doi:10.11835/j.issn.1000-582X.2024.214
引用格式:钟尧, 刘清蝉, 李昕泓, 等. 计及非负和低秩特性的用电数据缺失值插补[J]. 重庆大学学报, 2025,48(9): 1-11.

计及非负和低秩特性的用电数据缺失值插补

钟　尧[1],刘清蝉[1,2],李昕泓[2],林　聪[1,2],李腾斌[1],杨　超[2],付志红[2]
(1. 云南电网有限责任公司计量中心,昆明 650000; 2. 重庆大学 电气工程学院,重庆 400044)

摘要:随着智能电表的广泛应用,电网公司积累了大量原始用电数据,然而复杂工作环境下用电信息采集设备仍存在数据丢失的现象。在充分考虑高斯噪声影响的情况下,对存在缺失的原始用电数据进行填补。对独立用户数据序列重排得到原始用电数据矩阵,将其中的理想用电数据矩阵进行非负矩阵分解替代;分别选择F范数和核范数对高斯噪声和具有低秩特性的理想用电数据进行正则化约束以构建优化模型;最后,基于块坐标最小算法框架使用EM算法和直接法交替更新非负矩阵分解得到的矩阵因子,从而有效实现数据的准确插补。仿真分析和实验结果验证了算法的有效性和准确性。

关键词:用电数据;非负矩阵分解;范数;块坐标下降法;矩阵完备
中图分类号:U469.72+2　　**文献标志码:**A　　**文章编号:**1000-582X(2025)09-001-11

Complete electricity data reconstruction based on weighted nonnegative matrix factorization

ZHONG Yao[1], *LIU Qingchan*[1,2], *LI Xinhong*[2], *LIN Cong*[1,2], *LI Tengbin*[1], *YANG Chao*[2], *FU Zhihong*[2]
(1. Measurement Center of Yunnan Power Grid Co., Kunming 650000, P. R. China; 2. School of Electrical Engineering, Chongqing University, Chongqing 400044, P. R. China)

Abstract: With the widespread deployment of smart meters, power grids have accumulated vast amounts of raw electricity consumption data. However, data loss remains a challenge due to the complex operational environments of data acquisition equipment. This study addresses the problem of incomplete electricity consumption data by accounting for the influence of Gaussian noise and proposing a robust completion method. First, a electricity consumption data matrix is constructed by reorganizing the sequences of individual users, and the ideal electricity data matrix is approximated using nonnegative matrix factorization (NMF). Second, both the Frobenius norm and the nuclear norm are employed to regularize the Gaussian noise and promote low-rank characteristics of the ideal matrix, thereby formulating an optimization model. Finally, within a block coordinate descent framework, the EM algorithm and a direct updating method are applied alternately to update the matrix factors derived from NMF, enabling accurate and complete data reconstruction. Simulation and experimental results validate the proposed

收稿日期:2023-07-10　　**网络出版日期:**2024-09-26
基金项目:云南电网科技资助项目(YNKJXM20210147)。
Supproted by Yunnan Power Grid Technology Project (YNKJXM20210147).
作者简介:钟尧(1983—),男,高级工程师,主要从事电能计量及智能运维研究,(E-mail) 93336425@qq.com。
通信作者:李昕泓,女,(E-mail)1347240996@qq.com。

algorithm's effectiveness and accuracy.

Keywords: electricity consumption data; nonnegative matrix factorization; norm; block coordinate descent; matrix completion

近年来,随着数字化技术在智能电网中的广泛应用,电网公司在电力系统的运行中积累了大量数据,其中用户侧大数据占很大比重。用户侧大数据运用的基本特征在于数据量规模庞大、数据结构类型多和数据的交互性。基于这些基本特征,电力大数据的应用不仅可以促进电网向互动经济、安全可靠、清洁高效的现代能源互联网转变,还可以提升电力系统的运营效率和管理水平[1-2]。同时,由于数据挖掘技术的持续发展和逐渐成熟,电力大数据的应用也越来越普及,其蕴含的潜在价值也不断被挖掘并应用于电网建设中。直观反映电网各个节点的电能消耗与传输信息的应用,例如风电场功率预测、台区负荷预测、配电网低电压定位、配电网重过载风险评估、电网电能质量分析、电能计量设备在线监测、电网需求侧管理等[3-5]。

在复杂的工作环境下,智能电网的电力大数据因电能计量装置故障、传输过程通信故障、人为活动等随机因素存在着数据缺失的现象,如智能测控系统的损坏和异常、用电数据传输不稳定可能会导致采集数据缺失;线路维护或安检等配电网的正常活动可能会导致负荷数据缺失[6-7]。这些用电数据的缺失会影响电网数据质量,进一步影响后续电力大数据应用的准确性和有效性。因此,在对用电数据进行功能性应用之前,将缺失数据填补完整的处理极其重要。

数据缺失的机制有3种:1)完全随机缺失(missing completely at random,MCAR),数据的缺失与不完全变量、完全变量都是无关的;2)随机缺失(missing at random,MAR),数据的缺失仅仅依赖于完全变量;3)非随机缺失(not missing at random,NMAR),不完全变量中数据的缺失依赖于不完全变量本身。用电数据的缺失属于随机缺失,数据是否缺失与数据本身无关,而在于观测数据的方法手段。

随着人工智能技术在用电数据质量处理的应用逐渐增加,国内外学者提出了许多填补缺失数据的方法。Papageorgiou等[8-9]考虑用电数据内部有界噪声和异常值,使用"稀疏性感知"参数来建模和估计异常值,采用鲁棒去噪(greedy algorithm for robust denoising, GARD)的贪婪算法,在最小二乘优化标准和识别异常值的正交匹配追踪选择步骤之间交替。Mateos等[10]运用范数正则化估计对噪声和异常值建模,对具有稀疏性的异常值进行稳健的非参数回归。这种基于凸松弛的高效解算器等效于最小绝对收缩和选择算子Lasso的变分M型估计器。采用新型基于鲁棒样条曲线的平滑器对负荷曲线数据进行平滑的计算结果良好,但不确定集合使得鲁棒优化模型复杂度很高,求解也会变得困难。

有学者运用人工神经网络处理存在缺失的用电数据[11-14],如Suo等[11]提出了一种基于循环神经网络的插补法来填充多元时间序列中的缺失值,使用全局和特定变量循环神经网络基于历史信息执行插补并将它们融合,对每个时间节点使用回归层来估算某个变量的值;Liu等[12]基于长短期记忆网络和卷积神经网络开发了一种深度学习方法预测用电轨迹,根据已有历史数据对未知数据进行预测。但文献[11-12]疏于考虑原始数据中噪声和异常值的影响,导致实验结果与预期可能存在偏差。针对数据缺失和异常值问题,Meirat等[13]基于人工神经网络和模糊逻辑提出了一种人工智能处理模块,完成原始数据补全和异常校正,该方法在缺失基础上考虑异常值的影响,可以使计算结果更加合理准确。

