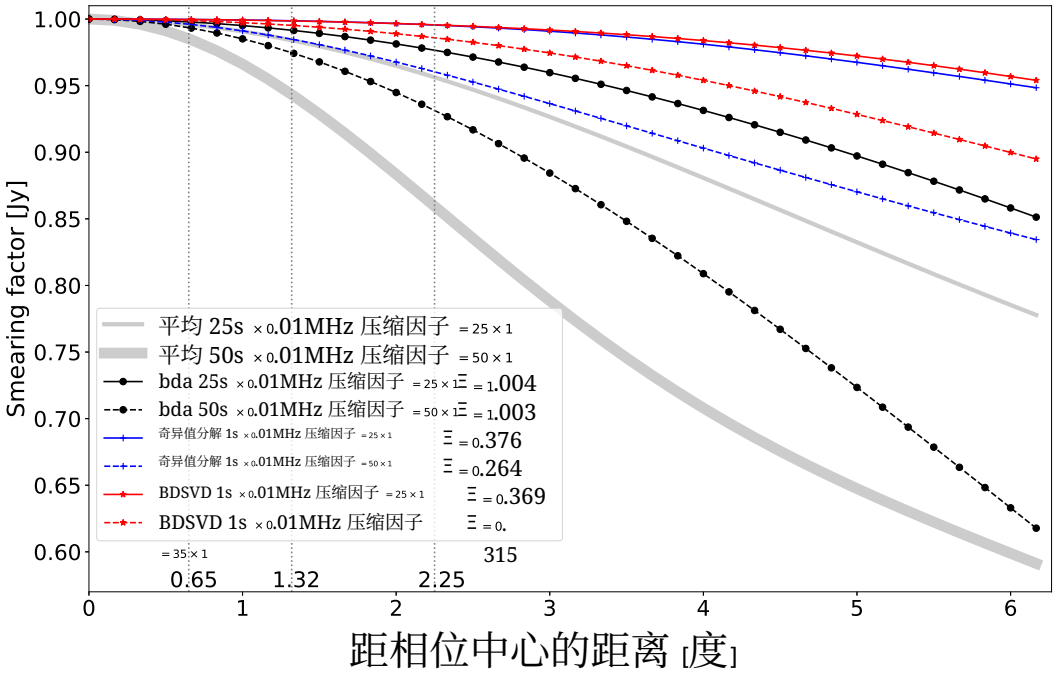


图 4. MeerKAT望远镜在 1.4 GHz 下观测 166 分 40 秒，总带宽为 0.8 MHz；展示了位于天空不同位置的 1 央斯基源的振幅损失程度及相关噪声惩罚。对于传统平均、采用 $CF = 25 \times 1$ 和 $CF = 50 \times 1$ 的 BDA，以及采用 $CF = 25 \times 1$、$CF = 35 \times 1$ 和 $CF = 50 \times 1$的 简单奇异值分解和 *BDSVD*，给出了随源距相位中心的距离变化的涂抹效应。相对于传统平均分箱，空间节省量为 SS，噪声惩罚为 Ξ。

形状与传统平均和奇异值分解相关方法的形状完全不同。

为量化点扩散函数的宽度，图 6 展示了观测中径向和切向点扩散函数分辨率的平均值（在半高全宽处测量），以及它们随距相位中心的距离的变化。*BDSVD*表现出卓越的能力，在半径大于 3 度时，仍能保持点扩散函数不失真，参数为 $CF = 25 \times 1$ 和 $CF = 35 \times 1$。对于 简单奇异值分解，也观察到类似结果，参数为 $CF = 25 \times 1$。然而，对于 简单奇异值分解，点扩散函数从半径 ~ 1。2 度处开始出现失真，参数为 $CF = 50 \times 1$。采用传统平均或 BDA 时，各压缩因子下的点扩散函数都很快开始出现失真，起始位置约为 ~ 1。2 度。这些结果证实，*BDSVD* 和 简单奇异值分解 能够在半径超过2。25 度的范围内保持点扩散函数不失真；这是 MeerKAT望远镜在 1.4 GHz 下能够达到的最大半径。

