缺失数据时间序列的鲁棒主导周期检测

Qingsong Wen[1],*Linxiao Yang*[2],*Liang Sun*[1]

[1]达摩院，阿里巴巴集团，美国贝尔维尤[2]达摩院，阿里巴巴集团，杭州，中国

arXiv:2303.03553v1 [cs.LG] 6 Mar 2023

摘要

周期性检测是时间序列分析中的一项重要任务，但由于时间序列数据具有多种特征，例如趋势突变、离群值、噪声，尤其是块状缺失数据，因此仍是一项具有挑战性的问题。本文提出了一种针对存在块状缺失数据的时间序列的鲁棒且有效的周期性检测算法。首先，我们设计了一种稳健趋势滤波器，以去除缺失数据情况下复杂趋势模式的干扰。随后，我们提出了一种能够有效处理缺失值和离群值的稳健自相关函数（ACF）。我们严格证明，当缺失块的长度小于周期长度的 $1/3$时，所提出的稳健自相关函数仍能良好工作。最后，通过结合时频信息，我们的算法能够准确地生成周期长度。实验结果表明，在真实世界时间序列数据集上，我们的算法优于现有的周期性检测算法。

关键词——周期性检测、季节性检测、缺失数据、鲁棒方法、时间序列

1. 引言

在物联网（IoT）、智能运维（AIOps）和云计算等现代应用中，许多时间序列信号具有重复循环或周期性这一特征 [1, 2, 3]。周期性检测与调整在许多真实世界时间序列应用中至关重要，例如异常检测 [4, 5, 6], 预测 [7, 8, 9], 分类和聚类 [10, 11], 以及分解 [12, 13, 14]。然而，由于物联网、智能运维和云计算中真实世界时间序列数据具有不同且复杂的特性 [2], , 准确的周期性检测仍然是一个具有挑战性的问题。下面我们简要强调实际周期性检测面临的三项挑战。首先，大多数时间序列是非平稳的，趋势变化甚至趋势突变都很常见。其次，时间序列数据通常包含噪声和离群值。第三，许多真实世界时间序列数据包含缺失值，甚至成块缺失值 [15]。

现有的周期性检测算法可分为两类：1）频域方法，依赖傅里叶变换后的周期图，例如费舍尔检验 [16, 17]；2）时域方法，依赖自相关函数（自相关函数），例如 [18, 19] 中的方法。为结合这两类方法的优点，近期提出了联合时频方法 [20,21, 7]。然而，上述方法无法直接处理含缺失数据的时间序列。为处理缺失值，研究者提出了基于Lomb-Scargle周期图的方法来检测周期性，如 [22, 23, 24] 所示。不幸的是，这些方法无法稳健地处理时间序列中的异常值和噪声。

本文提出了一种新颖的周期性检测方法，用于在存在缺失数据的情况下，稳健且准确地检测时间序列的主导周期。首先，为减轻趋势，尤其是趋势突变带来的副作用，我们设计了一种稳健滤波器，用于在存在异常值和缺失数据的情况下去除趋势分量。其次，为获得准确的周期长度，我们设计了一个稳健的自相关函数模块，该模块能够有效处理具有未知重尾噪声的脉冲随机噪声，以及块状缺失数据。我们严格证明，当缺失块的长度小于整个时间序列长度的 $1/3$时，我们的算法仍能成功运行。最后，我们结合时频信息，对周期长度进行最终的准确估计。与各种最先进算法相比，我们提出的稳健算法在真实世界数据集上的表现显著更好，尤其适用于包含缺失数据的时间序列。

2. 提出的周期性检测

2.1. 框架概述

我们考虑以下包含趋势和主导周期/季节性的复杂时间序列模型:

$$y_t = \tau_t + s_t + r_t, \quad t = 0, 1, \cdots, N-1 \tag{1}$$

其中， y_t 表示时间 t处的观测值， τ_t 是趋势分量， s_t 是周期长度为 T的主导周期/季节性分量， $r_t = a_t + n_t$ 表示包含噪声 n_t 和可能的异常值 a_t的剩余部分。请注意， y_t可能包含缺失数据。对于主导周期检测，我们旨在判断该时间序列是否具有周期性，并确定其主要周期长度。

