大数据情境下的数据完备化：挑战与对策*

陈松蹊　毛晓军　王　聪

摘要：随着数字经济时代的到来，数据作为一种重要的生产要素，深刻改变了管理决策范式。对具有超规模、跨领域、流信息的大数据的分析利用成为了赋能管理实践的重要因素，其中数据的质量与完备性是影响后续数据价值提炼的重要前提。然而受限于数据采集方式与过程、被采集主体行为模式特点等因素，数据常常呈现超高缺失率的特点。超高数据缺失会严重影响数据分析及所承载的管理决策效果。因而，预先对大数据进行有效完备化对保证后续分析决策效果具有重要意义。本文对大数据情境下的数据完备化问题进行了系统梳理，重点给出在超高维度、多源异构、时空关联的情境下的大数据完备化问题的主要挑战、求解思路及其对管理学研究的启示，以期为大数据完备化及赋能管理决策奠定理论和方法学基础。

关键词：数据完备化　超高维度　多源异构　时空关联　管理决策
DOI:10.19744/j.cnki.11-1235/f.2022.0015

一、引言

随着移动互联环境下新兴技术的快速发展，来自公共管理、电子商务、金融服务、医疗健康等应用领域的大数据不断涌现，深刻地改变了社会经济生活的面貌，推动我们所处的社会与经济向数字经济时代迈进。随着移动互联技术的深入、数据采集和存贮技术的飞跃发展，具有超大规模、超高维度、多源异构、流式产生特点的大数据日益可测可获，基于数据的管理决策逐渐成为科学研究和应用的主流（徐宗本等，2014）。近年来，对大数据的开发应用已上升至国家战略高度，2020年中共中央和国务院发布的《中共中央、国务院关于构建更加完善的要素市场化配置体制机制的意见》中，将数据与土地、劳动力、资本、技术等传统要素并列为生产要素之一。在这一环境下，领域情境、决策主体、理念假设、方法流程等决策要素受到冲击，催生了大数据决策范式的诞生（陈国青等，2020）。

数据作为大数据决策范式下的重要生产要素，其本身的完备与质量关乎后续决策效果。通过多种渠道采集而成的大数据尽管体量很大，但往往具有非常高的缺失比例，从而对利用其进行管理决策提出了新的挑战。如在线购物场景中，推荐系统常用于为用户推荐其感兴趣的商品或服务，以辅助其后续购物决策。用户历史评分数据常被用作推荐系统的输入，用于预测消费者对尚未购买商品的评分。然而，由于商品数量众多而用户接触到的商品非常有限，用户历史评分数据呈现高度缺失的特点。著名的在线视频公司Netflix曾举办过一个数据挖掘比赛，该比赛提供给选手的电影评分数据集具有大量缺失值（Feuerverger et al.，2012），如图1所示。该评分数据中共包含约48万观影者和1万8千部电影。然而每位观影者平均仅对约200部电影给出过评分，其他评分都是缺失的，缺失比例高达98.8%。

若直接使用具有超高缺失比例的数据训练推荐系统，难以对用户的真实偏好做出准确的预测，甚至会产生严重有偏差的推荐结果。这不仅会误导用户的购物决策，长此以往还会破坏用户对平台的信任（Kleinberg et al.，2018）。如能将该评分矩阵有效地进行补充，尽可能地恢复数据的原貌和内在结构，就可将该完整评分

*本文得到国家自然科学基金重大研究计划重点支持项目“面向管理决策大数据分析的理论与方法”（基金号：92046021）、国家自然科学基金重点项目“大数据驱动的管理决策模型与算法”（基金号：71532001）、国家自然科学基金青年项目“高维低秩矩阵完备化问题的研究”（基金号：12001109）、国家自然科学基金青年项目“基于动态适应性建模的个性化商品促销推荐方法与应用研究”（基金号：72101007）、上海市青年科技英才扬帆计划项目（19YF1402800）、上海市“科技创新行动计划”社会发展科技攻关项目（20dz1200600）的资助。王聪为本文通讯作者。

数据作为推荐系统模型的输入，进而为构建实时推荐系统、深层分析提供有效的准备。

超大规模的数据缺失问题，也给统计学研究带来了新的挑战。大量缺失数据的存在使得数据整体的不确定性增加，确定性成分更难把握。在小规模数据缺失的场景中，常对缺失数据进行删除处理，如Little和Rubin（2019）中介绍的complete-case analysis。然而在小数据集上的现有研究表明，缺失数据往往伴随选择偏差或隐性偏差，直接删除缺失数据，会造成数据资源的浪费，更可能加重由上述选择偏差导致的估计偏差。而对于超大规模缺失的大数据而言，数据删除方法会导致90%以上的数据被删除，显然是不可行的。因而，对大数据中的缺失数据进行完备化，尽可能地还原其固有的结构是大数据分析及进一步在其基础上进行管理决策的一个重要步骤。

尽管缺失数据填补是近30年统计学一个活跃的研究方向，形成了一套相关方法（Rubin，1987；Little，1988；Little and Rubin，1989；Allison，2000；Zhang，2003；Ibrahim et al.，2005；Reiter and Raghunathan，2007；Durrant，2009；Little and Rubin，2019）。但这些方法所能处理的缺失率鲜有能随着数据维度的变化而变化的，无法处理超大规模量级的缺失数据。此外，由于大数据具有超高维度、多源异质、流式产生等特点，对大数据完备化方法设计提出了挑战。因此，在对缺失数据进行完备化过程中，需充分考虑数据情境特点及其中的数据缺失机制，以设计简洁有效的数据完备化方法。本文将首先介绍数据完备化问题的一般性形式，进而考虑不同情境特点下的数据完备化方法设计问题，并给出在管理学领域的应用场景。

二、数据完备化问题的形式化定义

我们首先介绍大数据完备化问题的一般框架，再对不同数据缺失机制下的数据完备化方法进行梳理，并针对大数据的超高维度、多源异质、时空关联场景的特点分别展开探讨。由于大数据常常以矩阵形态存在，不失一般性，本文首先以矩阵形态考虑数据完备化问题，之后会扩展到更一般的流数据情况。

矩阵完备化研究的问题是如何根据较少的观测值精确地对原始矩阵进行还原。整个问题可以视为一个带有结构性假设的优化问题。在常见的矩阵完备化方法中，通常采用低秩结构假设，即高维矩阵的行或列是由少量行或列隐含生成。以上述Netflix电影评分矩阵高维矩阵为例，在低秩结构的假设下，可认为该矩阵只是由少量与电影类型及用户类型有关的隐变量生成。下面我们给出矩阵完备化的数学框架。

令$\boldsymbol{A}_0=(a_{0,ij})_{n_1\times n_2}$表示不可观测的真实矩阵，其具有$n_1$行$n_2$列。我们假设其具有低秩性质，即矩阵$\boldsymbol{A}_0$的秩$rank(A_0)$是一个比较小的整数。令$\boldsymbol{Y}=(y_{ij})_{n_1\times n_2}$是$\boldsymbol{A}_0$加上均值为0的噪音之后的可观测数据矩阵。在实际中，$\boldsymbol{Y}$只有小部分的元素可被观测到，其他为缺失元素。$\boldsymbol{Y}$的每个y_{ij}都可以写成$y_{ij}=a_{0,ij}+\varepsilon_{ij}$，其中$a_{0,ij}$表示$\boldsymbol{A}_0$对应位置的元素，$\varepsilon_{ij}$表示均值为0的干扰噪音。矩阵完备化问题的核心任务就是通过适当的完备化方法来得到真实矩阵$\boldsymbol{A}_0$的估计矩阵$\hat{\boldsymbol{A}}$。一般来说，$\hat{\boldsymbol{A}}$可以通过求解如下优化问题来获得：

$$\hat{\boldsymbol{A}}=\arg\min_{\boldsymbol{A}\in\mathcal{A}}\{L(\boldsymbol{A},\boldsymbol{Y})+R(\boldsymbol{A})\} \quad (1)$$