5.1.4 相对信噪比

第4.2 节理论讨论的信噪比如图7 所示，它是天空中源位置的函数。我们使用第5.1.2 节中的模拟，在应用不同压缩因子下的压缩方法时测量源振幅。

本节包含两项模拟，用于测量热噪声以及视场外源产生的噪声。使用上述高分辨率数据集构建 i) 含有 1 央斯基热噪声的空天空，以及 ii) 在距相位中心 20 度处模拟一个 10 央斯基的源，并分别对每项模拟应用不同的压缩方法。远场干扰通过尺寸为 $2^{10} \times 2^{10}$ 的图像进行测量。请注意，除远场源产生的信号外，该图像应为空。*BDSVD*和简单奇异值分解 的信噪比显著提升，在 2.5 度处达到约 ~ 1.5 dB，优于传统平均。然而，与传统平均相比，BDA 并未改善信噪比；这是因为 BDA 和传统平均的热噪声降低幅度相同，而与传统平均相比，BDA 并不会抑制视场外的源。

5.2 EVN

我们研究了在甚长基线干涉测量中应用简单奇异值分解和*BDSVD*，以在大幅压缩数据的同时，将去相关控制在一定水平以下。我们模拟了在 1.6 GHz 下进行的 15 分钟全 EVN（即 Badary、Effelsberg、Hartebeesthoek、Jodrell Bank、Medicina、Noto、Onsala、上海、Svetloe、Torun、Westerbork、Zelenchukskaya）观测。总带宽为50 kHz，划分为 10 个信道，每个信道宽度为5 kHz；15 分钟的观测以每0.01 秒的间隔采样，得到90 000 个时间步。为了应用简单奇异值分解和*BDSVD*，将90 000 个时间步缩放至 300×300，由此我们能够考察压缩因子 $CF = 10 \times 1$、$CF = 95 \times 1$ $CF = 150 \times 1$ 1 和 。我们模拟了一个振幅为 Jy 的单个点源，并测量其振幅，将其与传统平均和 BDA 的结果进行比较。结果如图8所示。可以观察到，当涂抹因子为1% 且 $CF = 10 \times 1$时，传统平均和 BDA 分别会产生 6 角分和 18 角分的视场。而采用$CF = 95 \times 1$时，*BDSVD* 能够成像超过 30 角分的视场，并具备更优的噪声降低能力。尽管这些结果在甚长基线干涉测量领域表现良好，

