多变量时间序列中基于克罗内克压缩感知的缺失数据预测算法

郭 艳 宋晓祥* 李 宁 钱 鹏

(陆军工程大学通信工程学院 南京 210007)

摘 要：针对现有算法在预测多变量时间序列中的缺失数据时不适用或只适用于缺失数据较少的情况，该文提出一种基于克罗内克压缩感知的缺失数据预测算法。首先，利用多变量时间序列的时域平滑特性和序列之间的潜在相关性从时空两个方面设计了稀疏表示基，从而将缺失数据预测问题建模成稀疏向量恢复问题。模型求解部分，根据缺失数据的位置特点设计了适合当前应用场景且与稀疏表示基相关性低的观测矩阵。接着，从稀疏表示向量是否足够稀疏和感知矩阵是否满足有限等距特性两个方面验证了模型的性能。最后，仿真结果表明，所提算法在数据缺失严重的情况下具有良好的性能。

关键词：多变量时间序列；缺失数据；克罗内克压缩感知；时域平滑特性；潜在相关性

中图分类号：TN911.7 **文献标识码**：A **文章编号**：1009-5896(2019)04-0858-07

DOI: 10.11999/JEIT180541

Missing Data Prediction Based on Kronecker Compressing Sensing in Multivariable Time Series

GUO Yan SONG Xiaoxiang LI Ning QIAN Peng

(*Institute of Communications Engineering, Army Engineering University, Nanjing* 210007, *China*)

Abstract: In view of the problem that the existing methods are not applicable or are only feasible to the case where only a low ratio of data are missing in multivariable time series, a missing data prediction algorithm is proposed based on Kronecker Compressed Sensing (KCS) theory. Firstly, the sparse representation basis is designed to largely utilize both the temporal smoothness characteristic of time series and potential correlation between multiple time series. In this way, the missing data prediction problem is modeled into the problem of sparse vector recovery. In the solution part of the model, according to the location of missing data, the measurement matrix is designed suitable for the current application scenario and low correlation with the sparse representation basis. Then, the validity of the model is verified from two aspects: Whether the sparse representation vector is sufficiently sparse and the sensing matrix satisfies the restricted isometry property. Simulation results show that the proposed algorithm has good performance in the case where a high ratio of data are missing.

Key words: Multivariable time series; Missing data; Kronecker Compressing Sensing (KCS); Temporal smoothness characteristic; Potential correlation

1 引言

现实生活中，我们往往需要布置多个传感器去测量同一目标不同特性[1]。例如，在环境监测系统中，需要布置多个传感器来检测不同气体的气体浓度；在个人医疗卫生体系，需要多个传感器去测得心脏、血压和睡眠质量等不同指标。本文将这些针对同一目标不同特性的多个时间序列称为多变量时间序列。多变量时间序列的普遍存在给以数据驱动的产业创造了巨大价值。然而，由于恶劣的工作条件或一些无法控制的因素，往往使得采集到的原始时间序列中存在缺失数据。缺失数据会在一定程度上影响原始数据的质量甚至导致建立错误的数据挖掘模型[2]。如何有效地预测缺失数据，从而挖掘其潜在价值成为亟待解决的问题。

目前的缺失数据预测方法多基于数据挖掘和统计方法。基于插值的方法是最简单的一类。其中，指数平滑和样条插值是数据插值的主要技术[3,4]。

收稿日期：2018-06-01；改回日期：2018-10-29；网络出版：2018-11-19
*通信作者： 宋晓祥 guoyan_1029@sina.com
基金项目：国家自然科学基金(61571463, 61371124, 61472445)；江苏省自然科学基金(BK20171401)
Foundation Items: The National Natural Science Foundation of China (61571463, 61371124, 61472445), The Jiangsu Province Natural Science Foundation (BK20171401)

基于模型的方法旨在从收集到的数据发现某种潜在的规律，从而预测缺失的数据[5]。文献[6]提出了旨在预测具有潜在季节性数据的季节性自回归滑动平均模型(SARIMA)，因此，如果数据没有严格的内部季节性，SARIMA将很难对数据进行建模。文献[7]探讨了基于递归神经网络(RNN)处理缺失数据的策略。然而，RNN需要一个大的训练数据集。故当有大量数据缺失时，RNN很难发现数据的潜在规律。基于统计学习的方法试图利用数据的统计特征去确定一个特殊的概率分布。然后，将最适合假定概率分布的值视为缺失的数据[8]。文献[9]使用内核概率主成分分析法(KPPCA)来预测交通流的缺失数据，性能良好。文献[10]利用基于时域动态矩阵分解(TDMF)的方法来预测多变量时间序列中的缺失数据。然而，TDMF的性能很容易受到参数的影响，且对于理想参数的设定至今没有科学的理论指导。

针对上述问题，本文提出了一种基于克罗内克压缩感知(Kronecker Compressed Sensing, KCS)的缺失数据预测算法。首先，分别利用时间序列的时域平滑特性和多变量时间序列中序列之间的潜在相关性设计了稀疏表示基，从而将缺失数据预测问题建模成稀疏向量恢复问题。其次，在模型求解部分，根据缺失数据的位置特点设计了适合于本文的应用场景且与稀疏表示基相关性低的观测矩阵。接着，本文从稀疏表示向量是否足够稀疏和感知矩阵是否满足有限等距特性[11]两个方面分析和验证了模型的性能。仿真结果表明，本文提出的算法可以有效地预测出多变量时间序列中的缺失数据。