聚类分析在数据挖掘中有重要作用[14-15],基于相似性将数据分类的统计分析方法来分类多组用电数据。袁忠军等[14]将聚类方法与神经网络相结合,提出了一种基于自组织特征映射网络的结构自适应的聚类神经网络;杨挺等[15]发现有效聚类可使用电数据矩阵呈良好低秩特性,通过低秩矩阵恢复算法完成缺失数据修补。由于单用户自身数据具备极强相似性,而上述求解过程采用的多用户聚类分析数据填充法,对用户类型繁杂、用电场景及行为差异较大的情形无法直接套用。同时,聚类前预填充的处理,不仅放大了非完备聚类的影响,也使得结果具有一定由预填充和聚类随机性所带来的误差,且计算复杂程度较高。杨挺等[15]以矩阵完备的数学思想处理存在缺失的用电数据,基于矩阵范数优化理论的低秩矩阵完备构建低秩矩阵恢复模型,

并且能够针对不同的噪声和异常进行滤除。与人工智能方法修复数据的思路不同，低秩矩阵填充的数学方法优势在于不需要先验知识的训练，且计算复杂程度更低。

基于低秩矩阵完备的数学方法，在低秩矩阵填充的框架上采用非负矩阵分解可以使得计算具有良好的收敛效果且结果更准确。Guan 等[16]提出了一种高效的非负矩阵分解求解器，应用 Nesterov 的最优梯度方法来交替优化，克服了可能不收敛、收敛速度慢、数值结果不稳定的问题。Dorffer 等[17]在 Guan 等[16]的基础上将 Nesterov 迭代的非负矩阵分解扩展到观察到的矩阵中缺少一些条目即非完备矩阵的情况，引入权重信息处理缺失数据。Giampouras 等[18]提出了针对低秩矩阵分解的交替最小化算法，运用正则化器交替更新由矩阵分解产生的 2 个矩阵因子，解决了运用非负矩阵分解进行正则化约束求解问题。

文中提出一种基于加权非负矩阵分解的低秩矩阵完备的用电数据质量处理算法，基于矩阵范数优化理论对高斯噪声和具有低秩特性的理想用电数据进行正则化约束以构建目标函数模型，并使用块坐标下降法交替更新。以伦敦某用电台区独立用户用电数据作为研究对象，仿真结果表明该算法不仅能够恢复缺失数据，得到理想用电并滤除高斯噪声，具有良好的收敛性和精确性。最后通过存在随机缺失的云南某台区独立用户用电数据验证该算法，图像曲线对比证明该算法可以有效地实现用电数据缺失值的插补，提升用电数据质量。

1　独立用户用电数据分析

将独立用户每天等间隔采集的用电量数据进行重排以构成原始用电数据矩阵，该矩阵部分元素存在缺失（如图 1 中电能为零的数据点），故需要通过对矩阵已知数据计算处理恢复出未知数据以填补缺失元素，即矩阵完备。就独立用户个体自身用电行为分析，该用户每天相同时间段内的用电情况较台区其余用户总体相同时间段内的用电行为而言具有显著的内在相似性，因此由独立用户用电数据构造的原始用电数据矩阵低秩特性的表现更为明显。文中根据不同用户的自身用电行为分析处理该用户用电数据并构造矩阵，再基于矩阵的低秩特性实现用电数据的缺失值插补。

以某低压台区电能表的采集数据为例，系统定时采样，间隔时间为 15 min。该低压台区某独立用户 30 d 消耗电能如图 1 所示，可知每天每间隔时间的用电数据，并且能够直观表现该用户的用电情况与用电趋势，若电能为 0 表示该时刻用电数据缺失。

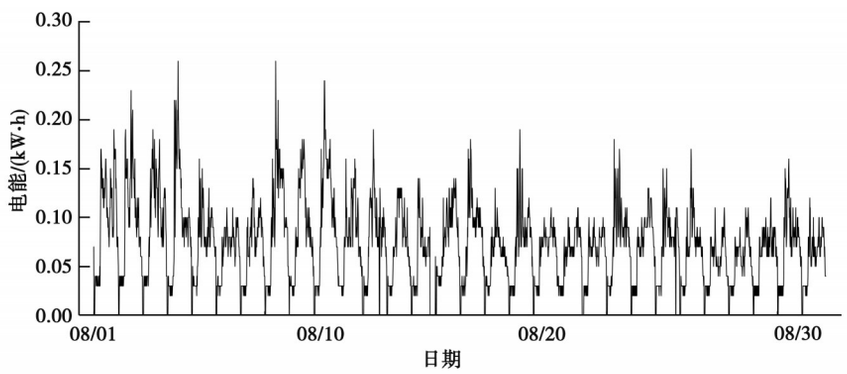


图 1　某用户 30 d 用电数据

Fig. 1　Electricity consumption data of a user for 30 days

将用电数据以 1 天 24 h 划分为序列（向量），重新排列构成独立用户的原始用电数据矩阵

$$\boldsymbol{X}=[\,\boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \cdots \quad \boldsymbol{x}_N\,]\in\mathbb{R}_{+}^{96\times N}, \tag{1}$$

式中：$\boldsymbol{x}_n=[x_1\ x_2\ x_3\cdots x_{96}]^{\mathrm{T}}$ 表示电能表采集的该用户第 n 天的用电数据；$N\in\mathbb{N}_{+}$ 表示采样天数。若当某一采样时刻的用电数据缺失，则该矩阵对应位置的元素 x_{mn} 置零。