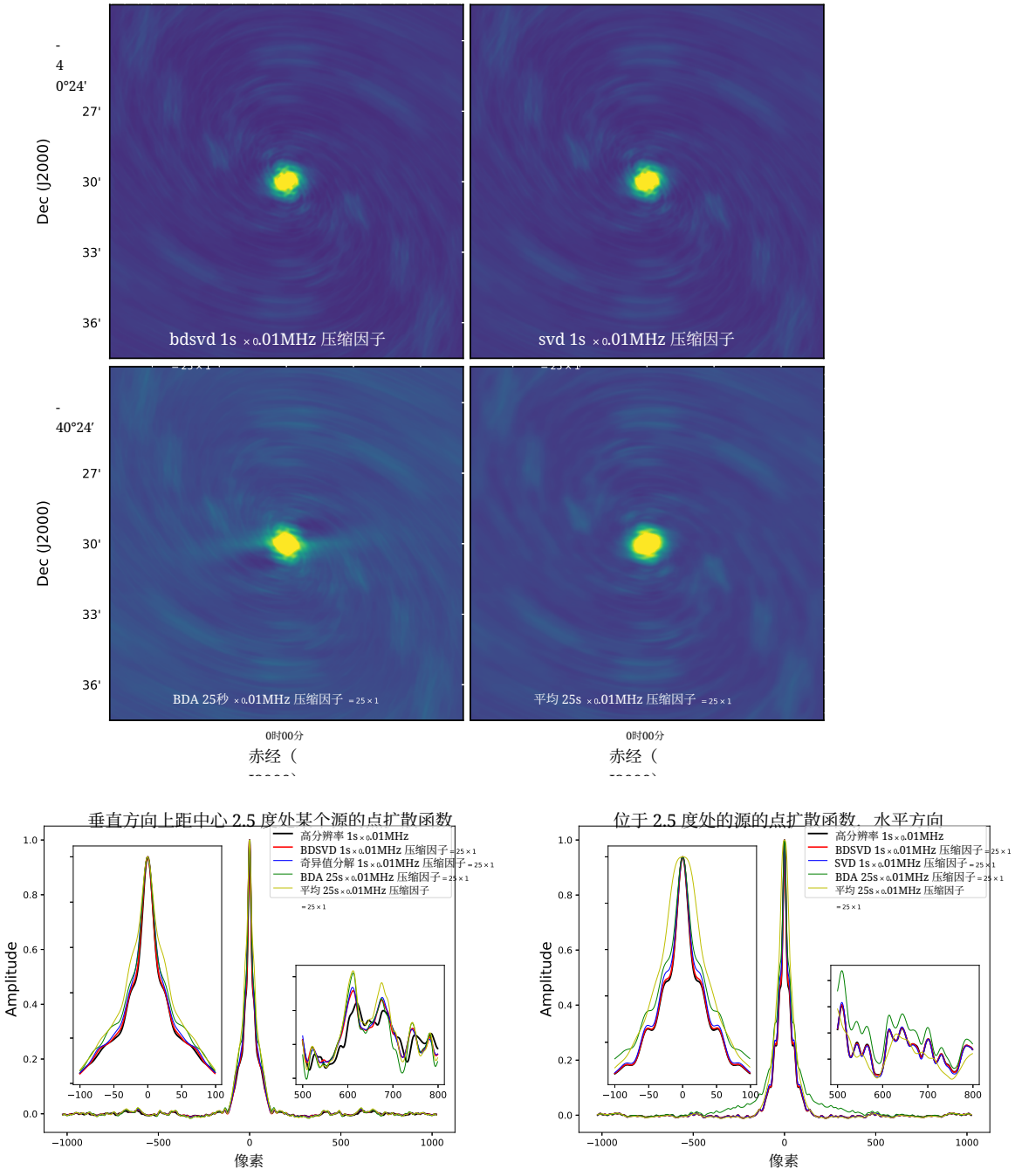


图 5. MeerKAT望远镜在 1.4 GHz 下观测 166 分 40 秒，总带宽为 0.8 MHz。当高分辨率数据集使用 $CF = 25 \times 1$ 压缩时，显示了位于 2.25 度处某个源的点扩散函数：采用*BDSVD* （左上）、简单奇异值分解 （右上）、BDA（左中）、传统平均（右中），以及归一化点扩散函数的垂直方向（左下）和水平方向（右下）。

但在真实数据上实现不同的压缩方法仍至关重要。

6 结论

造成数据量庞大的一个主要因素是阵列的长基线，因为它们会影响为避免天体物理信号去相关所需的数据采样程度。采用 BDA 时，采样率取决于基线，因为短基线可以采样得更加粗略，从而实现较高的压缩率。BDA 是一种成熟的射电干涉测量数据压缩技术。然而，该技术会产生不规则采样的数据。尽管 MS 格式通过许多频谱窗支持可变频率通道，因此在技术层面能够支持此类数据，但仍需要以特定方式重构数据，而这会降低数据处理性能。本研究提出了一种利用低秩矩阵近似的方法，能够在实现更高压缩率的同时，与 BDA 或传统平均相比，显著减小涂抹效应。更令人瞩目的是，即使.....

