W07 Как выбрать модель?

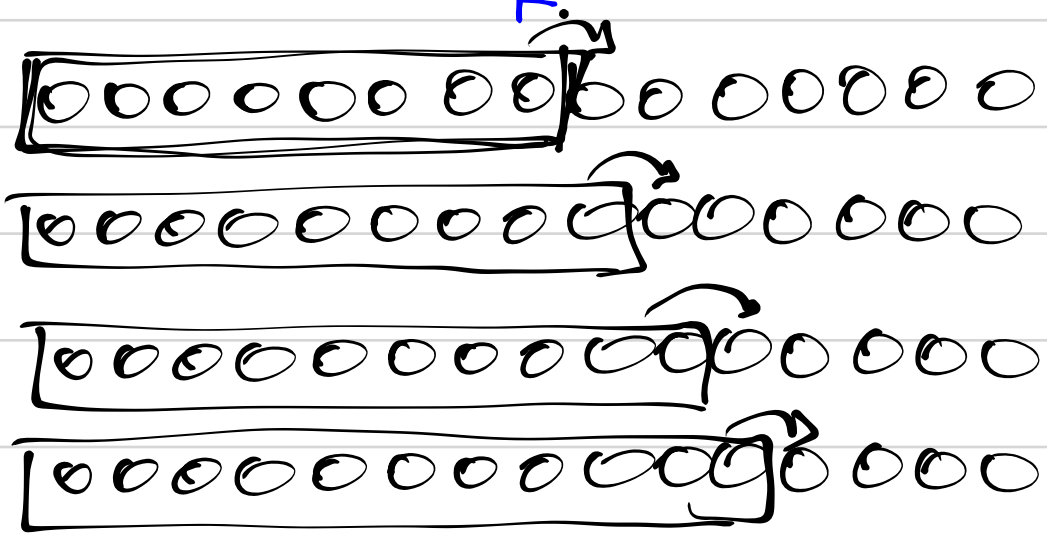
Идея! Кросс-валидация.

Модели = {Модель А, Модель Б}

$$MSE_M = E\left((y_{T+1} - \hat{y}^M_{T+1})^2\right)$$

[теоретическая MSE]

CV с растущим окном



$e_{F+1} = y_{F+1} - \hat{y}_{F+1|F}$

$e_{F+2} = y_{F+2} - \hat{y}_{F+2|F+1}$

$e_{F+3} = y_{F+3} - \hat{y}_{F+3|F+2}$

e_{F+4} ...

$$\hat{MSE}_M = \sum_{h=1}^{H} e^2_{F+h} / H$$

Идеальный способ!

* медленный (если много моделей/рядов)
* хочется много наблюдений!

(хотим: * H побольше, чтобы модель обучать на более длинном ряду и H побольше, чтобы точность оценивания MSE была).

* Способы, когда частная АС дорога:

* формальные статистические тесты
* информационные критерии (AIC)
* машинные подходы.

Автоматическая сезонная ARIMA.

ряд - месячный / кварт- ый

(1). Считаем по ряду "силу сезонности" (машинное)

Если сила сез-сти > порог = 0.8*

то применим
от $y_t \to \Delta_s y_t =$
$= y_t - y_{t-12}$

* может и другое !!

(2) Формальный статистический тест**

KPSS H_0*: (y_t) ~ стационарен

тот, со всем что можно возможно это уже от исходного ряда Δ_s

H_A*: (y_t) - не стац-н, но (Δy_t) стац

есть нюансы

Если H_0 отвергается то берем обычную разность

$(y_t) \to (\Delta y_t)$

это y_t после возможного преобразования на пред шаге

** и для шаг 2 один раз

перед 3-м шагом [мы подошли]
перед нами стационарный процесс.

Шаг 3 $\boxed{y_t \to \Delta^d \Delta_s^D y_t}$

$d \in \{0,1,2\}$

$D \in \{0,1\}$

оцениваем несколько моделей SARIMA-моделей (стац-х) для $\Delta^d \Delta_s^D y_t$

← шаг 2

например:

$y_t \sim SARIMA(0,d,1)(0,D,1)$

$y_t \sim SARIMA(1,d,0)(0,D,1)$

$\vdots$

$y_t \sim SARIMA(2,d,0)(2,D,0)$

← с шага 1

может не сойтись оптимизатор*

*возможно надо перебирать модели

$SARIMA(p,d,q)(P,D,Q)$

несез. сезон

$p+q \le 3$

$P+Q \le 3$

выбирается модель с min AIC.

$AIC = 2p - 2\ln L$

p - полное число своб-х пар-в в модели

$\ln L$ - логарифм. максимума правдоподобия

Детали

Main

Сила сезонности

$(y_t) \longrightarrow [\ STL\ ] \longrightarrow (trend_t) + (seas_t) + (rem_t)$

$$F_{seas} = \max\left(0,\ 1 - \frac{Var(rem)}{Var(seas + rem)}\right)$$

[$Var(a)$ = выборочная дисперсия ряда a

sample

$= \frac{\sum (a_i - \bar{a})^2}{n-1}$]

Шаг 2. детали про KPSS тест.

идея!

$y_1 \dots y_T$ — зависимые наблюдения

интуитивно: несут меньше информации чем независимые наблюдения

способы измерить?