将图 1 所示用电数据按式（1）重新排列，故可得以“1 天 24 h”为单位的该独立用户用电数据矩阵如图 2 所

示。反映在用电数据重排上，独立用户的数据具有相似的分布，由此可以初步判断独立用户自身可能具有相似的用电规律。

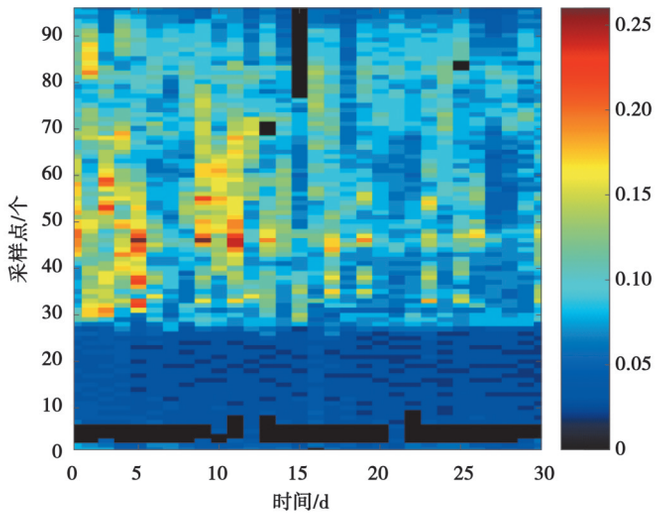


图2 某用户30 d用电数据重排

Fig. 2 Rearrangement of 30 day electricity consumption data for the certain user

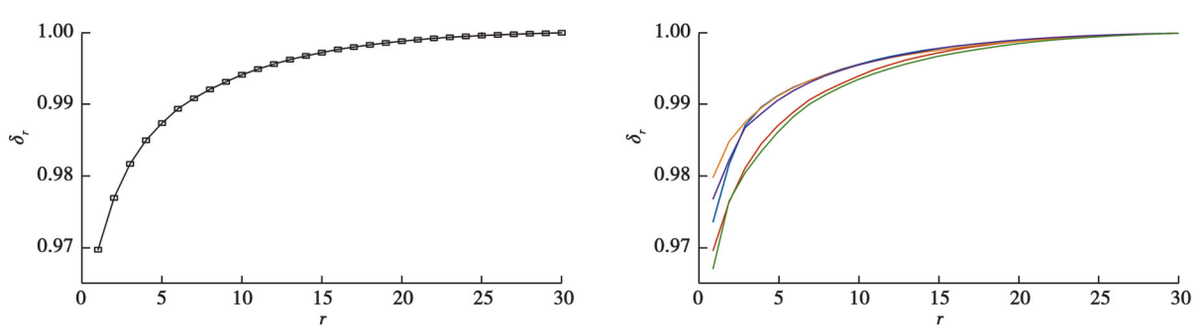
对原始用电数据矩阵$\boldsymbol{X}$进行SVD分解，

$$\boldsymbol{X}_{96\times N}=\boldsymbol{U}_{96\times 96}\boldsymbol{\Sigma}_{96\times N}\boldsymbol{V}^{\mathrm{T}}_{N\times N}, \tag{2}$$

式中：矩阵$\boldsymbol{U}$为左奇异矩阵；矩阵$\boldsymbol{\Sigma}$是半正定96×N阶对角矩阵；矩阵$\boldsymbol{V}$为右奇异矩阵。奇异值σ_i位于矩阵$\boldsymbol{\Sigma}$的主对角线上，且满足$\sigma_i>\sigma_j$，$1\leqslant i<j\leqslant\min(96,N)$。根据矩阵的低秩性判断条件

$$\delta_r=\frac{\sqrt{\sigma_1^2+\sigma_2^2+\cdots+\sigma_r^2}}{\sqrt{\sigma_1^2+\sigma_2^2+\cdots+\sigma_r^2+\cdots+\sigma_{R=\min(96,N)}^2}}, \tag{3}$$

式中，$0<\delta_r<1$恒成立，根据δ_r接近1时的r值可判定用电数据矩阵$\boldsymbol{X}$的低秩特性，即当$\delta_r\approx 1$时r越小，用电数据矩阵$\boldsymbol{X}$的低秩特性越好。对某用户某一组30 d用电数据进行低秩特性分析（见图2），其对应δ_r随r值的变化趋势如图3(a)所示，可知该用户30 d用电数据符合低秩。为进一步验证独立用户用电数据存在显著低秩，在该用户所有用电数据中随机选择5组30 d的数据按照上述方式进行SVD分解并判断低秩特性，其δ_r随r值的变化如图3(b)所示，可知每组数据均符合低秩。



(a) 一组数据低秩特性分析　　(b) 多组数据低秩特性分析

图3 用电数据低秩特性分析

Fig. 3 Low-rank analysis of electricity consumption data

上述结果表明，基于独立用户自身用电行为的内在相似性，由1天24 h用电数据构成的向量具有相似性，所以根据独立用户自身的用电行为就可以构造出符合低秩要求的矩阵，由独立用户用电数据构造的原始用电数据矩阵$\boldsymbol{X}$具有良好的低秩特性，从而可采用低秩矩阵完备的方法完成缺失数据的填补。

2　加权非负矩阵分解与完备

非负矩阵分解（nonnegative matrix factorization，NMF）将 1 个非负矩阵近似分解为 2 个低维的非负矩阵因子，寻找具有非负约束的线性模型使对数似然最大化。对于任意给定的一个非负矩阵 $\boldsymbol{X}\in\mathbb{R}_{+}^{mn}$，旨在找到 2 个非负矩阵 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 因子，使得 $\boldsymbol{X}\approx\boldsymbol{UV}$ 成立，选用误差服从高斯分布作为目标函数，则满足公式

$$\min_{U,V}\frac{1}{2}\left\|\boldsymbol{X}-\mathbf{UV}\right\|_{\mathrm{F}}^{2},\ \text{s.t. }\boldsymbol{U}\in\mathbb{R}_{+}^{m\times r},\ \boldsymbol{V}\in\mathbb{R}_{+}^{r\times n},\tag{4}$$

式中，$\left\|\bullet\right\|_{\mathrm{F}}$ 为矩阵的 Frobenius 范数，且满足 $r<\min\{m,n\}$。

若处理在非空子集 Ω 上部分已知的矩阵 $\boldsymbol{X}$ 时，NMF 满足公式

$$\min_{U,V}\frac{1}{2}\left\|P_{\Omega}(\boldsymbol{X})-P_{\Omega}(\mathbf{UV})\right\|_{\mathrm{F}}^{2},\ \text{s.t. }\boldsymbol{U}\in\mathbb{R}_{+}^{m\times r},\ \boldsymbol{V}\in\mathbb{R}_{+}^{r\times n},\tag{5}$$

式中：$P_{\Omega}(\bullet)$ 为矩阵的采样算子。

为求解式（5），可在 NMF 的基础上增加权重信息使其可以处理存在缺失的矩阵，即加权非负矩阵分解（weighted nonnegative matrix factorization，WNMF）。

加权情况下，WNMF 中包含一个二进制权重矩阵 $\boldsymbol{W}$，定义为

$$\boldsymbol{W}(i,j)=\begin{cases}1, & \text{if }(i,j)\in\Omega;\\ 0, & \text{其他。}\end{cases}\tag{6}$$

即非空子集 Ω 中，可观测元素（未缺失）的对应位置 $\boldsymbol{W}(i,j)$ 赋值为 1，不可观测（元素缺失）则置零。

使用权重矩阵将未知部分置零后，将存在缺失的矩阵 $\boldsymbol{X}$ 分解成 2 个非负矩阵 $\boldsymbol{U}$ 和 $\boldsymbol{V}$，式（5）变为

$$\min_{U>0,\,V>0}\frac{1}{2}\left\|\boldsymbol{W}\circ(\boldsymbol{X}-\mathbf{UV})\right\|_{\mathrm{F}}^{2},\tag{7}$$