2　KCS理论

压缩感知理论中，如果信号$\boldsymbol{s}\in R^N$稀疏或能在某个稀疏表示基$\boldsymbol{\varphi}\in R^{N\times N}$下稀疏表示，即$\boldsymbol{s}=\boldsymbol{\varphi x}$，就能够按照观测矩阵$\boldsymbol{\psi}\in \boldsymbol{R}^{M\times N}$，以低于奈奎斯特定律的速率对其采样，并通过观测值$\boldsymbol{y}=\boldsymbol{\psi\varphi x}=\boldsymbol{Ax}$以高概率恢复出原始信号[11]。为了将传统的压缩感知理论应用到高维的场景，文献[12]提出了KCS理论，从而提供了多维数据问题中感知矩阵的设计框架。在KCS理论中，假设J维数据为$\boldsymbol{s}\in R^{N_1\times N_2\times\cdots\times N_J}$。我们定义$\boldsymbol{s}$的第1维为$\boldsymbol{s}$的第1部分，如式(1)所示：

$$\boldsymbol{s}(:,n_2,\cdots,n_J)=[\boldsymbol{s}(1,n_2,\cdots,n_J),\cdots,\boldsymbol{s}(N_1,n_2,\cdots,n_J)] \tag{1}$$

式中，$n_i=1,2,\cdots,N_i$，$i=2,3,\cdots,J$。这个定义对$\boldsymbol{s}$的第i部分同样适用。KCS理论中，如果数据$\boldsymbol{s}$有J部分，就有J个稀疏表示基，假设它们分别为$\boldsymbol{\varphi}_1,\boldsymbol{\varphi}_2,\cdots,\boldsymbol{\varphi}_J$，那么$\boldsymbol{s}$的稀疏表示基就可以表示为$\boldsymbol{\varphi}_{\mathrm{KCS}}=\boldsymbol{\varphi}_1\otimes\boldsymbol{\varphi}_2\otimes\cdots\otimes\boldsymbol{\varphi}_J$。一旦数据$\boldsymbol{s}$被稀疏表示，就可以通过观测矩阵对其进行欠采样，如式(2)所示：

$$\boldsymbol{y}_{\mathrm{KCS}}=\boldsymbol{\psi}_{\mathrm{KCS}}\boldsymbol{\varphi}_{\mathrm{KCS}}\boldsymbol{x}_{\mathrm{KCS}} \tag{2}$$

式中，$\boldsymbol{y}_{\mathrm{KCS}}$是$\boldsymbol{s}_{\mathrm{KCS}}$的采样值，$\boldsymbol{\varphi}_{\mathrm{KCS}}\boldsymbol{x}_{\mathrm{KCS}}$是$\boldsymbol{s}$的向量展开式。$\boldsymbol{\psi}_{\mathrm{KCS}}$为

$$\boldsymbol{\psi}_{\mathrm{KCS}}=\begin{bmatrix}\boldsymbol{\psi}_1 & 0 & \cdots & 0\\ 0 & \boldsymbol{\psi}_2 & \cdots & \vdots\\ \vdots & \vdots & \ddots & 0\\ 0 & \cdots & 0 & \boldsymbol{\psi}_J\end{bmatrix} \tag{3}$$

已知$\boldsymbol{y}_{\mathrm{KCS}}$，$\boldsymbol{\psi}_{\mathrm{KCS}}$和$\boldsymbol{\varphi}_{\mathrm{KCS}}$，就可以通过求解一个凸优化的问题求出$\boldsymbol{x}_{\mathrm{KCS}}$。

3　模型建立

表1是一个多变量时间序列的例子。

表 1　多变量时间序列

数据源	t_1	t_2	t_3	t_4	…	t_K
$\boldsymbol{s}_1$	0.2	?	?	0.3		1.9
$\boldsymbol{s}_2$	?	0.4	0.5	?		?
$\boldsymbol{s}_3$	?	0.7	?	0.6		?
$\boldsymbol{s}_4$	0.8	?	?	?		2.1
⋮	⋮	⋮	⋮	⋮		⋮
$\boldsymbol{s}_N$	?	0.1	1.2	1.3		?

$\boldsymbol{S}'=\{\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_N\}$，$\boldsymbol{S}'\in R^{N\times K}$ $(N<K)$ 表示带有缺失数据的多变量时间序列矩阵。$\boldsymbol{s}_i\in R^K$表示从第i个数据源收集到的数据。t_j表示第j个采样时刻。s_{ij}表示第j个采样时刻在第i个数据源处的采样值。缺失的数据用“？”表示。此外，本文还定义了一个矩阵$\boldsymbol{W}\in R^{N\times K}$去表示$\boldsymbol{S}$中的数据是否缺失。

$$W_{ij}=\begin{cases}0, & s_{ij}\text{缺失}\\ 1, & \text{其它}\end{cases} \tag{4}$$

假设完整的多变量时间序列矩阵为$\boldsymbol{S}$，那么$\boldsymbol{S}'$可以看成是$\boldsymbol{S}$和$\boldsymbol{W}$的Hadamard积的运算结果，如式(5)：

