文章编号:1006-2467(2019)06-0734-07 DOI: 10.16183/j.cnki.jsjtu.2019.06.015

儿童孤独症的高秩矩阵填充模型与方法

李元超, 陈 峰
(上海交通大学 工业工程与管理系, 上海 200240)

摘 要: 针对儿童孤独症的临床数据填补,提出了一种基于交替方向乘子(ADMM)算法的高秩矩阵填充(HRMC)算法.该算法考虑到数据中的不同参数具有不同的重要性,在算法设计时将重要参数与非重要参数赋予不同权重.在案例分析中,利用生成的测试数据寻找最优的参数配置,并且将该算法应用到实际孤独症数据的填充中,填充结果与参数化填充方法和低秩矩阵填充方法的结果进行对比.结果显示:所提出的算法在矩阵填充精度方面显著优于其他算法,并可应用于实际的数据清洗与处理过程中.

关键词: 矩阵填充; 高秩矩阵; 交替方向乘子算法; 儿童孤独症
中图分类号: N 945 **文献标志码**: A

High-Rank Matrix Completion Method for Autism Spectrum Disorders

LI Yuanchao, *CHEN Feng*
(Industrial Engineering and Management, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: To solve the clinical data recover problem of autism spectrum disorders (ASDs), a high-rank matrix completion (HRMC) algorithm based on alternating direction method of multipliers (ADMM) was proposed. Under consideration of different parameters with different significance, the important parameters and unimportant ones were attached with various weights. In a case study, test data were generated to find the optimal parameters. Furthermore, the algorithm was applied on practical ASD clinical data. The results show that the algorithm performs better in comparison with other parameterized algorithms and normal matrix completion algorithm, which indicates that it can be applied in practical data cleaning and processing.

Key words: matrix completion; high-rank matrix; alternating direction method of multipliers (ADMM) algorithm; autism spectrum disorders (ASDs)

准确的儿童孤独症的影响因素分析对孤独症的早期发现与诊疗具有重要的作用.儿童孤独症(ASD)是广泛性发育障碍的一种亚型,以男性多见,男女比例约为 6∶1,起病于婴幼儿期,主要表现为不同程度的言语发育障碍、人际交往障碍、兴趣狭窄和行为方式刻板[1].孤独症的病因目前尚不明确,认为可能与遗传因素和后天因素都相关.又因为孤独症对患儿和患儿家庭的危害较大,所以孤独症的早

收稿日期:2017-10-17
基金项目:上海交通大学医工交叉合作项目(YG2013ZD05)
作者简介:李元超(1992-),男,陕西省榆林市人,硕士生,研究方向为统计优化与矩阵填充.
通信作者:陈 峰,男,副教授,博士生导师,E-mail:fchen@sjtu.edu.cn.

期诊断显得特别重要. 清晰准确地了解孤独症的影响因素可以有效地辅助孤独症的诊断，在孤独症的实际诊疗过程中发挥重要的作用，而临床数据中存在着大量的数据缺失现象. 一般而言，认为 10%～20%的数据缺失量即会对统计结果造成影响，使之偏离实际情况. 针对数据缺失现象，统计学家提出了很多填补缺失值的算法，如多重填补(MI)，随机森林(RF)，最大期望(EM)等.

矩阵填充(MC)方法是目前解决数据缺失的有效方法，由压缩感知发展而来. 所谓压缩感知，指的是利用信号的稀疏特性，在远远小于奈奎斯特(Nyquist)采样率的条件下，用随机采样获取信号的离散样本，然后通过非线性算法完美重建信号. 之后，Candès 等[2]将压缩感知的原理拓展到二维层面，使得原本的向量重建变为对矩阵中缺失元素的填补，并且称其为矩阵填充. 矩阵填充在处理大规模数据的表现方面尤其突出，时间复杂性低且精度高. 特别是在处理非结构化数据方面，如图形重建、数字识别和图形压缩等，矩阵填充方法表现出了极大的优势，而且在缺失类型为非随机缺失时，矩阵填充也表现良好.

近年来矩阵填充的新方法不断被提出，原有方法的改进报道也很常见. 常用的低秩矩阵填充方法有 SVT (Singular Value Thresholding) 算法、APGL (Accelerated Proximal Gradient Linesearch-like) 算法、IALM (Inexact Augmented Lagrange Multiplier)算法等. 许多新的正则化方式被应用到矩阵填充的凸优化建模中，包括交替最小化、交替最小二乘法和截断核范数等. Xu 等[3]将交替方向算法应用到非负矩阵填充中，Recht 等[4]提出了利用平行随机梯度下降法求解大规模矩阵填充的方法.