式中，$\circ$ 为矩阵的 Hadamard 积。

对于一个存在元素缺失的非完备矩阵，矩阵完备（matrix completion，MC）就是通过对其有效位置的元素进行采样，计算处理恢复出缺失的元素。文中使用低秩矩阵完备，对于在非空子集 Ω 上部分已知的低秩矩阵 $\boldsymbol{M}\in\mathbb{R}^{m\times n}$，由另一个低秩矩阵 $\boldsymbol{X}\in\mathbb{R}^{m\times n}$ 对其缺失元素进行填补，得到完备的矩阵。

$$\min_{X}\operatorname{rank}(\boldsymbol{X}),\ \text{s.t. }P_{\Omega}(\boldsymbol{X})=P_{\Omega}(\boldsymbol{M}),\tag{8}$$

式中，$P_{\Omega}(\bullet)$ 为矩阵的采样算子。

若要求矩阵完备前后恒为正，则须满足非负性，在 MC 的基础上进行非负矩阵完备（nonnegative matrix completion，NMC）。

$$\min_{X}\operatorname{rank}(\boldsymbol{X}),\ \text{s.t. }P_{\Omega}(\boldsymbol{X})=P_{\Omega}(\boldsymbol{M}),\ \boldsymbol{M}\in\mathbb{R}_{+}^{m\times n}。\tag{9}$$

为方便计算，可将秩最小的约束公式凸松弛转化为求解核范数的凸优化问题，对矩阵 $\boldsymbol{M}$ 进行非负矩阵完备的求解公式变为

$$\min_{X>0}\left\|\boldsymbol{X}\right\|_{*},\tag{10}$$

式中，$\left\|\bullet\right\|_{*}$ 为矩阵的核范数。

3　块坐标最小算法

坐标下降法（coordinate descent，CD）是一种非梯度的优化方法，即每次沿着单个维度方向进行搜索，得到一个当前维度最小值后再循环使用不同的维度方向，最终收敛得到最优解。块坐标下降法（block coordinate descent method，BCD）是在坐标下降法基础上的改进，可以同时更新多个变量，同时减少迭代次数，能够更好地解决高维问题。

将变量拆分成多个块：

$$f(\boldsymbol{x},\boldsymbol{y}),\ \{x_1,x_2,\cdots,x_N\}\in\boldsymbol{x},\ \{y_1,y_2,\cdots,y_N\}\in\boldsymbol{y}。\tag{11}$$

对式（11）使用块坐标下降法交替优化块：

$$\boldsymbol{x}^{k+1}=\arg\min_{x} f(\boldsymbol{x},\boldsymbol{y}^{k}),$$
$$\boldsymbol{y}^{k+1}=\arg\min_{y} f(\boldsymbol{x}^{k+1},\boldsymbol{y})。\tag{12}$$

在每次优化过程中,选择一个块内的某一个变量作为独立变量,同时将其他变量作为常数,由此目标函数转化为关于该变量的单变量函数。利用单变量函数的最优解作为该变量的最优值并反复迭代直至收敛,选择下一个块内的变量继续上述过程,直到所有块内的变量都求解完毕。

4 用电数据矩阵的完备

4.1 用电数据矩阵模型

对于独立用户每天间隔时间用电量构成的原始用电数据矩阵,该矩阵存在部分元素缺失,并且由观测数据和经验分析可知,用户用电量构成的用电数据矩阵一定是非负的。文中在充分考虑高斯噪声的情况下进行缺失值插补,根据低秩稀疏分解理论,将待修复的低秩原始用电数据矩阵$\boldsymbol{X}$加性分解[15],为

$$\boldsymbol{X}_{\Omega}=(\boldsymbol{U}+\boldsymbol{G})_{\Omega}。\tag{13}$$

式中:矩阵$\boldsymbol{U}$为接近真实用电数据的低秩完整的理想用电数据矩阵;$\boldsymbol{G}$为考虑高斯噪声的高斯分布的噪声矩阵;右下标Ω表示矩阵$\boldsymbol{X}$的某处数据是否缺失。

基于非负矩阵分解的思想,可将秩为r的低秩理想用电数据矩阵$\boldsymbol{U}$分解为2个低维矩阵[16],即$\boldsymbol{U}\approx\boldsymbol{A}_{m\times d}\boldsymbol{B}^{\mathrm{T}}{}_{d\times n}$,其中$r\leqslant d<<\min(m,n)$,且$\boldsymbol{A}\in\mathbb{R}_{+}$,$\boldsymbol{B}^{\mathrm{T}}\in\mathbb{R}_{+}$。则原始用电数据矩阵$\boldsymbol{X}$的分解写为

$$\boldsymbol{X}_{\Omega}=\left(\boldsymbol{A}\boldsymbol{B}^{\mathrm{T}}+\boldsymbol{G}\right)_{\Omega}。\tag{14}$$

基于矩阵范数优化理论,根据高斯噪声密度高、幅值小的特点,可选择凸函数矩阵的F范数对高斯噪声矩阵$\boldsymbol{G}$进行优化求解;根据理想用电数据矩阵$\boldsymbol{U}$的低秩非负性,可选择核范数对其进行优化求解[15]。求解算式为

$$\min_{\boldsymbol{G},\boldsymbol{U}>0}\|\boldsymbol{G}\|_{\mathrm{F}}^{2}+\lambda\|\boldsymbol{U}\|_{*}。\tag{15}$$

采用非负矩阵分解进行低秩矩阵完备,求解算式变为

$$\min_{\boldsymbol{G},\boldsymbol{A},\boldsymbol{B}^{\mathrm{T}}>0}\|\boldsymbol{G}\|_{\mathrm{F}}^{2}+\lambda\|\boldsymbol{A}\boldsymbol{B}^{\mathrm{T}}\|_{*}。\tag{16}$$

最后,矩阵$\boldsymbol{X}$中的缺失元素将用$\boldsymbol{U}$对应位置元素进行填补,实现用电数据的缺失值插补。

4.2 最优化求解

实际处理原始用电数据矩阵$\boldsymbol{X}$时,由于存在数据缺失使得矩阵中部分元素为0,这样的非完备矩阵不可直接使用非负矩阵分解[17]。因此,可以增加权重信息处理缺失数据,使用加权非负矩阵分解同时补全缺失数据。