$$\boldsymbol{S}'=\boldsymbol{S}\odot\boldsymbol{W} \tag{5}$$

下面将阐述如何设计相应的稀疏表示基从而对多变量时间序列中的缺失数据预测问题进行稀疏表示。

3.1　时域稀疏表示基

大多数时间序列信号都具有天然的时域平滑

性，即信号的值只有在少数时刻发生较大变化。因此，多变量时间序列矩阵$\boldsymbol{S}$的第i行$\boldsymbol{s}_i$的两个相邻采样值之差中应该只有少量值较大，而其他大部分可以忽略。为此，本文设计如式(6)，式(7)所示矩阵：

$$\boldsymbol{\Omega}_1=\begin{bmatrix}1&-1&0&\cdots\\0&1&-1&\cdots\\0&0&1&\cdots\\\vdots&\vdots&\vdots&\ddots\end{bmatrix} \tag{6}$$

$$\boldsymbol{\Omega}_2=\begin{bmatrix}2&-1&0&0&\cdots\\-1&2&-1&0&\cdots\\0&-1&2&-1&\cdots\\\vdots&\vdots&\vdots&\vdots&\ddots\end{bmatrix} \tag{7}$$

则$\boldsymbol{s}_i$在矩阵$\boldsymbol{\Omega}_1$下的投影向量为

$$\begin{aligned}\boldsymbol{\gamma}_{i1}&=\begin{bmatrix}1&-1&0&\cdots\\0&1&-1&\cdots\\0&0&1&\cdots\\\vdots&\vdots&\vdots&\ddots\end{bmatrix}\begin{bmatrix}s_{i1}\\s_{i2}\\\vdots\\s_{iK}\end{bmatrix}\\&=\begin{bmatrix}s_{i1}-s_{i2}\\s_{i2}-s_{i3}\\\vdots\\s_{iK}-s_{i1}\end{bmatrix}\end{aligned} \tag{8}$$

则$\boldsymbol{s}_i$在矩阵$\boldsymbol{\Omega}_2$下的投影向量为

$$\begin{aligned}\boldsymbol{\gamma}_{i2}&=\begin{bmatrix}2&-1&0&0&\cdots\\-1&2&-1&0&\cdots\\0&-1&2&-1&\cdots\\\vdots&\vdots&\vdots&\vdots&\ddots\end{bmatrix}\begin{bmatrix}s_{i1}\\s_{i2}\\\vdots\\s_{iK}\end{bmatrix}\\&=\begin{bmatrix}2s_{i1}-s_{i2}\\2s_{i2}-s_{i1}-s_{i3}\\\vdots\\2s_{iK}-s_{(K-1)i}\end{bmatrix}\end{aligned} \tag{9}$$

令$\boldsymbol{\Lambda}_1=\boldsymbol{\Omega}_1^{-1},\boldsymbol{\Lambda}_2=\boldsymbol{\Omega}_2^{-1}$。统一称$\boldsymbol{\Lambda}_1$和$\boldsymbol{\Lambda}_2$为$\boldsymbol{\Lambda}$，$\boldsymbol{\gamma}_{i1}$和$\boldsymbol{\gamma}_{i2}$为$\boldsymbol{\gamma}_i$。那么时间序列$\boldsymbol{s}_i$就可以稀疏表示为$\boldsymbol{s}_i=\boldsymbol{\Lambda}\boldsymbol{\gamma}_i$。

3.2 空间稀疏表示基

由于多变量时间序列中序列之间必然存在潜在的相关性且影响一个目标的主要成分不会因为缺失数据的存在发生变化[3,11,13]。基于此，本文利用主成分分析法从空间上设计$\boldsymbol{S}$的稀疏表示基。假设$\boldsymbol{S}$的转置的奇异值分解如式(10)所示：

$$\boldsymbol{S}^{\mathrm{T}}=\boldsymbol{U}\boldsymbol{\Sigma}_N\boldsymbol{V}^{\mathrm{T}} \tag{10}$$

式中，$\boldsymbol{V}\in R^{N\times N}$是一个正交矩阵。$\boldsymbol{\Sigma}_N\in R^{N\times N}$是一个对角矩阵，其对角线元素为$\boldsymbol{S}$的奇异值。$\boldsymbol{U}\in R^{K\times N}$则展示$\boldsymbol{S}$的动态特性。应用主成分分析法，可以求得矩阵$\boldsymbol{S}$的一个低秩逼近，如式(11)：

$$\boldsymbol{S}=\boldsymbol{V}\boldsymbol{\Sigma}_L\boldsymbol{U}_{\mathrm{PC}}^{\mathrm{T}} \tag{11}$$

式中，$\boldsymbol{\Sigma}_L$是由$\boldsymbol{\Sigma}_N$中前L个最大的奇异值组成的对角矩阵。这样，就可以将矩阵$\boldsymbol{S}$近似地稀疏表示为$\boldsymbol{S}\approx\boldsymbol{V}\boldsymbol{\theta}\left(\boldsymbol{\theta}=\boldsymbol{\Sigma}_L\boldsymbol{U}_{\mathrm{PC}}^{\mathrm{T}}\right)$。其中$\boldsymbol{V}$和$\boldsymbol{\theta}$分别是正交基和相应的系数。由于$\boldsymbol{S}$的主成分不会因为缺失数据的存在而发生变化，所以$\boldsymbol{S}'$可以稀疏表示为$\boldsymbol{S}'=\boldsymbol{V}\boldsymbol{\theta}_1,\boldsymbol{\theta}_1$是矩阵$\boldsymbol{S}'$对应的系数。

