Восстановление пропущенных значений временного ряда

Ничто не бьет человека в лоб с такой силой, как пропущенное им мимо ушей.

Ю. Слободенюк

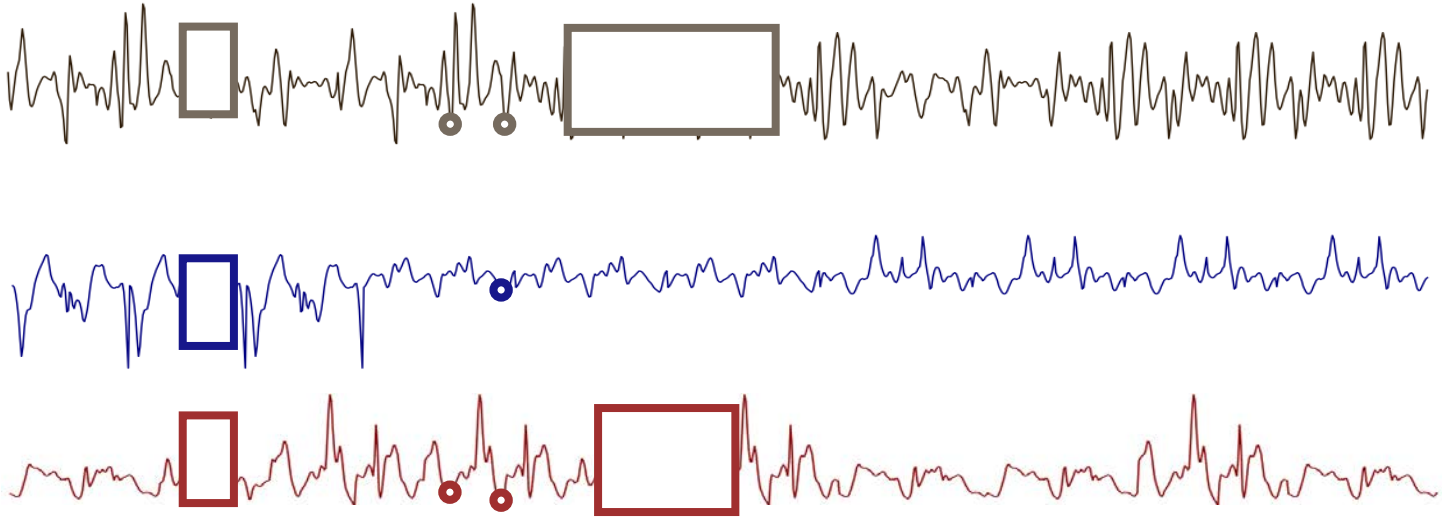
Содержание

- Постановка задачи
- Аналитические методы восстановления
- Нейросетевые методы восстановления
- Оценка точности восстановления

Восстановление ряда (в режиме офлайн)