矩阵填充的应用方向广泛，主要有协同滤波、系统识别、传感器网络和图像处理等方面. 经典的 Netflix 问题就是一个协同滤波矩阵填充问题. 根据 Netflix 问题，发展出了各种推荐系统的用户喜好预测问题. 曾广翔[5]进行了面向推荐系统的矩阵填充算法研究. Yeganli 等[6]将 SVT 算法应用于图形重建的过程中，有效地恢复了损坏图形中的纹理部分. Li 等[7]利用低秩矩阵填充进行了图形的修复，提出了一种鲁棒性较好的算法. Ji 等[8]研究了鲁棒视频去噪，提出基于矩阵填充的视频块的去噪算法，结果显著优于其他去噪算法. 在医学上，矩阵填充方法被广泛应用于医学影像处理方面，主要用于提取磁共振的影像的特征，去除影像中的噪声以及修复影像[9].

在待填充矩阵稀疏度低的情况下，低秩矩阵的填充效率低. 这里稀疏度指矩阵中零元素占总元素的比例. 而在日常应用中，待填充矩阵的稀疏性往往不够强烈，此时低秩矩阵填充的算法就变得十分迟钝，表现难以令人满意. 针对这种情况，统计学家提出了高秩矩阵填充的概念. 高秩矩阵填充采用子空间聚类相关的思路. 认为数据来自 k 个不同的未知线性空间，任务是将数据分别归类到自己所属的子空间中. 近年来，许多高维数据背景下的子空间聚类算法被提出. Elhamifar[10]将矩阵的自我表示模型用于高秩矩阵填充. Fan 等[11]将交替方向乘子(ADMM)应用于一般的矩阵填充，并且解决了若干典型图像修复问题和子空间聚类问题，得到了令人满意的效果.

本文将 ADMM 算法应用于矩阵填充，在现有的高秩矩阵填充的成果基础上，提出了一种考虑待填充矩阵中各因素的重要性的高秩矩阵填充(HRMC)算法. 其创新性在于结合矩阵填充的特征，根据数据中各要素的重要性在问题建模时赋予不同的权重参数，提高了算法的效率，在解决高秩矩阵填充问题方法表现较为突出. 在数值实验部分采用了生成的测试数据与实际孤独症就诊数据，算法精度均表现优秀，填充结果合乎预期.

1　HRMC 算法介绍

矩阵填充问题的一般描述如下：

$$\left.\begin{aligned}&\min\ \operatorname{rank}(\boldsymbol{X})\\&\text{s.t. } \boldsymbol{X}_{ij}=\boldsymbol{M}_{ij}\end{aligned}\right\}\tag{1}$$

式中：$\boldsymbol{X}$ 为待恢复矩阵；$\boldsymbol{M}$ 为 $\boldsymbol{X}$ 中已知元素的集合，且 $i,j\in\Omega$，Ω 为矩阵中已知元素的下标集合. 但是，该问题为 NP-hard，且时间复杂性为指数[2]，故求解起来十分困难. 此时，Candès 等[2]指出可以通过求解原问题的近似问题来求解：

$$\left.\begin{aligned}&\min\ \|\boldsymbol{X}\|_*\\&\text{s.t. } \boldsymbol{X}_{ij}=\boldsymbol{M}_{ij},\quad i,j\in\Omega\end{aligned}\right\}\tag{2}$$

式中：$\|\boldsymbol{X}\|_*$ 为核范数，$\|\boldsymbol{X}\|_*=\sum_{k=1}^{n}\sigma_k(\boldsymbol{X})$；$\sigma_k$ 为矩阵按从大到小排列的第 k 个奇异值. 令 $n=\max\{n_1,n_2\}$，$n_1\times n_2$ 为矩阵 $\boldsymbol{X}$ 的维度，则存在常数 G，当矩阵中被观测到的元素个数 m 满足以下条件时，矩阵能被完整恢复：

$$m\geqslant Gn^{6/5}r\lg n\tag{3}$$

式中：r 为矩阵的秩. 但是，这种矩阵填充的方法仅仅适用于低秩矩阵的填充，而在日常的应用中，往往会遇到需要恢复高秩矩阵甚至满秩矩阵的情况. 通常而言，高秩矩阵填充往往与子空间聚类(SSC)有

关，所谓子空间聚类，指有向量 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\in\mathbf{R}^n$，可以被聚类为最多 k 个子空间. SSC 的一般描述如下：

$$\left.\begin{aligned}&\{\boldsymbol{C},\boldsymbol{E}\}=\arg\min_{\boldsymbol{C},\boldsymbol{E}}\|\boldsymbol{C}\|_l+\lambda\|\boldsymbol{E}\|_l\\&\text{s.t. }\boldsymbol{X}=\boldsymbol{X}\boldsymbol{C}+\boldsymbol{E},\ \operatorname{diag}(\boldsymbol{C})=\boldsymbol{0}\end{aligned}\right\}\tag{4}$$

式中：$\boldsymbol{C}$ 为矩阵 $\boldsymbol{X}$ 的自我表示矩阵，将矩阵 $\boldsymbol{X}$ 表示成 $\boldsymbol{X}$ 本身的列的线性组合；$\boldsymbol{E}$ 为误差；l 为范数，可根据实际问题取某一个特定的范数；λ 为由原数据的噪声情况确定，原数据噪声特别大时，λ 取比较小的值，反之则取较大的值.