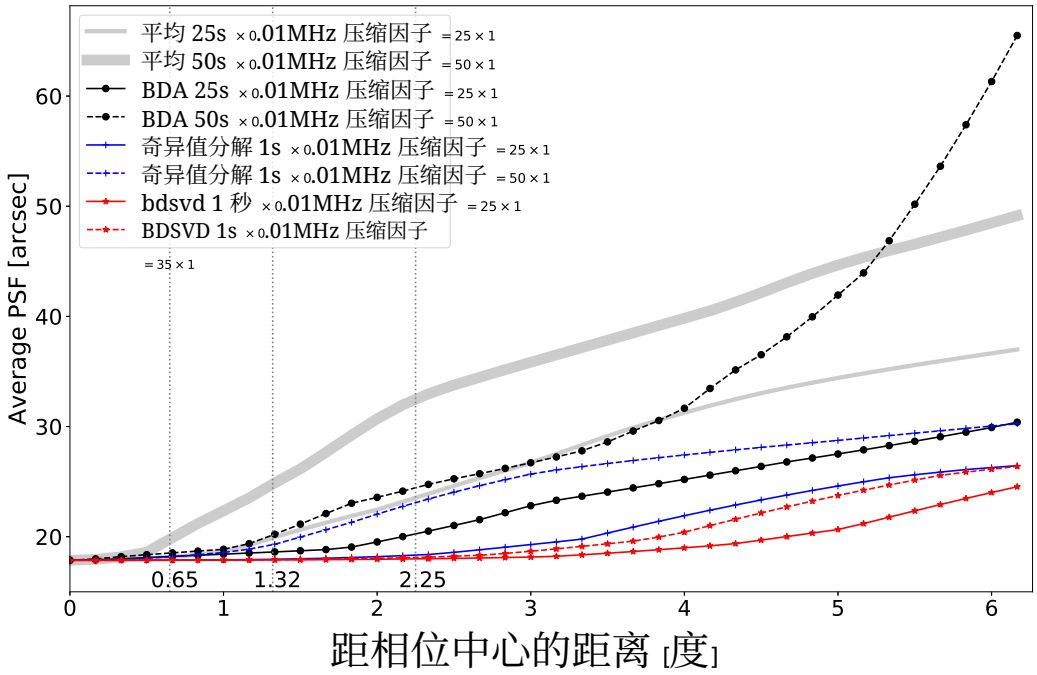


图 6. MeerKAT望远镜在 1.4 GHz 下观测 166 分 40 秒，总带宽为 0.8 MHz。图中显示了在不同天空位置测得的径向和切向点扩散函数分辨率的半高全宽（FWHM）平均值，分别采用传统平均和 BDA（$CF = 25 \times 1$ 和 $CF = 50 \times 1$），以及简单奇异值分解、*BDSVD*（$CF = 25 \times 1$、$CF = 35 \times 1$ 和 $CF = 50 \times 1$）。采用与奇异值分解相关的方法时，视场边缘的点扩散函数不会发生畸变。

更令人振奋的是，将低秩近似与基线相关形式体系相结合后，*BDSVD* 能有效消除 MeerKAT望远镜和 EVN 视场内的涂抹现象。不过，该方法有三个注意事项。

首先，奇异值分解计算开销很大，对于大型数据集，如果对每条基线分别执行，可能并不实用。幸运的是，我们表明，这些计算可以分布到多个计算机节点上，具体讨论见算法 3。

其次，简单奇异值分解和*BDSVD* 不会影响当前的校准算法或软件。由于压缩数据在校准前会先解压，因此仍然可以对可见度数据进行校准。这意味着时间和频率分辨率得以恢复，因而不会改变校准所需的解算间隔。我们还注意到，当前成像软件的功能得以保留。一个非常重要的结果是，压缩数据、压缩噪声与源之间的线性变换得以保留，这表明只要充分理解 $\boldsymbol{I}^{\mathrm{d}pqk}$，就可以在压缩域中对天空成像。这为一项潜在的研究开辟了方向，或可探索在奇异值分解空间中对数据成像。

第三，第 2.2 节讨论的实现方法能够将 BDA 数据压缩为磁盘和内存中均兼容 MS 的格式。相比之下，此处介绍的奇异值分解压缩方案会生成奇异值，这些奇异值必须在内存中“即时”扩展至完整分辨率。因此，这两种方法形成了有趣的对比：第一种方法减少了数据量，也减少了算法处理这些数据时执行的计算量（FLOPs）。但这需要付出实现复杂度增加的代价，尤其是在校准技术方面。

校准技术必须对跨越稀疏域中多个数据点的解算区间进行推理，而 BDA 数据就位于该稀疏域中。 简单奇异值分解 和 *BDSVD* 的压缩效果优于 BDA 和传统平均，但实现复杂度较低，代价是需要以完整数据分辨率处理数据。PyDATA 生态系统中的 BLOSC（Haenel 2014）等软件包已支持“即时”解压数据：它们将数据压缩成块，并将其精确解压至 CPU 的 L1 缓存。这降低了算法的内存访问压力；这很有益处，因为现代 CPU 的内存系统常常无法及时为其提供所需数据（Alted2010），而射电干涉数据的规模又非常庞大。在这种情况下，可以将奇异值传输到 CPU 的 L1 缓存，并将数据扩展至完整分辨率，供算法使用。未来还可以研究利用奇异值分解相关方法的完整归档潜力，实现在线压缩与解压策略。