其中，$\mathcal{A}$表示由$\boldsymbol{A}_0$的可能解构成的解空间集合，$\arg\min_{\boldsymbol{A}}$代表关于$\boldsymbol{A}$的极小化。人们通常假设矩阵问题的解被限定在用无穷范数表示的球内，也就是说$\mathcal{A}=\{\|\boldsymbol{A}\|_\infty\leq a\}$，其中$\|\boldsymbol{A}\|_\infty=\max_{ij}\{|a_{ij}|\}$代表矩阵的无穷范数。$L(\boldsymbol{A},\boldsymbol{Y})$表示一个损失函数，即用于衡量矩阵$\boldsymbol{A}$与$\boldsymbol{Y}$的差异的函数，通常取平方损失函数或者绝对值损失函数形式。惩罚项$R(\boldsymbol{A})$为一正则化项，用于对矩阵$\boldsymbol{A}$的结构进行一定的规约，比如对高秩的解进行惩罚鼓励低秩的解，并解决一些参数的过拟合问题等。

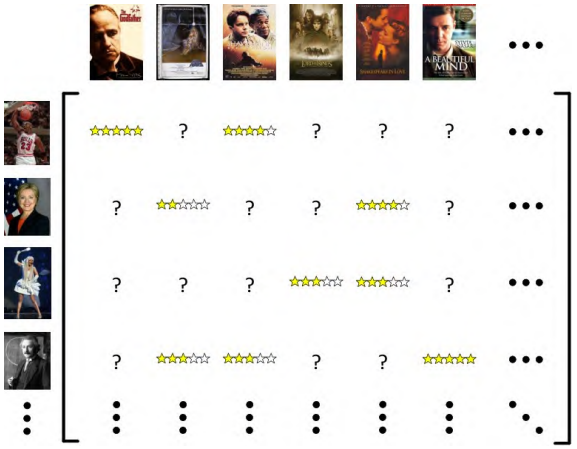

图1 在线观影系统用户评分矩阵示例
注：其中？代表评分缺失，星级代表评分高低

在上述优化问题中，损失函数是用来评价模型的预测值和真实值不同程度的函数。通常情况下损失函数越小，

模型的性能也越好。其中常见的一类损失函数是平方损失函数，经常应用于回归问题。最小化平方损失函数又称最小二乘法，其几何意义是高维空间中的一个向量在低维子空间的投影。与此同时，对于常见的正态分布参数估计问题，通过极大似然估计求解也可以等价于一个最小化平方损失函数的问题。

惩罚项是对损失函数的补充调节，为了使得填充后的高维矩阵具有低秩结构，一个自然的想法是直接使用矩阵的秩本身作为惩罚函数，即将矩阵的秩，也就是矩阵非0奇异值的个数纳入到上述优化函数的正则化项中。然而已有研究表明，这样的方式是NP-难的（Chistov and Grigoriev，1984），难以在多项式时间内得到有效的计算结果。Candès和Recht（2009）及Recht（2011）提出了用核范数$\|\boldsymbol{A}\|_*$作为惩罚函数来解决矩阵完备化问题。具体而言，核范数是矩阵奇异值的和。数学意义上，矩阵的秩本身是非凸的，而核范数则是矩阵的秩的凸近似，是凸的。因此使用$\|\boldsymbol{A}\|_*$会使得整个优化问题变得更容易计算，不再是NP-难，能够在多项式时间内进行求解。

在实际中，我们需要通过已有的数据来构造损失函数$L(\boldsymbol{A},\boldsymbol{Y})$。若所有数据的观测值直接可得，则对于平方损失函数而言，一个直接的选择是$L(\boldsymbol{A},\boldsymbol{Y})=\frac{1}{n_1 n_2}\|\boldsymbol{Y}-\boldsymbol{A}\|_F^2$，其中$\|\boldsymbol{M}\|_F=trace(\boldsymbol{M}\boldsymbol{M}^T)$表示矩阵的Frobenius范数。然而，由于$\boldsymbol{Y}$中仅有部分数据可被观测到，我们无法直接使用以上$L(\boldsymbol{A},\boldsymbol{Y})$形式，而需要结合问题的特点，构建损失函数的形式。

如上所述，矩阵数据完备化问题可表示为如问题（1）所示的一个优化问题。由于不同场景下造成大数据缺失的机制不尽相同，数据缺失呈现不同的形态，人们需结合问题特点进行分析并采用有针对性的数据完备化方式加以解决。以下我们将从大数据的3个典型特点（即超高维度、多源异质、时空关联）出发，讨论在这3种情境下数据完备化问题的特点及对应的挑战，并结合作者近年来的研究，阐述相关的领域情境、概念内涵、问题建模、求解路径以及管理决策意义。

三、超高维度缺失数据完备化问题

超高维度是大数据的一个突出特点。如在电子商务环境中，常常包括上亿级别的用户及商品，从而使得用户商品评分矩阵呈现超高维度的特点。而用户所接触及评论的商品数量非常有限，从而产生大量缺失的点评数据。为实现超高维度缺失数据的完备化工作，需对数据缺失机制进行分析以具体化（1）中损失函数$L(\boldsymbol{A},\boldsymbol{Y})$及惩罚项$R(\boldsymbol{A})$的形式设定。

对于$R(\boldsymbol{A})$而言，为了使得填充后的高维矩阵具有低秩结构，通常情况下我们使用核范数$\|\boldsymbol{A}\|_*$作为惩罚函数来解决矩阵完备化问题。对于$R(\boldsymbol{A})$而言，由于在高维缺失的情况下，有大量的数据无法被观测到。需要构建一个只由0/1元素组成的观测示性矩阵$\boldsymbol{T}=(t_{ij})$，其中如果$t_{ij}=1$，则y_{ij}可被观测到，反之则令$t_{ij}=0$。对于示性矩阵$\boldsymbol{T}$，假设其对应的观测概率矩阵为$\Theta=(\theta_{ij})$，其中θ_{ij}代表t_{ij}取1的概率，即t_{ij}服从以θ_{ij}为“成功”概率的伯努利分布，具体示例如图2所示。

根据不同的数据缺失机制，θ_{ij}的表示形式各不相同，从而使得$L(\boldsymbol{A},\boldsymbol{Y})$的设定形式不尽相同，以下将分别就完全随机缺失、随机缺失、非随机缺失机制下的数据矩阵完备化问题的特点、优化问题设定及求解方法进行介绍。

（一）完全随机缺失机制（Missing Completely At Random）

在完全随机缺失（Missing Completely At Random，简称MCAR）的情况下，一个元素是否被观测到的概率与y_{ij}以及数组中观测到的任何其他变量都无关，其中均匀缺失机制（Uniformly Missing Mechanism）是一种特殊情形。在均匀缺失机制下，$\boldsymbol{Y}$中每个元素具有相同的边缘缺失概率，即Θ矩阵中所有$\theta_{ij}\equiv\theta$。这曾经是高维

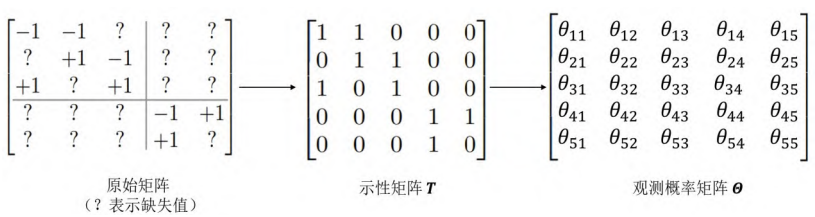

图2 原始矩阵、示性矩阵、观测概率矩阵示例

矩阵完备化中常采用的一种缺失机制假设，在数据矩阵完备化的最早文献中被普遍使用（Candès and Recht，2009；Keshavan et al.，2010；Recht 2011；Rohde and Tsybakov，2011；Koltchinskii et al.，2011）。在均匀缺失机制下即使具体的观测概率θ未知，可以使用$L(\boldsymbol{A},\boldsymbol{Y})=\frac{1}{n_1n_2}\left\|\boldsymbol{T}\circ(\boldsymbol{A}-\boldsymbol{Y})\right\|_F^2$作为损失函数，其中$\circ$的运算为矩阵之间Hadamard算子，用于表示矩阵对应位置元素相乘所得到的新矩阵。此时，对问题（1）中的损失函数和正则化项部分进行替换，可以得到用于刻画数据完备化的优化问题：

$$\hat{\boldsymbol{A}}=\arg\min_{\|A\|_\infty\le a}\left\{\frac{1}{n_1n_2}\left\|\boldsymbol{T}\circ(\boldsymbol{A}-\boldsymbol{Y})\right\|_F^2+\lambda\|\boldsymbol{A}\|_*\right\} \tag{2}$$

