

Восстановление пропущенных значений временного ряда



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

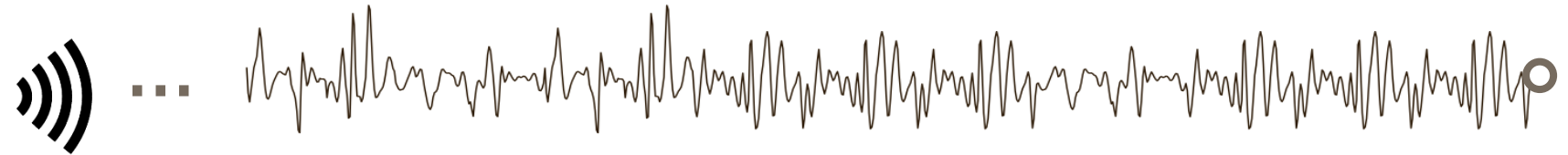
Ю. Слободенюк

Содержание

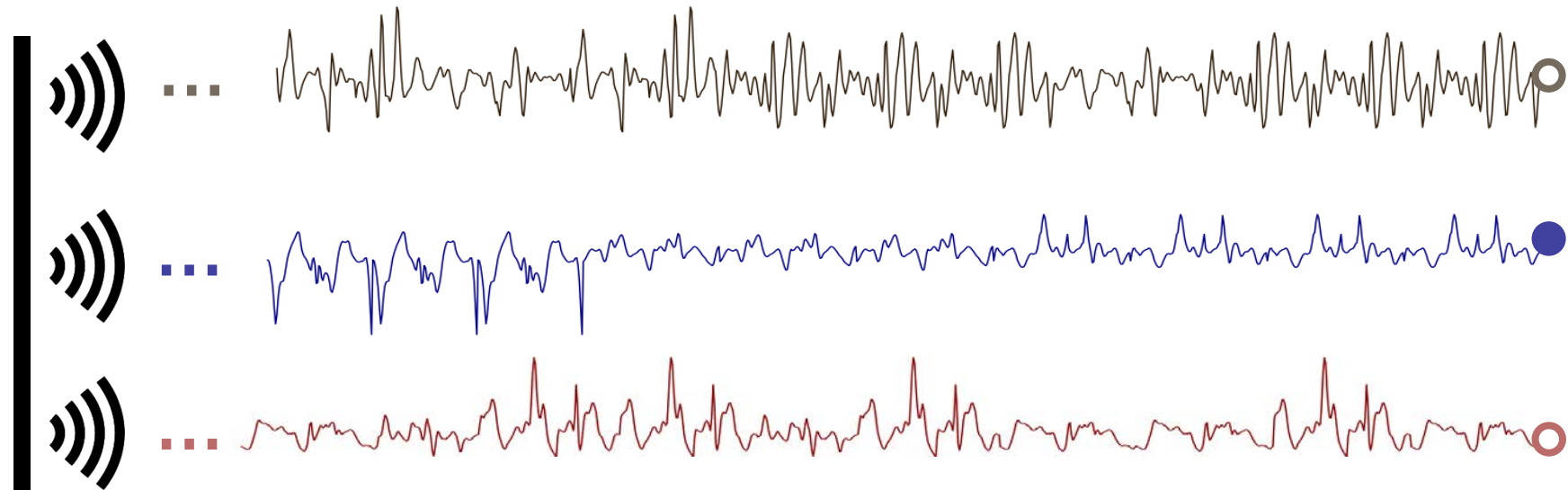
- Постановка задачи
- Классификация методов восстановления
- Аналитические методы восстановления
- Нейросетевые методы восстановления
- Оценка точности восстановления

Восстановление потокового ряда в режиме реального времени

Одномерный ряд



Многомерный ряд



Восстановление ряда (в режиме офлайн)

Одномерный ряд



Многомерный ряд



Механизмы формирования пропущенных значений

- ***MCAR (Missing Completely At Random)***, совершенно случайный пропуск

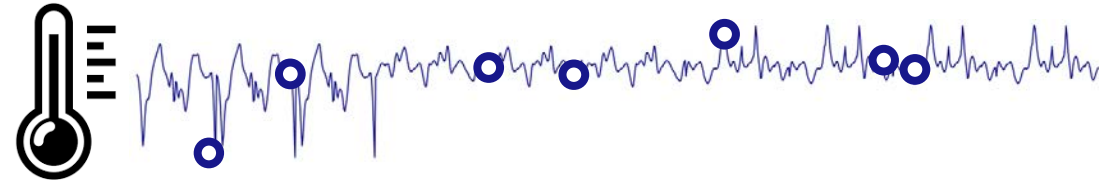
Вероятность пропуска не зависит от имеющихся и пропущенных данных

- ***MAR (Missing At Random)***, случайный пропуск

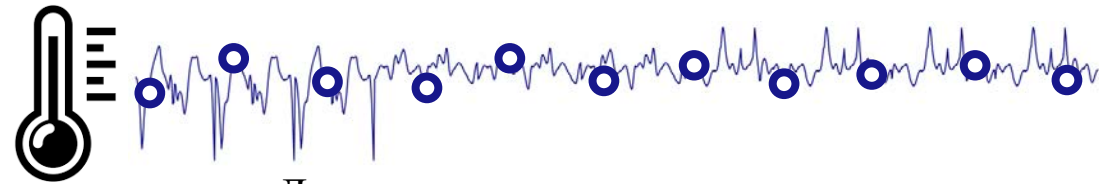
Вероятность пропуска не зависит от пропущенных данных, но зависит от имеющихся данных

- ***MNAR (Missing Not At Random)***, неслучайный пропуск

Вероятность пропуска зависит от пропущенных данных



Датчик может выходить из строя независимо от дня недели, времени суток, температуры и др.



Датчик может выходить из строя в выходные дни



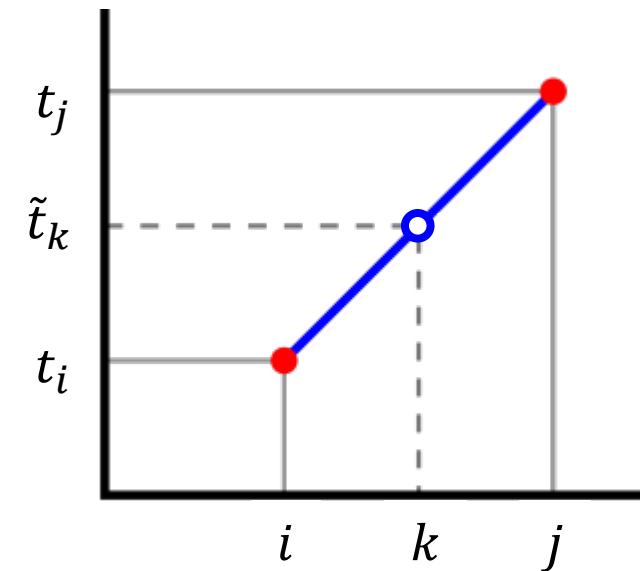
Датчик может выходить из строя в дни с очень низкой/высокой температурой

Классификация методов восстановления

- *Интерполяция* – заполнение пропусков на основе функции(-й), синтезируемой на основе известных значений
- *Сглаживание* – заполнение пропусков на основе среднего значения или др. статистических мер
- *Машинное обучение*
 - Аналитические алгоритмы (обучение без учителя)
 - Нейросетевые методы