我们提出的周期性检测算法包含三个步骤：1）稳健去趋势滤波器；2）稳健自相关函数；3）时频组合。图 1 展示了所提出算法的示意图，以下各小节将对此进行详细阐述。

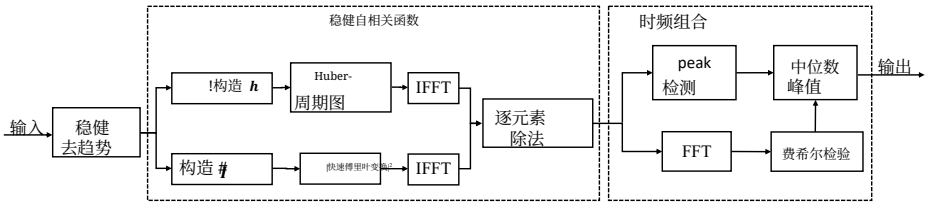


图1：针对含缺失数据的时间序列提出的稳健主导周期检测算法框架。

2.2 缺失数据下的稳健去趋势

大多数真实世界时间序列通常存在缺失数据、变化的趋势和离群值，这使得自相关函数难以捕捉周期信息。为了在存在缺失数据的情况下去除趋势分量，并受到 RobustTrend 滤波器 [25], 的启发，我们设计了一种适用于缺失数据的新型稳健趋势滤波器，其目标函数为

$$||\mathbf{W}(\mathbf{y}-\boldsymbol{\tau})||_1+\lambda_1||\mathbf{D}^{(1)}\boldsymbol{\tau}||_1+\lambda_2||\mathbf{D}^{(2)}\boldsymbol{\tau}||_1, \tag{2}$$

其中， $D^{(1)}\in\mathbb{R}^{(N-1)\times N}$和 $D^{(2)}\in\mathbb{R}^{(N-2)\times N}$分别是用于捕捉突发趋势变化和缓慢趋势变化的一阶和二阶差分矩阵，其形式如下:

$$\mathbf{D}^{(1)}=\begin{bmatrix}1 & -1 & & \\ & 1 & -1 & \\ & & \ddots & \\ & & 1 & -1\end{bmatrix},\ \mathbf{D}^{(2)}=\begin{bmatrix}1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & & \ddots & \\ & & & 1 & -2 & 1\end{bmatrix},$$

其中，$\mathbf{W}=\text{diag}(w_t)$是一个对角矩阵，当 y_t 缺失时， w_t 取0; 否则取1。请注意，我们设计的趋势滤波器与RobustTrend滤波器 [25]有两点不同。第一，损失函数中引入了$\mathbf{W}$ 矩阵，因此无需进行缺失值插补即可提取趋势。第二，考虑到最小绝对偏差（LAD）损失对异常值具有与原RobustTrend滤波器的Huber损失相似的稳健性，且易于通过交替方向乘子法（ADMM）求解，我们将原RobustTrend滤波器中的Huber损失替换为LAD损失。具体而言，为了从式（2）中获得趋势分量，我们可以将其改写为等价形式:

$$\begin{aligned}\min\quad & ||\mathbf{e}||_1\\ \text{s.t.}\quad & \mathbf{A}\boldsymbol{\tau}-\mathbf{b}=\mathbf{e}\end{aligned} \tag{3}$$

其中 $\mathbf{A}=[\mathbf{W}^T,\ \mathbf{D}^{(1)T},\ \mathbf{D}^{(2)T}]^T$, $\mathbf{b}=[(\mathbf{W}\mathbf{y})^T,\ \mathbf{0}^T]^T$。然后可以得到增广拉格朗日函数:

$$L_\rho(\mathbf{x},\mathbf{y},\mathbf{u})=||\mathbf{y}||_1+\mathbf{u}^T(\mathbf{A}\mathbf{x}-\mathbf{b}-\mathbf{y})+\frac{\rho}{2}||\mathbf{A}\mathbf{x}-\mathbf{b}-\mathbf{y}||_2^2$$

其中，$\mathbf{u}$ 是对偶变量， ρ 是惩罚参数。接下来， .
该解可以通过 ADMM [26] 获得，如下所示:

$$\mathbf{x}_{t+1}=\arg\min_{\mathbf{x}}L_\rho(\mathbf{x},\mathbf{y},\mathbf{u})=(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T(\mathbf{b}+\mathbf{y}_t-\mathbf{u}_t/\rho)$$

$$\mathbf{y}_{t+1}=\arg\min_{\mathbf{y}}L_\rho(\mathbf{x},\mathbf{y},\mathbf{u})=S_{1/\rho}(\mathbf{A}\mathbf{x}_{t+1}-\mathbf{b}+\mathbf{u}_t/\rho)$$

$$\mathbf{u}_{t+1}=\arg\min_{\mathbf{u}}L_\rho(\mathbf{x},\mathbf{y},\mathbf{u})=\mathbf{u}_t+\rho(\mathbf{A}\mathbf{x}_{t+1}-\mathbf{y}_{t+1}-\mathbf{b})$$