在受到射频干扰（RFI）污染的真实数据中，对奇异值分解进行解压缩时，功率会分布到所有可见度中。有两种方法可以避免这种情况：要么在应用奇异值分解之前从数据中去除射频干扰，要么在压缩过程中使用奇异值分解去除射频干扰。后者需要作为未来工作进行深入研究，因为去除射频干扰相当于去除较高奇异值分量。去除较高奇异值分量存在问题，因为目前尚不清楚在这一过程中，微弱射电信号会从何处开始受到抑制。

真实数据通常会有少量条目因某些原因被标记。一个值得探究的未来研究方向是，考察这些被标记的条目会如何影响奇异值。我们预计，实际情况会介于以下两种情形之间：

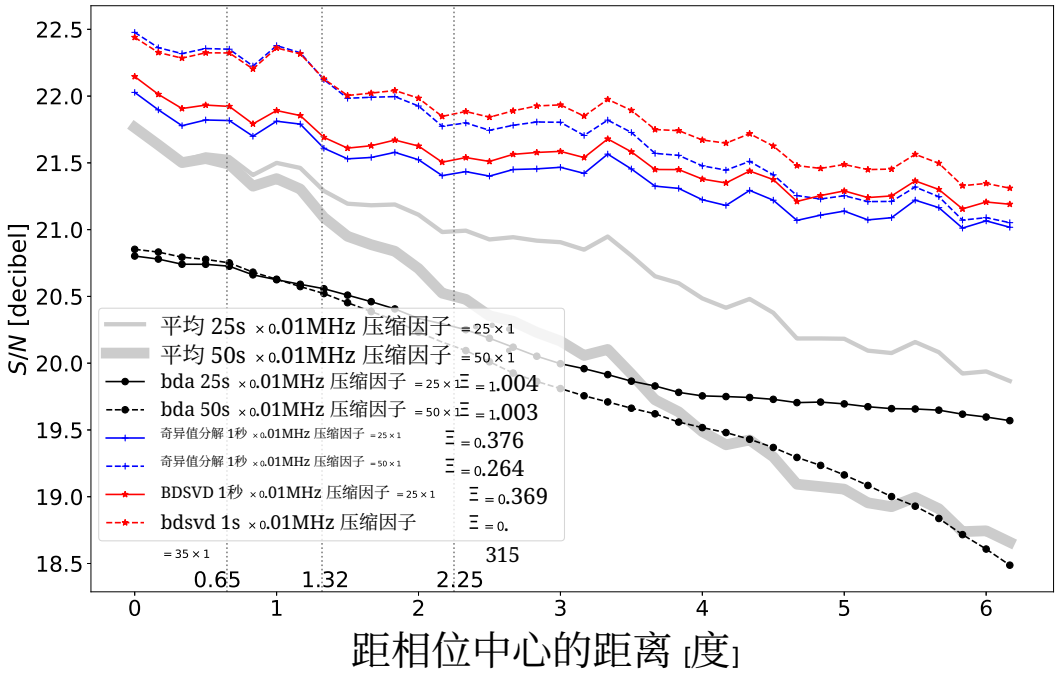


图 7. MeerKAT望远镜在 1.4 GHz 下观测 166 分钟 40 秒，总带宽为 0.8 MHz。根据方程 49 定义的信噪比，在不同天空位置测量压缩因子 $CF = 25\times1$、$CF = 35\times1$ 和 $CF = 50\times1$，采用传统平均、BDA、简单奇异值分解 和 *BDSVD*。与 SVD 相关的方法可提供更高的信噪比。

- 如果被标记的真实数据条目属于同一个邻域，且被赋予相同的值，那么邻域内的相似性就会增加，因此，与未进行标记的情况相比，奇异值分解会得到更低的秩（相应地，压缩因子会更大）；并且
- 如果被标记的真实数据项并非来自同一邻域（例如，随机标记的数据），那么与未进行标记的情况相比，SVD 的秩会更高（因此压缩因子会更小）。

为解决这一问题，一种可能的方案是在存在标记的位置对模型数据进行修复，然后在应用奇异值分解之前，使用修复后的数据填补未压缩数据中被标记的条目。

致谢

本研究得到南非科学与技术部和国家研究基金会的南非研究主席计划支持。MeerKAT望远镜由南非射电天文台运营；该天文台是国家研究基金会的设施，而国家研究基金会是科学与创新部的机构。MA感谢罗德大学提供的支持。