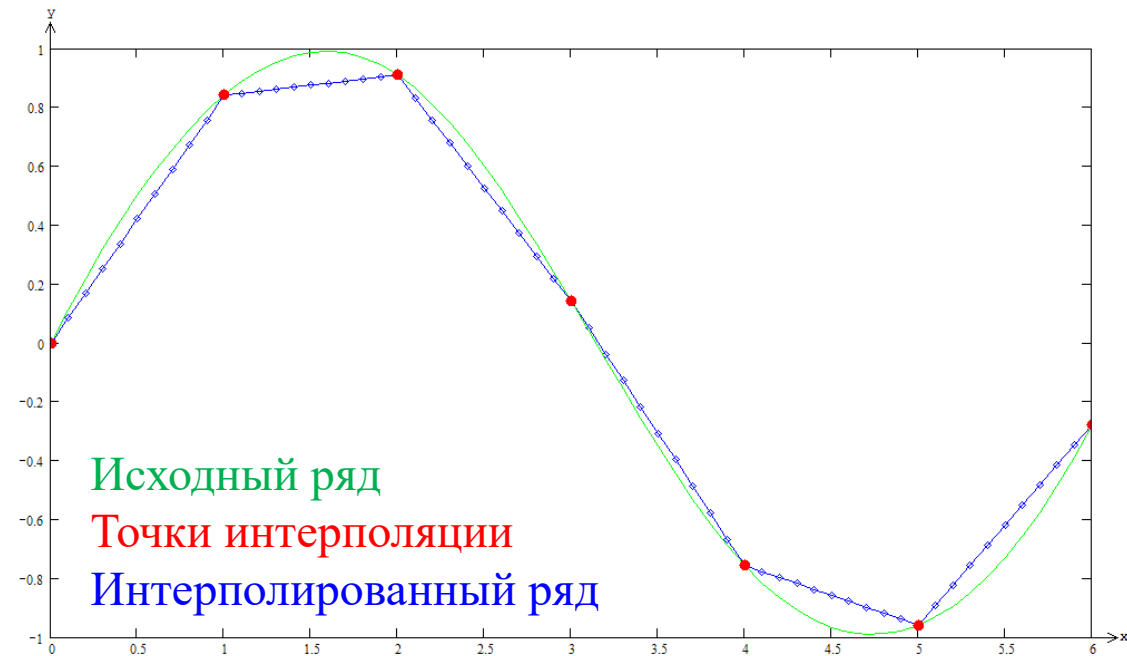
Линейная интерполяция

- На соседних точках $[i, k, j]$ ($t_k = \text{NULL}$) функция $t(k)$ заменяется прямой, проходящей через точки (i, t_i) и (j, t_j) , которая задается уравнением $\frac{t-t_i}{t_j-t_i} = \frac{k-i}{j-i}$



$$\tilde{t}_k \approx t_i + \frac{t_j - t_i}{j - i} (k - i)$$

- Приближенное представление ряда в виде кусочно-линейной функции

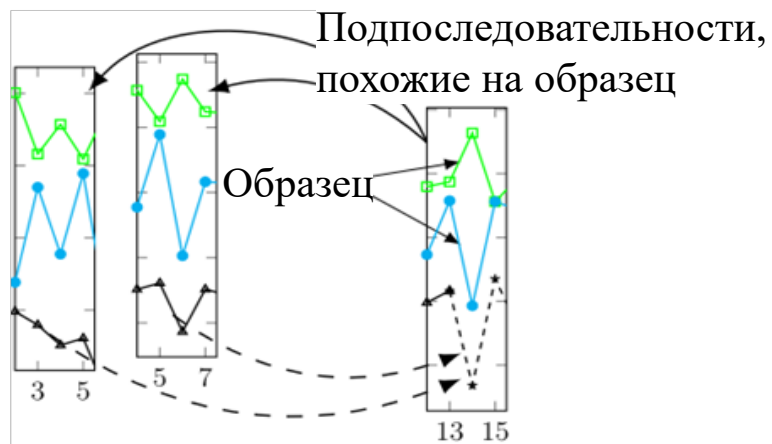
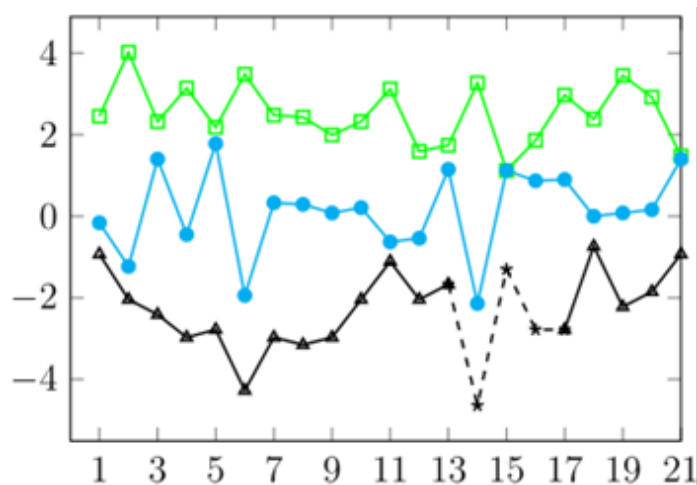


Содержание

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

Восстановление на основе шаблонов

▲ Базовый ряд □ Опорные ряды -*- Пропуски
 ● ряды



Основные идеи

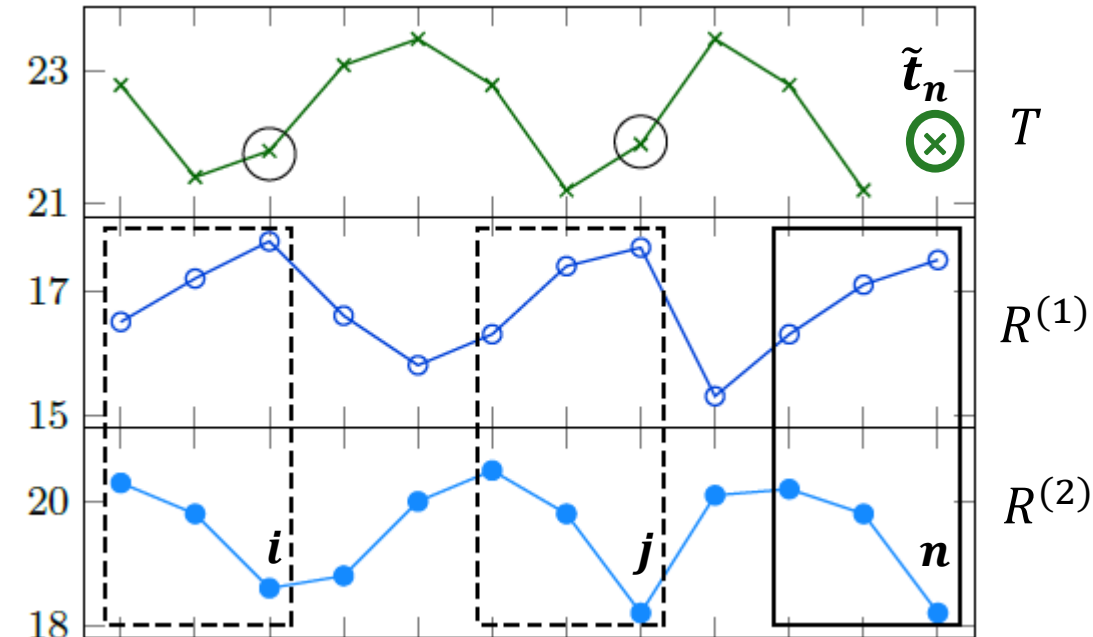
- Базовый ряд (base time series) может иметь пропуски
- Опорные ряды (reference time series) не имеют пропусков
- Высокое сходство между базовым и опорными рядами
- При обнаружении пропуска для каждого опорного ряда определяется образец – набор точек, захватывающих момент (для офлайн) или предшествующих моменту (для онлайн) пропуска
- В опорных рядах выполняется поиск по образцу
- Точки найденных шаблонов используются для синтезирования пропусков (усреднение и др.)

Параметры

- Длина образца влияет на точность и производительность: маленькая длина – потеря точности (особенно для нециклических рядов), большая длина – возрастание затрат на поиск
- Способ поиска шаблона, мера схожести, число искомых шаблонов

TKCM (Top- k Case Matching)

Wellenzohn K. *et al.* Continuous imputation of missing values in streams of pattern-determining time series. EDBT 2017.