其中，λ是一个调节参数，用于平衡损失函数与促进低秩的正则化项之间的相对权重。

在均匀随机缺失下，Candès和Recht（2009）在观测值没有噪音的情况下给出如下经典的理论结果：对于一个$n_1\times n_2$的秩为r的矩阵A_0，当该矩阵满足特定的不连贯条件（Incoherence Condition）且数据均匀缺失的情形下，人们只需观测到$c(n_1+n_2)r\log^2(n_1+n_2)$个矩阵元素就可以接近1的概率对高维矩阵进行完备化。当观测值有噪音的时候，Candès和Plan（2010）及Koltchinskii等（2011）研究了在不同噪音情形下具有均匀缺失机制的高维矩阵完备化问题，对于填充数据矩阵误差的上界及最优收敛速度进行了分析。Mazumder等（2010）设计了针对问题（2）进行优化求解的softImpute算法并且提供了相应的R包[①]可供研究者直接使用。我们将softImpute算法应用到维度高达480000×18000的Netflix比赛数据上，该算法可仅用3.3个小时左右的时间拟合得到一个秩为95的矩阵，对应的均方误差能够仅为0.9497，可达到较好的完备化效果。然而，均匀缺失机制通常不能反映实际问题中的缺失机制，很多时候我们需要考虑其他的数据缺失机制情形。

（二）随机缺失机制（Missing At Random）

另一类常用的数据缺失机制是随机缺失机制（Missing At Random，简称MAR），即y_{ij}是否被观测到的概率只与一些可观测到的协变量有关，而与其具体取值y_{ij}无关，即观测概率矩阵$\boldsymbol{\Theta}$中的元素可表示为协变量x_{ij}的函数，即$\theta_{ij}=\theta(x_{ij})$。在MAR情形下，可采用$L(\boldsymbol{A},\boldsymbol{Y})=\frac{1}{n_1n_2}\left\|\boldsymbol{T}\circ\boldsymbol{\Theta}^{\circ(-1/2)}\circ(\boldsymbol{A}-\boldsymbol{Y})\right\|_F^2$作为损失函数，其中$\boldsymbol{\Theta}^{\circ(-1/2)}=(\theta_{ij}^{-1/2})$。由MAR性质可得此时损失函数$L(\boldsymbol{A},\boldsymbol{Y})$是$E\|\boldsymbol{Y}-\boldsymbol{A}\|_F^2$的无偏估计。

在实际构建矩阵完备化优化问题时，数据矩阵重构的具体形式又与观测概率矩阵$\boldsymbol{\Theta}$是否已知有关。在绝大多数情况下，观测概率矩阵$\boldsymbol{\Theta}$的先验知识并不可得。换言之，我们需先构建$\boldsymbol{\Theta}$的估计$\hat{\boldsymbol{\Theta}}$，再代入上述损失函数$L(\boldsymbol{A},\boldsymbol{Y})$中。此时数据完备化优化问题可表示为：

$$\hat{\boldsymbol{A}}=\arg\min_{\|A\|_\infty\le a}\left\{\frac{1}{n_1n_2}\left\|\boldsymbol{T}\circ\hat{\boldsymbol{\Theta}}^{\circ(-1/2)}\circ(\boldsymbol{A}-\boldsymbol{Y})\right\|_F^2+\lambda\|\boldsymbol{A}\|_*\right\} \tag{3}$$

由此可见，对$\hat{\boldsymbol{\Theta}}$建模的质量直接决定了最终可得的矩阵$\hat{\boldsymbol{A}}$的性质。下面我们将总结几类常见的$\hat{\boldsymbol{\Theta}}$建模方法，包括结合协变量信息的Logistic模型、低秩模型、不依赖具体模型设定的非参数模型。

1.利用协变量信息的缺失机制建模

在协变量信息$\boldsymbol{X}$已知的情况下，可将数据观测概率θ_{ij}表示为协变量的函数，即$\theta_{ij}=\Pr(t_{ij}=1|x_{ij})=\theta(x_{ij})$。以电影推荐场景为例，若用户及电影的特征已知，如已知用户性别、年龄、职业等，同时知晓电影类型、导演等信息，则用户评分是否可以被观测到可表示为这些协变量的函数。具体而言，可采用Logistic模型对观测概率矩阵$\boldsymbol{\Theta}$进行建模（Mao et al.，2019）：

$$\theta_{ij}=\Pr\left(t_{ij}=1|\boldsymbol{X}\right)=\frac{e^{x_i^T\gamma_j}}{1+e^{x_i^T\gamma_j}} \tag{4}$$

其中，$\boldsymbol{\gamma}=(\gamma_j)$表示协变量$\boldsymbol{X}$的系数向量。对于这里的参数$\gamma_j$，我们可以通过极大似然估计来做参数估计。

在得到$\boldsymbol{\Theta}$的估计$\hat{\boldsymbol{\Theta}}$后，我们可进一步对评分矩阵$\boldsymbol{A}_0$建立列空间分解的半参数模型$\boldsymbol{A}_0=\boldsymbol{X}\boldsymbol{\beta}_0+\boldsymbol{B}_0$来改变问题（3）的形式，其中$\boldsymbol{B}_0$是一个低秩矩阵。为了满足模型的可识别性，Mao等（2019）假设协变量$\boldsymbol{X}$的列空间与低秩矩阵$\boldsymbol{B}_0$的列空间正交。通过使用额外的协变量$\boldsymbol{X}$和这个正交性质，Mao等（2019）把通常使用的迭代算法变成了只需要求解一具有解析解的奇异值分解（Singular Value Decomposition，SVD）算法，从而大大地降低了计算

复杂度。与此同时，该研究也给出了完备化矩阵的均方误差的上界，并刻画了使用额外协变量$\boldsymbol{X}$所带来的理论优势。具体的奇异值分解算法可以参考Cai等（2010）。Mao等（2019）将该方法应用于实际数据MovieLens100K[②]进行完备化。该数据包含由943个影评人对1682部电影给出100000个评分，及额外的影评人和电影协变量信息。通过使用额外的协变量信息，模型完备化效果可得到一定的提升。

2.低秩模型

在缺少协变量信息对θ_{ij}建模的情况下，也可考虑以低秩缺失机制实现对$\boldsymbol{\Theta}$的稳健估计（Mao et al.，2021）。即假设缺失机制矩阵$\boldsymbol{\Theta}$具有低秩性质，$\boldsymbol{\Theta}$可由一个高维低秩的隐矩阵$\boldsymbol{M}=(m_{ij})$经过联接函数族$\mathcal{F}=\{f\}$映射得到，即$\boldsymbol{\Theta}=f(\boldsymbol{M})$。这时候对于观测到的矩阵$\boldsymbol{Y}$可以分解出两个低秩矩阵，具体参见图3所示，其中$\boldsymbol{A}_0$代表完整的真实评分矩阵，具有低秩性，$\boldsymbol{T}$为0-1示性矩阵，联接函数$f$背后的隐矩阵$\boldsymbol{M}$也具有低秩性。

对于缺失机制$\boldsymbol{\Theta}$的低秩估计$\hat{\boldsymbol{\Theta}}$，可通过对隐矩阵$\boldsymbol{M}$做均值分解的方法来克服可能存在的概率的高估问题（Mao et al.，2021）。具体而言，首先对$\boldsymbol{M}$做均值分解$\boldsymbol{M}=\mu\boldsymbol{J}+\boldsymbol{Z}$，其中$\mu$是$\boldsymbol{M}$的所有元素的均值，$\boldsymbol{J}$是元素全为1的矩阵，而$\boldsymbol{Z}$是剩下的元素和为0的矩阵。进一步地，在特定的约束条件下最大化如下带核范数惩罚的似然函数问题：

$$f(\mu,\boldsymbol{Z}|\lambda)=\sum_{i,j}\left\{t_{ij}\log\left(f\left(\mu+z_{ij}\right)\right)+(1-t_{ij})\log\left(1-f\left(\mu+z_{ij}\right)\right)\right\}-\lambda\|\boldsymbol{Z}\|_*$$