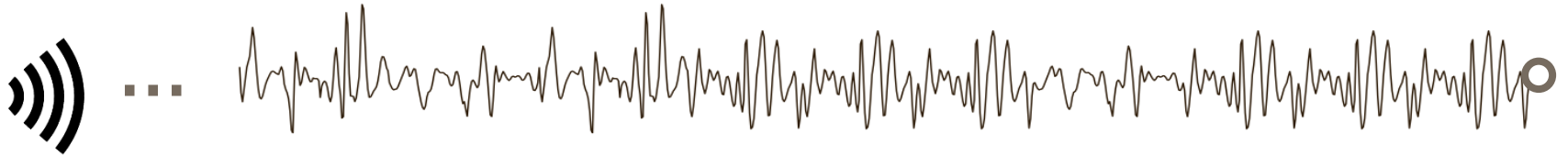
Одномерный ряд

Многомерный ряд

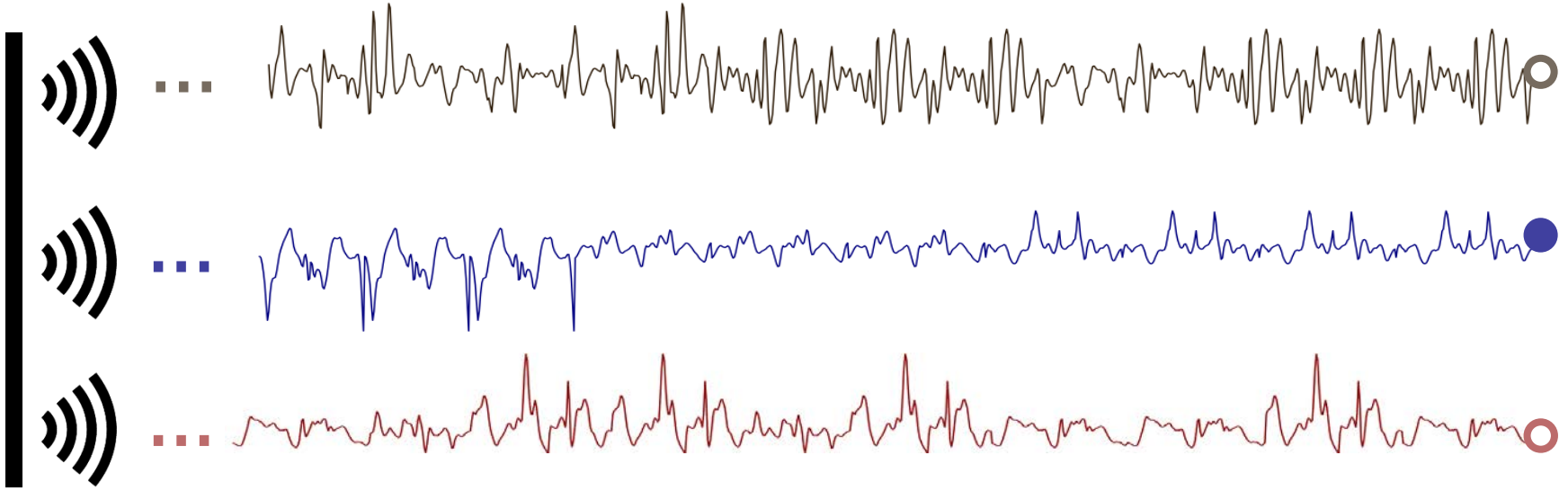


Восстановление потокового ряда в режиме реального времени

Одномерный ряд



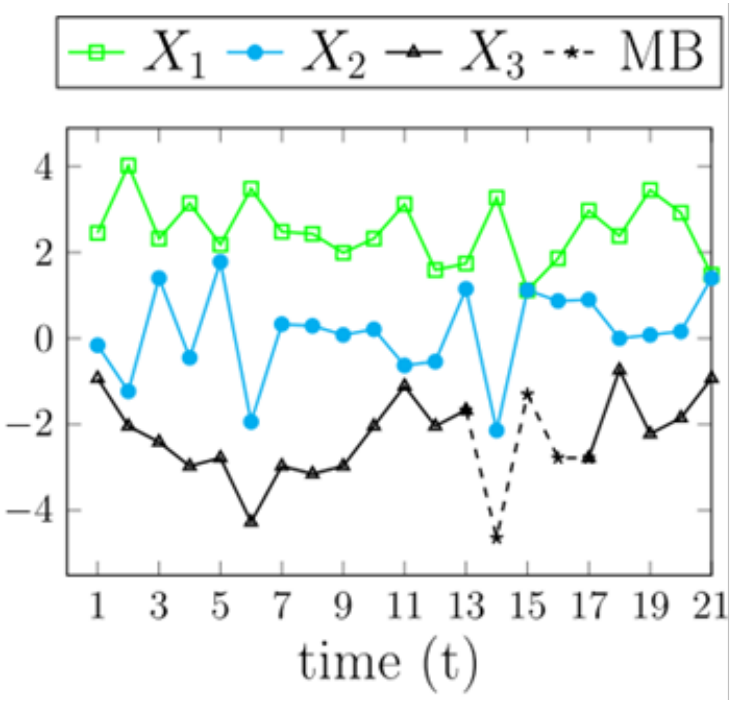
Многомерный ряд

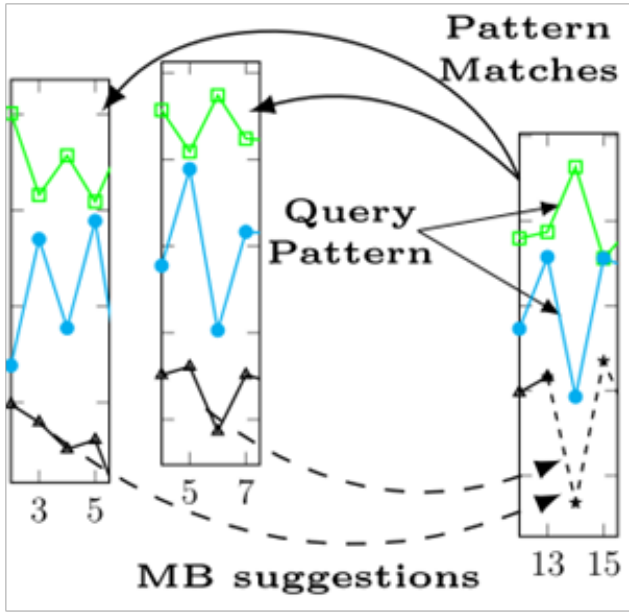


Содержание

- Постановка задачи
- **Аналитические методы восстановления**
- Нейросетевые методы восстановления
- Оценка точности восстановления

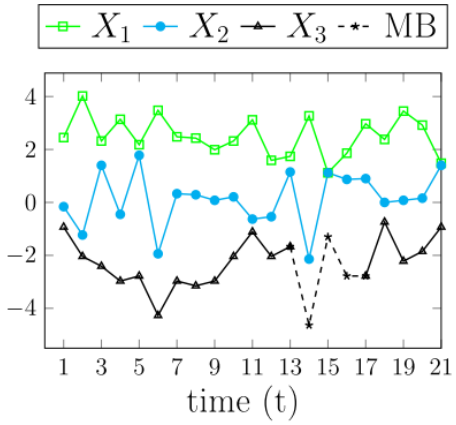
Восстановление на основе шаблонов

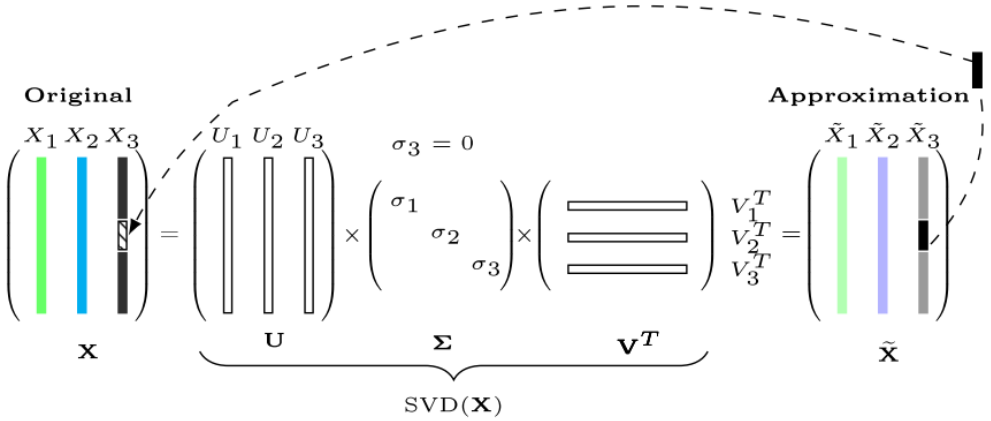




- **Imputation**
  - A high degree of similarity exists between series
  - When a block is missing in a base series, an algorithm would leverage the similarity to any number of reference series
  - The observed values in the reference series are treated as a query pattern
  - Any blocks matching that pattern may reveal candidate replacement values in the base series
- **Parametrization**
  - The length of the query pattern greatly impacts the accuracy/efficiency trade-off: too small – loss of accuracy (esp. for non-cyclic time series), too big – comparison in pattern search becomes too costly
- **Algorithms**
  - TKCM, DynaMMo, STMVL

Восстановление на основе матричного разложения





- **Reduction** the data dimensionality by SVD (Singular Value Decomposition): $X = U \cdot \Sigma \cdot V^T$
  - The Σ matrix exposes the linearly independent dimensions of the data and presents them sorted by importance ($\sigma 1 > \sigma 2 > \sigma 3$)
  - Nullify the smallest cells in the diagonal in Σ (e.g., $\sigma 3$; to be parametrized since it impacts the accuracy/efficiency trade-off)
- **Imputation**
  - Multiple the matrices after the reduction back and use the results to fill the original missing block
  - Can be iterative based on an objective function that minimizes the distance between the input and the approximated matrices w.r.t. some norm (Frobenius, nuclear, etc.)
- **Algorithms**
  - SVDImpute, SoftImpute, SVT, CDRec, GROUSE, SPIRIT, ROSL, TRMF, TeNMF

Содержание

- Постановка задачи
- Аналитические методы восстановления
- **Нейросетевые методы восстановления**
- Оценка точности восстановления

BRITS (Bidirectional Recurrent Imputation for Time Series)

Cao W. *et al.* BRITS: Bidirectional recurrent imputation for time series. NeurIPS 2018.