本文结合矩阵填充的思路，利用 ADMM 算法来实现. ADMM 算法通过求解一系列结构相似的子问题来优化未知变量和参数[12]. ADMM 算法的一般描述如下：

$$\left.\begin{aligned}&\{\boldsymbol{x},\boldsymbol{z}\}=\arg\min_{x,z}f(\boldsymbol{z})+g(\boldsymbol{z})\\&\text{s.t. }\boldsymbol{a}\boldsymbol{x}+\boldsymbol{b}\boldsymbol{z}=\boldsymbol{c}\end{aligned}\right\}\tag{5}$$

先求得其增广拉格朗日乘子函数：

$$L_\mu(\boldsymbol{x},\boldsymbol{z},\boldsymbol{y})=f(\boldsymbol{x})+g(\boldsymbol{z})+\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{a}\boldsymbol{x}+\boldsymbol{b}\boldsymbol{z}-\boldsymbol{c})+\frac{\mu}{2}\|\boldsymbol{a}\boldsymbol{x}+\boldsymbol{b}\boldsymbol{z}-\boldsymbol{c}\|_2^2\tag{6}$$

然后进行迭代更新，并且重复更新过程，直到整个算法收敛为止.

因为矩阵是不完整的，所以要解决的问题较 SSC 的研究问题增加了一个约束条件. 在实际应用中很难知道待填充的矩阵的秩，故引入了参数 α_{p} 和 λ. 其中：α_{p} 影响填充矩阵的秩，对较高秩的待填充矩阵，α_{p} 应该取比较小的值，反之则取一个较大的值. 一般而言，待研究的结构化数据构成的目标矩阵中往往存在关键因素和非关键因素，为了提高算法的运行速度和效率，对关键因素与非关键因素分别赋权值. 如此可以使算法将更多资源投入关键因素的缺失项填充过程中，提升效率. 注意到约束 $\boldsymbol{X}=\boldsymbol{X}\boldsymbol{C}+\boldsymbol{E}$ 中，对于不确定的 $\boldsymbol{X}$ 和 $\boldsymbol{C}$，它是非凸的，而对每一个确定的 $\boldsymbol{X}$ 和 $\boldsymbol{C}$ 来说，该约束都是凸的，如此则可用 ADMM 算法求解.

本文研究的高秩矩阵填充整个问题描述如下：

$$\left.\begin{aligned}&\{\boldsymbol{C},\boldsymbol{E}\}=\arg\min_{\boldsymbol{C},\boldsymbol{E}}\|\boldsymbol{C}\|_l+\lambda\|\boldsymbol{E}\|_l+\sum\alpha_{\mathrm{p}}\|\boldsymbol{Y}_{\mathrm{p}}\|_*\\&\text{s.t. }\boldsymbol{X}=\boldsymbol{X}\boldsymbol{C}+\boldsymbol{E},\ \boldsymbol{X}_{ij}=\boldsymbol{M}_{ij},\ \operatorname{diag}(\boldsymbol{C})=\boldsymbol{0}\end{aligned}\right\}\tag{7}$$

为了方便用 ADMM 求解，式(7)可以表示为

$$\left.\begin{aligned}&\{\boldsymbol{X},\boldsymbol{C},\boldsymbol{E}\}=\arg\min_{\boldsymbol{X},\boldsymbol{C},\boldsymbol{E}}\|\boldsymbol{C}\|_l+\sum\alpha_{\mathrm{p}}\|\boldsymbol{Y}_{\mathrm{p}}\|_*+\lambda\|\boldsymbol{E}\|_l\\&\text{s.t. }\boldsymbol{X}=\boldsymbol{X}\boldsymbol{A}+\boldsymbol{E},\ \boldsymbol{X}_{ij}=\boldsymbol{M}_{ij},\ \operatorname{diag}(\boldsymbol{C})=\boldsymbol{0}\\&\quad\boldsymbol{Y}=\boldsymbol{X},\ \boldsymbol{C}=\boldsymbol{A}\end{aligned}\right\}\tag{8}$$

式中：$\boldsymbol{Y}$ 和 $\boldsymbol{A}$ 均为辅助矩阵；$\boldsymbol{Y}_{\mathrm{p}}$ 为原矩阵中对应权值 α_{p} 的因素的集合. 增广拉格朗日乘子函数

$$\begin{aligned}L_\mu=&\|\boldsymbol{C}\|_l+\sum\alpha_{\mathrm{p}}\|\boldsymbol{Y}_{\mathrm{p}}\|_*+\lambda\|\boldsymbol{E}\|_l+\\&\operatorname{tr}(\boldsymbol{Q}_1^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{X}\boldsymbol{A}-\boldsymbol{E}))+\\&\operatorname{tr}(\boldsymbol{Q}_2^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{X}))+\operatorname{tr}(\boldsymbol{Q}_3^{\mathrm{T}}(\boldsymbol{C}-\boldsymbol{A}))+\\&\frac{\mu}{2}(\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{A}-\boldsymbol{E}\|_F^2+\|\boldsymbol{Y}-\boldsymbol{X}\|_F^2+\\&\|\boldsymbol{C}-\boldsymbol{A}\|_F^2)\end{aligned}\tag{9}$$