从而同时得到μ和$\boldsymbol{Z}$的估计量$\hat{\mu}$和$\hat{\boldsymbol{Z}}$。这里我们可以采用Chen等（2016）提出的交替方向乘子法（Alternating Direction Method of Multipliers，简称ADMM）来完成。在同时获得$\hat{\mu}$和$\hat{\boldsymbol{Z}}$之后，就可以分别得到$\boldsymbol{M}$和$\boldsymbol{\Theta}$的估计，$\hat{\boldsymbol{M}}=\hat{\mu}\boldsymbol{J}+\hat{\boldsymbol{Z}}$和$\hat{\boldsymbol{\Theta}}=\mathcal{F}(\hat{\boldsymbol{M}})$。通过进一步结合一些截短方法，我们可以使最终得到的概率矩阵估计$\hat{\boldsymbol{\Theta}}$更加光滑，避免出现一些极小值。将$\hat{\boldsymbol{\Theta}}$带入式（3）中可以进一步得到最终的评分矩阵$\hat{\boldsymbol{A}}$的估计。理论研究表明，在真实缺失机制为均匀缺失的情况下，即便我们通过低秩模型来做了缺失概率矩阵估计$\hat{\boldsymbol{\Theta}}$，最终我们的目标矩阵估计$\hat{\boldsymbol{A}}$依然可以以概率1得到最优收敛速度；另一方面，在非均匀缺失的低秩模型下，只要最小缺失概率$\theta_L=\min\{\theta_{ij}\}$满足一定条件，我们依旧可以以概率1得到评分矩阵估计$\hat{\boldsymbol{A}}$的误差上界的最优估计。对于最终评分矩阵$\hat{\boldsymbol{A}}$的估计的目标函数（3），可以采用由Beck和Teboulle（2009）提出的快速迭代收缩阈值算法（Fast Iterative Shrinkage-Thresholding Algorithm）。Mao等（2021）将该方法应用到实际数据Yahoo Webscope[③]上。该数据包含了由15400个乐评人对1000首歌曲给出的300000个评分。通过引入低秩缺失机制，该方法相较于采用均匀缺失机制的完备化方法效果提升了约25%。

3.非参数模型

尽管上述对$\boldsymbol{\Theta}$的估计方式可在对应缺失假设下取得一定的效果，但其对数据矩阵完备化的效果严重依赖于缺失模型假设是否正确，其在实际应用中难以被验证。而对于最终完成高维矩阵完备化的这个目标来说，并不需要一定给出正确的缺失概率$\boldsymbol{\Theta}$。这是因为最终我们是通过解决优化问题（1）来得到$\boldsymbol{A}_0$的估计，而缺失概率$\boldsymbol{\Theta}$的估计只是中间的一步副产品。如果我们找到一个合适的权重矩阵$\boldsymbol{W}$来替代$\boldsymbol{\Theta}^{\circ(-1)}$，这里$\boldsymbol{\Theta}^{(-1)}=(\theta_{ij}^{-1})$，比如说使得$\hat{\boldsymbol{W}}$和$\boldsymbol{\Theta}^{\circ(-1)}$在总体上的误差足够接近，使得对于最终估计$\boldsymbol{A}_0$带来的概率矩阵部分的误差可以忽略不计，那么我们还是可以得到好的$\boldsymbol{A}_0$的估计。理想情况下，若示性矩阵$\boldsymbol{T}$的生成概率$\boldsymbol{\Theta}=(\theta_{ij})$已知，则只需要直接选取权重矩阵为$\boldsymbol{W}=(\theta_{ij}^{-1})$即可。在生成概率$\boldsymbol{\Theta}$未知的情况下，通过观察我们有$E(\boldsymbol{T}\circ\boldsymbol{\Theta}^{\circ(-1)})=\boldsymbol{J}$，其中$\boldsymbol{J}$是一个所有元素全部为1的矩阵。Wang等（2021）考虑找合适的权重矩阵$\boldsymbol{W}$使得度量$\|\boldsymbol{T}\circ\boldsymbol{W}-\boldsymbol{J}\|$足够小。进一步的，为了克服权重矩阵$\boldsymbol{W}$总共有$n_1n_2$个参数带来的过拟合问题，Wang等（2021）

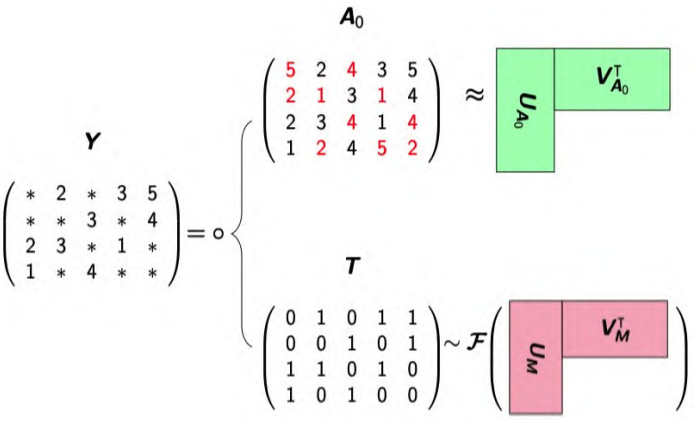

图3 低秩缺失机制示例

通过求解(5)式带有约束的优化问题,来求解$\boldsymbol{W}$矩阵:

$$\min_{\boldsymbol{W}\geq 1}\|\boldsymbol{T}\circ\boldsymbol{W}-\boldsymbol{J}\|+\kappa\|\boldsymbol{T}\circ\boldsymbol{W}\|_F^2 \tag{5}$$

其中,κ是一个调节参数。由此可得权重估计矩阵$\hat{\boldsymbol{W}}$。这里得到的权重矩阵$\hat{\boldsymbol{W}}$不仅不依赖于缺失机制,甚至不依赖于观测矩阵$\boldsymbol{Y}$。所以该方法比较稳健。在得到权重矩阵$\hat{\boldsymbol{W}}$之后,类似问题(3),可通过如下的风险函数对高维矩阵$\boldsymbol{A}_0$进行填充:

$$\hat{\boldsymbol{A}}=\arg\min_{\|\boldsymbol{A}\|_\infty\leq a}\left\{\frac{1}{n_1n_2}\left\|\boldsymbol{T}\circ\hat{\boldsymbol{W}}^{\wedge(1/2)}\circ(\boldsymbol{A}-\boldsymbol{Y})\right\|_F^2+\lambda\|\boldsymbol{A}\|_*\right\} \tag{6}$$

对于上述问题,也可采用Beck和Teboulle(2009)提出的快速迭代收缩阈值算法进行求解。Wang等(2021)将该方法应用到实际数据Coat Shopping Dataset[④]和Yahoo Webscope[⑤]上。Coat Shopping Dataset包含了由290个用户对300种商品给出的约7000个评分信息。通过引入不依赖于缺失机制的非参数模型,该方法与采用均匀缺失机制和一些特殊的秩一(rank-one)缺失机制的完备化方法相比效果都有所提升。

(三)非随机缺失机制(Missing Not At Random)

另一种常见的数据缺失机制为非随机缺失(Missing Not At Random,简称MNAR),即数据缺失与否取决于其具体取值y_{ij},这有违于之前所描述的MAR机制。如Yahoo!进行的一项调查显示,在5400名参与者中,有64.85%认为他们对歌曲的喜好程度会影响他们公开评分的意愿[⑥],即用户评分矩阵的缺失情况并非是随机于y_{ij}的值,而是依赖于y_{ij}。在此情境下为实现对$\boldsymbol{Y}$的无偏估计,可采用逆倾向性得分(Inverse Propensity Score,简称IPS方法)即使用$P_{ij}=1/E(\theta_{ij}|y_{ij})$对每一维观测值进行逆概率加权(Schnabel et al.,2016),进而数据完备化问题可表示为式(7)所示的形式:

$$\hat{\boldsymbol{A}}=\arg\min_{\|\boldsymbol{A}\|_\infty\leq a}\left\{\frac{1}{n_1n_2}\left\|\frac{\boldsymbol{T}\circ(\boldsymbol{A}-\boldsymbol{Y})}{\boldsymbol{P}}\right\|_F^2+\lambda\|\boldsymbol{A}\|_*\right\} \tag{7}$$

由此,非随机缺失机制下的数据完备化方法可以分为以下两个步骤,其一是估计逆倾向性得分矩阵$\boldsymbol{P}$,其二是根据估计出的在逆倾向性得分进行数据完备化。在逆倾向性得分估计准确的情形下,对$\boldsymbol{Y}$的还原可视为是无偏的。然而逆倾向性得分是否无偏本身在实际应用中并无法进行验证。而且尽管IPS统计量具有无偏性,其在实际应用中常表现出较大的方差变异。由此,相关研究进一步设计了双稳健统计量用于对缺失数据矩阵进行加权(Wang et al.,2019)。