4.2.1 EM算法求解

将理想用电数据矩阵$\boldsymbol{U}$用2个矩阵$\boldsymbol{A}$和$\boldsymbol{B}^{\mathrm{T}}$相乘来代替并且填补$\boldsymbol{X}$的缺失数据后,矩阵的约束可以通过F范数双线性分解和核范数正则化模型来实现[18],迭代公式如下:

$$\min_{\boldsymbol{A}>0,\boldsymbol{B}^{\mathrm{T}}>0}\frac{1}{2}\left\|\boldsymbol{W}\circ\boldsymbol{X}+(\boldsymbol{1}_{m\times n}-\boldsymbol{W})\circ(\boldsymbol{A}_{t}\boldsymbol{B}_{t}^{\mathrm{T}})-\boldsymbol{A}_{t+1}\boldsymbol{B}_{t+1}^{\mathrm{T}}\right\|_{F}^{2}+\frac{\lambda}{2}\left\|\boldsymbol{A}_{t+1}\boldsymbol{B}_{t+1}^{\mathrm{T}}\right\|_{*},\tag{17}$$

式中,$\boldsymbol{1}_{m\times n}$为所有元素均为1的$m\times n$阶矩阵。

根据核范数的变分定义[19]:

$$\|\boldsymbol{Z}\|_{*}=\|\boldsymbol{A}\boldsymbol{B}^{\mathrm{T}}\|_{*}=\frac{1}{2}\left(\|\boldsymbol{A}\|_{\mathrm{F}}^{2}+\|\boldsymbol{B}\|_{\mathrm{F}}^{2}\right),\tag{18}$$

得到最终计算模型为

$$\min_{\boldsymbol{A}>0,\boldsymbol{B}^{\mathrm{T}}>0}\frac{1}{2}\left\|\boldsymbol{W}\circ\boldsymbol{X}+(1_{m\times n}-\boldsymbol{W})\circ(\boldsymbol{A}_t\boldsymbol{B}_t^{\mathrm{T}})-\boldsymbol{A}_{t+1}\boldsymbol{B}_{t+1}^{\mathrm{T}}\right\|_{\mathrm{F}}^2+\frac{\lambda}{2}\left(\left\|\boldsymbol{A}_{t+1}\right\|_{\mathrm{F}}^2+\left\|\boldsymbol{B}_{t+1}\right\|_{\mathrm{F}}^2\right)。\tag{19}$$

进行数次迭代后,将较为准确的数据填入存在缺失的原始数据矩阵,使该矩阵完备,得到与预期相符合的实验结果。

求解上述目标加权框架模型,可以使用块坐标最小算法,采用EM间接方式[20]。

1)E-step:用完备矩阵$\boldsymbol{Y}$代替非完备矩阵$\boldsymbol{X}$

$$\boldsymbol{Y}=\boldsymbol{W}\circ\boldsymbol{X}+(\boldsymbol{1}_{m\times n}-\boldsymbol{W})\circ(\boldsymbol{A}_t\boldsymbol{B}_t^{\mathrm{T}})。\tag{20}$$

2)M-step:对矩阵$\boldsymbol{Y}$运用非负矩阵分解的方法更新矩阵$\boldsymbol{A}$和$\boldsymbol{B}$。

更新矩阵$\boldsymbol{A}$

$$\boldsymbol{A}_{t+1}=\arg\min_{\boldsymbol{A}_t}f(\boldsymbol{A}_t)=\arg\min_{\boldsymbol{A}_t}\frac{1}{2}\left\|\boldsymbol{Y}-\boldsymbol{A}_t\boldsymbol{B}_t^{\mathrm{T}}\right\|_{\mathrm{F}}^2+\frac{\lambda}{2}\left\|\boldsymbol{A}_t\right\|_{\mathrm{F}}^2,\tag{21}$$

$$\boldsymbol{A}_{t+1}=\max\left(\frac{\boldsymbol{Y}\boldsymbol{B}_t}{\boldsymbol{B}_t^{\mathrm{T}}\boldsymbol{B}_t+\lambda\boldsymbol{I}_{d\times d}},0\right),\tag{22}$$

式中,$\boldsymbol{I}_{d\times d}$为$d\times d$阶单位矩阵。

同理,更新矩阵$\boldsymbol{B}$

$$\boldsymbol{B}_{t+1}=\arg\min_{\boldsymbol{B}_t}f(\boldsymbol{B}_t)=\arg\min_{\boldsymbol{B}_t}\frac{1}{2}\left\|\boldsymbol{Y}-\boldsymbol{A}_{t+1}\boldsymbol{B}_t^{\mathrm{T}}\right\|_{\mathrm{F}}^2+\frac{\lambda}{2}\left\|\boldsymbol{B}_t\right\|_{\mathrm{F}}^2,\tag{23}$$

$$\boldsymbol{B}_{t+1}=\max\left(\frac{\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{A}_{t+1}}{\boldsymbol{A}_{t+1}^{\mathrm{T}}\boldsymbol{A}_{t+1}+\lambda\boldsymbol{I}_{d\times d}},0\right)。\tag{24}$$

4.2.2　直接法求解

同上述非负矩阵分解,用$\boldsymbol{A}$和$\boldsymbol{B}^{\mathrm{T}}$相乘替代理想用电数据矩阵$\boldsymbol{U}$,原始用电数据矩阵$\boldsymbol{X}$的缺失部分仍将其保留,分别选择F范数和核范数对高斯噪声矩阵$\boldsymbol{G}$和具有低秩特性的理想用电数据矩阵$\boldsymbol{U}$进行正则化约束以构建优化模型,再根据核范数的变分定义,得到最终模型为

$$\min\frac{1}{2}\left\|\boldsymbol{W}\circ(\boldsymbol{X}-\boldsymbol{A}\boldsymbol{B}^{\mathrm{T}})\right\|_{\mathrm{F}}^2+\frac{\lambda}{2}\left(\left\|\boldsymbol{A}\right\|_{\mathrm{F}}^2+\left\|\boldsymbol{B}\right\|_{\mathrm{F}}^2\right)。\tag{25}$$

求解上述目标加权框架模型,同样使用块坐标最小算法,并采用直接求解方法交替更新矩阵。首先将$\boldsymbol{A}$和$\boldsymbol{W}$按行展开,即$\boldsymbol{A}^{\mathrm{T}}=[\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_m]$和$\boldsymbol{W}^{\mathrm{T}}=[\hat{\boldsymbol{w}}_1,\hat{\boldsymbol{w}}_2,\cdots,\hat{\boldsymbol{w}}_m]$,同时将$\boldsymbol{X}$按行和列展开分别为$\boldsymbol{X}^{\mathrm{T}}=[\hat{\boldsymbol{x}}_1,\hat{\boldsymbol{x}}_2,\cdots,\hat{\boldsymbol{x}}_m]$和$\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n]$,对应单行向量$\boldsymbol{a}_i$的优化问题可以改写为

$$\boldsymbol{a}_i^{k+1}=\arg\min_{\boldsymbol{a}_i^k}\frac{1}{2}\left\|\mathrm{diag}(\hat{\boldsymbol{w}}_i)\cdot(\hat{\boldsymbol{x}}_i-\boldsymbol{B}_k\boldsymbol{a}_i^k)\right\|_2^2+\frac{\lambda}{2}\left\|\boldsymbol{a}_i^k\right\|_2^2,\tag{26}$$