由上所述，分别从时空两个角度考虑设计了稀疏表示基$\boldsymbol{\Lambda}$和$\boldsymbol{V}$。将$\boldsymbol{\Lambda}$和$\boldsymbol{V}$进行Kronecker积运算就可以得到多变量时间序列的稀疏表示基，如式(12)：

$$\boldsymbol{\varphi}_{\mathrm{KCS}}=\boldsymbol{\Lambda}\otimes\boldsymbol{V} \tag{12}$$

令$\boldsymbol{S}_{\mathrm{vec}}=\mathrm{vec}(\boldsymbol{S})$，通过设计时域稀疏表示基和空间稀疏表示基就可以得到

$$\boldsymbol{S}_{\mathrm{vec}}=\boldsymbol{\varphi}_{\mathrm{KCS}}\boldsymbol{\alpha}_{\mathrm{KCS}} \tag{13}$$

只要求解得到$\boldsymbol{\alpha}_{\mathrm{KCS}}$，就可以通过式(13)得到完整的多变量时间序列。

4 稀疏表示向量求解模型

由上知，矩阵$\boldsymbol{W}$的第i列表示时间序列$\boldsymbol{s}_i$中的数据是否缺失，而缺失数据预测就是利用那些未缺失的数据作为测量值来恢复原时间序列。基于此，对$\boldsymbol{s}_i$设计如式(14)所示的观测矩阵：

$$\boldsymbol{\psi}_i=\begin{bmatrix}1&0&0&0&\cdots\\0&0&1&0&\cdots\\0&0&0&1&\cdots\\\vdots&\vdots&\vdots&\vdots&\ddots\end{bmatrix} \tag{14}$$

如果矩阵$\boldsymbol{\psi}_i\in R^{M_i\times K}$在$(m,n)$处的值为1，表明第$m$个测量值是在第$n$个采样时刻得到的。这样，就可以得到多变量时间序列矩阵的观测矩阵为

$$\boldsymbol{\psi}_{\mathrm{KCS}}=\begin{bmatrix}\boldsymbol{\psi}_1&0&\cdots&0\\0&\boldsymbol{\psi}_2&\cdots&\vdots\\\vdots&\vdots&\ddots&0\\0&\cdots&0&\boldsymbol{\psi}_N\end{bmatrix} \tag{15}$$

令$\boldsymbol{y}_i\in R^{M_i}$表示时间序列$\boldsymbol{s}_i$的测量值，$\boldsymbol{y}=(\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_N)$，$\boldsymbol{y}_{\mathrm{KCS}}=\mathrm{vec}(\boldsymbol{y})$。则多变量时间序列矩阵$\boldsymbol{S}$的观测向量就可以表示为

$$\boldsymbol{y}_{\mathrm{KCS}}=\boldsymbol{\psi}_{\mathrm{KCS}}\boldsymbol{\varphi}_{\mathrm{KCS}}\boldsymbol{\alpha}_{\mathrm{KCS}} \tag{16}$$

已知$\boldsymbol{y}_{\mathrm{KCS}}$，$\boldsymbol{\psi}_{\mathrm{KCS}}$和$\boldsymbol{\varphi}_{\mathrm{KCS}}$，就可以通过现有的稀疏表示向量求解算法求解一个凸优化问题求解$\boldsymbol{\alpha}_{\mathrm{KCS}}$。已有的解决稀疏向量恢复问题的算法有很多，主要包括基追踪(Basis Pursuit, BP)[14]、正交匹配追踪

(Orthogonal Matching Pursuit, OMP)[15]和加权最小二乘算法(Iteratively ReWeighted, IRW)[16]以及基于高斯广义近似消息传递的稀疏贝叶斯学习算法(GAMP-SBL)[17]。其中，相比于其他恢复算法，基于稀疏贝叶斯学习的算法在计算精度上展现了很好的性能[17]。文献[17]提出的基于GAMP的低复杂度稀疏贝叶斯算法，在保证以往的基于稀疏贝叶斯学习算法计算精度的同时大大降低了计算复杂度。因此，本文后续部分将选择GAMP-SBL作为恢复算法。

5　模型性能分析

要精确地求解稀疏向量恢复问题必须满足两个条件：(1)时间序列在稀疏表示基下的稀疏表示向量足够稀疏；(2)感知矩阵满足有限等距特性(Restricted Isometry Property, RIP)。下面将从这两个方面对所提模型的性能进行分析。

