Пропуски, аномалии и структурные сдвиги

Заполнение пропусков

Заполнение пропусков: план

- Линейная интерполяция.

Заполнение пропусков: план

- Линейная интерполяция.
- Модели для заполнения пропусков.

Заполнение пропусков: план

- Линейная интерполяция.
- Модели для заполнения пропусков.
- Использование STL-разложения.

Линейная интерполяция

Идея

Заполним пропуски так, чтобы восстановленные значения идеально ложились на прямую (образовывали арифметическую прогрессию),

$$\Delta y_t^{imp} = const.$$

Линейная интерполяция

Идея

Заполним пропуски так, чтобы восстановленные значения идеально ложились на прямую (образовывали арифметическую прогрессию),

$$\Delta y_t^{imp} = const.$$

Пример:

10, NA, NA, 100.

Линейная интерполяция

Идея

Заполним пропуски так, чтобы восстановленные значения идеально ложились на прямую (образовывали арифметическую прогрессию),

$$\Delta y_t^{imp} = const.$$

Пример:

10, NA, NA, 100.

10, 40, 70, 100

Модели для заполнения пропусков

1. Оцениваем модель, допускающую пропуски в данных. ARIMA подходит! И автоматическая ARIMA тоже!

Модели для заполнения пропусков

1. Оцениваем модель, допускающую пропуски в данных. ARIMA подходит! И автоматическая ARIMA тоже!
2. Пропущенные значения y_t заменяем на условное математическое ожидание, полагая оценённые параметры модели равным истинными,

$$y_t^{imp} = \mathbb{E}(y_t \mid \text{данные}).$$

Используется фильтр Калмана.

Модели для заполнения пропусков

1. Оцениваем модель, допускающую пропуски в данных. ARIMA подходит! И автоматическая ARIMA тоже!
2. Пропущенные значения y_t заменяем на условное математическое ожидание, полагая оценённые параметры модели равным истинными,

$$y_t^{imp} = \mathbb{E}(y_t \mid \text{данные}).$$

Используется фильтр Калмана.

Модели для заполнения пропусков

1. Оцениваем модель, допускающую пропуски в данных. ARIMA подходит! И автоматическая ARIMA тоже!
2. Пропущенные значения y_t заменяем на условное математическое ожидание, полагая оценённые параметры модели равным истинными,

$$y_t^{imp} = \mathbb{E}(y_t \mid \text{данные}).$$

   Используется фильтр Калмана.

Возможность оценивать модель на данных с пропусками сильно зависит от реализации.

Использование STL-разложения

1. Раскладываем ряд с пропусками на составляющие:

$$y_t = \text{trend}_t + \text{seasonal}_t + \text{remainder}_t = \text{seasonal}_t + \text{deseason}_t.$$

STL восстанавливает сезонную компоненту без пропусков!

Использование STL-разложения

1. Раскладываем ряд с пропусками на составляющие:

   $$y_t = \text{trend}_t + \text{seasonal}_t + \text{remainder}_t = \text{seasonal}_t + \text{deseason}_t.$$

   STL восстанавливает сезонную компоненту без пропусков!
2. Восстанавливаем пропущенные значения десезонированного ряда линейной интерполяцией.

Использование STL-разложения

1. Раскладываем ряд с пропусками на составляющие:

   $$y_t = \text{trend}_t + \text{seasonal}_t + \text{remainder}_t = \text{seasonal}_t + \text{deseason}_t.$$

   STL восстанавливает сезонную компоненту без пропусков!
2. Восставливаем пропущенные значения десезонированного ряда линейной интерполяцией.
3. Пропущенные значения y_t заменяем на сумму восстановленных десезонированных значений и сезонной составляющей,

   $$y_t^{imp} = \text{seasonal}_t + \text{deseason}_t^{imp}.$$

Зачем заполнять пропуски?

- Иногда заполнение пропусков — основная задача.

Зачем заполнять пропуски?

- Иногда заполнение пропусков — основная задача.
- Возможность использовать больше алгоритмов прогнозирования для восстановленного ряда.

Зачем заполнять пропуски?

- Иногда заполнение пропусков — основная задача.
- Возможность использовать больше алгоритмов прогнозирования для восстановленного ряда.
- Возможность использовать восстановленный ряд как предиктор.

Заполнение пропусков: итоги

- Линейная интерполяция: просто и быстро!

Заполнение пропусков: итоги

- Линейная интерполяция: просто и быстро!
- Использование ARIMA или более сложных моделей.

Заполнение пропусков: итоги

- Линейная интерполяция: просто и быстро!
- Использование ARIMA или более сложных моделей.
- STL-разложение и восстановление компонент.

Заполнение пропусков: итоги

- Линейная интерполяция: просто и быстро!
- Использование ARIMA или более сложных моделей.
- STL-разложение и восстановление компонент.
- Вариации у каждого алгоритма.

Байесовские структурные модели

Байесовские структурные модели: план

- Как добавить предикторы в ETS?

Байесовские структурные модели: план

- Как добавить предикторы в ETS?
- Идея байесовского подхода.

