Методы восстановления регрессии. Прогнозирование временных рядов.

К. В. Воронцов vokov@forecsys.ru
А. А. Романенко alexromsput@gmail.com

10 апреля 2015

Содержание

Метод наименьших квадратов

- X — объекты (часто $\mathbb{R}^n$); Y — ответы (часто $\mathbb{R}$, реже $\mathbb{R}^m$);
  $X^\ell = (x_i, y_i)_{i=1}^\ell$ — обучающая выборка;
  $y_i = y(x_i)$, $y\colon X \to Y$ — неизвестная зависимость;
- $a(x) = f(x, \alpha)$ — модель зависимости,
  $\alpha \in \mathbb{R}^p$ — вектор параметров модели.
- Метод наименьших квадратов (МНК):

$$Q(\alpha, X^\ell) = \sum_{i=1}^{\ell} w_i \bigl(f(x_i, \alpha) - y_i\bigr)^2 \to \min_\alpha,$$

  где w_i — вес, степень важности i-го объекта.

  $Q(\alpha^*, X^\ell)$ — *остаточная сумма квадратов*
  (residual sum of squares, RSS).

Метод максимума правдоподобия

Модель данных с некоррелированным гауссовским шумом:

$$y(x_i) = f(x_i,\alpha) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0,\sigma_i^2), \quad i = 1,\dots,\ell.$$

Метод максимума правдоподобия (ММП):

$$L(\varepsilon_1,\dots,\varepsilon_\ell|\alpha) = \prod_{i=1}^{\ell} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_i^2}\varepsilon_i^2\right) \to \max_{\alpha};$$

$$-\ln L(\varepsilon_1,\dots,\varepsilon_\ell|\alpha) = \mathrm{const}(\alpha) + \frac{1}{2}\sum_{i=1}^{\ell}\frac{1}{\sigma_i^2}\bigl(f(x_i,\alpha) - y_i\bigr)^2 \to \min_{\alpha};$$

Теорема

Решения МНК и ММП, совпадают, причём веса объектов обратно пропорциональны дисперсии шума, $w_i = \sigma_i^{-2}$.

Многомерная линейная регрессия

$f_1(x), \dots, f_n(x)$ — числовые признаки;
Модель многомерной линейной регрессии:

$$f(x, \alpha) = \sum_{j=1}^{n} \alpha_j f_j(x), \qquad \alpha \in \mathbb{R}^n.$$

Матричные обозначения:

$$\underset{\ell \times n}{F} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}, \quad \underset{\ell \times 1}{y} = \begin{pmatrix} y_1 \\ \dots \\ y_\ell \end{pmatrix}, \quad \underset{n \times 1}{\alpha} = \begin{pmatrix} \alpha_1 \\ \dots \\ \alpha_n \end{pmatrix}.$$

Функционал квадрата ошибки:

$$Q(\alpha, X^\ell) = \sum_{i=1}^{\ell} \bigl(f(x_i, \alpha) - y_i\bigr)^2 = \|F\alpha - y\|^2 \to \min_{\alpha}.$$

Нормальная система уравнений

Необходимое условие минимума в матричном виде:

$$\frac{\partial Q}{\partial \alpha}(\alpha) = 2F^{\mathsf{T}}(F\alpha - y) = 0,$$

откуда следует *нормальная система* задачи МНК:

$$F^{\mathsf{T}}F\alpha = F^{\mathsf{T}}y,$$

где $\underset{n\times n}{F^{\mathsf{T}}F}$ — ковариационная матрица набора признаков $f_1, \dots, f_n$.

Решение системы: $\alpha^* = (F^{\mathsf{T}}F)^{-1}F^{\mathsf{T}}y = F^{+}y$.

Значение функционала: $Q(\alpha^*) = \|P_F y - y\|^2$,
где $P_F = FF^{+} = F(F^{\mathsf{T}}F)^{-1}F^{\mathsf{T}}$ — *проекционная матрица*.

Сингулярное разложение

Произвольная $\ell{\times}n$-матрица представима в виде *сингулярного разложения* (singular value decomposition, SVD):

$$F = VDU^{\mathsf{T}}.$$

Основные свойства сингулярного разложения:

1. $\ell{\times}\ell$-матрица $V = (v_1, \ldots, v_l)$ ортогональна, $V^{\mathsf{T}}V = I_\ell$, столбцы v_j — собственные векторы матрицы FF^{T};
2. $n{\times}n$-матрица $U = (u_1, \ldots, u_n)$ ортогональна, $U^{\mathsf{T}}U = I_n$, столбцы u_j — собственные векторы матрицы $F^{\mathsf{T}}F$;
3. $l{\times}n$-матрица D диагональна, $D = \operatorname{diag}\bigl(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}\bigr)$, $\lambda_j \geqslant 0$ — собственные значения матриц $F^{\mathsf{T}}F$ и FF^{T}.