5.1　稀疏表示基的稀疏信号能力

利用KCS理论精确地预测出缺失数据，必须保证$\boldsymbol{\alpha}_{\mathrm{KCS}}$是可压缩的。压缩感知理论表明只要$\boldsymbol{\alpha}_{\mathrm{KCS}}$中只包含一些非常大的元素，而其他的元素可以被忽略，就认为其是可压缩的。令$\boldsymbol{x}_i$表示时间序列$\boldsymbol{s}_i$的稀疏表示向量，则$\boldsymbol{x}_i$对应$\boldsymbol{\alpha}_{\mathrm{KCS}}$中的第$(i-1)K+1$个到第$iK$个元素。本文中，以$\sum_{j=1}^{n}\left(x_i^j\right)^2\Big/\sum_{j=1}^{K}\left(x_i^j\right)^2$为标准衡量稀疏表示基的稀疏信号能力，其中$x_i^j$表示稀疏向量$\boldsymbol{x}_i$中第$j$大的元素；$\sum_{j=1}^{n}\left(x_i^j\right)^2$表示$\boldsymbol{x}_i$中前$n$个最大元素的能量；$\sum_{j=1}^{K}\left(x_i^j\right)^2$表示向量$\boldsymbol{x}_i$的总能量。对于给定的$n$，$\sum_{j=1}^{n}\left(x_i^j\right)^2\Big/\sum_{j=1}^{K}\left(x_i^j\right)^2$越大，稀疏表示基的稀疏信号能力越强，稀疏表示向量的可压缩性越好。下面用3个真实数据集中的数据来检验稀疏表示基$\boldsymbol{\varphi}_{\mathrm{KCS1}}$和$\boldsymbol{\varphi}_{\mathrm{KCS2}}$的稀疏信号能力。MOTES数据集是一个中等数据集，包含了在英特尔伯克利研究实验室部署的54个传感器历时一个月采集而来的时间序列数据[18]；GSA数据集是一个大数据集，数据集的气体传感器阵列是在加州大学圣地亚哥分校生物电路研究所化学信号实验室的一个气体传递平台上收集的[19]；SST数据集是一个中等的数据集，数据集中的数据是由热带大气海洋工程每小时的温度测量组成[20]。

表2展示了当稀疏表示矩阵分别为$\boldsymbol{\varphi}_{\mathrm{KCS1}}$和$\boldsymbol{\varphi}_{\mathrm{KCS2}}$时，$n$等于100, 200, 300, 500的$\sum_{j=1}^{n}\left(x_i^j\right)^2\Big/\sum_{j=1}^{K}\left(x_i^j\right)^2$的大小。从表中可以看出，当$n$=100，选用$\boldsymbol{\varphi}_{\mathrm{KCS1}}$作为稀疏表示基时，$\sum_{j=1}^{n}\left(x_i^j\right)^2\Big/\sum_{j=1}^{K}\left(x_i^j\right)^2$的值介于0.8759到0.9897之间，选用$\boldsymbol{\varphi}_{\mathrm{KCS2}}$作为稀疏表示基，$\sum_{j=1}^{n}\left(x_i^j\right)^2\Big/\sum_{j=1}^{K}\left(x_i^j\right)^2$的值介于0.5227到0.7428之间。由此可见，稀疏表示基$\boldsymbol{\varphi}_{\mathrm{KCS1}}$的稀疏信号能力要强于$\boldsymbol{\varphi}_{\mathrm{KCS2}}$。

表 2　稀疏表示基稀疏性能比较

n	100	200	300	500
MOTES_1	0.9897	0.9952	0.9973	0.9992
MOTES_2	0.7428	0.8731	0.9301	0.9786
GSA_1	0.9687	0.9995	0.9998	0.9999
GSA_2	0.5227	0.6841	0.9018	0.9721
SST_1	0.8759	0.9429	0.9722	0.9935
SST_2	0.7390	0.8785	0.9393	0.9849

5.2　稀疏表示基和观测矩阵的低相关性

要确定一个感知矩阵是否满足RIP性是非常困难的，文献[21]中得出可以通过计算稀疏表示基和观测矩阵之间非相关性的大小来间接反映RIP性，非相关性越好，RIP性越好。给定$(\boldsymbol{\varphi},\boldsymbol{\psi})$，它们之间非相关性的大小可以计算为

$$I(\boldsymbol{\varphi},\boldsymbol{\psi})=\min\|\boldsymbol{\theta}_i\|_0,1\le i\le M \tag{17}$$

其中，$\boldsymbol{\theta}_i$表示矩阵$\boldsymbol{\psi}$的第i个行向量在由稀疏表示基$\boldsymbol{\varphi}$各列张成的空间中的投影向量即

$$\boldsymbol{\theta}_i=(\boldsymbol{\varphi}^{\mathrm{T}}\boldsymbol{\varphi})^{-1}\boldsymbol{\varphi}^{\mathrm{T}}\boldsymbol{\psi}_i^r \tag{18}$$

其中，$\boldsymbol{\psi}_i^r$表示观测矩阵$\boldsymbol{\psi}$的第i个行向量。$I(\boldsymbol{\varphi},\boldsymbol{\psi})$越大，$\boldsymbol{\varphi}$和$\boldsymbol{\psi}$的非相关性越好。

数据可能随机缺失(randomly)、均匀缺失(evenly)和连续缺失(continuously)，因此就对应有3种不同的观测矩阵$\boldsymbol{\psi}_R$，$\boldsymbol{\psi}_E$和$\boldsymbol{\psi}_C$。表3中，我们计算了当稀疏表示基分别为$\boldsymbol{\varphi}_{\mathrm{KCS1}}$和$\boldsymbol{\varphi}_{\mathrm{KCS2}}$时，不同的观测矩阵和稀疏表示基之间非相关性的大小。从表3中可以看出无论数据以哪种缺失类型缺失，稀疏表示基和观测矩阵之间的非相关性都很好，即感知矩阵的RIP性很好。另外，可以发现当观测矩阵相同时，稀疏表示基分别为$\boldsymbol{\varphi}_{\mathrm{KCS1}}$和$\boldsymbol{\varphi}_{\mathrm{KCS2}}$产生的非相关性大小十分接近。然而，从表2中得出，稀疏表示基$\boldsymbol{\varphi}_{\mathrm{KCS1}}$的稀疏信号能力要强于$\boldsymbol{\varphi}_{\mathrm{KCS2}}$。因