其中， $S_\kappa(x)$ 是软阈值函数，表示为 $S_\kappa(x)=(1-\kappa/|x|)_+x$。

2.3. 基于 M-周期图的鲁棒自相关函数

2.3.1. 所提出鲁棒自相关函数的结构

令 $\{x_t\}$ 表示去趋势时间序列，即 $x_t=y_t-\tau_t$。然后，可以通过找到自相关函数 $\{r_k\}$ 的最大值所在位置来估计其周期，其中 $\{r_k\}$ 定义为 $r_k=\mathbb{E}(x_{t+k}x_t)$。

自相关函数通常使用 $r_k\approx\frac{1}{|\mathbb{Q}_k|}\sum_{t\in\mathbb{Q}_k}x_{t+k}x_t$ 估计，其中集合 定义为， 表示 的大小。然而，这种估计量不能直接用于含有缺失值的时间序列。为解决这一问题，我们提出一种新的无偏估计量。为了更好地说明我们的方法，令 为一个二进制序列，用于指示是否观测到 。具体而言，当观测到 时，令 ; 当 缺失时，令 。我们提出的自相关函数估计量为ˆ ˆ，其中集合为ˆ 。不难看出，该估计量是自相关的无偏估计量，但直接计算它的计算复杂度为 $O(N^2)$ 阶。为降低复杂度，我们通过补零来定义序列 $\{h_t\}$，即若 $I_t=1$，则 $h_t=x_t$；若 $I_t=0$，则 $h_t=0$。容易看出， $h_t=x_tI_t$。因此，我们提出的自相关函数估计量可以改写为

$$\begin{aligned}r_k\approx&\frac{1}{|\hat{\mathbb{Q}}_k|}\sum_{t\in\hat{\mathbb{Q}}_k}x_{t+k}x_t=\frac{1}{|\hat{\mathbb{Q}}_k|}\sum_{t\in\hat{\mathbb{Q}}_k}I_{t+k}x_{t+k}I_tx_t\\ =&\frac{1}{|\hat{\mathbb{Q}}_k|}\sum h_{t+k}\hat{h}_{k-(t+k)}=\frac{1}{|\hat{\mathbb{Q}}_k|}(\mathbf{h}*\hat{\mathbf{h}})_k\end{aligned} \tag{4}$$

其中 $h_t=h_{-t}$，运算“*”表示线性卷积。回顾一下，ˆ $\mathbb{Q}_k=\{t|I_t=1,I_{t+k}=1\}=\{t|I_tI_{t+k}=1\}$，ˆ ˆ ˆ 我们有 $|\mathbb{Q}_k|=\sum I_nI_{t+k}=\sum I_{-t}I_{t+k}=(\mathbf{I}*\mathbf{I})_k$，其中ˆ$I_t=I_{-t}$。此外，我们可以利用基于循环卷积定理的快速傅里叶变换/快速傅里叶逆变换，将 $(\mathbf{h}*\mathbf{h})_k$ 和 $(\hat{\mathbf{I}}*\mathbf{I})_k$ 的计算复杂度从 $O(N^2)$ 降至 $O(N\log N)$。因此，我们高效且稳健的自相关函数为

$$r_k\approx\frac{(\mathbf{h}*\hat{\mathbf{h}})_k}{(\hat{\mathbf{I}}*\mathbf{I})_k}=\frac{\text{IFFT}(|\text{FFT}(\bar{\mathbf{h}})|^2)_k}{\text{IFFT}(|\text{FFT}(\bar{\mathbf{I}})|^2)_k}, \tag{5}$$

其中，$\mathbf{h}$和$\mathbf{I}$的长度通过填充 N 个零而加倍，并分别记为 $\bar{}\mathbf{h}=[\mathbf{h}^T,\ \mathbf{0},\ \cdots,\ 0]^T$和 $\bar{}\mathbf{I}=[\mathbf{I}^T,\ \mathbf{0},\ \cdots,\ 0]^T$。