式中：$\boldsymbol{Q}_1^{\mathrm{T}}$、$\boldsymbol{Q}_2^{\mathrm{T}}$ 和 $\boldsymbol{Q}_3^{\mathrm{T}}$ 均为拉格朗日乘子；μ 为单调不减的罚因子；F 为范数. 利用 ADMM 求解上述问题的流程步骤如图 1 所示.

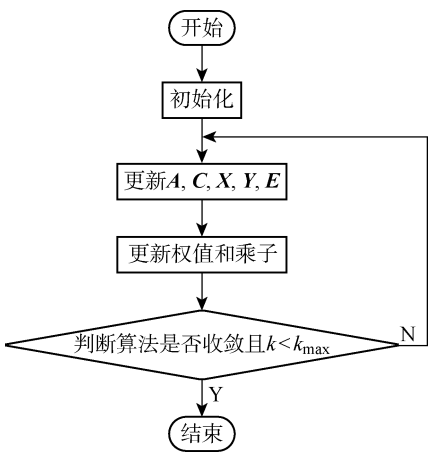

图 1 HRMC 算法流程

Fig. 1 HRMC algorithm process

输入：原始矩阵 $\boldsymbol{X}$，矩阵中已知值 $\boldsymbol{M}$，参数 α_{p} 和 λ.

初始化：$\boldsymbol{C}^{(0)}=\boldsymbol{A}^{(0)}=\boldsymbol{0}$，$\boldsymbol{Y}^{(0)}=\boldsymbol{X}^{(0)}=\boldsymbol{E}^{(0)}=\boldsymbol{0}$，$\boldsymbol{Q}_1^{(0)}=\boldsymbol{0}$，$\boldsymbol{Q}_2^{(0)}=\boldsymbol{0}$，$\boldsymbol{Q}_3^{(0)}=\boldsymbol{0}$，$\mu=0.1$，$\mu_{\max}=10^7$，步长 $\rho=1.01$，控制计算精度 $\varepsilon=10^{-6}$，最大迭代次数 $k_{\max}=10\,000$，$k=0$.

当算法未收敛且 $k<k_{\max}$ 时，执行如下迭代步骤：

(1) 更新 $\boldsymbol{A}^{(k+1)}$，即

$$\boldsymbol{A}^{(k+1)}=\arg\min\frac{\mu}{2}\times\left(\left\|\boldsymbol{X}^{(k)}-\boldsymbol{X}^{(k)}\boldsymbol{A}^{(k)}-\boldsymbol{E}^{(k)}+\frac{\boldsymbol{Q}_1}{\mu}\right\|_F^2+\left\|\boldsymbol{C}^{(k)}-\boldsymbol{A}^{(k)}+\frac{\boldsymbol{Q}_3}{\mu}\right\|^2\right)\tag{10}$$

(2) 更新 $\boldsymbol{C}^{(k+1)}$，即

$$\left.\begin{aligned}&\boldsymbol{C}^{(k+1)}=\arg\min\|\boldsymbol{C}\|_l+\frac{\mu}{2}\left\|\boldsymbol{C}^{(k)}-\boldsymbol{A}^{(k+1)}+\frac{\boldsymbol{Q}_3}{\mu}\right\|_F^2\\&\operatorname{diag}(\boldsymbol{C})=\boldsymbol{0}\end{aligned}\right\}\tag{11}$$

(3) 更新 $\boldsymbol{X}^{(k+1)}$，即

$$\left.\begin{aligned}\boldsymbol{X}^{(k+1)} &= \arg\min \frac{\mu}{2}\times\\ &\left(\left\|\boldsymbol{X}^{(k)}-\boldsymbol{X}^{(k)}\boldsymbol{A}^{(k+1)}-\boldsymbol{E}^{(k)}+\frac{\boldsymbol{Q}_1}{\mu}\right\|_F^2+\right.\\ &\left.\left\|\boldsymbol{Y}^{(k)}-\boldsymbol{X}^{(k)}+\frac{\boldsymbol{Q}_2}{\mu}\right\|_F^2\right)\\ \boldsymbol{X}_{ij} &= \boldsymbol{M}_{ij}\end{aligned}\right\} \tag{12}$$

(4) 更新 $\boldsymbol{Y}^{(k+1)}$，即

$$\boldsymbol{Y}^{(k+1)} = \arg\min \alpha\|\boldsymbol{Y}^{(k)}\|_* + \frac{\mu}{2}\left\|\boldsymbol{Y}^{(k)}-\boldsymbol{X}^{(k+1)}+\frac{\boldsymbol{Q}_2}{\mu}\right\|_F^2 \tag{13}$$

此时 $\boldsymbol{Y}$ 为 $\boldsymbol{Y}_{\mathrm{p}}$ 的集合.