表 3　稀疏表示基和观测矩阵非相关性的大小

NK	2000	3000	4000	6000	8000
$I(\boldsymbol{\psi}_R,\boldsymbol{\varphi}_{\mathrm{KCS1}})$	1998	2995	3997	5996	7993
$I(\boldsymbol{\psi}_R,\boldsymbol{\varphi}_{\mathrm{KCS2}})$	1998	2999	4000	5999	8000
$I(\boldsymbol{\psi}_E,\boldsymbol{\varphi}_{\mathrm{KCS1}})$	1999	2996	3997	5999	7995
$I(\boldsymbol{\psi}_E,\boldsymbol{\varphi}_{\mathrm{KCS2}})$	1999	2999	3999	6000	8000
$I(\boldsymbol{\psi}_C,\boldsymbol{\varphi}_{\mathrm{KCS1}})$	1997	2996	3997	5995	7995
$I(\boldsymbol{\psi}_C,\boldsymbol{\varphi}_{\mathrm{KCS2}})$	1998	2999	3998	5999	7998

此，接下的仿真实验中如无特殊说明，我们将选择φ_{KCS1}作为稀疏表示基。

6 仿真与分析

本节将使用上文所述的数据集进行仿真。根据不同的数据缺失率，从完整的数据集中随机删除一些数据来模拟缺失的数据。缺失率定义为缺失数据的数量与数据总量之比。本文采用均方根误差(RMSE)和平均运行时间(ART)作为性能评价标准。定义为

$$\mathrm{RMSE}=1\Big/\sum_{ij}(1-W_{ij})\cdot\left(\sum_{ij}(1-W_{ij})(s_{ij}-\hat{s}_{ij})^2\right)^{1/2} \tag{19}$$

其中，s_{ij}表示实际值，$\hat{s}_{ij}$表示相应的预测值，W_{ij}用来表示s_{ij}是否缺失。

为了分析不同方法的计算复杂度，每种方法重复执行100次计算其以秒为单位的平均运行时间(ART)。

$$\mathrm{ART}=T/100 \tag{20}$$

6.1 所提算法与其它算法性能比较

为了对算法的性能进行有效评估，本节将比较所提算法(KCS-GAMP-SBL)和其他4个多变量时间序列缺失数据预测算法的性能。其中包括：(1)基于插值的方法：SI[4]; (2)基于模型的方法：RNN[7]; (3)基于统计学习的方法：KPPCA[9]和TDMF[10]。

表4以均方根误差为标准分别比较了各种方法在MOTES, GSA和SST数据集上的仿真结果。从表中可看出，相比于其他算法，不管在哪个数据集上，KCS-GAMP-SBL算法都具有更小的均方根误差。这是因为本文的算法在设计过程中充分利用了时间序列的时域平滑特性和多个时间序列之间的相关信息，算法是针对多变量时间序列的本质特性进行建模而不依赖于极少的观测数据。仿真结果表明，当缺失数据较多时，所提方法在解决多变量时间序列中的缺失数据问题上是适用并且非常有效的。

表5中，在缺失率为80%的条件下，以ART为标准比较了各个算法的计算复杂度。从表中可以看出，SI算法需要的运算时间最短，这是因为SI算法仅仅对数据进行简单的插值运算，故而运算时间也就最少。所提方法在建模过程中需要将数据扩展到多维进行运算，故而运算时间相对较长。但是整个算法运行过程中，只进行一次建模且都是简单的扩维运算且算法的求解过程只是求解一个1维的稀疏向量恢复问题，故而运算时间依旧非常可观。

表 4 各方法在不同数据缺失率下的性能比较

		数据缺失率(%)				
		20	50	80	90	95
MOTES	KCS-GAMP-SBL	0.0266	0.0356	0.0840	0.1101	0.1509
	TDMF	0.0287	0.0871	0.1962	0.2604	0.3684
	SI	0.0862	0.1604	1.0172	1.7662	3.5204
	RNN	0.0392	0.0951	0.8875	0.9174	1.0529
	KPPCA	0.0278	0.0608	0.1826	0.2769	0.4516
GSA	KCS-GAMP-SBL	0.0357	0.0460	0.1985	0.2769	0.3304
	TDMF	0.0548	0.1278	0.3557	0.4764	0.5688
	SI	0.0935	0.1455	0.8348	2.9442	4.6315
	RNN	0.0526	0.0945	1.0858	1.1639	1.7984
	KPPCA	0.0498	0.1012	0.3892	0.4879	0.5872
SST	KCS-GAMP-SBL	0.0181	0.0327	0.0956	0.1225	0.1767
	TDMF	0.0242	0.0544	0.1788	0.2374	0.2961
	SI	0.0485	0.0851	0.7451	1.3802	2.0596
	RNN	0.0262	0.0488	0.3417	0.5896	1.0421
	KPPCA	0.0226	0.0644	0.1736	0.2705	0.2909

6.2 稀疏表示基的选择对算法均方根误差的影响

表6展示了不同的稀疏表示基对均方根误差大小的影响。本文已经在前文中理论分析了稀疏表示

表 5 不同算法平均运行时间(ART)的比较(s)

	SI	RNN	KPPCA	TDMF	KCS-GAMP-SBL
MOTES	0.7350	3.1832	11.5781	8.4362	3.0802
GSA	1.5361	7.0320	24.8685	20.4093	15.8633
SST	0.7216	2.5216	10.8683	8.9381	2.9032