除了处理缺失数据，我们还考虑减轻式 (5) 中 $\bar{}\mathbf{h}$ 的异常值影响。注意， $|\text{FFT}(\bar{}\mathbf{h})|^2$ 是 $\bar{}\mathbf{h}$的传统周期图。为使其对异常值具有稳健性，我们引入 Huber-周期图 [7] ，定义如下:

$$(|\text{FFT}(\bar{\mathbf{h}})|^2)_k\approx\frac{N'}{4}\left|\hat{\boldsymbol{\beta}}_M(k)\right|^2=\frac{N'}{4}\left|\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^2}\gamma(\boldsymbol{\phi}\boldsymbol{\beta}-\bar{\mathbf{h}})\right|^2, \tag{6}$$

其中， $N'=2N$， ˆ$\boldsymbol{\beta}_M(k)$ 是谐波回归量 $\boldsymbol{\phi}_t=[\cos(2\pi kt/N'),\sin(2\pi kt/N')]$的稳健估计， $\boldsymbol{\phi}$ 具有矩阵形式 $[\boldsymbol{\phi}_0,\boldsymbol{\phi}_1,\cdots,\boldsymbol{\phi}_{N'-1}]^T$，而 $\gamma(\mathbf{x})$ 是稳健化目标函数，定义为 Huber 损失之和: $\gamma(\mathbf{x})=\sum_{t=0}^{N'-1}\gamma^{hub}(x_t)$。

基于上述对缺失数据和异常值的处理，我们的稳健自相关函数的最终结构如图 1 中间部分所示。

2.3.2. 缺失数据下的理论约束 *g Data*

接下来，我们考察稳健自相关函数估计器所能承受的最大缺失块长度。显然，利用式 (4) 计算自相关函数需要 $Q_k = |\mathbb{Q}_k| > 0$。而 Q_k 的值取决于缺失条目的位置和数量，这使得分析变得困难。实际中，缺失条目最糟糕的情况是集中成簇。下面，我们对所有缺失条目集中在一个块中的情形进行理论分析，即当 $m \le n \le m+l-1$ 时 $I_k = 0$，否则为 $I_k = 1$，其中 m 表示该块的起始位置，l 表示该块的长度。这里假设该块不包含序列的起点和终点，即 $m > 0$ 且 $m+l-1 < N-1$。注意，所提出的方法当且仅当 $Q_k > 0$ 时才能运行。回顾一下，当 $k \ge 0$ 时，Q_k 为 $\sum_{t=k}^{N-1} I_t I_{t-k}$，否则为 $\sum_{t=0}^{N+k} I_t I_{t-k}$。形式化地，我们有如下命题。

命题：对于任意满足 $1 < k < N-1$ 和 $l < N$ 的 k、m 和 l，$m > 0$，$m+l-1 < N-1$，若 $Q_k > 0$，则始终有 $k \ge m$。

证明：根据 Q_k 的定义，我们有 $Q_k \ge 0$，因此只需证明不存在满足 $Q_k = 0$ 的 k。不失一般性，我们仅考虑 $k > 0$ 的情况。令 $g_t = I_{t-k}$，则对于 $m+k \le t \le m+l+k-1$，有 $g_t = 0$。假设存在 k 和 m 使得 $Q_k = 0$，则对所有 $n \in [k, N-1]$ 都有 $I_n I_n = 0$，于是我们有

$$[k, N-1] \subset [m, m+l-1] \cup [m+k, m+l+k-1] \quad (7)$$

注意，公式（7）意味着以下三种可能性：

$$[k, N-1] \subset [m, m+l-1] \quad (8)$$

$$[k, N-1] \subset [m+k, m+l+k-1] \quad (9)$$

$$[k, N-1] \subset [m, m+l-1] \cup [m+k, m+l+k-1]$$
$$[m, m+l-1] \cup [m+k, m+l+k-1] = [m, m+l+k-1] \quad (10)$$

显然，公式（8）要求 $m+k \le m+l-1$ 或，这与和矛盾。注意，公式（9）意味着 $N-1 \le m+l-1$ 和公式（10）要求 $k+1 \ge m$ 或 $m > N-l$，这与和矛盾。注意，公式（10）意味着 $m \le k$、$m+k \le m+l$ 和 $N-1 \le m+l+k-1$，可将其改写为