Байесовские структурные модели: план

- Как добавить предикторы в ETS?
- Идея байесовского подхода.
- Пример структурной модели.

Как добавить предикторы в ETS?

- Классическая ETS модель не позволяет включать предикторы.

Как добавить предикторы в ETS?

- Классическая ETS модель не позволяет включать предикторы.
- А что мешает их туда добавить и получить новую модель?

Как добавить предикторы в ETS?

- Классическая ETS модель не позволяет включать предикторы.
- А что мешает их туда добавить и получить новую модель? В новой модели может оказаться слишком много параметров.

Как добавить предикторы в ETS?

- Классическая ETS модель не позволяет включать предикторы.
- А что мешает их туда добавить и получить новую модель? В новой модели может оказаться слишком много параметров. Качество прогнозов может быть плохим.

Как добавить предикторы в ETS?

- Классическая ETS модель не позволяет включать предикторы.
- А что мешает их туда добавить и получить новую модель?
  В новой модели может оказаться слишком много параметров.
  Качество прогнозов может быть плохим.
- Спасительная идея — регуляризация.

Как добавить предикторы в ETS?

- Классическая ETS модель не позволяет включать предикторы.
- А что мешает их туда добавить и получить новую модель?
  В новой модели может оказаться слишком много параметров.
  Качество прогнозов может быть плохим.
- Спасительная идея — регуляризация.
  Рассматриваем модель с большим числом параметров и дополнительной информацией, что параметры небольшие.

Байесовский подход!

- Трактуем все параметры как ненаблюдаемые случайные величины, $\theta = (a, b, c)$.

Байесовский подход!

- Трактуем все параметры как ненаблюдаемые случайные величины, $\theta = (a, b, c)$.
- Модель задаёт распределение ряда при заданных параметрах,

$$y_t = a + u_t + bu_{t-1}, \quad u_t \sim \mathcal{N}(0; c).$$

Байесовский подход!

- Трактуем все параметры как ненаблюдаемые случайные величины, $\theta = (a, b, c)$.
- Модель задаёт распределение ряда при заданных параметрах,

$$y_t = a + u_t + bu_{t-1}, \quad u_t \sim \mathcal{N}(0; c).$$

- Добавляем информацию в виде априорного распределения,

$$a \sim \mathcal{N}(0; 100), \quad b \sim \mathcal{N}(0, 1), \quad \ln c \sim \mathcal{N}(0; 4).$$

Байесовский подход!

- Трактуем все параметры как ненаблюдаемые случайные величины, $\theta = (a, b, c)$.
- Модель задаёт распределение ряда при заданных параметрах,

$$y_t = a + u_t + bu_{t-1}, \quad u_t \sim \mathcal{N}(0; c).$$

- Добавляем информацию в виде априорного распределения,

$$a \sim \mathcal{N}(0; 100), \quad b \sim \mathcal{N}(0, 1), \quad \ln c \sim \mathcal{N}(0; 4).$$

- Алгоритмы MCMC (Markov Chain Monte Carlo) позволяют сгенерировать большую выборку из апостериорное распределение

$$(a, b, c \mid y_1, y_2, \ldots, y_T).$$

Немного про MCMC

- Выборка из апостериорного распределения $(\theta \mid y)$ позволяет считать всё!

$$(a_1, b_1, c_1), (a_2, b_2, c_2), \ldots, (a_S, b_S, c_S).$$

Немного про MCMC

- Выборка из апостериорного распределения $(\theta \mid y)$ позволяет считать всё!

$$(a_1, b_1, c_1), (a_2, b_2, c_2), \ldots, (a_S, b_S, c_S).$$

- Имея значения параметров можно симулировать будущие траектории.

Немного про MCMC

- Выборка из апостериорного распределения $(\theta \mid y)$ позволяет считать всё!

$$(a_1, b_1, c_1), (a_2, b_2, c_2), \ldots, (a_S, b_S, c_S).$$

- Имея значения параметров можно симулировать будущие траектории.
- MCMC позволяет работать с моделями фантастической сложности.

Немного про MCMC

- Выборка из апостериорного распределения $(\theta \mid y)$ позволяет считать всё!

$$(a_1, b_1, c_1), (a_2, b_2, c_2), \ldots, (a_S, b_S, c_S).$$

- Имея значения параметров можно симулировать будущие траектории.
- MCMC позволяет работать с моделями фантастической сложности.
- MCMC работает медленно.

Немного про MCMC

- Выборка из апостериорного распределения $(\theta \mid y)$ позволяет считать всё!

$$(a_1, b_1, c_1), (a_2, b_2, c_2), \ldots, (a_S, b_S, c_S).$$

- Имея значения параметров можно симулировать будущие траектории.
- MCMC позволяет работать с моделями фантастической сложности.
- MCMC работает медленно.
- Генерируемая выборка только в пределе похожа на выборку из апостериорного закона.

Конструктор структурных моделей

Байесовские структурные модели: итоги

- Трактуем параметры как случайные величины.

Байесовские структурные модели: итоги

- Трактуем параметры как случайные величины.
- Конструируем из кубиков настоящего монстра с предикторами.