表 6　稀疏表示基的选择对算法均方根误差(RMSE)的影响

	数据缺失率(%)				
	20	50	80	90	95
MOTES_1	0.0266	0.0356	0.0930	0.1201	0.1509
MOTES_2	0.0312	0.0422	0.1151	0.1387	0.1954
GSA_1	0.0357	0.0460	0.1985	0.2769	0.3304
GSA_2	0.0412	0.0681	0.2313	0.3343	0.4068
SST_1	0.0181	0.0327	0.0996	0.1275	0.1767
SST_2	0.0199	0.0340	0.1265	0.1534	0.2463

基φ_{KCS1}和φ_{KCS2}稀疏信号能力的强弱以及不同的稀疏表示基与观测矩阵之间的非相关性大小。由表3得到当不同的稀疏表示基与观测矩阵之间的非相关性大小十分接近。由表2得出稀疏表示基φ_{KCS1}具有更强的稀疏信号能力。从表6中可以看出，仿真结果与前面的理论分析相一致。也就是说，我们选择φ_{KCS1}作为稀疏表示基时，算法具有更好的性能。

6.3　数据缺失类型对算法性能的影响

图1–图3以均方根误差为衡量标准展示了数据缺失类型对算法性能的影响。从图中可以发现，当数据均匀缺失时，均方根误差最小。这是因为如果数据是均匀缺失的，那么观测到的数据也是均匀的，我们就可以获得每一段时间内的有用信息。值得注意的是，即使在数据连续缺失的情况下，算法产生的均方根误差依旧很小。这是因为本文设计的算法充分利用了KCS原理将缺失预测问题建模成了稀疏向量恢复问题。只要观测值的数量大于稀疏表示向量的稀疏度，就能利用这些观测值以高概率恢复出原始时间序列。

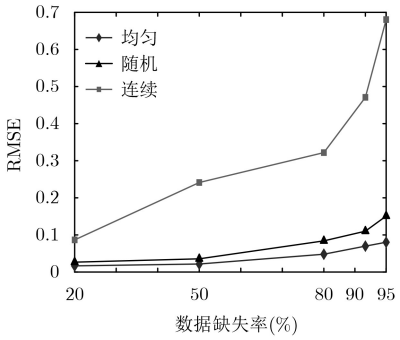

图 1 MOTES中数据缺失类型的影响

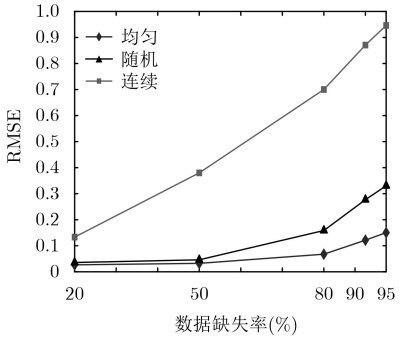

图 2 GSA中数据缺失类型的影响

6.4　数据集的特性对算法性能的影响

为了研究数据集的特性对算法性能的影响，本文从GSA数据集和SST数据集中任意删除一些数据列，使得3个数据集的维数相同，比较它们在相同情况下的均方根误差的大小。由图4中可以发现，算法在MOTES数据集上具有最好的性能，SST数据集上次之。回顾表2，可以发现当稀疏表示基为φ_{KCS1}时，MOTES的稀疏性最好，GSA和SST的稀疏性非常接近。但图4中，算法明显在SST上的性能更好。由此，我们有理由相信，算法的性能不仅仅与信号在稀疏表示基下的稀疏向量的稀疏程度有关，还与数据集中多个时间序列的之间的相关程度有关。

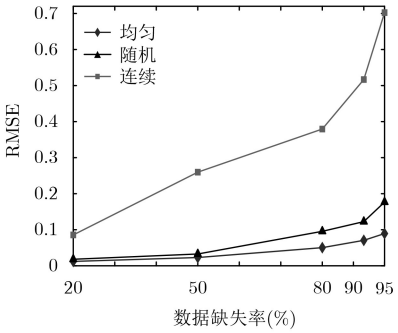

图 3 SST中数据缺失类型的影响

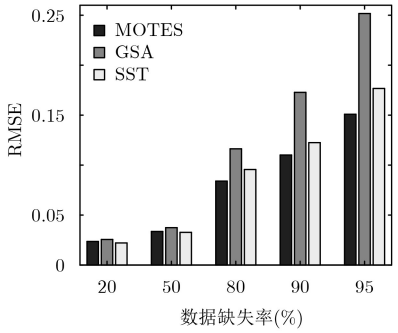

图 4 数据集特性对算法性能的影响

7　结束语

本文充分利用多变量时间序列的特点设计了一种基于克罗内克压缩感知的多变量时间序列缺失数据预测算法。通过应用KCS理论将缺失数据预测问题建模成了稀疏向量恢复问题，从而仅仅利用较少的观测值就可以同时预测出多个时间序列中的缺失数据。本文的建模过程对解决一般的1维或多维时间序列中的缺失数据预测问题具有很好的参考意义。仿真结果表明本文算法在解决多变量时间序列中的缺失数据问题上是非常适用且有效的。

参 考 文 献

[1] SOWMYA R and SUNEETHA K R. Data mining with big data[C]. International Conference on Intelligent Systems and Control, Coimbatore, India, 2017: 246–250. doi: 10.1109/ISCO.2017.7855990.