上述不同缺失机制下超高维度大数据完备化方法可应用于电子商务、内容服务等诸多领域。如在电子商务情境下,推荐系统预测用户偏好以实现个性化推荐的重要实现方式为预测用户对商品的评分,即可视为对用户评分矩阵的完备化问题。由于用户及产品都呈现超高维度的特点,在进行矩阵完备化过程中需根据不同的缺失机制设计相应的优化问题,以实现对用户偏好的还原进一步展开个性化推荐。

四、多源异质场景下数据完备化问题

多源异质是大数据的另一突出特点。体量庞大的大数据通常由多种来源的数据汇集而成,不同源的数据的概率分布或模型通常是不同的,因而汇集而成的大数据呈现了异质性的特点。在这种情况下的缺失数据完备化问题需充分考虑数据的多源异质特点。如在智慧城市监测过程中,由于传感器记录时间粒度不够精细、仪器故障等问题,常常会出现数据缺失问题。而且由于数据是由多地部署的传感器采集汇集而成,数据具有很强的多源异质特点,在处理其数据缺失时应格外关注。具体而言,数据的多源异质性既包含数据分布相同但参数不同的情形,也包括数据分布不同的情形。以下我们将分别对两种多源异质情形进行讨论。

(一)数据分布相同但参数不同的情形

这是一种较为温和的多源异质情形,即不同来源的数据具有相同分布族,但分布参数不同。在现实中一种常见数据场景是二元数值问题,以视频推荐系统和新闻推荐系统为例,通常观众对于特定视频或者新闻可

以表达“点赞”或者“踩”的态度，这类数据可以抽象成二元取值数据{1，-1}。其所对应的推荐系统也就是二元的推荐系统。其他的二元数值数据场景还包括政治选举数据和市场调查数据等。Davenport等（2014）在问题（1）的框架下研究了观测值y_{ij}是二元数值{1，-1}的情形下的矩阵完备化问题。其考虑的二元数值的模型为：

$$y_{ij}=\begin{cases}1 & \text{以概率} \quad f(a_{0,ij})\\ -1 & \text{以概率} \quad 1-f(a_{0,ij})\end{cases} \tag{8}$$

其中，$a_{0,ij}$表示观测概率矩阵对应的参数矩阵$\boldsymbol{A}_0$中的元素，每一维$a_{0,ij}$取值可以不同，从而反映出数据异质性特点。这时我们所关心的真实矩阵等价于参数矩阵$\boldsymbol{A}_0$。注意到真实的参数矩阵$\boldsymbol{A}_0$与最终的观测值$\boldsymbol{Y}$通过一个联接函数f来联系。如第三节所示，常见的联接函数f可以取成Logit或者Probit函数。进一步地，我们考虑用对应的负对数似然函数来作为损失函数$L(\boldsymbol{A},\boldsymbol{Y})$，即：

$$L(\boldsymbol{A},\boldsymbol{Y})=\sum_{(i,j)}t_{ij}\left\{\mathrm{I}_{[y_{ij}=1]}\log\left(f(a_{ij})\right)+\mathrm{I}_{[y_{ij}=-1]}\log\left(1-f(a_{ij})\right)\right\} \tag{9}$$

其中，$\boldsymbol{T}=(t_{ij})$是对应的示性矩阵，对应的惩罚项同样使用核范数$R(\boldsymbol{A})=\|\boldsymbol{A}\|_*$以使得结果具有低秩性。Davenport等（2014）将该方法应用到实际数据MovieLens100K⑦上。为了使得观测到的评分变成二元数值，Davenport等（2014）根据已有评分的均分3.5作为划分，大于等于3.5的评分映射成+1，小于3.5的评分映射成-1，从而形成二元数值{1，-1}。通过采用上述的最小化负对数似然函数损失函数和核范数惩罚，相较于经典的均匀缺失机制下的矩阵完备化方法，该方法将准确率从60%提升到了73%。

更一般地，Fan等（2019）在问题（1）的框架下提出了基于广义高维迹回归模型。对应地，他们考虑使用指数分布族对应的负对数似然函数来作为损失函数$L(\boldsymbol{A},\boldsymbol{Y})$。具体而言，在问题（1）的框架下，基于指数族分布特征构建损失函数，即：

$$L(\boldsymbol{A},\boldsymbol{Y})=\frac{1}{N}\sum_{(i,j)}t_{ij}\left\{b(a_{ij})-y_{ij}a_{ij}\right\} \tag{10}$$

其中，$N=\sum t_{ij}$是观测到的元素个数，$b(\cdot)$是一个已知的跟具体分布函数有关的联接函数。比如对于常见的高斯分布，由其对应的指数族分布的表达式，$b(a_{ij})=\sigma^2a_{ij}^2/2$，其中$\sigma^2$是已知的方差常数；对于取值为0或1的Bernoulli分布，我们有$b(a_{ij})=\log(1+\exp(a_{ij}))$；对于Poisson分布，有$b(a_{ij})=\exp(a_{ij})$。在Fan等（2019）的工作里，他们使用同样的核范数惩罚$R(\boldsymbol{A})=\|\boldsymbol{A}\|_*$来使得最终的参数矩阵$\hat{\boldsymbol{A}}$具有近似低秩的性质。Fan等（2019）将该方法应用到S&P500的股票收益率预测和图像分类的经典数据集CIFAR10⑧上。在S&P500的股票收益率预测问题上，该方法采用核范数作为惩罚项，普遍比不带惩罚项的方法得到的效果好。在图像分类问题上，该方法采用了卷积神经网络（Convolutional neural network，CNN）加上核范数惩罚的方法，比对应的卷积神经网络加上L1范数惩罚项效果更好。

（二）数据分布不同的情形

这是一种更一般的刻画数据多源异质性的情形，即各来源数据的概率分布与模型各不相同。比如我们在多任务学习的框架下想要同时解决分类问题和回归预测问题，其中分类问题的数据可以来自条件Bernoulli分布，回归预测问题则可以来自Gaussian分布。比如连续值数据可以是Gaussian分布，0/1取值数据可以用Bernoulli分布或Logistic模型，多值离散数据可以用Multi-probit分布模型等条件分布。

Alaya和Klopp（2019）考虑了基于指数分布族的损失函数$L(\boldsymbol{A},\boldsymbol{Y})$构建。他们假设观测到的矩阵$\boldsymbol{Y}$的数据元素来自$\boldsymbol{S}$个不同的概率分布，即数据$\boldsymbol{Y}$和其对应的真实参数矩阵$\boldsymbol{A}_0$可以分成$\boldsymbol{S}$块，分别记为$\boldsymbol{Y}=[\boldsymbol{Y}^{(1)},\cdots\cdots,\boldsymbol{Y}^{(S)}]$和$\boldsymbol{A}_0=[\boldsymbol{A}_0^{(1)},\cdots\cdots,\boldsymbol{A}_0^{(S)}]$，其中$\boldsymbol{A}_0^{(S)}=(a_{0,ij}^{(s)})$，$s=1,\cdots\cdots,S$。具体来说，假设每个数据$y_{ij}^{(s)}$属于参数可取不同值的指数分布。在该模型的假设下，实际场景里的数据可以来自于不同的来源和任务。针对每个分布，即便分布形式一样，其中的具体参数也可以完全不同，比如同样都是高斯分布，不同的任务可以有不同的均值μ和方差σ^2。基于这一前提假设，Alaya和Klopp（2019）考虑以加权平均方式的损失函数来同时完成$\boldsymbol{S}$个不同任务，基于指数分布族的特征构建矩阵完备化问题的损失函数：

$$L(\boldsymbol{A},\boldsymbol{Y})=\frac{1}{N}\sum_{s=1}^{S}\sum_{(i,j)}t_{ij}^{(s)}\left\{b_s\left(a_{ij}^{(s)}\right)-y_{ij}^{(s)}a_{ij}^{(s)}\right\} \tag{11}$$

其中每个s代表不同的任务和数据来源，$\boldsymbol{T}^{(s)}=(t_{ij}^{(s)})$是每个不同来源数据分别对应的示性矩阵，$N=\sum t_{ij}^{(s)}$是总的观测值。此时，数据异质性特通过不同的联接函数$b_s(\cdot)$体现，即代表不同的数据的分布及任务。对于不同源数据之间共享的特征，我们则是通过公共的惩罚项$R(\boldsymbol{A})=\|\boldsymbol{A}\|_*$来约束进行同步学习，使得多源异质数据$\boldsymbol{A}_0$能够共享低秩的结构信息。在这个框架下，Alaya和Klopp（2019）建立了预测误差的上界。Alaya和Klopp（2019）将该方法应用到模拟数据集上，该方法比分别单独估计每个来源的矩阵完备化的准确率更高。