(5) 更新 $\boldsymbol{E}^{(k+1)}$，即

$$\boldsymbol{E}^{(k+1)} = \arg\min \lambda\|\boldsymbol{E}^{(k)}\|_* + \frac{\mu}{2}\left\|\boldsymbol{X}^{(k+1)}-\boldsymbol{X}^{(k+1)}\boldsymbol{A}^{(k+1)}-\boldsymbol{E}^{(k)}+\frac{\boldsymbol{Q}_1}{\mu}\right\|_F^2 \tag{14}$$

(6) 更新乘子：

$$\left.\begin{aligned}\boldsymbol{Q}_1 &= \boldsymbol{Q}_1+\mu(\boldsymbol{X}^{(k+1)}-\boldsymbol{X}^{(k+1)}\boldsymbol{A}^{(k+1)}-\boldsymbol{E}^{(k+1)})\\ \boldsymbol{Q}_2 &= \boldsymbol{Q}_2+\mu(\boldsymbol{Y}^{(k+1)}-\boldsymbol{X}^{(k+1)})\\ \boldsymbol{Q}_3 &= \boldsymbol{Q}_3+\mu(\boldsymbol{C}^{(k+1)}-\boldsymbol{A}^{(k+1)})\end{aligned}\right\} \tag{15}$$

(7) 更新参数：

$$\mu = \min(\rho\mu,\mu_{\max}) \tag{16}$$

ADMM 的收敛性已被 Stephen Boyd 等人证明，且本文在实际求解中，算法均可正常收敛.

在上面的算法描述中，各个迭代步骤都有相似的表达形式. 本文可以通过如下方法求 $\boldsymbol{A}^{(k+1)}$. 即

$$\boldsymbol{A}^{(k+1)} = ((\boldsymbol{X}^{(k)})^{\mathrm{T}}\boldsymbol{X}^{(k)}+\boldsymbol{I})^{-1}\left[(\boldsymbol{X}^{(k)})^{\mathrm{T}}\left(\boldsymbol{X}^{(k)}-\boldsymbol{E}^{(k)}+\frac{\boldsymbol{Q}_1}{\mu}\right)+\boldsymbol{C}^{(k)}+\frac{\boldsymbol{Q}_3}{\mu}\right] \tag{17}$$

同理，在步骤(2)中，$\boldsymbol{C}^{(k+1)}$ 的求法是：

$$\left.\begin{aligned}\boldsymbol{C}^{(k+1)} &= \Phi_{1/\mu}\left(\boldsymbol{A}^{(k+1)}-\frac{\boldsymbol{Q}_3}{\mu}\right)\\ \boldsymbol{C}^{(k+1)} &= \boldsymbol{C}^{(k+1)}-\mathrm{diag}(\boldsymbol{C}^{(k+1)})\end{aligned}\right\} \tag{18}$$

式中：$\Phi_\tau(*)$ 是一个近邻收缩阈值算子[13]，其具体定义为

$$\Phi_\tau(\omega) = \max\{|\omega|-\tau,0\}\mathrm{sgn}(\omega) \tag{19}$$

式中：sgn(*)为符号函数.

步骤(3)中，

$$\left.\begin{aligned}\boldsymbol{X}^{(k+1)} &= \left(\left(\boldsymbol{E}^{(k)}-\frac{\boldsymbol{Q}_1}{\mu}\right)(\boldsymbol{I}-\boldsymbol{A}^{(k+1)})^{\mathrm{T}}+\boldsymbol{Y}^{(k)}+\frac{\boldsymbol{Q}_2}{\mu}\right)((\boldsymbol{I}-\boldsymbol{A}^{(k+1)})(\boldsymbol{I}-\boldsymbol{A}^{(k+1)})^{\mathrm{T}}+\boldsymbol{I})^{-1}\\ \boldsymbol{X}_{ij}^{(k+1)} &= \boldsymbol{M}_{ij}\end{aligned}\right\} \tag{20}$$

步骤(4)中，

$$\boldsymbol{Y}^{(k+1)} = \boldsymbol{\Theta}_{\alpha/\mu}\left(\boldsymbol{X}^{(k+1)}-\frac{\boldsymbol{Q}_2}{\mu}\right) \tag{21}$$

式中：$\boldsymbol{\Theta}$ 为奇异值阈值算子[14].