[2] JAQUES N, TAYLOR S, SANO A, *et al*. Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction[C]. International Conference on Affective Computing & Intelligent Interaction, San Antonio, USA, 2017: 202–208. doi: 10.1109/ACII.2017.8273601.

[3] BALOUJI E, SALOR Q, and ERMIS M. Exponential smoothing of multiple reference frame components with GPUs for real-time detection of time-varying harmonics and interharmonics of EAF currents[C]. IEEE Industry Applications Society Meeting, Cincinatti, USA, 2017: 1–8. doi: 10.1109/IAS.2017.8101815.

[4] KOZERA R and WILKOLAZKA M. Natural spline interpolation and exponential parameterization for length estimation of curves[C]. International Conference of Numerical Analysis & Applied Mathematics, Rhodes, Greece, 2017: 1–140.

[5] LAO Wenchao, WANG Ying, CHEN Peng, *et al*. Time series forecasting via weighted combination of trend and seasonality respectively with linearly declining increments and multiple sine functions[C]. International Joint Conference on Neural Networks, Beijing, China, 2014: 832–837. doi: 10.1109/IJCNN.2014.6889609.

[6] LIPPI M, BERTINI M, and FRASCONI P. Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning[J]. *IEEE Transactions on Intelligent*, *Transportation*, *System*, 2013, 14(2): 871–882. doi: 10.1109/TITS.2013.2247040.

[7] STRAUMAN A S, BIANCHI F M, and MIKALSEN K Ø. Classification of postoperative surgical site infections from blood measurements with missing data using recurrent neural networks[C]. International Conference on Biomedical & Health Informatics, Las Vegas, USA, 2018: 307–310. doi: 10.1109/BHI.2018.8333430.

[8] LI Li, LI Yuebiao, and LI Zhiheng. Efficient missing data imputing for traffic flow by considering temporal and spatial dependence[J]. *Transportation Research Part C*, 2013, 34(9): 108–120.

[9] LI Yuebiao, LI Zhiheng, LI Li, *et al*. Comparison on PPCA, KPPCA and MPPCA based missing data imputing for traffic flow[C]. Proceedings of IEEE Conference on Intelligent Transportation System, Wuhan, China, 2013: 1535–1540.

[10] SHI Weiwei, ZHU Yongxin, and HUANG Tian. Effective prediction of missing data on apache spark over multivariable time series[J]. *IEEE Transactions on Big Data*, 2018. doi: 10.1109/TBDATA.2017.2719703.

[11] HADI A and WAHIDAH I. Delay estimation using compressive sensing on WSN IEEE 802.15.4[C]. International Conference on Control, Electronics, Renewable Energy and Communications, Bandung, Indonesia, 2017: 192–197. doi: 10.1109/ICCEREC.2016.7814975.

[12] DUARTEAND M F and BARANIUK R G. Kronecker compressive sensing[J]. *IEEE Transactions on Image Processing*, 2012, 21(2): 494–504. doi: 10.1109/TIP.2011.2165289.

[13] ZHOU Haifei, TAN Liangsheng, GE Fei, *et al*. Traffic matrix estimation: Advanced-Tomogravity method based on a precise gravity model[J]. *International Journal of Communication Systems*, 2015, 28(10): 1709–1728. doi: 10.1002/dac.2787.

[14] CHEN S S, DONOHO D L, and SAUNDERS M A. Atomic decomposition by basis pursuit[J]. *SIAM Review*, 2001, 43(1): 129–159. doi: 10.1137/S003614450037906X.

[15] TROPP J A and GILBERT A C. Signal recovery from random measurements via orthogonal matching pursuit[J]. *IEEE Transactions on Information Theory*, 2007, 53(12): 4655–4666. doi: 10.1109/TIT.2007.909108.

[16] CHARTRAND R and YIN W. Iteratively reweighted algorithms for compressive sensing[C]. IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, 2008: 3869–3872. doi: 10.1109/ICASSP.2008.4518498.

[17] AL-SHOUKAIRI M, SCHNITER P, and RAO B D. A GAMP based low complexity sparse Bayesian learning algorithm[J]. *IEEE Transactions on Signal Processing*, 2018, 66(2): 294–308. doi: 10.1109/TSP.2017.2764855.

[18] SAMUEL M. Intel lab data[OL]. http://db.csail.mit.edu, 2016.

[19] FONOLLOSA J, SHEIK S, HUERTA R, *et al*. Reservoir computing compensates slow response of chemo sensor arrays exposed to fast varying gas concentrations in continuous monitoring[J]. *Sensors & Actuators B Chemical*, 2015, 215: 618–629.

[20] Tropical Atmosphere Ocean. NOAA/Pacific Marine Environmental Laboratory[OL]. http://www.pmel.noaa.gov/tao/proj_over/proj_over.html, 2016.

[21] WU Xiaopei and LIU Mingyan. In-situ soil moisture sensing: Measurement scheduling and estimation using compressive sensing[C]. ACM/IEEE, International Conference on Information Processing in Sensor Networks, Beijing, China, 2012: 1–11. doi: 10.1109/IPSN.2012.6920949.

郭 艳：女，1971年生，教授，研究方向为大数据、信号处理、压缩感知.

宋晓祥：男，1993年生，硕士生，研究方向为大数据、压缩感知.

李 宁：男，1967年生，副教授，研究方向为信号处理、认知无线电.

钱 鹏：男，1991年生，博士生，研究方向为压缩感知、无源目标定位.