- Найти k наиболее похожих шаблонов
 - шаблон не содержит момент пропуска
 - шаблоны не пересекаются между собой
 - отличие между шаблонами и образцами минимально

$$\Delta = \sum_{i \in A}^k \delta(R_{i,\ell}, R_{n-\ell+1,\ell}) \rightarrow \min$$

$$\delta(R_{i,\ell}, R_{n-\ell+1,\ell}) = \sqrt{\sum_{p=1}^d \sum_{q=0}^{\ell} (r_{i-q}^{(p)} - r_{n-q}^{(p)})^2}$$

- Восстановить усреднением значений в якорных точках

$$\tilde{t}_n = \frac{1}{k} \sum_{i \in A}^k t_i$$

Число шаблонов:	$k = 2$
Длина образца:	$\ell = 3$
Число опорных рядов:	$d = 2$
Якорные точки шаблонов:	i, j
Множество якорных точек:	$A, A = k$

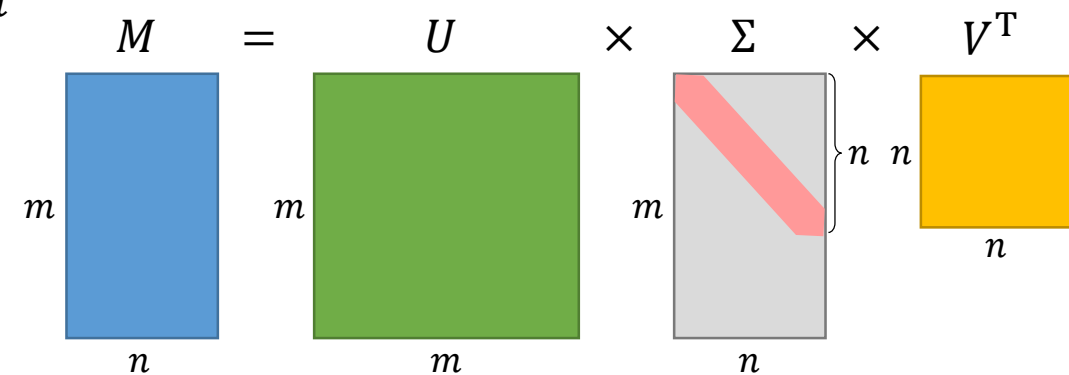
Сингулярное разложение (SVD, Singular Value Decomposition)

- Сингулярное разложение матрицы $M \in \mathbb{R}^{m \times n}$

$$M = U \cdot \Sigma \cdot V^T$$
- Матрица левых сингулярных векторов $U \in \mathbb{R}^{m \times m}$
 унитарная: $UU^T = E$
- Матрица правых сингулярных векторов $V \in \mathbb{R}^{n \times n}$
 унитарная: $VV^T = E$
- Матрица сингулярных элементов $\Sigma \in \mathbb{R}^{m \times n}$

$$\forall i \neq j \Sigma_{ij} = 0,$$

$$\forall i = j \Sigma_{ij} - \text{сингулярное число матрицы } M$$
- $\sigma \in \mathbb{R}_+$ – сингулярное число матрицы M , если:
 - $\exists u \in \mathbb{R}^m, v \in \mathbb{R}^n \ \|u\| = \|v\| = 1$
 (левый и правый сингулярные векторы);
 - $Mv = \sigma u \wedge M^T u = \sigma v$



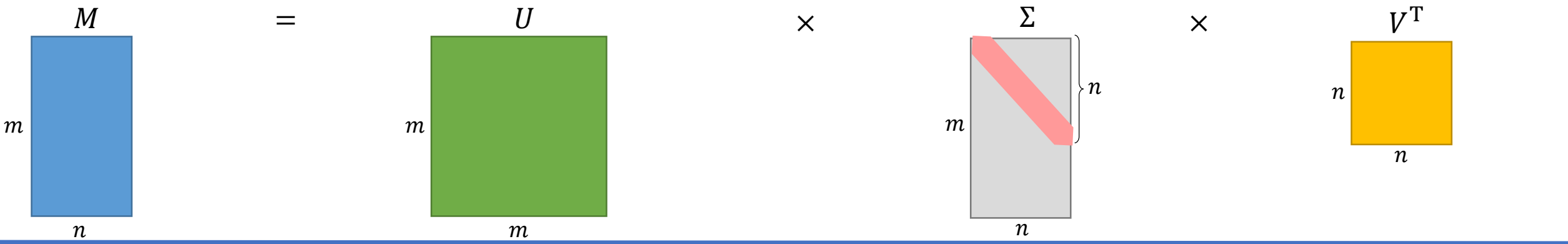
SVD: пример

$$\begin{pmatrix} 12 & 4 & 8 \\ 16 & 20 & 10 \\ 6 & 2 & 14 \\ 18 & 10 & 6 \\ 8 & 12 & 6 \end{pmatrix} = \begin{pmatrix} 0.33848 & 0.351065 & -0.397266 & 0.194769 & -0.752615 \\ 0.652286 & -0.352317 & 0.367604 & -0.524378 & -0.200726 \\ 0.275838 & 0.836263 & 0.350746 & -0.0898933 & 0.305735 \\ 0.500512 & -0.0815744 & -0.670821 & 0 & 0.54114 \\ 0.365178 & -0.215952 & 0.366092 & 0.824022 & 0.0835092 \end{pmatrix} \times \begin{pmatrix} 41.4183 & 0 & 0 \\ 0 & 11.805 & 0 \\ 0 & 0 & 8.07244 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \times \begin{pmatrix} 0.678058 & 0.0336574 & -0.734238 \\ 0.587628 & -0.62488 & 0.514022 \\ 0.441509 & 0.779995 & 0.443483 \end{pmatrix}$$

Нормализованные
собственные вектора
матрицы $M \times M^T$

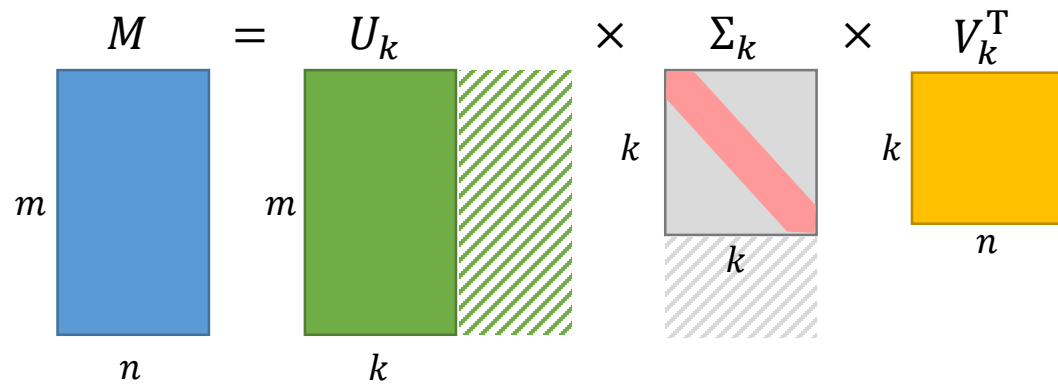
$\sigma_i = \sqrt{\lambda_i}$,
 $\lambda_1, \dots, \lambda_n$ – собственные значения
матрицы $M \times M^T$