Байесовские структурные модели: итоги

- Трактуем параметры как случайные величины.
- Конструируем из кубиков настоящего монстра с предикторами.
- MCMC работает медленно.

Байесовские структурные модели: итоги

- Трактуем параметры как случайные величины.
- Конструируем из кубиков настоящего монстра с предикторами.
- MCMC работает медленно.
- С помощью MCMC можно сгенерировать большую выборку из апостериорного распределения параметров.

Как оценить эффект воздействия?

Оценивание эффекта: план

- Основная идея оценивания.

Оценивание эффекта: план

- Основная идея оценивания.
- Чем хорош байесовский подход?

Оценивание эффекта: идея

- Возьмём любую модель.

Оценивание эффекта: идея

- Возьмём любую модель.
- Разделим выборку на обучающую и тестовую по точке воздействия.

Оценивание эффекта: идея

- Возьмём любую модель.
- Разделим выборку на обучающую и тестовую по точке воздействия.
- Оценим модель по обучающей выборке и получим прогноз.

Оценивание эффекта: идея

- Возьмём любую модель.
- Разделим выборку на обучающую и тестовую по точке воздействия.
- Оценим модель по обучающей выборке и получим прогноз.
- Разница прогноза и фактических значений на тестовой выборке оценивает эффект воздействия.

Нюансы идеи

- Модель должна хорошо прогнозировать.

Нюансы идеи

- Модель должна хорошо прогнозировать.
  Если сезонность сильна, то даже наивная сезонная модель подойдёт.

Нюансы идеи

- Модель должна хорошо прогнозировать.
  Если сезонность сильна, то даже наивная сезонная модель подойдёт.
- Разумно протестировать идею в точке, где воздействия ещё нет.

Нюансы идеи

- Модель должна хорошо прогнозировать.
  Если сезонность сильна, то даже наивная сезонная модель подойдёт.
- Разумно протестировать идею в точке, где воздействия ещё нет.
- Можно оценивать эффект воздействия на конкретное значение y_{T+h}, а можно оценивать кумулятивный эффект на $y_{T+1} + y_{T+2} + \ldots + y_{T+h}$.

Нюансы идеи

- Модель должна хорошо прогнозировать.
  Если сезонность сильна, то даже наивная сезонная модель подойдёт.
- Разумно протестировать идею в точке, где воздействия ещё нет.
- Можно оценивать эффект воздействия на конкретное значение y_{T+h}, а можно оценивать кумулятивный эффект на $y_{T+1} + y_{T+2} + \ldots + y_{T+h}$.
- Помните о доверительных интервалах.

А зачем тут байесовский подход?

- Позволяет оценивать более сложные модели с предикторами.

А зачем тут байесовский подход?

- Позволяет оценивать более сложные модели с предикторами.
  Возможность более точно оценить эффект воздействия.

А зачем тут байесовский подход?

- Позволяет оценивать более сложные модели с предикторами.
  Возможность более точно оценить эффект воздействия.
- Позволяет построить доверительный интервал для кумулятивного эффекта.

А зачем тут байесовский подход?

- Позволяет оценивать более сложные модели с предикторами.
  Возможность более точно оценить эффект воздействия.
- Позволяет построить доверительный интервал для кумулятивного эффекта.
  Проблема: прогнозы $\hat{y}_{T+1}, \ldots, \hat{y}_{T+h}$ нетривиально коррелированы между собой. В частотном подходе сложно получить явную формулу для доверительного интервала.

А зачем тут байесовский подход?

- Позволяет оценивать более сложные модели с предикторами.
  Возможность более точно оценить эффект воздействия.
- Позволяет построить доверительный интервал для кумулятивного эффекта.
  Проблема: прогнозы $\hat{y}_{T+1}, \ldots, \hat{y}_{T+h}$ нетривиально коррелированы между собой. В частотном подходе сложно получить явную формулу для доверительного интервала.
  Решение: в байесовском подходе апостериорная выборка параметров модели позволяет сгенерировать множество гипотетических будущих траекторий без воздействия.

А зачем тут байесовский подход?

- Позволяет оценивать более сложные модели с предикторами.
  Возможность более точно оценить эффект воздействия.
- Позволяет построить доверительный интервал для кумулятивного эффекта.
  Проблема: прогнозы $\hat{y}_{T+1}, \ldots, \hat{y}_{T+h}$ нетривиально коррелированы между собой. В частотном подходе сложно получить явную формулу для доверительного интервала.
  Решение: в байесовском подходе апостериорная выборка параметров модели позволяет сгенерировать множество гипотетических будущих траекторий без воздействия.
- Графики будущих траекторий!

Оценивание эффекта воздействия: итоги

- Делим выборку на обучающую и тестовую в точке воздействия.

Оценивание эффекта воздействия: итоги

- Делим выборку на обучающую и тестовую в точке воздействия.
- Смотрим на разницу прогноза модели и фактических значений.

Оценивание эффекта воздействия: итоги

- Делим выборку на обучающую и тестовую в точке воздействия.
- Смотрим на разницу прогноза модели и фактических значений.
- Байесовских подход дарит доверительные интервалы для кумулятивного эффекта.