Robin等（2020）同样也考虑了上述的问题框架，更具体地，他们对于具体参数矩阵$\boldsymbol{A}_0$进行了更加细致地建模，类似于Mao等（2019）的思路，将$\boldsymbol{A}_0$分解成主效应和相互效应两个部分$\boldsymbol{A}_0=\alpha\boldsymbol{U}+\boldsymbol{L}$。对应地，用来约束多源异质数据$\boldsymbol{A}_0$的惩罚项$R(\boldsymbol{A})$则变为（12）式，进一步可通过求解整体优化问题来寻求最优完备化方式。

$$R(\boldsymbol{A})=|\alpha|_0+\|\boldsymbol{L}\|_* \tag{12}$$

多源异质情境下的数据完备化方法可以广泛应用于多种领域。如在面向制造企业车间执行层的生产信息化管理系统（MES）整合了包括RFID、条码设备、传感器等多种渠道采集的数据。由于不同采集设备的数据分布形态各不相同，且可能以不同的频率产生故障，从而造成采集到的数据中的缺失情况呈现多源异质的特点。应用上述数据完备化方式可对其中蕴含的多源异质特点进行充分建模，从而实现更优的数据完备化效果以供后续分析决策使用。

五、时空关联场景下的缺失数据完备化问题

流式产生是大数据的另一突出特点，即大数据以一定的时间颗粒度产生及被记录下来。若在此情境下发生数据缺失问题将具有强时空关联性的特点。如在金融大数据领域，常见的数据来源包括股价、交易记录、高频交易信息、分析师预测、新闻、社交媒体用户情绪数据等。而机构/散户对于某一公司/股票的关注情况常常并不连续，造成大量信息缺失。但这些缺失信息之间呈现出强时序性的特点。在设计相关数据完备化方法时，应对其特点充分加以考虑。

在此类数据完备化问题中，为实现对时空维度的刻画，通常在二维矩阵表示的数据形态中引入新的用于表征时间或空间的维度，从而形成张量（Tensor）数据。张量指的是多维（或者K维）阵列数据。特别地，一维张量（K=1）对应的是向量数据；二维张量（K=2）对应的是矩阵数据。通常人们将$K\geqslant3$的张量称为高阶张量。如在考虑时间动态性的推荐系统里，除了已有的用户—商品的二维评分矩阵，还会考虑额外的时间标签信息。又如，对于一些交通网络数据，也能获得额外的时间或者空间信息形成张量形式。相应地，如果观测值带有缺失的情况下，我们需要考虑张量完备化来完成对应的高维数据完备化。为了符号简洁和讨论方便，本文只对三阶张量形态的数据完备化进行介绍，更高阶的张量模型可以做类似推广。如果不对张量的维度做特殊的结构假设，我们可以将矩阵完备化方法直接推广到张量完备化里来。

常用的张量分解方法为Kiers（2000）给出CANDECOMP/PARAFAC分解，简称CP分解；其中CANDECOM是canonical decomposition的缩写，该方法在Carroll和Chang（1970）中提出；PARAFAC是parallel factors的缩写，在Harshman（1970）中提出。对于秩为r的张量$\boldsymbol{A}_0$，根据张量秩的定义，我们可以将它表示成r个秩为1的张量之和，即：

$$\boldsymbol{A}_0=[\![\boldsymbol{U}_0,\boldsymbol{V}_0,\boldsymbol{W}_0]\!]=\sum_{i=1}^{r}\boldsymbol{u}_{0i}\circ\boldsymbol{v}_{0i}\circ\boldsymbol{w}_{0i} \tag{13}$$

其中，$\boldsymbol{U}_0=[\boldsymbol{u}_{01},\boldsymbol{u}_{02},\cdots,\boldsymbol{u}_{0r}]\in\mathbb{R}^{n_1\times r}$，$\boldsymbol{V}_0=[\boldsymbol{v}_{01},\boldsymbol{v}_{02},\cdots,\boldsymbol{v}_{0r}]\in\mathbb{R}^{n_2\times r}$，$\boldsymbol{W}_0=[\boldsymbol{w}_{01},\boldsymbol{w}_{02},\cdots,\boldsymbol{w}_{0r}]\in\mathbb{R}^{n_3\times r}$，$[\![\cdots]\!]$为CP分解的表示符号，具体如图4所示。

对于观测到的带有缺失值的流数据，我们能对应地产生一个由{0，1}元素组成的指示符张量$\boldsymbol{T}=(t_{ijk})$，其中t_{ijk}是y_{ijk}的缺失指示，即t_{ijk}=0表示缺失；t_{ijk}=1为非缺失。进而可以直接将矩阵数据完备化问题的形式推广到张

量形式。在一般的情况下，损失函数和正则化项部分可分别表示为：

$$L(\boldsymbol{A},\boldsymbol{Y})=\frac{1}{n_1n_2n_3}\left\|\boldsymbol{T}\circ\left(\llbracket\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}\rrbracket-\boldsymbol{Y}\right)\right\|_F^2 \tag{14}$$

$$R(\boldsymbol{A})=\lambda\left(\|\boldsymbol{U}\|_F^2+\|\boldsymbol{V}\|_F^2+\|\boldsymbol{W}\|_F^2\right) \tag{15}$$

在张量完备化问题中，为了减少计算的复杂度，我们通常假设要完备化的张量的秩r是已知的。对于带有时空属性维度的张量，Zhou等（2015）考虑在惩罚项$R(\boldsymbol{A})$上继续加上一些带有时空属性的特殊结构约束。特别地，他们考虑如下惩罚项：

$$\lambda\left(\|\boldsymbol{U}\|_F^2+\|\boldsymbol{V}\|_F^2+\|\boldsymbol{W}\|_F^2\right)+\alpha\left(\llbracket\boldsymbol{FU},\boldsymbol{V},\boldsymbol{W}\rrbracket+\llbracket\boldsymbol{U},\boldsymbol{GV},\boldsymbol{W}\rrbracket+\llbracket\boldsymbol{U},\boldsymbol{V},\boldsymbol{HW}\rrbracket\right) \tag{16}$$

其中，$\boldsymbol{F}$和$\boldsymbol{G}$是空间约束矩阵，$\boldsymbol{H}$是时间约束矩阵，λ和α是两个不同的调节参数。不同于矩阵完备化问题，在张量完备化中，需对具有时空属性的维度做特殊的结构约束（如AR模型、Toeplitz矩阵等），使得该完备化不是简单的矩阵完备化的拓展，而是得到一个具有时空性质的张量。

考虑时空关联性的数据完备化方法在管理实践中具有广阔的应用前景。如在对大气环境进行长期监测以应用于宏观政策分析时，监测数据中的缺失情形呈现时空关联的特点，需在完备化过程中加以考虑。通过加入上述对特殊时空属性的结构约束，可保证数据完备化结果体现了时空关联情形，更好地保障完备化效果，以供后续环境政策分析决策使用。

六、讨论与总结

随着移动互联环境下新兴技术的快速发展，多维度、跨领域的大规模数据日益可测可获，不仅深刻地改变了社会经济生活的面貌，也孕育着管理决策理论与方法的重大变革，推动管理决策研究向大数据驱动范式转变。然而，超高比例的数据缺失现象常常制约着数据价值挖掘及后续管理决策的进行。为提升数据质量及完备性，需结合问题情境特点设计精准高效的数据完备化方法。

在实际应用中根据问题特点选择合适的方法进行数据完备化对后续分析及管理决策制定至关重要。在进行方法选择时，可从以下两方面考虑。首先，我们可以从实际数据特点出发。如果实际数据是维数大于等于3维的张量数据，我们优先考虑流式数据完备化方法，进一步地，如果一些数据维度有特定的信息，比如包含时间或者空间等信息，则可以考虑应用具有时空性质的流数据完备化方法。如果是一般的矩阵数据，则需要首先对数据分布进行判断。如对二元数据，可以采用二元数值矩阵完备化方法。对于连续型数据，可以采用平方损失函数的矩阵完备化方法。如果数据来自不同的分布，则可以应用指数分布族等混合型分布的矩阵完备化方法。其次，选择不同完备化方法的另一个主要影响因素是数据缺失机制。在实际应用中相对比较难以验证实际的缺失机制是否符合模型假设，因而我们建议可分别采取比较经典的缺失机制，比如完全随机缺失机制中的均匀缺失，随机缺失机制中的低秩缺失机制来得到初步结果。如果初步结果相差不大，则可以采用这些得到的结果，如果结果差别很大，说明缺失机制较为复杂。建议可以采用非参数模型的缺失机制，通过构建平衡权重的方法来完成矩阵完备化。