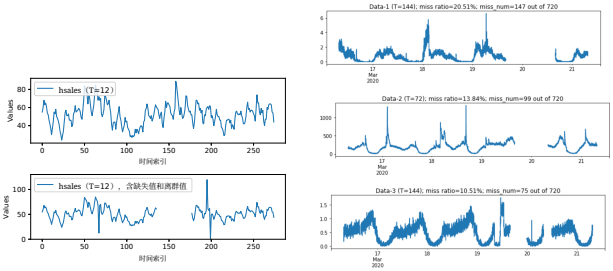
$$\max(m, N-m-l) \le k \le l \quad (11)$$

我们证明，如果 $l < \frac{N}{3}$，则无法找到满足公式（11）的 m 和 k。不难看出，对于所有 $0 < m < N-l$，$\max(m,\ N-m-l)$ 的最小值在 $m = \frac{N-l}{2}$ 处取得，并且等于 $\frac{N-l}{2}$。由于 $l < \frac{N}{3}$，我们可以看到，l 始终小于 $\frac{N-l}{2}$，即 $l < \frac{N-l}{2} \le \max(m,\ N-m-l)$，这意味着不存在满足公式（11）的 k。证明到此完成。

2.4. 最终时频组合

在频域中，为了检测主导周期，费舍尔检验 [16] 将 g-统计量定义为 $g = \max_k P_k / \sum_{j=1}^{N} P_j$，$k = 1, 2, \cdots, N$，其中 $P_k = \mathrm{FFT}\{r_t\}$ 是基于稳健自相关函数 r_t，通过维纳–辛钦定理 [27] 计算得到的周期图。因此，该g-统计量同样对异常值和缺失数据具有稳健性。其分布-

g统计量的分布给出一个 p值，用于判断时间序列是否具有周期性 [16]。如果该值小于预定义的阈值 α，我们就拒绝原假设 H_0，并得出时间序列具有周期性的结论，其周期长度候选值为 N/k，其中 $k = \mathrm{argmax}_k P_k$。如果费舍尔检验通过，我们将进一步在时域中细化周期长度候选值，方法类似于 [20, 7]。



（a）一个 CRAN 周期性时间序列及其含离群值和缺失值的变体。（b）来自云计算行业的三个典型周期性时间序列，含离群值和缺失值。

图 2：具有代表性的困难周期性时间序列。

具体而言，我们首先通过峰值检测 [28]汇总稳健自相关函数的峰值。然后，计算峰高超过预定义阈值的这些峰值之间的中位数距离。此外，根据周期图的分辨率，即索引为 k 处的 P_k 峰值对应于范围 $[\frac{N}{k}, \frac{N}{k-1}]$ 内的周期长度，只有当自相关函数峰值的中位数距离位于 $R_k = \left[\frac{1}{2}\left(\frac{N}{k+1} + \frac{N}{k}\right) - 1, \cdots, \frac{1}{2}\left(\frac{N}{k} + \frac{N}{k-1}\right) + 1\right]$ 内时，才将其作为最终周期长度。需要注意的是，这种组合是必要的。一方面，周期图的分辨率有限，并且可能存在谱泄漏 [20],，这会导致费舍尔检验得到的候选值不准确。另一方面，仅依赖自相关函数可能会产生假阳性结果，因为自相关函数无法判断是否存在主导周期。

3. 实验与讨论

3.1. 周期性检测比较

基线算法与数据集：

我们将所提出的算法与以下基线算法进行比较：ACF-Med（自相关函数峰值的中位数距离）、费舍尔检验、Lomb–Scargle 周期图 [22, 23, 24], 以及最先进的时频方法 RobustPeriod [7]。值得注意的是，还有其他类型的时频算法 [20, 21],，但它们的性能不如 RobustPeriod [7]。除 Lomb–Scargle 和我们提出的算法外，对于其他算法，若数据存在缺失值，我们会先进行线性插值，因为如果直接在含缺失值的时间序列上运行，这些算法的性能会显著下降。

对于数据集，我们考虑了 CRAN 中公开的单周期时间序列 [29, 19],，其中包含 82 条来自不同场景（如物联网、销售、太阳黑子等）的真实世界时间序列，长度为 16 至 3024，周期长度为 2 至 52。我们还在数据集中加入异常值和块状缺失数据，以评估所有算法的鲁棒性。离群值比例（OR）设为 0.01 或 0.05，离群值幅度设为原始时间序列标准差的 5 倍；块缺失比例（MR）设为 0.05 或 0.30。由于实际应用中的离群值比例通常较低，我们将其上限设为 0.05；而由于块状缺失可能导致缺失率相对较高，我们将其上限设为 0.3。示例如图2(a)所示。此外，我们采用了一家领先云计算公司的 3个典型真实世界数据集，如图2(b)所示。数据集长度为 5 天，采样分辨率为 10 分钟，其真实周期长度分别为 144（一天）、72（半天）和 144（一天）。评估时，精确率按正确估计周期长度的时间序列数量占时间序列总数的比例计算。