式中,diag($\boldsymbol{w}_i$)表示由列向量$\hat{\boldsymbol{w}}_i$构成的对角矩阵。

采用类似式(22)方式,经过对式(26)简单求偏导可得矩阵$\boldsymbol{A}$的单行更迭显式公式:

$$\boldsymbol{a}_i^{k+1}=\max\left(\frac{\boldsymbol{B}_k^{\mathrm{T}}\mathrm{diag}(\hat{\boldsymbol{w}}_i)\hat{\boldsymbol{x}}_i}{\boldsymbol{B}_k^{\mathrm{T}}\mathrm{diag}(\hat{\boldsymbol{w}}_i)\boldsymbol{B}_k+\lambda\boldsymbol{I}_{d\times d}},0\right)。\tag{27}$$

依次计算$\boldsymbol{a}_1$~$\boldsymbol{a}_m$,得到整个矩阵$\boldsymbol{A}$的迭代更新。

同理,将$\boldsymbol{B}^{\mathrm{T}}$和$\boldsymbol{W}$按列展开,即$\boldsymbol{B}^{\mathrm{T}}=[\boldsymbol{b}_1,\boldsymbol{b}_2,\cdots,\boldsymbol{b}_n]$和$\boldsymbol{W}=[\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_n]$,对应单列向量$\boldsymbol{b}_j$的优化问题可以改写为

$$\boldsymbol{b}_i^{k+1}=\arg\min_{\boldsymbol{b}_i^k}\frac{1}{2}\left\|\mathrm{diag}(\boldsymbol{w}_i)\cdot(\boldsymbol{x}_i-\boldsymbol{A}_{k+1}\boldsymbol{b}_i^k)\right\|_{\mathrm{F}}^2+\frac{\lambda}{2}\left\|\boldsymbol{b}_i^k\right\|_{\mathrm{F}}^2,\tag{28}$$

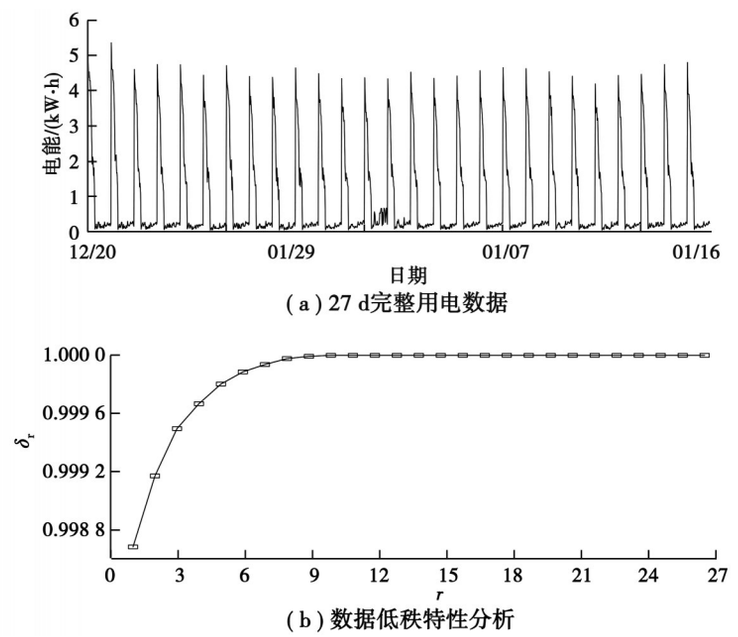
$$\boldsymbol{b}_i^{k+1}=\max\left(\frac{\boldsymbol{A}_{k+1}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{w}_i)\boldsymbol{x}_i}{\boldsymbol{A}_{k+1}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{w}_i)\boldsymbol{A}_{k+1}+\lambda\boldsymbol{I}_{d\times d}},0\right)。\tag{29}$$

依次计算$\boldsymbol{b}_1$~$\boldsymbol{b}_n$,得到整个矩阵$\boldsymbol{B}$的迭代更新。

5 仿真分析与验证

5.1 仿真分析

图4(a)为伦敦某用电台区某用户27 d中每天每个采样间隔时间的完整用电数据,采样间隔时间为30 min,每天采样48次,该数据中存在高斯噪声,但不存在数据缺失。将该用户用电数据以“1天24 h”划分为向量并重排形成矩阵并对该矩阵进行低秩分析,其对应δ_r随r值的变化趋势如图4(b)所示,可知该用电数据矩阵呈现良好的低秩特性。