Нормализованные
собственные вектора
матрицы $M^T \times M$

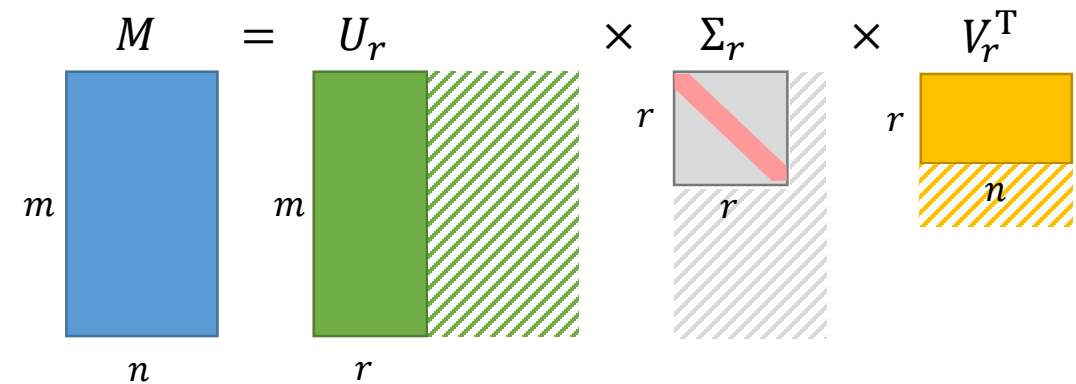


Сокращенное сингулярное разложение (Reduced SVD)

«Тонкое» разложение (Thin SVD)
 $k = \min(m, n) \ll \max(m, n)$

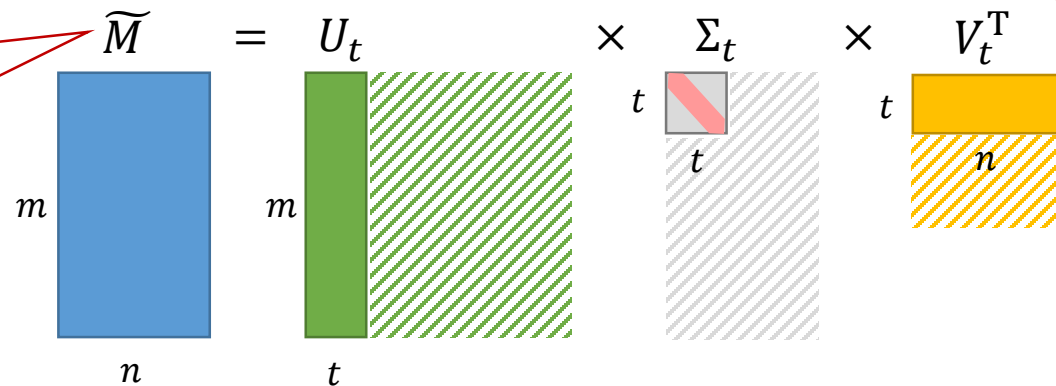


Компактное разложение (Compact SVD)
 $r \ll \min(m, n)$



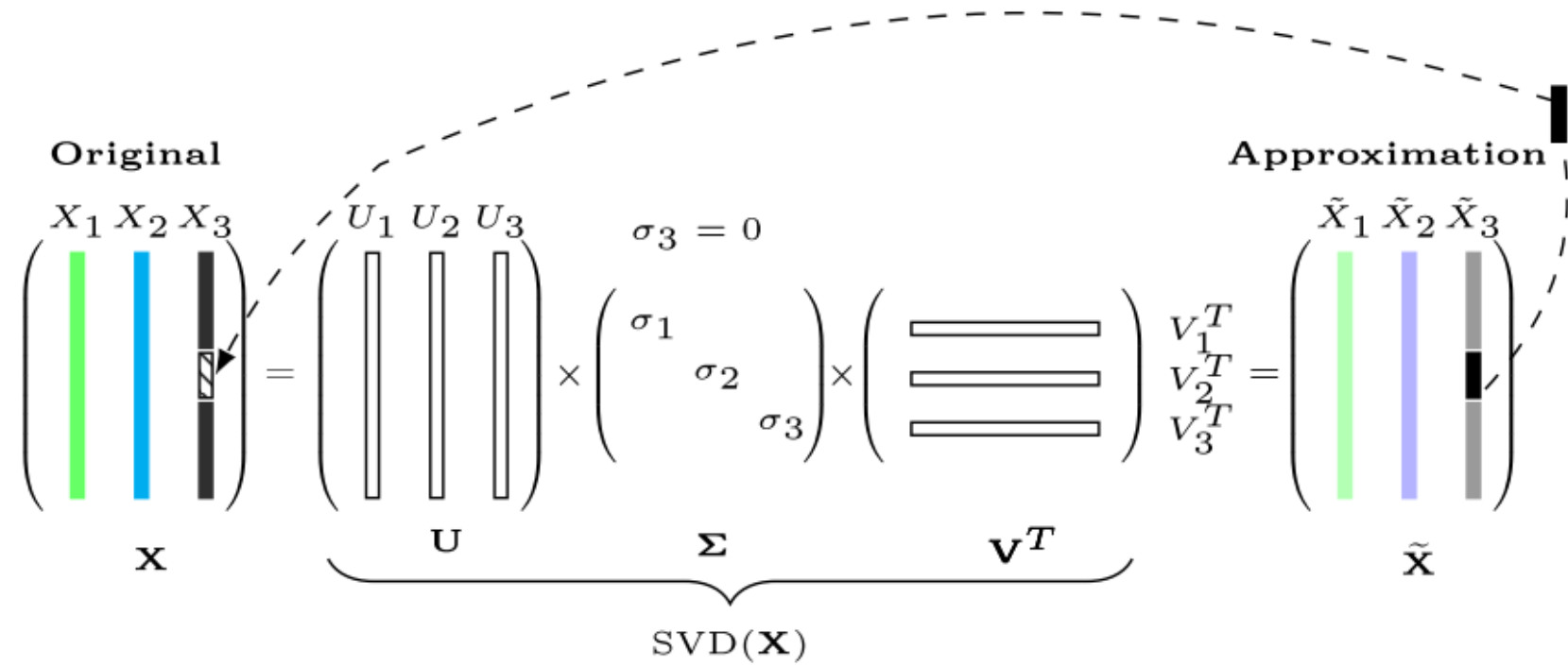
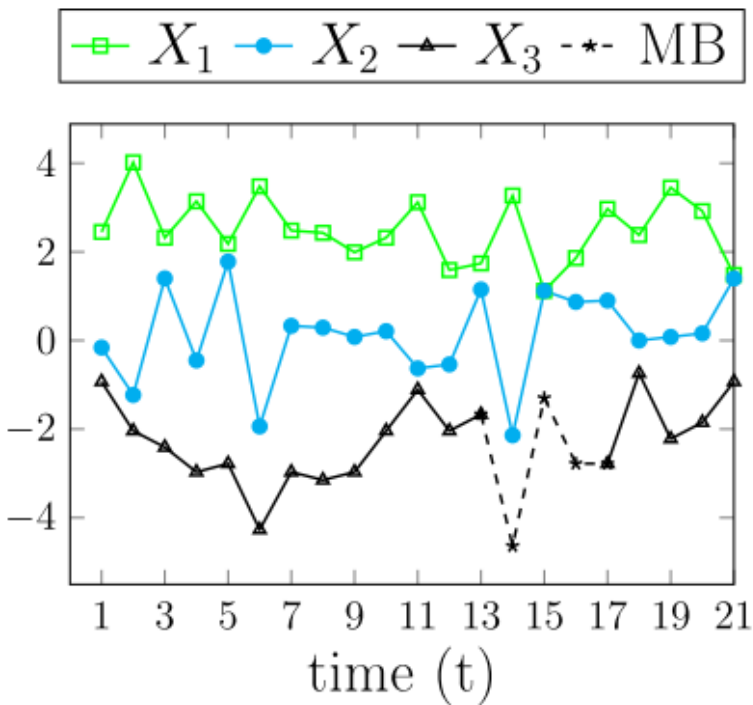
Усеченное разложение (Truncated SVD): $t \ll r$

Приближение исходной матрицы матрицей меньшего ранга



Вычисления только для t наибольших сингулярных значений (остальные отсекаются)