表 1：不同周期性检测算法在公开 CRAN 数据上的精确率比较。MR 和 OR 分别表示缺失率和离群值比例。最佳结果以粗体突出显示。

《算法》	包含或不包含缺失值和离群值的 CRAN 数据集								
	MR=0			MR=0.05			MR=0.30		
	OR=0	OR=0.01	异常值率=0.05	OR=0	异常值率=0.01	异常值率=0.05	OR=0	异常值率=0.01	异常值率=0.05
ACF-Med	0.57	0.55	0.24	0.56	0.50	0.23	0.52	0.50	0.27
费舍尔检验	0.52	0.44	0.32	0.49	0.44	0.30	0.42	0.33	0.24
Lomb–Scargle	0.45	0.13	0.09	0.45	0.13	0.10	0.37	0.11	0.10
RobustPeriod	0.62	0.60	0.59	0.59	0.57	0.50	0.47	0.45	0.43
所提方法	**0.63**	**0.62**	**0.60**	**0.60**	**0.60**	**0.59**	**0.56**	**0.52**	**0.51**

表2：在3个典型真实世界数据集（含缺失数据）上对比不同周期性检测算法。

《算法》	Data-1 (T=144)	Data-2 (T=72)	Data-3 (T=144)
ACF-Med	144	72	143
费舍尔检验	144	73	143
Lomb–Scargle	140	71	140
RobustPeriod	142	(72,144)	144
所提出的	**144**	**72**	**144**

不同算法的性能比较：
我们在表1中汇总了公开 CRAN 数据集上的周期检测精确率结果，涵盖有缺失数据和异常值以及无缺失数据和异常值的情况。可以看出，在没有缺失数据和异常值时，RobustPeriod 与我们所提出的算法表现相近，且优于其他方法。当没有异常值、但缺失数据逐渐增多时（即 OR

=0 =0 0 05, 、

0 3

MR、.....），Lomb–Scargle 的性能下降幅度相对小于 ACF-Med、费舍尔检验和 RobustPeriod。这是因为 Lomb–Scargle 无需插补即可直接处理缺失数据。然而，Lomb–Scargle 对异常值并不鲁棒；存在异常值时，其性能会远差于其他方法。当缺失数据比例固定、但异常值逐渐增多时（即 OR=0、0.05、0.3），RobustPeriod 的性能下降幅度远小于 ACF-Med、费舍尔检验和 Lomb–Scargle，这表明它对异常值具有鲁棒性。当缺失数据和异常值同时存在时，现有算法的性能都会显著下降，尤其是 ACF-Med、费舍尔检验和 Lomb–Scargle 方法。这是因为这些方法严重依赖传统的自相关函数或快速傅里叶变换，而块状缺失数据和异常值会严重扭曲自相关函数或快速傅里叶变换的结果。总体而言，我们所提出的算法在所有测试场景中表现最佳，尤其是在缺失数据和异常值比例较高时，性能优于其他方法。

表2总结了图2(b)所示的3个典型真实数据集的检测结果。可以看出，RobustPeriod 可能会产生一些假阳性，这是因为它旨在进行通用的多周期检测。有趣的是，尽管 Lomb–Scargle 可以直接处理缺失数据，但它对较大的离群值和噪声并不鲁棒。总体而言，我们所提出的算法取得了最佳结果。

3.2 组件研究

首先，我们通过实验展示所设计的稳健趋势滤波器在缺失数据情况下的优势。图3描绘了在存在缺失数据、趋势变化和异常值的时间序列中，所提出的无需插补的稳健趋势滤波器与采用 [30] 线性插值的常规 Hodrick–Prescott（HP）趋势滤波器的对比。由于 HP 滤波器无法处理含有缺失数据的时间序列，我们使用线性插值对时间序列进行修复，并将其输入