数据可用性

本文所依据的数据可应合理请求向通讯作者获取。

参考文献

Alted F., 2010, 《科学与工程计算》, 12, 68 Ansah-Narh T., Akoto-Danso A., Proven-Adzri E., Takyi B., 2020, 《仪器学报》, 15, P04020 Atemkeng M., Smirnov O., Tasse C., Foster G., Jonas J., 2016, 《皇家天文学会月刊》, 462, 2542 Atemkeng M., Smirnov O., Tasse C., Foster G., Keimpema A., Paragi Z., Jonas J., 2018, 《皇家天文学会月刊》, 477, 4511 Atemkeng M., S mirnov O., Tasse C., Foster G., Makhathini S., 2020, 《皇家天文学会月刊》, 499, 292 Bobin J., Starck J.-L., Ottensamer R., 2008, 《IEEE信号处理精选主题期刊》, 2, 718 Cotton W., 2009, OBIT开发备忘录系列 Demmel J., Dumitriu I., Holtz O., 2007, 《数值数学》, 108, 59 Dewdney P. E., Hall P. J., Schilizzi R. T., Lazio T. J. L., 2009, 《IEEE会议录》, 97, 1482 Gu M., Eisenstat S. C., 1996, 《SIAM科学计算期刊》, 17, 848 Haenel V., 2014, CoRR, abs/1404.6383 Hamaker J., Bregman J., Sault R., 1996, 《天文学与天体物理学增刊系列》, 117, 137 Iheanetu K., Girard J., Smirnov O., Asad K., de Villiers M., Thorat K., Makhathini S., Perley R., 2019, 《皇家天文学会月刊》, 485, 4107 Johnston S., et al., 2008, 《实验天文学》, 22, 151 Jonas J. L., 2009, 《IEEE会议录》, 97, 1522 Kartik V. S., Carrillo R. E., Thiran J.-P., Wiaux Y., 2017, 《皇家天文学会月刊》, 468, 2382 Kemball A., Wieringa M., 2000, 网址：http://casa. nrao. edu/Memos/229. html Lee J., Kim S., Lebanon G., Singer Y., Bengio S., 2016, 《机器学习研究会议录》第

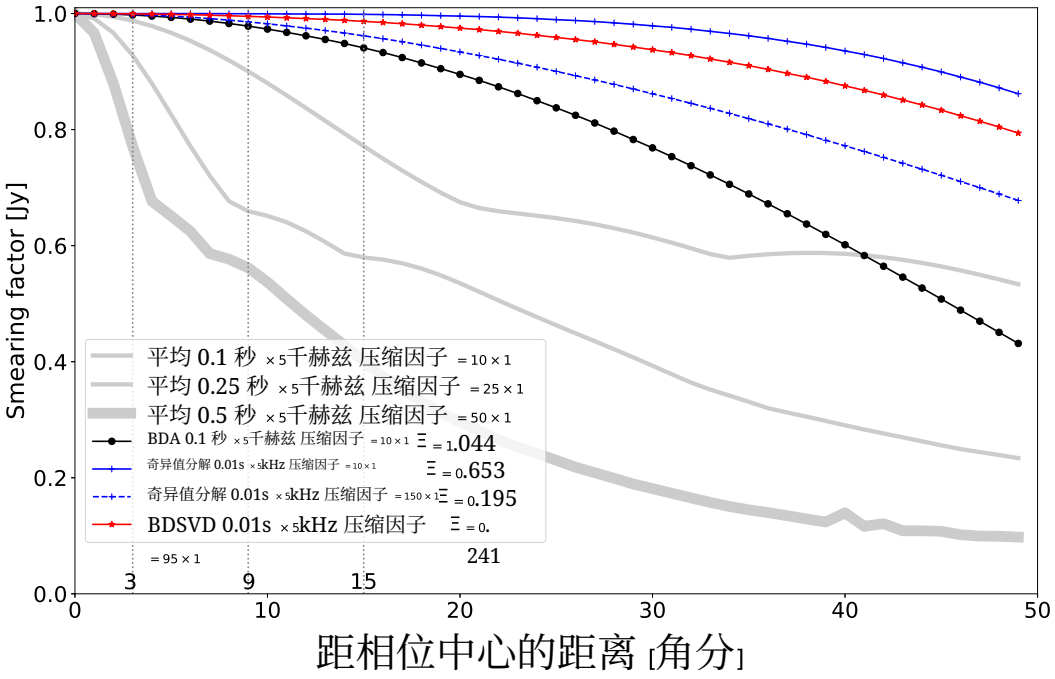


图 8. EVN 望远镜在 1.6 GHz 下观测 15 分钟，总带宽为 50 kHz；展示了不同天空位置处 1 央斯基源振幅损失的涂抹程度及其对应的噪声惩罚。对于传统平均、BDA、简单奇异值分解 和 *BDSVD*，涂抹程度随源距相位中心的距离变化，分别对应 $CF = 10\times1$、$CF = 25\times1$、$CF = 50\times1$、$CF = 95\times1$ 和 $CF = 150\times1$。噪声惩罚 Ξ 是相对于传统平均的分箱给出的。