Восстановление на основе матричного разложения



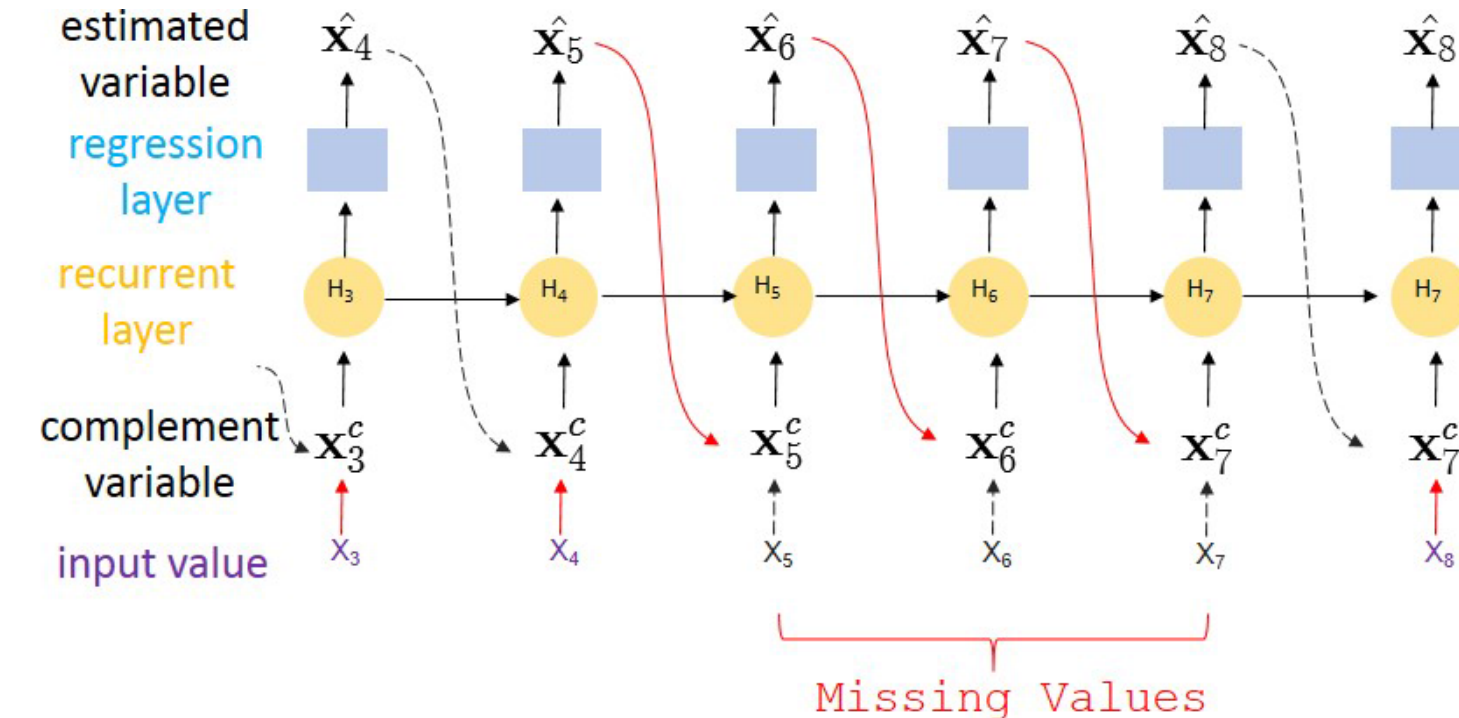
Содержание

- Постановка задачи
- Классификация методов восстановления
- 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.

https://proceedings.neurips.cc/paper_files/paper/2018/file/734e6bfcd358e25ac1db0a4241b95651-Paper.pdf

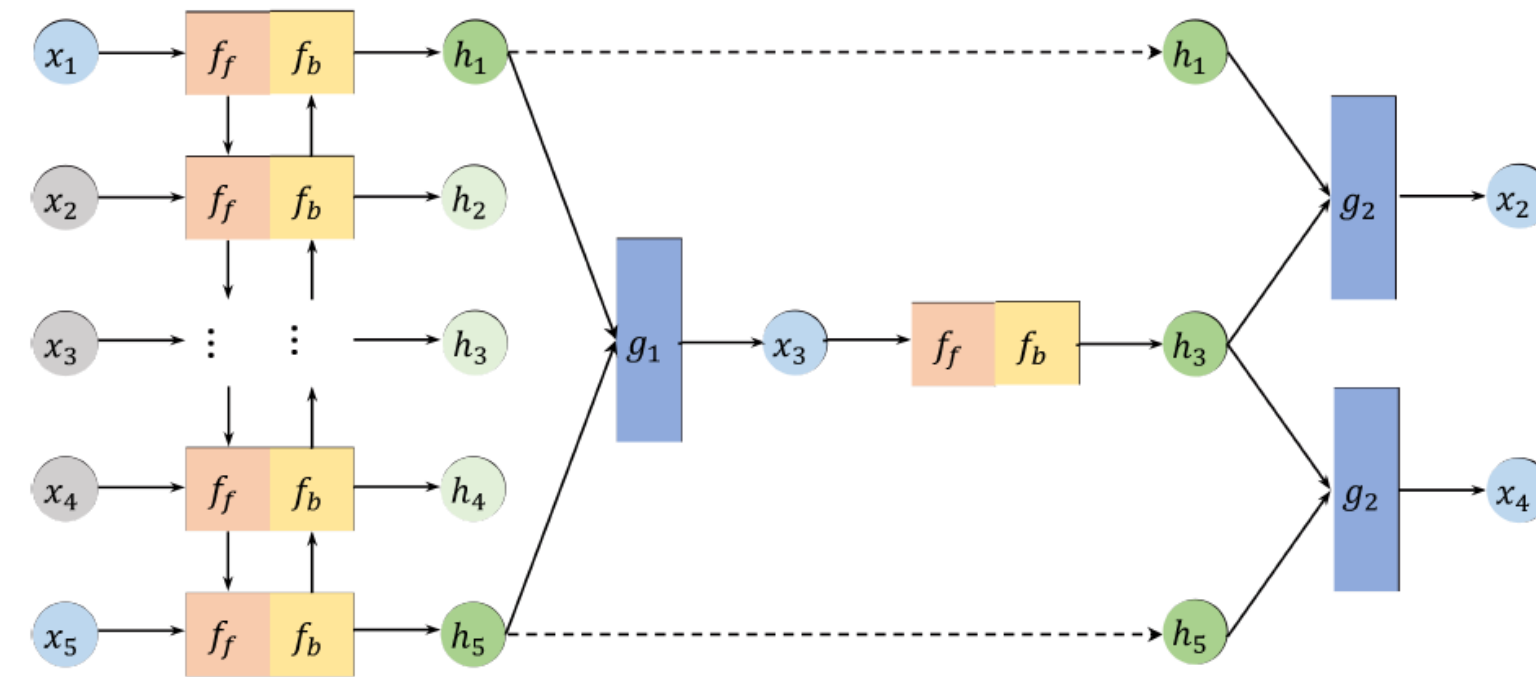


- Два слоя рекуррентных нейронов для обработки входной подпоследовательности и ее реверс-копии
- Количество нейронов в слое совпадает с длиной подпоследовательности
- Каждый нейрон прогнозирует следующую точку, учитывая все предшествующие ей точки
- Пропущенная i -я точка восстанавливается как среднее прогнозов обоих слоев по $(i - 1)$ -й точке, которое далее передается на вход $(i + 1)$ -го нейрона

NAOMI (Non-AutoRegressive Multiresolution Imputation)

Liu Y. *et al.* NAOMI: Non-autoregressive multiresolution sequence imputation. NeurIPS 2019. 11236–11246.

https://proceedings.neurips.cc/paper_files/paper/2019/file/50c1f44e426560f3f2cdbc3e19e39903-Paper.pdf