关于完备化后的数据矩阵的统计学性质及在管理实践中的应用也是统计学领域近期的关注方向。其一，完备化好的矩阵可直接用于管理决策。如在电子商务、内容推荐等领域广泛应用的推荐系统，在对用户—商品评分矩阵进行补全后，可直接采用对完备化后的评分值进行排序的方式展开Top-N推荐（Kang et al.，2016）。其二，可对完备化好的矩阵进行后续统计推断、机器学习等任务。如

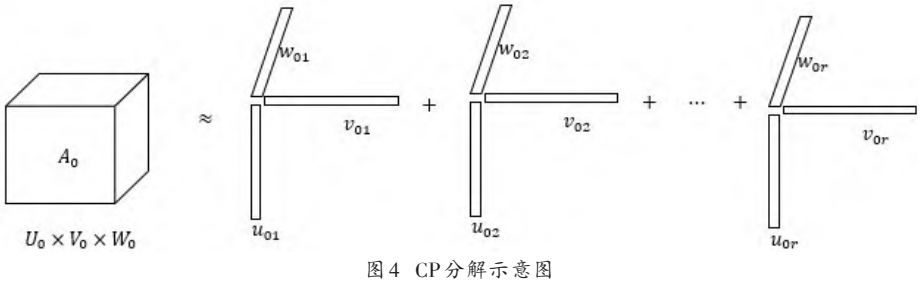

图4 CP分解示意图

Chen等(2019)分别对采用凸和非凸方法进行完备化后的矩阵构造了对应的纠偏统计量,使得纠偏后的矩阵能够对缺失数据和低秩因子等构建置信区间和置信区域。Xia和Yuan(2021)通过数据分裂构建具有渐近正态性质的矩阵估计,从而对线性形式的参数提供置信区间的估计和假设检验。通过这些方法,在对矩阵完成完备化后,我们可以进一步地针对完备化好的矩阵应用一些传统统计方法进行推断。

综上,本文在系列工作的基础上对大数据情境下数据完备化问题进行了梳理。针对大数据时代数据所呈现的超高维度、多源异质、时空关联的三类典型情境,分别总结了其情境特点、数据完备化挑战、求解思路及管理意义。后续研究可进一步探索融合多种情境特点的大数据完备化问题的建模形式、求解路径,并进一步思考相关方法在管理实践中的具体应用及价值测算。

(作者单位:陈松蹊,北京大学光华管理学院、北京大学统计科学中心;毛晓军,上海交通大学数学科学学院;王聪,北京大学光华管理学院)

注释

①https://cran.r-project.org/web/packages/softImpute/index.html.

②⑦https://grouplens.org/datasets/movielens/100k/.

③⑤⑥http://research.yahoo.com/Academic Relations.

④http://www.cs.cornell.edu/~schnabts/mnar/.

⑧https://www.cs.toronto.edu/~kriz/cifar.html.

参考文献

(1)陈国青、曾大军、卫强、张明月、郭迅华:《大数据环境下的决策范式转变与使能创新》,《管理世界》,2020年第2期。

(2)徐宗本、冯芷艳、郭迅华、曾大军、陈国青:《大数据驱动的管理与决策前沿课题》,《管理世界》,2014年第11期。

(3)Allison, P. D., 2000, "Multiple Imputation for Missing Data: A Cautionary Tale", *Sociological Methods & Research*, 28(3), pp.301~309.

(4)Alaya, M. Z. and Klopp, O., 2019, "Collective Matrix Completion", *Journal of Machine Learning Research*, 20, pp.148:1~148:43.

(5)Beck, A. and Teboulle, M., 2009, "A Fast Iterative Shrinkage-thresholding Algorithm for Linear Inverse Problems", *SIAM Journal on Imaging Sciences*, 2(1), pp.183~202.

(6)Candès, E. J. and Recht, B., 2009, "Exact Matrix Completion Via Convex Optimization", *Foundations of Computational Mathematics*, 9(6), pp.717~772.

(7)Candes, E J., Plan, Y., 2010, "Matrix Completion with Noise", *Proceedings of the IEEE*, 98(6), pp.925~936.

(8)Cai, J. F., Candès, E. J. and Shen, Z., 2010, "A Singular Value Thresholding Algorithm for Matrix Completion", *SIAM Journal on Optimization*, 20(4), pp.1956~1982.

(9)Carroll, J. D. and Chang, J. J., 1970, "Analysis of Individual Differences in Multidimensional Scaling Via an N-way Generalization of 'Eckart-Young' Decomposition", *Psychometrika*, 35(3), pp.283~319.

(10)Chen, C., He, B., Ye, Y. and Yuan, X., 2016, "The Direct Extension of ADMM for Multi-block Convex Minimization Problems is Not Necessarily Convergent", *Mathematical Programming*, 155(1~2), pp.57~79.

(11)Chen, Y., Fan, J., Ma, C. and Yan, Y., 2019, "Inference and Uncertainty Quantification for Noisy Matrix Completion", *Proceedings of the National Academy of Sciences*, 116(46), pp.22931~22937.

(12)Chistov, A. L. and Grigoriev, D. Y., 1984, "Complexity of Quantifier Elimination in the Theory of Algebraically Closed Fields", *International Symposium on Mathematical Foundations of Computer Science*, Springer, Berlin, Heidelberg, pp.17~31.

(13)Davenport, M. A., Plan, Y., Van Den Berg, E. and Wootters, M., 2014, "1-bit Matrix Completion", *Information and Inference: A Journal of the IMA*, 3(3), pp.189~223.

(14)Durrant, G. B., 2009, "Imputation Methods for Handling Item-nonresponse in Practice: Methodological Issues and Recent Debates", *International Journal of Social Research Methodology*, 12(4), pp.293~304.

(15)Fan, J., Gong, W. and Zhu, Z., 2019, "Generalized High-dimensional Trace Regression Via Nuclear Norm Regularization", *Journal of Econometrics*, 212(1), pp.177~202.

(16)Feuerverger, A., He, Y. and Khatri, S., 2012, "Statistical Significance of the Netflix Challenge", *Statistical Science*, 27(2), pp.202~231.

(17)Harshman, R. A., 1970, "Foundations of the PARAFAC Procedure: Models and Conditions for an 'Explanatory' Multimodal Factor Analysis", *UCLA Working Papers in Phonetics*, 16, pp.1~84.

(18)Ibrahim, J. G., Chen, M. H., Lipsitz, S. R. and Herring, A. H., 2005, "Missing-data Methods for Generalized Linear Models: A Comparative Review", *Journal of the American Statistical Association*, 100(469), pp.332~346.

(19)Kang, Z., Peng, C. and Cheng, Q., 2016, "Top-n Recommender System Via Matrix Completion", *Thirtieth AAAI Conference on Artificial Intelligence*, pp.179~185.

(20)Keshavan, R. H., Montanari, A. and Oh, S., 2010, "Matrix Completion from Noisy Entries", *Journal of Machine Learning Research*, 11, pp.2057~2078.

(21) Kiers, H. A. L., 2000, "Towards a Standardized Notation and Terminology in Multiway Analysis", *Journal of Chemometrics: A Journal of the Chemometrics Society*, 14(3), pp.105~122.

(22) Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. and Mullainathan, S., 2018, "Human Decisions and Machine Predictions", *The Quarterly Journal of Economics*, 133(1), pp.237~293.

(23) Koltchinskii, V., Lounici, K. and Tsybakov, A. B., 2011, "Nuclear-norm Penalization and Optimal Rates for Noisy Low-rank Matrix Completion", *The Annals of Statistics*, 39(5), pp.2302~2329.

(24) Little, R. J. A., 1988, "Missing-data Adjustments In Large Surveys", *Journal of Business & Economic Statistics*, 6(3), pp.287~296.

(25) Little, R. J. A. and Rubin, D. B., 1989, "The Analysis of Social Science Data with Missing Values", *Sociological Methods & Research*, 18(2~3), pp.292~326.