步骤(5)中，$\boldsymbol{E}^{(k+1)}$ 求法比较特殊，因为根据不同的应用背景，$\boldsymbol{E}$ 的范数可以取不同类型，本文取 2.1 范数，具体的计算方法为

$$\boldsymbol{E}^{(k+1)} = \frac{1}{\mu}(2\lambda+\mu)\times\left(\boldsymbol{X}^{(k+1)}-\boldsymbol{X}^{(k+1)}\boldsymbol{A}^{(k+1)}+\frac{\boldsymbol{Q}_1}{\mu}\right) \tag{22}$$

2　数值实验

本文利用了人工生成的测试数据和来自上海市精神卫生中心孤独症门诊的患儿就诊数据进行案例分析. 在中国，孤独症患者的数据十分珍贵. 因孤独症未得到社会和家庭的足够重视，前来就诊的患儿较少，而且患儿家长在录入基本信息和填写有关量表的时候，由于自身的文化程度或其他原因，往往会有漏填、错填的现象出现，故原始数据中有不少缺失项需要处理. 又因数据珍贵，若删除缺失观测，原来的患者分布统计会有较大偏差，而且孤独症各个量表的得分，患者的基本情况各项相关性较弱，故整个数据矩阵的秩较高.

数值实验流程：首先采用生成的测试矩阵测试；然后利用实际的孤独症数据，抽取有代表性的无缺失样本进行测试，并且与常用的参数化填补方式和低秩矩阵填充算法做对比；最后利用全部的实际数据进行矩阵填充实验，观察算法的效率和填充是否合理.

2.1　测试数据填充实验

本文采用生成测试矩阵的方式来测试 HRMC 的表现，分别生成低秩矩阵，高秩矩阵和满秩矩阵，对于每一种矩阵均采用不同的缺失度(缺失度为缺失数据量占矩阵总数据量的比例)来测试. 假设测试数据矩阵 $\boldsymbol{X}(D\times N)$ 来自 k 个不同的子空间 $\{S_j\in\mathbf{R}^{d_j}\}_{j=1}^k$，根据 Yang 等[15] 的研究结果，2 个独立随机高斯矩阵(矩阵中的元素均符合标准正态分布)$\boldsymbol{A}$ 和 $\boldsymbol{B}$，$\boldsymbol{A}^{(j)}\in\mathbf{R}^{D\times d_j}$，$\boldsymbol{B}^{(j)}\in\mathbf{R}^{d_j\times n_j}$ 符合条件的测试数据矩阵 $\boldsymbol{X}=\{\boldsymbol{A}^{(j)}\boldsymbol{B}^{(j)}\}_{j=1}^k$，$\boldsymbol{X}$ 的秩满足

$$\mathrm{rank}(\boldsymbol{X}) = \min\left\{D,N,\sum_{j=1}^k d_j\right\}$$

矩阵填充的误差可以表示为

$$E_{\mathrm{MC}} = \frac{\|P_{\bar{\Omega}}(\boldsymbol{X}-\boldsymbol{X}_{\mathrm{ini}})\|_F}{\|P_{\bar{\Omega}}(\boldsymbol{X}_{\mathrm{ini}})\|_F} \tag{23}$$

式中：$\boldsymbol{X}$ 为填充后的矩阵；$\boldsymbol{X}_{\mathrm{ini}}$ 为填充前的矩阵；$\bar{\Omega}$ 为缺失元素的标号；$P_{\bar{\Omega}}$ 为投影算子. 测试矩阵配置如表 1 所示. 矩阵元素的缺失机制采用随机缺失

(MAR),即矩阵中的元素是否缺失符合伯努利分布,缺失率为 δ,令 $P(\boldsymbol{\Psi}) = \delta$,$\boldsymbol{\Psi}$ 表示矩阵中元素缺失这一事件. 注意到矩阵是完全随机生成的,所有列的权重均等,故 α_{p} 相等. 在不同的缺失度下,令 α_{p} 和 λ 分取不同值,进行矩阵填充实验,观察不同参数配置下的结果,确定一个合理的参数范围.

实验结果如表 2 所示. 表中:r_{MC} 为矩阵填充的秩;E_{MC} 为矩阵填充误差.