- Энкодер преобразует (x_1, \dots, x_5) в скрытые состояния
- Декодер предсказывает x_3 , используя скрытые состояния h_1, h_5
- После предсказания скрытые состояния обновляются
- Затем x_2 восстанавливается на основе x_1 и x_3
- Процесс повторяется, пока не восстановлены все пропуски

- Энкодер и Декодер представляют собой два слоя рекуррентных нейронов, направленных друг к другу
- Энкодер для каждой из двух крайних точек подпоследовательности формирует ее скрытое состояние
- Декодер, используя полученные от Энкодера значения скрытого состояния крайних точек, восстанавливает значение точки, находящейся в середине подпоследовательности
- Далее Энкодер+Декодер так же рекурсивно обрабатывает части входной подпоследовательности слева и справа от ее серединной точки

GAIN (Generative Adversarial Imputation Networks)

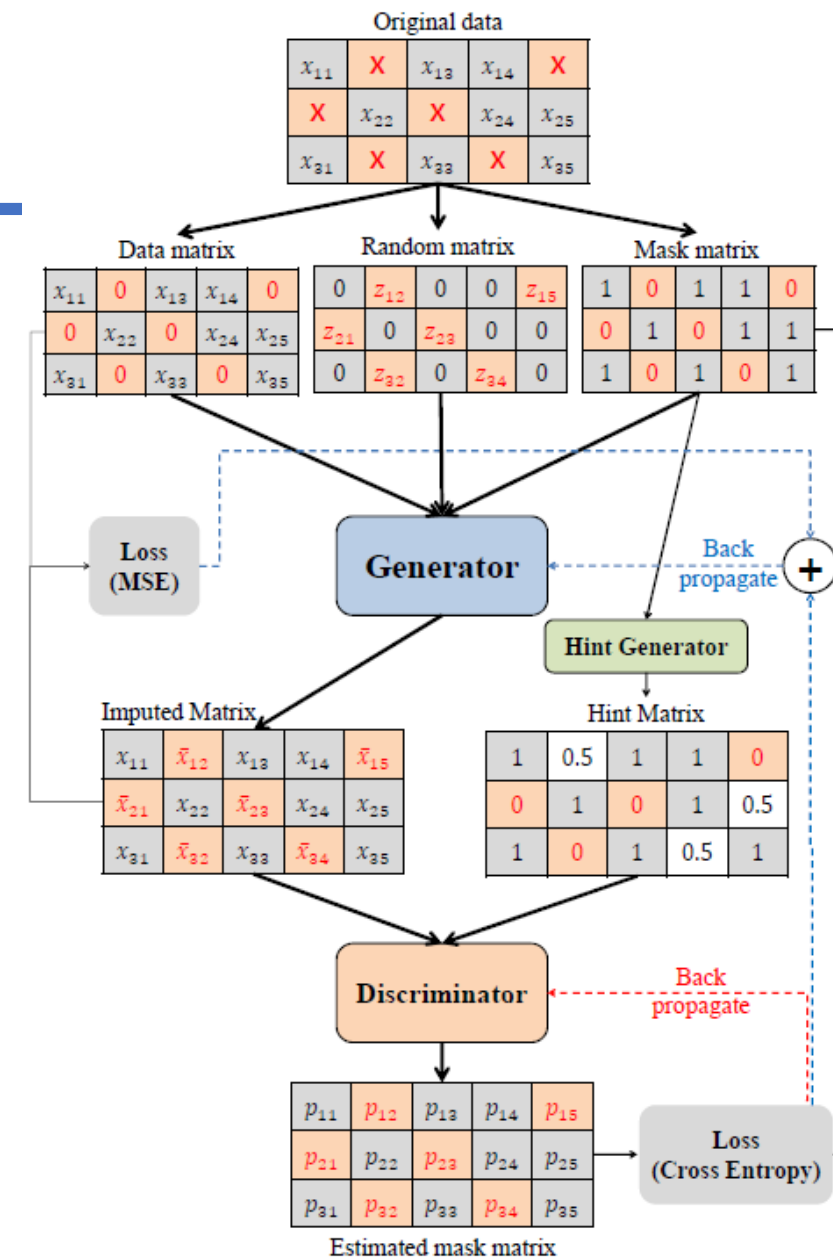
Yoon J. *et al.* GAIN: Missing data imputation using generative adversarial nets. ICML 2018. Proc. of Machine Learning Research. 2018. 80, 5675–5684 <https://arxiv.org/abs/1806.02920>

• Схема GAN

- Генератор принимает на вход вектор случайных чисел и продуцирует синтетическую подпоследовательность меньшей длины
- Дискриминатор оценивает вероятность факта, является ли некая входная подпоследовательность ряда реальной или синтетической
- Обучение Генератора и Дискриминатора осуществляется совместно:
 - сперва производится обновление весов нейронов Дискриминатора для максимизации точности классификации реальных и синтетических данных;
 - далее выполняется обновление весов нейронов Генератора для минимизации расхождения реальных и синтетических данных, полученных Генератором

• Генератор получает на вход исходную подпоследовательность, в которой пропущенные значения заменены на случайный шум, и битовую матрицу пропусков

• Дискриминатор получает на вход исходную подпоследовательность с пропусками и ее версию с восстановленными значениями и определяет степень правдоподобия восстановленных значений



Содержание

- Постановка задачи
- Классификация методов восстановления
- Регрессионные методы восстановления
- 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
R², квадрат коэффициента 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)

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

- Полезна, когда важно минимизировать большие ошибки (большие ошибки штрафуются сильнее, чем маленькие)
- Неустойчива к выбросам. Затрудняет интерпретацию результатов (ошибка в квадрате)
- Если важно минимизировать среднюю ошибку без учета величины ошибок, то лучше использовать MAE. Если важно минимизировать процентную ошибку, то лучше использовать MAPE или SMAPE

Средняя абсолютная ошибка (Mean Absolute Error)

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

- Полезна, когда важно минимизировать среднюю ошибку без учета величины ошибок и когда выбросы не являются серьезной проблемой
- Преимущества по сравнению с MSE
 - Более устойчива к выбросам: не штрафует большие ошибки сильнее, чем маленькие
 - Облегчает интерпретацию результатов: измеряет ошибку в тех же единицах, что и исходные данные
- Недостатки по сравнению с MSE
 - Может быть менее чувствительна к изменениям в данных и не давать достаточно большого веса большим ошибкам

Коэффициент детерминации

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

- Измеряет долю дисперсии спрогнозированной/восстановленной части ряда в общей дисперсии ряда. Соответствует нормированной среднеквадратичной ошибке
- Если $R^2 \approx 1$, то модель/алгоритм хорошо объясняет данные. Если $R^2 \approx 0$, то качество прогноза сопоставимо с константным предсказанием
- Завышенное качество при малом количестве данных или наличии в данных выбросов и/или неслучайных ошибок (систематические ошибки измерения и др.). Неадекватное сравнение моделей при использовании ими разных наборов переменных

Средняя абсолютная процентная ошибка (Mean Absolute Percentage Error)

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

- Метрика с простой интерпретацией: ошибка прогноза составляет MAPE от фактических значений
- Полезна, когда прогнозные значения имеют разный масштаб или положительные и отрицательные отклонения прогноза от фактических значений влияют на результат одинаково
- Неустойчива к выбросам. Может давать некорректные результаты, когда фактические значения близки к нулю

Корень средней квадратичной ошибки (Root Mean Square Error)

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

- Среднеквадратическое отклонение прогноза от фактического значения. Чем меньше RMSE, тем лучше качество прогноза
- Неустойчива к выбросам. Может давать некорректные результаты, когда фактические значения близки к нулю
- Одна из наиболее часто используемых метрик

Симметричная MAPE (SMAPE, Symmetric MAPE)

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

- Средняя относительная ошибка прогноза в процентах. Чем меньше SMAPE, тем лучше качество прогноза
- Полезна, когда прогнозные значения имеют разный масштаб или положительные и отрицательные отклонения прогноза от фактических значений влияют на результат одинаково
- Неустойчива к выбросам. Может давать некорректные результаты, когда фактические значения близки к нулю