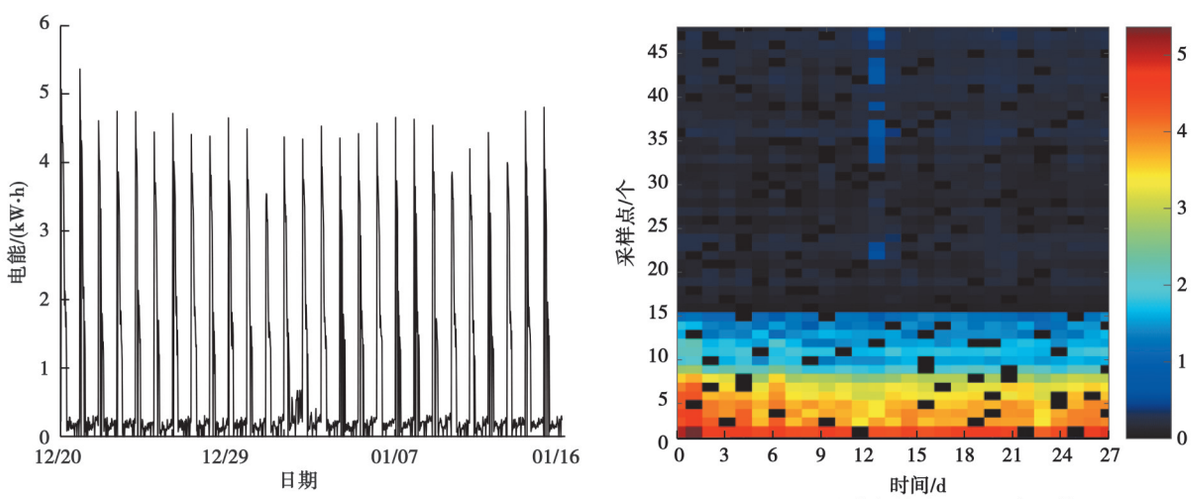
(a) 27 d完整用电数据

(b) 数据低秩特性分析

图4 用户27 d完整用电数据及其低秩分析

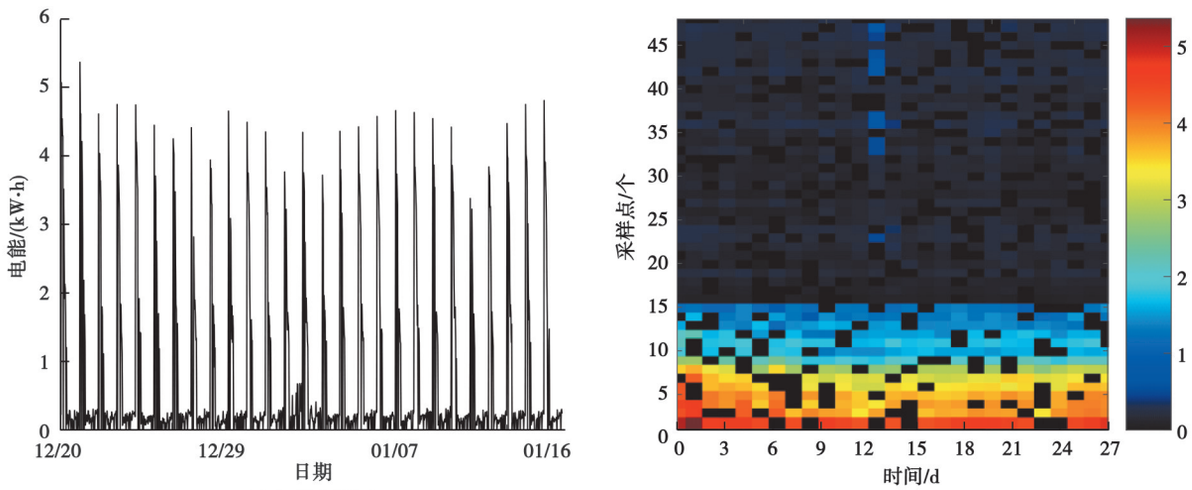
Fig. 4 27 day complete electricity consumption data of a user and its low-rank analysis

首先,随机剔除10%和20%的用电数据使之分别存在10%和20%的缺失,并将其作为待插补的原始用电数据矩阵,如图5所示;然后,充分考虑高斯噪声对用电数据质量的影响使用,加权非负矩阵分解对原始用电数据矩阵进行低秩完备;最后,使用EM算法和直接法迭代求解并比较其收敛性和准确性。



(a) 用电数据缺失10%

(b) 用电数据缺失10%矩阵重排



(c) 用电数据缺失20%　　(d) 用电数据缺失20%矩阵重排

图 5　部分剔除后的非完备用电数据

Fig. 5　Partially excluded incomplete electricity consumption data

根据迭代次数可以判断收敛性，平均范数相对值 η_{err} 可以判断准确性，为

$$\eta_{\mathrm{err}}=\frac{\left\| \boldsymbol{X}-\boldsymbol{A}\boldsymbol{B}^{\mathrm{T}} \right\|_{\mathrm{F}}}{\left\| \boldsymbol{X} \right\|_{\mathrm{F}}}。\tag{30}$$

在此计算模型下，使用 EM 算法和直接法完成用电数据的缺失值插补，得到计算结果的评估指标如表 1 所示。无论是剔除 10% 还是 20% 的原始数据，2 种算法均收敛且均能实现用电数据缺失值插补，EM 算法的平均迭代次数都高于直接法，说明直接法的收敛速度明显优于 EM 算法。2 种算法的平均范数相对值都很小，说明用电数据缺失值插补结果都较准确，但 EM 算法的平均范数相对值略大于直接法，直接法的准确性更高。

表 1　EM 算法和直接法仿真计算结果

Table 1　EM algorithm and direct method simulation results

算法	平均迭代次数		平均 η_{err}	
	剔除数据 10%	剔除数据 20%	剔除数据 10%	剔除数据 20%
EM 算法	445.72	325.05	0.016 109	0.021 924
直接法	307.02	247.76	0.015 091	0.020 891

根据表 1 结果可知，直接法比 EM 算法收敛性和准确性都更好。

5.2　实验验证

通过云南电网公开的某台区用户用电数据（存在缺失）测试，验证文中提出的用电数据缺失值插补方法。选取该台区某一个用户 5 个月的原始用电数据，采样间隔时间为 15 min，每天采样次数为 96 次，可知每天每个采样间隔时间的用电。用文中所提出的基于加权非负矩阵分解的算法框架进行求解，并在计算过程中使用 EM 算法和直接法，结果如图 6 所示。其中，黑色曲线为原始用电数据（电能为 0 的采样点表示此处用电数据缺失），红色和蓝色曲线分别为 EM 算法和直接法求解得到的接近真实用电数据值的理想用电数据。

由图可知，红蓝 2 条曲线均不存在缺失，且均与黑色曲线拟合十分相近，符合原始用电数据与剔除高斯噪声得到的理想用电数据相近的特点，故 2 种算法均有效可行。最后，将原始用电数据中的缺失数据用理想用电数据中对应数据进行填补，得到用电数据缺失值插补的最终结果。

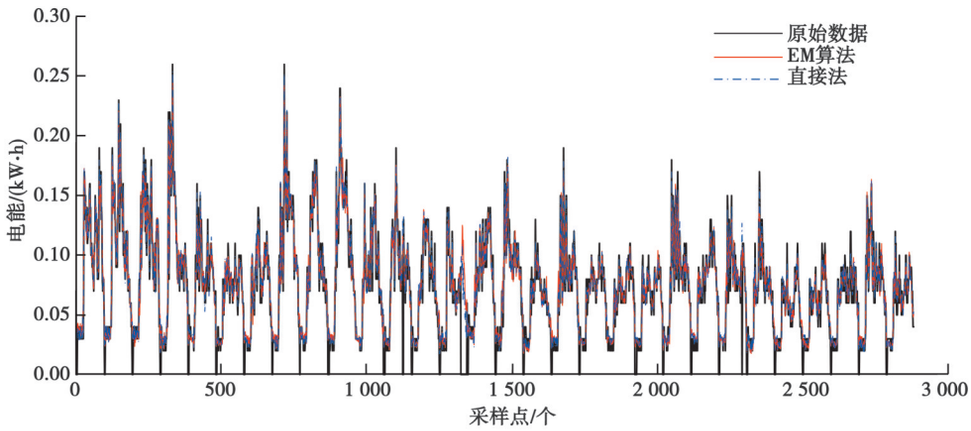


图 6　用电数据缺失值插补结果验证

Fig. 6　Verification of interpolation results for missing values in electricity data