NAOMI (Non-AutOregressive Multiresolution Imputation)

Liu Y. *et al.* NAOMI: Non-autoregressive multiresolution sequence imputation. NeurIPS 2019. 11236–11246. https://dl.acm.org/doi/10.5555/3454287.3455295.

GAIN (Generative Adversarial Imputation Networks)

Yoon J. *et al*. GAIN: Missing data imputation using generative adversarial nets. ICML 2018. Proc. of Machine Learning Research. 2018. 80, 5675–5684.

E^2GAN (End-to-End Generative Adversarial Network)

Luo Y. *et al.* E^2GAN: End-to-End generative adversarial network for multivariate time series imputation. IJCAI 2019. 3094–3100. https://doi.org/10.24963/ijcai.2019/429

Содержание

- Постановка задачи
- Аналитические методы восстановления
- Нейросетевые методы восстановления
- **Оценка точности восстановления**

Метрики точности восстановления

- Средняя квадратичная ошибка (MSE, Mean Squared Error)
- Средняя абсолютная ошибка (MAE, Mean Absolute Error)
- Коэффициент детерминации (R^2, квадрат коэф-та корреляции)
- Средняя абсолютная процентная ошибка (MAPE, Mean Absolute Percentage Error)
- Корень средней квадратичной ошибки (RMSE, Root Mean Square Error)
- Симметричная MAPE (SMAPE, Symmetric MAPE)
- Средняя абсолютная масштабированная ошибка (MASE, Mean absolute scaled error)

Средняя квадратичная ошибка (Mean Squared Error)

$$MSE = \frac{1}{h}\sum_{i=1}^{h}(t_i - \tilde{t}_i)^2$$

- аб

Средняя абсолютная ошибка (Mean Absolute Error)

$$MAE = \frac{1}{h}\sum_{i=1}^{h}|t_i - \tilde{t}_i|$$

- аб

Коэффициент детерминации

$$R^2 = 1 - \frac{\sum_{i=1}^{h}(t_i - \tilde{t}_i)^2}{\sum_{i=1}^{h}(t_i - \bar{t})^2}$$

- аб

Средняя абсолютная процентная ошибка (Mean Absolute Percentage Error)

$$MAPE = 100\% \cdot \frac{1}{h} \sum_{i=1}^{h} \frac{|t_i - \tilde{t}_i|}{|t_i|}$$

- аб

Корень средней квадратичной ошибки (Root Mean Square Error)

$$RMSE = \sqrt{\frac{1}{h}\sum_{i=1}^{h}(t_i - \tilde{t}_i)^2}$$

- аб

Симметричная MAPE

$$SMAPE = \frac{1}{h}\sum_{i=1}^{h} 2 \cdot \frac{|t_i - \tilde{t}_i|}{|t_i| + |\tilde{t}_i|}$$

- аб

Средняя абсолютная масштабированная ошибка (Mean absolute scaled error)

$$MASE = \frac{\frac{1}{h}\sum_{i=1}^{h}|t_i - \tilde{t}_i|}{\frac{1}{h-1}\sum_{i=2}^{h}|t_i - prev(t_i)|}$$

- Измеряет MAE в единицах фактических значений ряда, нормированную на MAE наивного прогноза (предыдущее значение ряда)
- Позволяет сравнивать качество восстановления/прогноза *разных* рядов
- Если $MASE < 1$, то прогноз лучше, чем наивный прогноз, $MASE = 1$ – не лучше, $MASE > 1$ – хуже. Например,
  - $MASE = 0.5$: MAE прогноза в 2 раза меньше, чем MAE наивного прогноза
  - $MASE = 2$: MAE прогноза в 2 раза больше, чем MAE наивного прогноза

Сценарии оценки точности восстановления

Литература

1. Khayati M., Lerner A., Tymchenko Z., Cudre-Mauroux P. Mind the gap: An experimental evaluation of imputation of missing values techniques in time series. Proc. VLDB Endow. 2020. 13(5), 768–782. https://doi.org/10.14778/3377693377383
2. Khayati M., Arous I., Tymchenko Z., Cudre-Mauroux P. ORBITS: Online recovery of missing values in multiple time series streams. Proc. VLDB Endow. 2020. 14(3). 294–306. https://dl.acm.org/doi/10.14778/3430915.3430920