表 1 测试矩阵配置

Tab. 1 Parameters of test matrix

矩阵类型	D	k	d_j	n_j	j	r
低秩	100	2	5	50	2	10
高秩	25	5	4	50	4	20
满秩	25	5	5	50	5	25

表 2 测试矩阵填充情况

Tab. 2 Result of matrix completion

δ	α_{p}	λ	r_{MC}			E_{MC}			k		
			低秩	高秩	满秩	低秩	高秩	满秩	低秩	高秩	满秩
0.25	0.01	0.01	87	20	25	0.059 8	0.398 3	0.373 6	1 010	1 211	1 191
	0.1	0.1	10	20	25	0.030 1	0.088 7	0.128 4	909	1 148	1 136
	0.1	1	10	20	25	0.020 8	0.010 8	0.074 2	854	1 082	1 068
	1	1	10	20	25	0.040 4	0.108 0	0.228 3	855	840	837
	1	10	10	20	25	0.030 4	0.072 6	0.224 0	836	887	918
	10	10	99	25	25	0.811 1	0.289 1	0.371 5	76	809	825
0.50	0.01	0.01	93	20	25	0.205 6	0.822 7	0.770 0	1 032	1 260	1 229
	0.1	0.1	47	20	25	0.014 7	0.394 5	0.558 0	979	1 182	1 144
	0.1	1	10	20	25	0.001 5	0.420 8	0.619 0	868	1 095	1 077
	1	1	36	20	25	0.010 8	0.522 8	0.654 8	809	934	941
	1	10	10	20	25	0.001 8	0.542 0	0.672 4	897	1 006	962
	10	10	99	25	25	0.969 0	0.700 5	0.757 6	45	834	791
0.75	0.01	0.01	96	20	25	0.961 7	1.658 5	1.472 5	1 142	1 321	1 294
	0.1	0.1	80	20	25	0.461 2	1.162 2	1.194 0	1 047	1 215	1 188
	0.1	1	66	20	25	0.596 2	1.116 8	1.170 1	979	1 191	1 175
	1	1	73	20	25	0.434 6	1.251 7	1.306 0	959	1 065	1 071
	1	10	34	25	25	0.524 6	1.264 4	1.304 7	983	1 036	1 037
	10	10	70	25	25	0.607 4	1.346 3	1.375 2	972	919	900

由表 2 可以看出,随机生成的测试数据实验结果较好,可以正常收敛,且矩阵填充误差均在可接受范围内. 该算法在不同的缺失率下均可使填充后矩阵的秩与原矩阵相等. 在随机缺失的前提下,δ 分别为 0.25 和 0.5 时基本可以正常填充,δ 到达 0.75 时填充效果就有了明显的下降. 这与低秩矩阵填充的经验不符. 原因是低秩矩阵中有效的元素本身较少,大部分元素均与有效元素相关,即使是高度缺失的矩阵中仍然保留有足够完整恢复原矩阵的信息;而高秩矩阵中本身的有效元素较多,高度缺失的情况下有效信息的损毁过于严重,以至于难以将原矩阵完整恢复.

2.2 实际孤独症数据子样本填充

为了检验 HRMC 算法在实际应用时的有效性,本文从原始数据中取部分数据组成一个有代表性的样本,该样本可以为 321×13 的矩阵. 该矩阵所有元素完整无缺失,而且能模拟真实数据缺失情况,缺失机制采用非随机缺失(NMAR),即任意人为成

块挖掉矩阵的数据，而不是矩阵中的每一个元素都有相等的几率被划去. 经过缺失处理，矩阵中共有 1 544 个值被划去，缺失度为 36.9%，训练结果如表 3 所示. 注意到实际的孤独症的备选影响因素有主次之分，故根据影响因素的重要性对 α_p 分别赋值，α_p^M 适用于主要影响因素，α_p^m 则适用于次要影响因素.

从表 3 中可以看出，当 $\alpha_p^M=1$，$\alpha_p^m=0.1$，$\lambda=1$ 时，矩阵填充误差达到最小，且收敛速度也最快，表明 HRMC 算法对缺失类型为 NMAR 的数据可以有效地填充. 为了说明 HRMC 算法的有效性，本文利用现在较为流行的处理缺失数据的方法—多重插补(MI)和被广泛应用的低秩矩阵填充(SVT)算法与之对比，缺失模式分别采用 MAR 和 NMAR，结果如表 4 所示.

由表 4 可见，HRMC 在 3 种算法中表现最为优秀，MI 紧随其后，SVT 最差，而且随着缺失度增加，矩阵填充误差随之上升. 其原因是 MI 的原理为贝叶斯预测，对矩阵的秩不敏感，而 SVT 对数据的低秩性和稀疏性要求较高. 而 MI 在缺失模式为 NMAR 时，填充误差急剧上升，HRMC 则不会. 这说明 MI 在处理 MAR 数据表现较为出色，但无法很好地处理 NMAR 数据，HRMC 则对数据的缺失类型不敏感. 若是能寻找一个原始数据的稀疏表示，则一般低秩矩阵填充算法也可应用. 综上所述，HRMC 在处理实际孤独症数据时表现良好，可以应用到大规模的数据处理中.