6　结束语

针对用电数据存在采集缺失和背景高斯噪声干扰的复合数据质量问题，提出了一种基于加权非负矩阵分解的低秩矩阵完备的用电数据填充方法。在独立用户自身原始用电数据符合矩阵完备低秩要求的情况下，充分考虑高斯噪声的影响，基于矩阵范数优化理论引入正则化项以构建目标函数模型，同时使用块坐标下降法交替更新，实验结果表明该方法可以实现含有高斯噪声的用电数据随机缺失的填充，且具有良好的收敛性和准确性，其中直接法比EM算法效果更佳。

后续研究中，在高斯噪声的基础上将同时考虑尖峰脉冲噪声和结构稀疏的异常数据，完善算法模型，使其能够更好地适用于用电数据缺失值插补。

参考文献

[1] 林成, 宋伟杰, 廖志戈, 等. 大数据背景下用户侧用电数据在电力系统的应用[J]. 科技创新与应用, 2020, 10(16): 167-168.
Lin C, Song W J, Liao Z G, et al. Application of user-side electricity consumption data in power system under the background of big data[J]. Technology Innovation and Application, 2020, 10(16): 167-168. (in Chinese)

[2] 陈永淑. 大数据技术在电力系统的应用[J]. 信息技术与信息化, 2020(1): 43-45.
Chen Y S. Application of big data technology in power system[J]. Information Technology and Informatization, 2020(1): 43-45. (in Chinese)

[3] 马俊明, 张珍芬. 浅谈电力大数据在电网建设中的运用[J]. 中国新通信, 2020, 22(1): 99.
Ma J M, Zhang Z F. Discussion on the application of power big data in power grid construction[J]. China New Telecommunications, 2020, 22(1): 99. (in Chinese)

[4] 彭小圣, 邓迪元, 程时杰, 等. 面向智能电网应用的电力大数据关键技术[J]. 中国电机工程学报, 2015, 35(3): 503-511.
Peng X S, Deng D Y, Cheng S J, et al. Key technologies of electric power big data and its application prospects in smart grid[J]. Proceedings of the CSEE, 2015, 35(3): 503-511. (in Chinese)

[5] 王喜宾, 文俊浩, 廖臣, 等. 智能电网需求侧个性化推荐系统[J]. 重庆大学学报, 2022, 45(1): 38-49.
Wang X B, Wen J H, Liao C, et al. Personalized recommendation system in the demand side of smart grid[J]. Journal of Chongqing University, 2022, 45(1): 38-49. (in Chinese)

[6] 邓东林, 徐尤, 陈剑, 等. 智能用电数据的采集与预处理[J]. 电力大数据, 2019, 22(3): 81-86.
Deng D L, Xu Y, Chen J, et al. Acquisition and preprocessing of smart electric appliance network power data[J]. Power Systems and Big Data, 2019, 22(3): 81-86. (in Chinese)

[7] 郑旭东. 用电信息采集系统电能计量数据异常的原因浅析[J]. 科技创新导报, 2019, 16(23):81-81, 83.
Zheng X D. Analysis on the causes of abnormal electric energy metering data in electricity consumption information acquisition system[J]. Science and Technology Innovation Herald, 2019, 16(23):81-81, 83. (in Chinese)

［8］Papageorgiou G, Bouboulis P, Theodoridis S. Robust linear regression analysis: a greedy approach[J]. IEEE Transactions on Signal Processing, 2015, 63(15): 3872-3887.

［9］Papageorgiou G, Bouboulis P, Theodoridis S. Robust non-linear regression analysis: a greedy approach employing kernels[J]. Journal of Machine Learning Research, 2015, 1: 1-48.

［10］Mateos G, Giannakis G B. Robust nonparametric regression via sparsity control with application to load curve data cleansing[J]. IEEE Transactions on Signal Processing, 2012, 60(4): 1571-1584.

［11］Suo Q L, Yao L Y, Xun G X, et al. Recurrent imputation for multivariate time series with missing values[C]//2019 IEEE International Conference on Healthcare Informatics (ICHI). June 10-13, 2019. Xi'an, China. IEEE, 2019: 1-3.

［12］Liu M, Liu D P, Sun G Y, et al. Deep learning detection of inaccurate smart electricity meters: a case study[J]. IEEE Industrial Electronics Magazine, 2020, 14(4): 79-90.

［13］Meira de Andrade P H, Villanueva J M M, de Macedo Braz H D. An outliers processing module based on artificial intelligence for substations metering system[J]. IEEE Transactions on Power Systems, 2020, 35(5): 3400-3409.

［14］袁忠军, 陈刚 . 基于结构自适应神经网络用电量时间特征的聚类分析[J]. 重庆大学学报(自然科学版), 2007, 30(8): 44-48.
Yuan Z J, Chen G. Clustering analysis for time feature of user power consumption based on structural self-adaptation ANN[J]. Journal of Chongqing University (Natural Science Edition), 2007, 30(8): 44-48. (in Chinese)

［15］杨挺, 孙兆帅, 季浩, 等 . 基于矩阵范数优化理论的用电数据质量提升算法[J]. 中国电机工程学报, 2022,42(10):3501-3512.
Yang T, Sun Z S, Ji H, et al. Electricity consumption data quality improvement algorithm based on matrix norm optimization theory[J]. Proceedings of the CSEE, 2022, 42(10): 3501-3512. (in Chinese)

［16］Guan N Y, Tao D C, Luo Z G, et al. NeNMF: an optimal gradient method for nonnegative matrix factorization[J]. IEEE Transactions on Signal Processing, 2012, 60(6): 2882-2898.

［17］Dorffer C, Puigt M, Delmaire G, et al. Fast nonnegative matrix factorization and completion using nesterov iterations[M]// Latent Variable Analysis and Signal Separation. Cham: Springer International Publishing, 2017: 26-35.

［18］Giampouras P V, Rontogiannis A A, Koutroumbas K D. Alternating iteratively reweighted least squares minimization for low-rank matrix factorization[J]. IEEE Transactions on Signal Processing, 2019, 67(2): 490-503.

［19］Cabral R, De la Torre F, Costeira J P, et al. Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition[C]//2013 IEEE International Conference on Computer Vision. December 1-8, 2013, Sydney, NSW, Australia. IEEE, 2013: 2488-2495.

［20］Hastie T, Mazumder R, Lee J D, et al. Matrix completion and low-rank SVD via fast alternating least squares[J]. Journal of Machine Learning Research: JMLR, 2015, 16: 3367-3402.

（编辑　詹燕平）