(26) Little, R. J. A. and Rubin, D. B., 2019, *Statistical Analysis with Missing Data*, John Wiley & Sons.

(27) Mao, X., Chen, S. X. and Wong, R. K. W., 2019, "Matrix Completion with Covariate Information", *Journal of the American Statistical Association*, 114(525), pp.198~210.

(28) Mao, X., Wong, R. K. W. and Chen, S. X., 2021, "Matrix Completion under Low-Rank Missing Mechanism", *Statistica Sinica*, 31(4), pp.2005~2030.

(29) Mazumder, R., Hastie, T. and Tibshirani, R., 2010, "Spectral Regularization Algorithms for Learning Large Incomplete Matrices", *Journal of Machine Learning Research*, 11, pp.2287~2322.

(30) Recht, B., 2011, "A Simpler Approach to Matrix Completion", *Journal of Machine Learning Research*, 12(12), pp.3413~3430.

(31) Reiter, J. P. and Raghunathan, T. E., 2007, "The Multiple Adaptations of Multiple Imputation", *Journal of the American Statistical Association*, 102(480), pp.1462~1471.

(32) Rohde, A. and Tsybakov, A. B., 2011, "Estimation of High-dimensional Low-rank Matrices", *The Annals of Statistics*, 39(2), pp.887~930.

(33) Robin, G., Klopp, O., Josse, J., Moulines, É. and Tibshirani, R., 2020, "Main Effects and Interactions in Mixed and Incomplete Data Frames", *Journal of the American Statistical Association*, 115(531), pp.1292~1303.

(34) Rubin, D. B., 1987, *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley

(35) Schnabel, T., Swaminathan, A., Singh, A., Chandak, N. and Joachims, T., 2016, "Recommendations as Treatments: Debiasing Learning and Evaluation", *International Conference on Machine Learning*, PMLR, pp.1670~1679.

(36) Wang, J., Wong, R. K. W., Mao, X. and Chan, K. C. G., 2021, "Matrix Completion with Model-free Weighting", *In International Conference on Machine Learning*, PMLR, pp.10927~10936.

(37) Wang, X., Zhang, R., Sun, Y. and Qi, J., 2019, "Doubly Robust Joint Learning for Recommendation on Data Missing Not at Random", *International Conference on Machine Learning*, PMLR, pp.6638~6647.

(38) Xia, D. and Yuan, M., 2021, "Statistical Inferences of Linear Forms for Noisy Matrix Completion", *Journal of the Royal Statistical Society: Series B(Statistical Methodology)*, 83(1), pp.58~77.

(39) Zhang, P., 2003, "Multiple Imputation: Theory and Method", *International Statistical Review/Revue International Statistical Review*, pp.581~592.

(40) Zhou, H., Zhang, D., Xie, K. and Chen, Y., 2015, "Spatio-temporal Tensor Completion for Imputing Missing Internet Traffic Data", *IEEE 34th International Performance Computing and Communications Conference(IPCCC)*, *IEEE*, pp.1~7.

(上接第163页)*Industrial Marketing Management*, Vol.43, No.6, pp.938~950.

(71) Walsh, J. P., 1995, "Managerial and Organizational Cognition: Notes from a Trip Down Memory Lane", *Organization Science*, Vol.6, No.3, pp.280~321.

(72) Wang, T. and Chen, Y., 2018, "Capability Stretching in Product Innovation", *Journal of Management*, Vol.44, No.2, pp.784~810.

(73) Wassmer, U., Li, S. and Madhok, A., 2017, "Resource Ambidexterity through Alliance Portfolios and Firm Performance", *Strategic Management Journal*, Vol.38, No.2, pp.384~394.

(74) Wen, J., Qualls, W. J. and Zeng, D., 2020, "Standardization Alliance Networks, Standard-Setting Influence, and New Product Outcomes", *Journal of Product Innovation Management*, Vol.37, No.2, pp.138~157.

(75) Yang, W., Gao, Y., Li, Y., Shen, H. and Zheng, S., 2017, "Different Roles of Control Mechanisms in Buyer-supplier Conflict: An Empirical Study from China", *Industrial Marketing Management*, Vol.65, pp.144~156.

(76) Yin, R. K., 2009, *Case Study Research: Design and Methods*, Beverly Hills, CA: Sage Publications Inc.

(77) Yin, R. K., 2013, "Validity and Generalization in Future Case Study Evaluations", *Evaluation*, Vol.19, No.3, pp.321~332.

(78) Yoo, Y., Boland, R. J., Lyytinen, K. and Majchrzak, A., 2012, "Organizing for Innovation in the Digitized World", *Organization Science*, Vol.23, No.5, pp.1398~1408.

(79) Yoo, Y., Henfridsson, O. and Lyytinen, K., 2010, "The New Organizing Logic of Digital Innovation: An Agenda for Information Systems Research", *Information Systems Research*, Vol.21, No.4, pp.724~735.

(80) Zittrain, J., 2006, "The Generative Internet", *Harvard Law Review*, Vol.119, No.7, pp.1974~2040.

Missing Data Completion in the Big Data Era: Challenges and Solutions

Chen Song Xi[a,b], *Mao Xiaojun*[c] *and Wang Cong*[a]

(a. Guanghua School of Management, Peking University, Beijing; b. Center for Statistical Science, Peking University, Beijing; c. School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai)

Summary: With the advent of the digital economy era, data has been regarded as an important production factor and has profoundly changed the paradigm of managerial decision-making. The analysis and utilization of big data have become an important factor in enabling managerial practices. The quality and completeness of the data is an important prerequisite for subsequent data value creation. However, due to factors such as the data collection method, collection process, and behavior patterns of the subjects collected, the big data often exhibits the characteristics of ultra-high missing rate, which seriously affects the subsequent effects of data analytics and decision-making. Therefore, the effective completion of big data is significant to subsequent data analytics and decision-making.

Given that most big data are in matrix format, this article systematically studies the big data completion problem first from the perspective of matrix completion. The core task of the matrix completion problem is to obtain an estimated matrix $\hat{\boldsymbol{A}}$ of the true matrix $\boldsymbol{A}_0$ through the observed matrix $\boldsymbol{Y}$ with a high missing rate. To make the problem tractable, some structural assumption is usually imposed on the estimated matrix, e.g., assuming that the observed matrix is generated by a small number of rows or columns. Hence, this article proposes to formulate the data completion problem as an optimization problem to obtain the estimated $\hat{\boldsymbol{A}}$ by minimizing the combination of loss function $L(\boldsymbol{A},\boldsymbol{Y})$ and the regularization term $R(\boldsymbol{A})$, i.e., $\boldsymbol{A}=\arg\min_{A\in\mathcal{A}}\{L(\boldsymbol{A},\boldsymbol{Y})+R(\boldsymbol{A})\}$

Subsequently, the article describes the matrix completion optimization problems in three typical contexts of big data, i.e., big data with ultra-high dimension, multi-source heterogeneity, and temporal-spatial correlation. In each context, the specification of the loss function and the regularization is introduced and the solution paths are described with concrete managerial examples. In the big data with ultra-high dimension context, the low-rank regularizer is usually imposed on the structure of the estimated matrix. For the loss function part, three types of missing mechanisms, i.e., missing completely at random, missing at random and missing not at random, are considered and the loss functions are derived accordingly. In the heterogeneous data context, two typical scenarios are studied, i.e., mild heterogeneity with data from the same distribution of different parameters, and strong heterogeneity, where data come from different distributions. The article articulates the state-of-the-art optimization problem specification, solution methods as well as potential completion outcomes for both settings. For the temporal-spatial correlated case, the tensor completion problem can be formulated to take account of the temporal or spatial information, and thus the loss function, as well as the regularization parts, are designed accordingly. Tensor factorization methods can be used to solve the high-dimensional optimization problem.

The article also provides a guideline to select different data completion methods based on the characteristics of the data, such as missing mechanism, dimensionality, source diversification, etc. Furthermore, this article also discusses how follow-up data analytics can be carried out after completion.

This article contributes to related literature as well as managerial practices in three folds. First, a unified problem formulation framework with easy extension is proposed. Second, various state-of-the-art solutions are systematically summarized with their applicability, pros and cons, which provides a good reference for relevant scholars and managerial practitioners. Third, the effective usage of the data completion methods will contribute to managerial decision-making.

Keywords: data completion; ultra-high dimension; multi-source heterogeneity; temporal-spatial correlation; managerial decision-making

JEL Classification: C81