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

Метрики точности восстановления: резюме

Метрика	Применение и недостатки
MSE	<ul style="list-style-type: none"> Важно минимизировать большие ошибки Неустойчива к выбросам, может давать завышенную оценку качества модели
MAE	<ul style="list-style-type: none"> Важно минимизировать ср. ошибку, независимо от ее размера. Менее чувствительна к выбросам, чем MSE Менее чувствительна к изменениям в данных, чем MSE. Может не давать большого веса большим ошибкам
R ²	<ul style="list-style-type: none"> Оценка соответствия модели данным в целом. Сравнение качества нескольких моделей Неустойчива к выбросам. Отсутствие возможности сравнения моделей с разными наборами переменных
MAPE	<ul style="list-style-type: none"> Важно минимизировать процентную ошибку. Сравнение качества моделей на разных временных рядах Неустойчива к выбросам, может давать некорректные результаты, когда фактические значения близки к нулю
RMSE	<ul style="list-style-type: none"> Важно минимизировать большие ошибки. Сравнение качества моделей на одном временном ряде Неустойчива к выбросам, может давать некорректные результаты, когда фактические значения близки к нулю
SMAPE	<ul style="list-style-type: none"> Важно минимизировать процентную ошибку. Сравнение качества моделей на одном временном ряде Неустойчива к выбросам, может давать некорректные результаты, когда фактические значения близки к нулю
MASE	<ul style="list-style-type: none"> Сравнение качества модели с качеством наивного прогноза. Сравнение качества моделей на разных рядах Неустойчива к выбросам

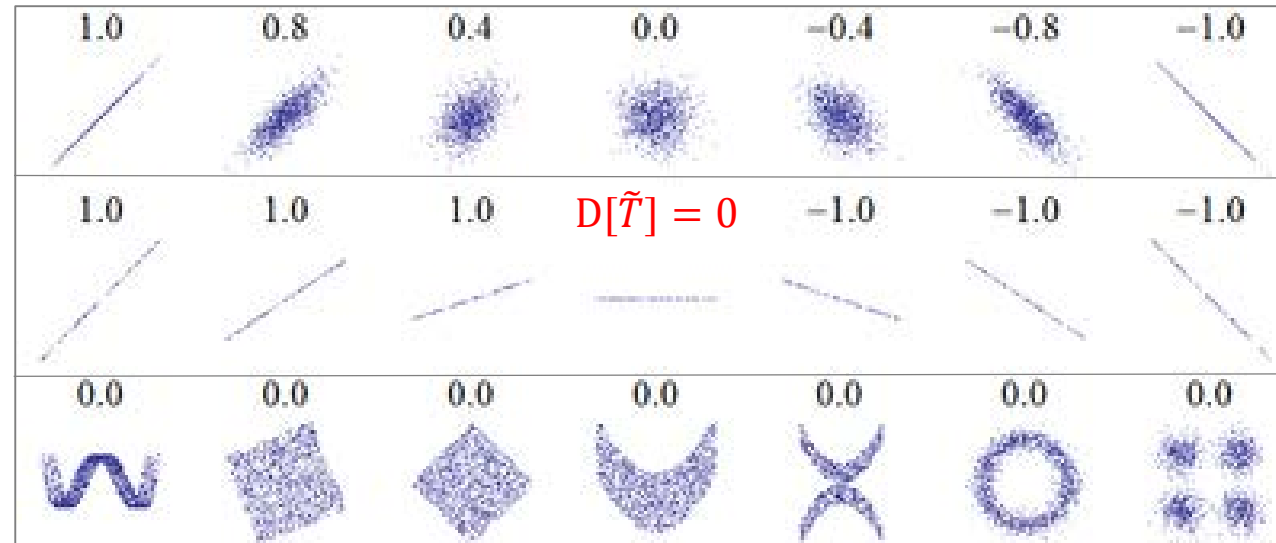
Дополнительный показатель: корреляции Пирсона и Спирмена

$$Pearson(T, \tilde{T}) = \frac{\sum_{i=1}^n t_i \tilde{t}_i - n\mu_T \mu_{\tilde{T}}}{n\sigma_T \sigma_{\tilde{T}}}$$

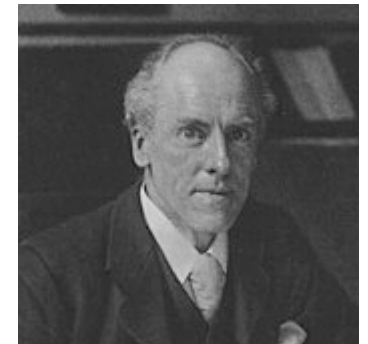
$$Spearman(T, \tilde{T}) = Pearson(r(T), r(\tilde{T}))$$

Шкала Чеддока силы корреляции

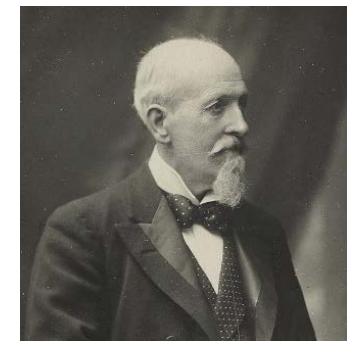
- +1.0** идеальная положительная
- +0.9** очень сильная положительная
- +0.7** сильная положительная
- +0.4** умеренная положительная
- +0.1** слабая положительная
- 0.0** отсутствие корреляции
- 0.1** слабая отрицательная
- 0.4** умеренная отрицательная
- 0.7** сильная отрицательная
- 0.9** очень сильная отрицательная
- 1.0** идеальная отрицательная



1-я строка: линейная зависимость
 2-я строка: линейная зависимость с наклоном
 3-я строка: нелинейная зависимость