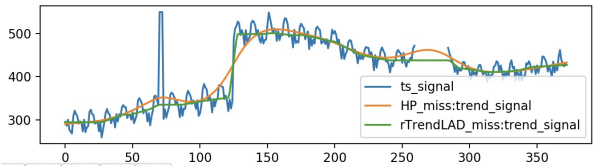
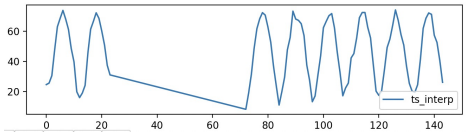
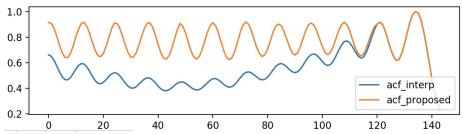


图3：所提出的无需插补的稳健趋势滤波器与采用线性插值处理缺失数据、趋势变化和异常值的 HP 趋势滤波器的对比（彩色版本效果更佳）。



(a) 经过线性插值的时间序列



(b) 缺失数据下所提出的稳健自相关函数与自相关函数的对比

采用线性插值（彩色版本效果更佳）。

图4：所提出的无需插补的稳健自相关函数与采用线性插值的自相关函数在缺失数据下的对比。

HP 滤波器。插补后，HP 滤波器仍会受到离群值和趋势突变的影响。相比之下，我们的方法（记为“rTrendLAD miss”）无需插补即可处理缺失数据，同时对离群值和趋势变化具有稳健性。

接下来，我们展示所提出的稳健自相关函数估计器在存在缺失数据时的优越性。我们针对长度为 144、周期为 12 的时间序列进行实验。我们将第24至第71个点设为缺失，因此缺失率为 1/3。图4(a)显示了经过线性插值的不完整时间序列。图4(b)展示了所提出的方法在不完整时间序列上得到的自相关函数，以及标准自相关函数在线性插值处理后的结果。可以看出，经过线性插值的时间序列所得到的自相关函数受到严重影响，因此难以找到用于定位自相关函数峰值的阈值。相比之下，我们提出的稳健自相关函数估计器能够提供可靠的自相关函数估计，从而实现准确的周期性检测。

4. 结论与未来工作

本文提出了一种新颖的周期性检测算法，用于检测主导周期。对于缺失数据场景下的时间序列，我们的算法能够稳健地处理趋势变化和离群值。未来，我们计划扩展该框架，以处理具有多个周期的时间序列，并将其应用于更多实际应用中。

5. 参考文献

[1] Heonho Kim、Unil Yun、Bay Vo、Jerry Chun-Wei Lin 和 Witold Pedrycz，“面向智能系统的时间序列数据周期性分析”，*IEEE Systems Journal*，第 15 卷，第 4 期，第 4958–4969 页，2020 年。[2] Qingsong Wen、Linxiao Yang、Tian Zhou 和 Liang Sun，“鲁棒时间序列分析及应用：工业视角”，载于*KDD 2022*，2022 年，第 4836–4837 页。[3] Shuo Zhang、XiaoFei Chen、JiaYuan Chen、Qiao Jiang 和 Hejiao Huang，“物联网高采集频率场景下周期性多变量时间序列的异常检测”，载于数据挖掘国际会议研讨会，2020 年，第 543–552 页。[4] Jingkun Gao、Xiaomin Song、Qingsong Wen、Pichao Wang、Liang Sun 和 Huan Xu，“RobustTAD：通过分解和卷积神经网络进行鲁棒时间序列异常检测”，*KDD Workshop MileTS*，2020 年。[5] Ramona Tolas、Raluca Portase、Andrei Iosif 和 Rodica Po-tolea，“周期性检测算法及其在物联网数据中的应用”，载于第 *20* 届国际并行与分布式计算研讨会，2021 年，第 81–88 页。[6] Chaoli Zhang、Tian Zhou、Qingsong Wen 和 Liang Sun，“TFAD：一种结合时频分析的分解式时间序列异常检测架构”，载于*CIKM*，2022 年。[7] Qingsong Wen、Kai He、Liang Sun、Yingying Zhang、Min Ke 和 Huan Xu，“RobustPeriod：用于鲁棒多重周期性检测的时频挖掘”，载于*ACM* 数据管理国际会议（*SIGMOD*），2021 年，第 2328–2337 页。[8] Qingyang Xu、Qingsong Wen 和 Liang Sun，“季节性时间序列预测的两阶段框架”，载于 *ICASSP*，2021 年，第 3530–3534 页。[9] Tian Zhou、Ziqing Ma、Qingsong Wen、Xue Wang、Liang Sun 和 Rong Jin，“Fedformer：用于长期序列预测的频率增强分解式 Transformer”，载于国际机器学习会议，2022 年，第 27268–27286 页。[10] X. Wang、K. Smith-Miles 和 R. Hyndman，“基于特征的时间序列数据聚类”，《数据挖掘与知识发现》，第 13 卷，第 335–364 页，2006 年 9 月。[11] Michail Vlachos、Christopher Meek、Zografoula Vagena 和 Dimitrios Gunopulos，“识别在线搜索查询中的相似性、周期性和突发”，载于*ACM SIGMOD* 数据管理国际会议，2004 年，第 131–142 页。[12] Robert B Cleveland、William S Cleveland、Jean E McRae 和 Irma Terpenning，“STL：一种基于局部加权回归的季节性趋势分解方法”，官方统计学杂志，第 6 卷，第 1 期，第 3–73 页，1990 年。[13] Qingsong Wen、Jingkun Gao、Xiaomin Song、Liang Sun、Huan Xu 和 Shenghuo Zhu，“RobustSTL：一种适用于长时间序列的鲁棒季节性趋势分解算法”，载于*AAAI*人工智能会议，2019 年，第 5409–5416 页。[14] Linxiao Yang、Qingsong Wen、Bo Yang 和 Liang Sun，“一种鲁棒且高效的多尺度季节性趋势分解方法”，载于*ICASSP*，2021 年。