* информация Фишера
* долгосрочная дисперсия (long run)
* эффективное число наблюдений

идеальный мир

разовых независимых пока

$y_1 \dots y_T \sim$ неав с $\mathbb{E}(y_t) = \mu$ и $Var(y_t) = \lambda^2$

95% асимт для μ: $\left[\bar{y} - 1.96 \cdot \sqrt{\frac{\lambda^2}{T}}, \bar{y} + 1.96 \cdot \sqrt{\frac{\lambda^2}{T}}\right]$

$Var(\bar{y}) = \frac{\lambda^2}{T}$ $\quad$ $Var(y_t) = \lambda^2$

$y_1 \dots y_T$ — случ-ый процесс

$Var(y_t) = \gamma_0$ $\quad$ $Var(\bar{y}) = \frac{Var(y_1 + \dots + y_T)}{T^2} =$

$= \frac{T \cdot \gamma_0 + 2(T-1) \cdot \gamma_1 + \dots + 2 \cdot \gamma_{T-1}}{T^2}$ $\quad$ $\gamma_k = Cov(y_t, y_{t-k})$

$\neq \frac{\gamma_0}{T}$

def [долгосрочная дисперсия λ^2] — это число, что [случ-го процесса] $\left[Var(\bar{y}) = \frac{\lambda^2}{T} + o\left(\frac{1}{T}\right)\right]$

def T_{eff} – эффективное число наблюдений

опр-ся по формуле:

$$Var(\bar{y}) = \frac{\sigma^2}{T_{eff}}$$

Упр. $y_t = u_t + u_{t-1}$ $\quad u_t \sim N(0;16)$ б. шум

а) $Var(y_t)$? б) долгоср-ая дисперсия?

в) $\lim\limits_{T\to\infty} \frac{T_{eff}}{T}$? $= \frac{1}{2}$

$Var(y_t) = 32$ $\quad Var\left(\frac{y_1 + y_2 + \dots y_T}{T}\right) =$

$= \frac{T\cdot 32 + 2\cdot(T-1)\cdot 16 + 0}{T^2} =$

$= \frac{64}{T} - \frac{32}{T^2} = \frac{64}{T} + o\left(\frac{1}{T}\right)$ $\quad \lambda^2 = 64$

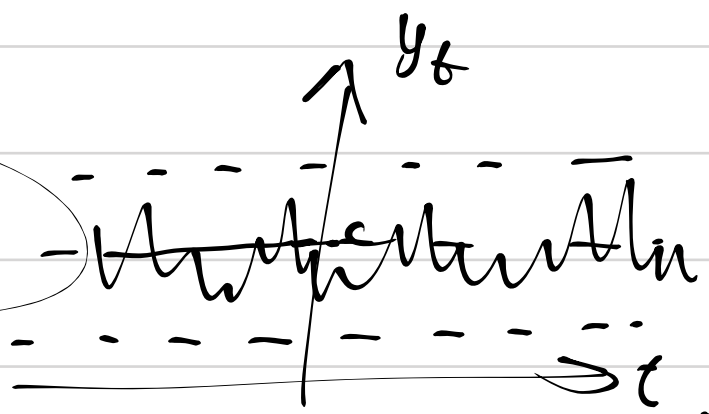
KPSS тест [с константой]

Предположим: $y_t = c + \tau_t + x_t$, где

(x_t) ~ стац-ый процесс с $E(x_t) = 0$.

независимый с τ_t

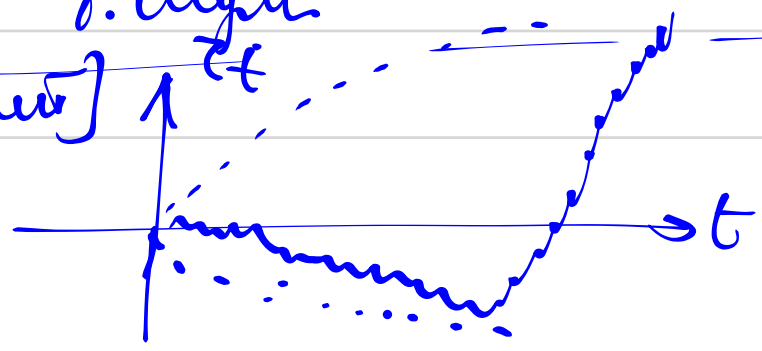
H_0: $\tau_t \equiv 0$. $\quad (y_t)$ – стац-н.



H_A: $\tau_t = u_1 + u_2 + u_3 + \dots + u_t$, где

(u_t) – б. шум

[τ_t – не стац-ный]



Форма ее процедура:

$$\hat{c} = \bar{y} = \frac{y_1 + y_2 + \dots + y_T}{T}$$

$S_t = y_1 + y_2 + \dots + y_t - t\bar{y}$ ↑ накопленная сумма ошибок прогноза привязанной модели

$$KPSS = \frac{\sum_{t=1}^{T} S_t^2}{T^2 \cdot \hat{\lambda}^2}$$

$\hat{\lambda}^2$ – оценка долгосрочной дисперсии.

$KPSS \xrightarrow[T \to \infty]{\text{по верна}}$ особое распределение [KPSS^c]