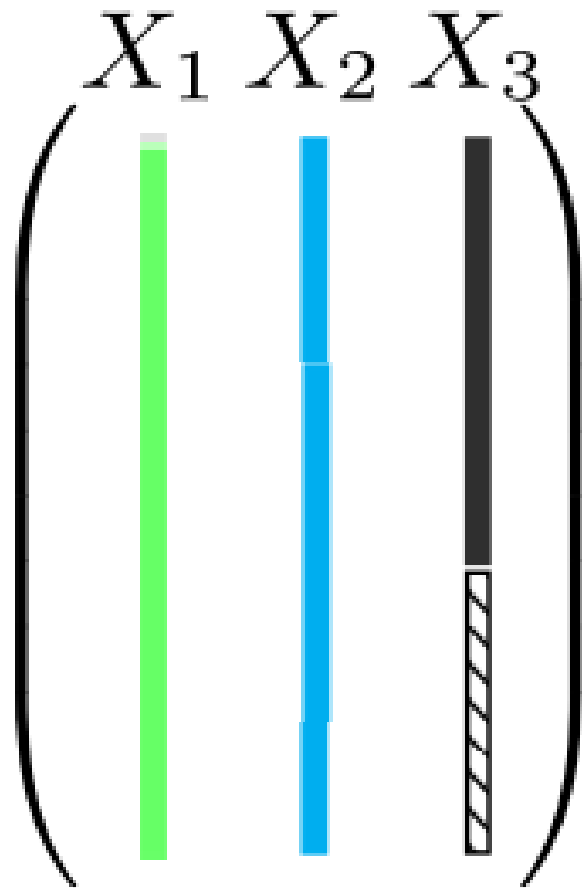


Карл Пирсон
 (Karl Pearson)
 1857-1936



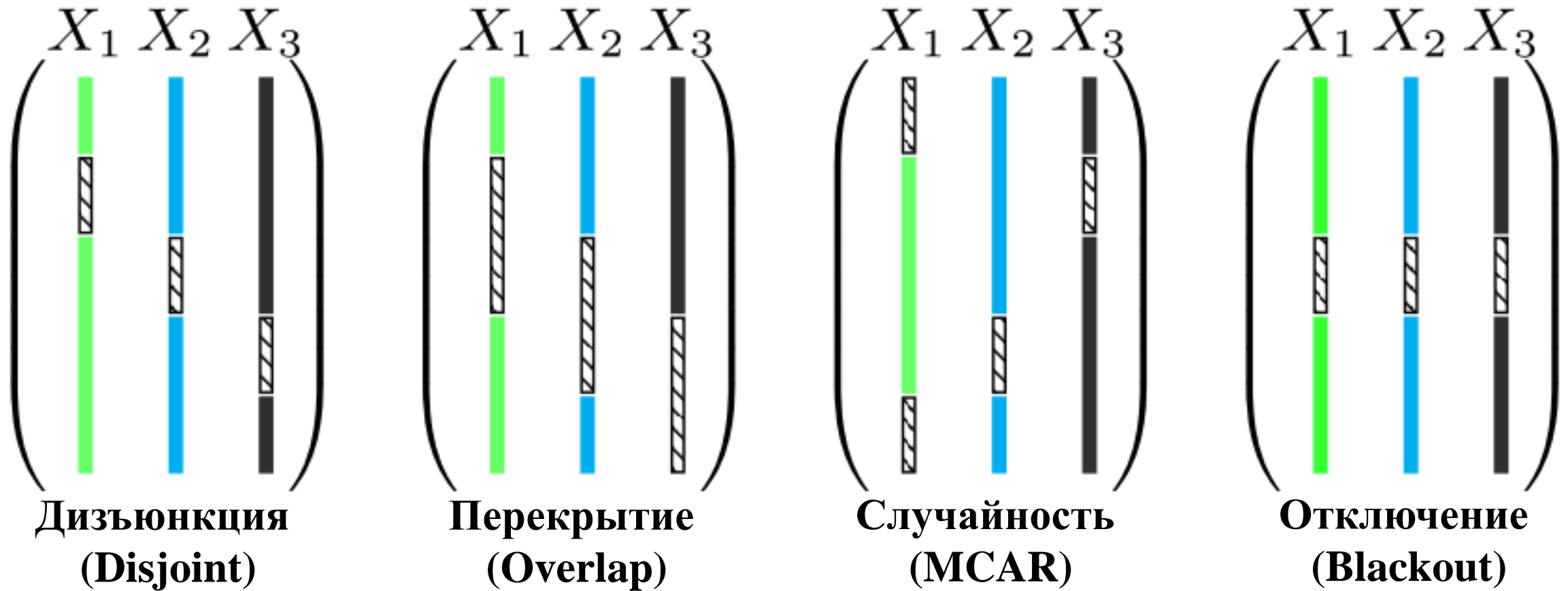
Чарльз Спирмен
 (Charles Spearman)
 1863-1945

Сценарии оценки точности восстановления: режим реального времени



- Гиперпараметры:
 - длина блока пропущенных значений (доля длины ряда)
 - количество временных рядов-координат, в которых имеются блоки пропущенных значений (доля от размерности)
- Обучающая и тестовая выборки не должны пересекаться

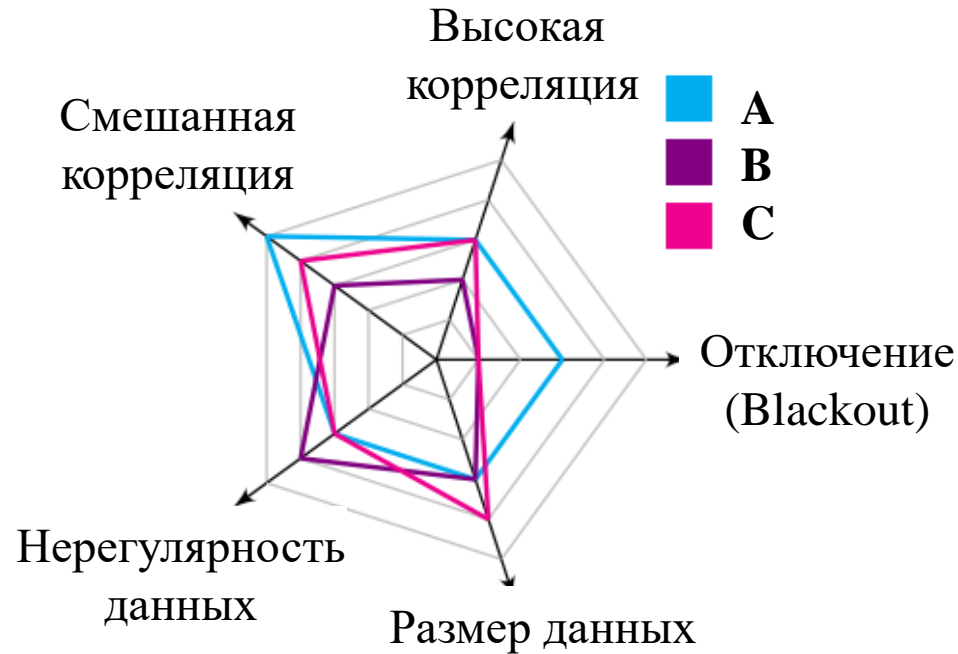
Сценарии оценки точности восстановления: режим офлайн



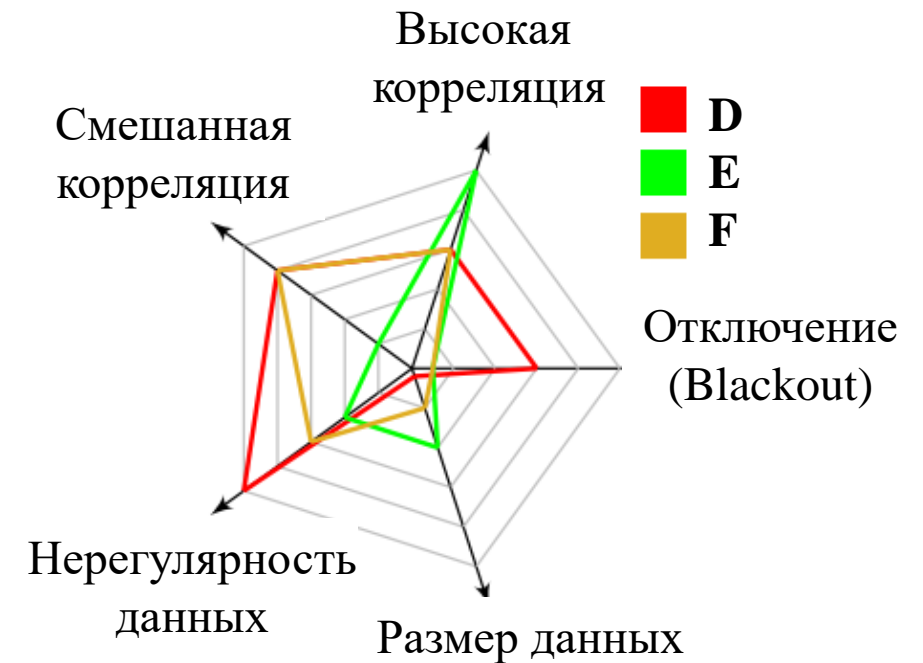
Гиперпараметры: длина блока, количество координат

Подбор алгоритма восстановления

Точность восстановления



Быстрота восстановления



- **Высокая корреляция:** между рядами-координатами
- **Смешанная корреляция:** много случаев однократной высокой/низкой и положительной/отрицательной корреляции
- **Нерегулярность данных:** флуктуации, выбросы, пики и др.
- **Отключение:** режим Blackout
- **Размер данных:** количество и длина рядов

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/10.14778/33773693377383>
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>
3. Fang C., Wang C. Time series data imputation: A survey on deep learning approaches. <https://doi.org/10.48550/arXiv.2011.11347>