[15] Xiuwen Yi、Yu Zheng、Junbo Zhang 和 Tianrui Li，“ST-MVL：填补地理传感时间序列数据中的缺失值”，载于人工智能国际联合会议（*IJCAI*），2016年。[16] Ronald Aylmer Fisher，“调和分析中的显著性检验”，《伦敦皇家学会会刊。*A*辑：数学与物理学论文集》，第125卷，第796期，第54–59页，1929年。[17] S. Wichert、K. Fokianos 和 K. Strimmer，“ 识别微阵列时间序列数据中周期性表达的转录本”，《生物信息学》，第20卷，第1期，第5–20页，2004年。[18] Wang, J.、Chen, T. 和 Huang, B.，“离散时间循环平稳信号的循环周期估计”，*IEEE*信号处理汇刊，第54卷，第1期，第83–94页，2006年。[19]Maximilian Toller、Tiago Santos 和 Roman Kern，“SAZED：时间序列数据中无参数、领域无关的季节长度估计”，《数据挖掘与知识发现》，2019年。[20] Michail Vlachos、Philip Yu 和 Vittorio Castelli，“论周期检测与结构周期相似性”，载于*SIAM*国际数据挖掘会议，2005年，第449–460页。[21] Abdullah Almasri，“一种检验周期性的新方法”，《统计学通讯——理论与方法》，第40卷，第7期，第1196–1217页，2011年。[22] Feiyan Hu、Alan F Smeaton 和 Eamonn Newman，“针对含缺失值和不规则采样数据的生活日志数据周期检测”，载于*IEEE BIBM*，2014年，第16–23页。[23] Earl F Glynn、Jie Chen 和 Arcady R Mushegian，“使用Lomb–Scargle周期图检测不等间隔基因表达时间序列中的周期模式”，《生物信息学》，第22卷，第3期，第310–316页，2006年。[24] Nicholas R Lomb，“不等间隔数据的最小二乘频率分析”，《天体物理学与空间科学》，第39卷，第2期，第447–462页，1976年。[25] Qingsong Wen、Jingkun Gao、Xiaomin Song、Liang Sun 和 Jian Tan，“RobustTrend：采用一阶与二阶差分联合正则化的Huber损失时间序列趋势滤波”，载于*IJCAI*，2019年，第3856–3862页。[26] Stephen Boyd、Neal Parikh、Eric Chu、Borja Peleato、Jonathan Eckstein 等，“通过交替方向乘子法实现分布式优化与统计学习”，《机器学习基础与趋势》，第3卷，第1期，第1–122页，2011年。[27] Norbert Wiener，“广义调和分析”，《数学学报》，第55卷，第117–258页，1930年。[28] Felix Scholkmann、Jens Boss 和 Martin Wolf，“一种用于在噪声周期信号和准周期信号中自动检测峰值的高效算法”，《算法》，第5卷，第4期，第588–603页，2012年。[29] Kurt Hornik，“R综合档案网络”，《*Wiley*跨学科评论：计算统计学》，第4卷，第4期，第394–398页，2012年。[30] Robert J Hodrick 和 Edward C Prescott，“战后美国商业周期：一项实证研究”，《货币、信贷与银行学杂志》，第1–16页，1997年。