表 3　实际孤独症数据非随机缺失填充结果

Tab. 3　Result of practical ASD data completion in NMAR

α_p^M	α_p^m	λ	r_{MC}	E_{MC}	k
0.01	0.001	0.01	13	0.711 521	4 273
0.10	0.010	0.10	13	0.695 746	5 214
0.10	0.010	1.00	13	0.696 262	3 224
1.00	0.100	1.00	13	0.572 314	2 658
1.00	0.100	10.00	13	0.572 482	5 211
10.00	0.100	10.00	13	0.873 494	10 708

表 4　HRMC 与其他算法对比结果

Tab. 4　Comparison of HRMC and other algorithms

δ	HRMC	MI	SVT
0.25(MAR)	0.448	0.544	1.223
0.50(MAR)	0.603	0.576	1.523
0.75(MAR)	1.145	1.289	1.845
随意挖去(NMAR)	0.377	0.649	1.697

2.3　实证分析

经过必要的数据前处理，原始的孤独症患者数据(包括基本情况和各量表作答情况)被整理为一个 499×143 的矩阵，矩阵中共有 71 357 个数据，其中 15 497 个数据缺失，缺失度为 21.72%；共有 12 433 个值为 0，稀疏度为 17.4%，为稠密矩阵. 对该矩阵应用 HRMC 算法进行矩阵填充，结果很好，缺失值的填充结果均符合各项参数的取值范围，而且速度极快，大约 1 000 步迭代后即会收敛.

3　结语

本文研究了高秩矩阵的矩阵填充问题，提出了一种基于 ADMM 算法的高秩矩阵填充算法(HRMC). 该算法利用对偶最小化的方法，使得矩阵填充在矩阵本身的秩比较高(甚至满秩)，矩阵中的无效元素比较少的时候仍然能以较高的概率恢复原矩阵. 本文利用生成的测试数据和截取的实际孤独症数据进行测试，并且将 HRMC 与一般的参数化方法和低秩矩阵填充方法的表现做对比，结果是 HRMC 的精确度要显著优于其他算法. 不足之处在于全部实际孤独症数据的填充由于原始数据的缺失，故无法精确衡量填充结果. 另外，HRMC 算法对数据归一化要求高，若数据的各个参数取值波动巨大时，该算法收敛速度较慢. HRMC 算法不仅在医疗数据的填补方面，在图形修复、模式识别等领域应该也有不错的表现和广阔的应用前景.

参考文献：

[1] LAI M C, LOMBARDO M V, BARON-COHEN S. Autism[J]. **Lancet**, 2014, 383(9920): 896-910.

[2] CANDÉS E J, RECHT B. Exact matrix completion via convex optimization [J]. **Foundations of Computational Mathematics**, 2009, 9(6): 717-772.

[3] XU Y Y, YIN W T, WEN Z W, *et al*. An alternating direction algorithm for matrix completion with nonnegative factors[J]. **Frontiers of Mathematics in China**, 2012, 7(2): 365-384.

[4] RECHT B, RÉ C. Parallel stochastic gradient algorithms for large-scale matrix completion[J]. **Mathematical Programming Computation**, 2013, 5(2): 201-226.

[5] 曾广翔. 面向推荐系统的矩阵填充算法研究[D]. 合肥：中国科学技术大学，2015.
ZENG Guangxiang. Research of matrix completion algorithm for recommendation system [D]. Hefei: University of Science and Technology of China, 2015.

[6] YEGANLI S F, YU R. Image inpainting via singular value thresholding[C]// **Signal Processing and Communications Applications Conference (SIU)**. Haspolat: IEEE, 2013: 1-4.

[7] LI W, ZHAO L, LIN Z J, *et al*. Non-local image inpainting using low-rank matrix completion [J]. **Computer Graphics Forum**, 2014, 34(6): 121-122.

[8] JI H, LIU C, SHEN Z, *et al*. Robust video denoising using low rank matrix completion[C]// **IEEE Computer Society Conference on Computer Vision and Pattern Recognition**. San Francisco: IEEE, 2012: 1791-1798.

[9] ROOZGARD A, BARZIGAR N, VERMA P, *et al*. 3D medical image denoising using 3D block matching and low-rank matrix completion[C]// **2013 Asilomar Conference on Signals, Systems and Computers**. CA, USA: IEEE, 2013: 253-257.

[10] ELHAMIFAR E. High-rank matrix completion and clustering under self-expressive models [C]// **Advances in Neural Information Processing Systems**. Barcelona: Curran Associates, Inc. 2016: 73-81.

[11] FAN J C, CHOW T W S. Sparse subspace clustering for data with missing entries and high-rank matrix completion [J]. **Neural Networks**, 2017, 93: 36-44.

[12] BOYD S. Distributed optimization and statistical learning via the alternating direction method of multipliers [J]. **Foundations and Trends in Machine Learning**, 2010, 3(1): 1-122.

[13] BECK A, TEBOULLE M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems [J]. **SIAM Journal on Imaging Sciences**, 2009, 2(1): 183-202.

[14] CAI J F, CANDÉS E J, SHEN Z W, *et al*. A singular value thresholding algorithm for matrix completion [J]. **SIAM Journal on Optimization**, 2010, 20(4): 1956-1982.

[15] YANG C Y, ROBINSON D, VIDAL R. Sparse subspace clustering with missing entries [J]. **Proceedings of the 32nd International Conference on Machine Learning, PMLR**, 2015, 37: 2463-2472.

（本文编辑：钱宝珍）