《机器学习研究会议录》 Offringa A., De Bruyn A., Biehl M., Zaroubi S., Bernardi G., Pandey V., 2010, 《皇家天文学会月刊》, 405, 155 Shishkin S. L., Shalaginov A., Bopardikar S. D., 2019, 《数值线性代数及其应用》, 26, e2246 Smirnov O. M., 2011, 《天文学与天体物理学》, 527, A106 Stewart G. W., 1998, 《矩阵算法: 第1卷: 基本分解》. SIAM Tasse C., et al., 2018, 《天文学与天体物理学》, 611, A87 Wijnholds S., Willis A., Salvini S., 2018, 《皇家天文学会月刊》, 476, 2029 Zarka P., et al., 2015, 载于2015年国际天线理论与技术会议（ICATT）。第1–6页 van Haarlem M. P., et al., 2013, 《天文学与天体物理学》, 556, A2

附录 A: 追踪噪声

给定一次观测（假设 $M < N$）

$$\underset{M\times N}{Y} = \bar{Y} + \epsilon = \underset{M\times M}{U}\,\underset{M\times M}{\Lambda}\,\underset{M\times N}{V}, \tag{A1}$$

with

$$\mathrm{vec}(\epsilon) \sim N(0,\Sigma). \tag{A2}$$

阈值处理奇异值后，我们还能追踪噪声属性吗？我们可以利用该分解追踪奇异值的协方差。为此，我们注意到，由于 U和 V 是酉矩阵，因此有

$$\Lambda = U^{\dagger}(\bar{Y}+\epsilon)V^{\dagger} = \bar{\Lambda} + \epsilon_{\Lambda}, \tag{A3}$$

其中 $\bar{\Lambda} = U^{\dagger}\bar{Y}V^{\dagger}$ 以及 $\epsilon_{\Lambda} = U\dagger\epsilon V\dagger$。现在，SVD 空间中的协方差为

$$\begin{aligned}
\Sigma_{\Lambda} &= \mathbb{E}[\mathrm{vec}(\epsilon_{\Lambda})\mathrm{vec}(\epsilon_{\Lambda}^{\dagger})] && \text{(A4)}\\
&= \mathbb{E}[(V\otimes U^{\dagger})\mathrm{vec}(\epsilon)\mathrm{vec}(\epsilon)^{\dagger}(V\otimes U^{\dagger})^{\dagger}] && \text{(A5)}\\
&= (V\otimes U^{\dagger})\mathbb{E}[\mathrm{vec}(\epsilon)\mathrm{vec}(\epsilon)^{\dagger}](V\otimes U^{\dagger})^{\dagger} && \text{(A6)}\\
&= (V\otimes U^{\dagger})\Sigma(V\otimes U^{\dagger})^{\dagger}, && \text{(A7)}
\end{aligned}$$

其中我们使用了恒等式 $\mathrm{vec}(ABC) = (C\dagger\otimes A)\mathrm{vec}(B)$，以及以下事实:

$$\mathbb{E}[\mathrm{vec}(\epsilon)\mathrm{vec}(\epsilon)^{\dagger}] = \Sigma. \tag{A8}$$

请注意，根据 Kronecker 积的混合乘积性质，只要 ϵ 是独立同分布的（即 Σ 是单位矩阵的倍数），奇异值分解空间中的噪声也将是独立同分布的。在这种情况下，由于奇异值彼此独立，阈值处理后的不确定性很容易追踪。当 Σ_{Λ} 中存在非对角元素时，情况会稍微复杂一些。当 Σ 为对角矩阵时，Σ_{Λ} 很可能以对角元素为主，但这一点需要验证。如果确实如此，我们大概可以放心地将其视为对角矩阵。

本文由作者使用 TEX/LATEX 文件排版。