Решение МНК через сингулярное разложение

Псевдообратная F^+, вектор МНК-решения α^*,
МНК-аппроксимация целевого вектора $F\alpha^*$:

$$F^+ = (UDV^\top VDU^\top)^{-1}UDV^\top = UD^{-1}V^\top = \sum_{j=1}^{n} \frac{1}{\sqrt{\lambda_j}} u_j v_j^\top;$$

$$\alpha^* = F^+ y = UD^{-1}V^\top y = \sum_{j=1}^{n} \frac{1}{\sqrt{\lambda_j}} u_j (v_j^\top y);$$

$$F\alpha^* = P_F y = (VDU^\top)UD^{-1}V^\top y = VV^\top y = \sum_{j=1}^{n} v_j (v_j^\top y);$$

$$\|\alpha^*\|^2 = \|D^{-1}V^\top y\|^2 = \sum_{j=1}^{n} \frac{1}{\lambda_j} (v_j^\top y)^2.$$

Проблема мультиколлинеарности

Если имеются $\lambda_j \to 0$, то

- МНК-решение α^* неустойчиво и неинтерпретируемо: $\|\alpha\| \to \infty$;
- ответы на новых объектах $y' = F'\alpha^*$ неустойчивы;
- в то время как на обучении, казалось бы, «всё хорошо»: $Q(\alpha^*) = \|F\alpha^* - y\|^2 \to 0$;
- мультиколлинеарность влечёт *переобучение*.

Три стратегии устранения мультиколлинеарности:

- Регуляризация: $\|\alpha\| \to \min$;
- Преобразование признаков: $f_1, \dots, f_n \to g_1, \dots, g_m, \quad m \ll n$;
- Отбор признаков: $f_1, \dots, f_n \to f_{j_1}, \dots, f_{j_m}, \quad m \ll n$.

Регуляризация (гребневая регрессия)

Штраф за увеличение нормы вектора весов $\|\alpha\|$:

$$Q_\tau(\alpha) = \|F\alpha - y\|^2 + \frac{1}{2\sigma}\|\alpha\|^2,$$

где $\tau = \frac{1}{\sigma}$ — неотрицательный *параметр регуляризации*.

Вероятностная интерпретация: априорное распределение вектора α — гауссовское с ковариационной матрицей σI_n.

Модифицированное МНК-решение (τI_n — «гребень»):

$$\alpha^*_\tau = (F^\intercal F + \tau I_n)^{-1} F^\intercal y.$$

Преимущество сингулярного разложения:
можно подбирать параметр τ, вычислив SVD только один раз.

Регуляризованный МНК через сингулярное разложение

Вектор регуляризованного МНК-решения α^*_τ
и МНК-аппроксимация целевого вектора $F\alpha^*_\tau$:

$$\alpha^*_\tau = U(D^2 + \tau I_n)^{-1} D V^{\mathsf{T}} y = \sum_{j=1}^{n} \frac{\sqrt{\lambda_j}}{\lambda_j + \tau} u_j (v_j^{\mathsf{T}} y);$$

$$F\alpha^*_\tau = VDU^{\mathsf{T}}\alpha^*_\tau = V \operatorname{diag}\left(\frac{\lambda_j}{\lambda_j + \tau}\right) V^{\mathsf{T}} y = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \tau} v_j (v_j^{\mathsf{T}} y);$$

$$\|\alpha^*_\tau\|^2 = \|D^2(D^2 + \tau I_n)^{-1} D^{-1} V^{\mathsf{T}} y\|^2 = \sum_{j=1}^{n} \frac{1}{\lambda_j + \tau} (v_j^{\mathsf{T}} y)^2.$$

$F\alpha^*_\tau \neq F\alpha^*$, но зато решение становится гораздо устойчивее.

Выбор параметра регуляризации τ

Контрольная выборка: $X^k = (x_i', y_i')_{i=1}^k$;

$$\underset{k\times n}{F'} = \begin{pmatrix} f_1(x_1') & \dots & f_n(x_1') \\ \dots & \dots & \dots \\ f_1(x_k') & \dots & f_n(x_k') \end{pmatrix}, \quad \underset{k\times 1}{y'} = \begin{pmatrix} y_1' \\ \dots \\ y_k' \end{pmatrix}.$$

Вычисление функционала Q на контрольных данных T раз потребует $O(kn^2 + knT)$ операций:

$$Q(\alpha_\tau^*, X^k) = \|F'\alpha_\tau^* - y'\|^2 = \Bigl\| \underbrace{F'U}_{k\times n} \operatorname{diag}\Bigl(\frac{\sqrt{\lambda_j}}{\lambda_j+\tau}\Bigr) \underbrace{V^\mathsf{T} y}_{n\times 1} - y' \Bigr\|^2.$$

Зависимость $Q(\tau)$ обычно имеет характерный минимум.

Регуляризация сокращает «эффективную размерность»

Сжатие (shrinkage) или *сокращение весов* (weight decay):

$$\|\alpha_\tau^*\|^2 = \sum_{j=1}^n \frac{1}{\lambda_j + \tau}(v_j^\intercal y)^2 \quad < \quad \|\alpha^*\|^2 = \sum_{j=1}^n \frac{1}{\lambda_j}(v_j^\intercal y)^2.$$

Почему говорят о *сокращении эффективной размерности*?

Роль размерности играет след проекционной матрицы:

$$\operatorname{tr} F(F^\intercal F)^{-1}F^\intercal = \operatorname{tr}(F^\intercal F)^{-1}F^\intercal F = \operatorname{tr} I_n = n.$$

При использовании регуляризации:

$$\operatorname{tr} F(F^\intercal F + \tau I_n)^{-1}F^\intercal = \operatorname{tr}\operatorname{diag}\left(\frac{\lambda_j}{\lambda_j+\tau}\right) = \sum_{j=1}^n \frac{\lambda_j}{\lambda_j+\tau} < n.$$

Лассо Тибширани — другой подход к регуляризации
LASSO — Least Absolute Shrinkage and Selection Operator

$$\begin{cases} Q(\alpha) = \|F\alpha - y\|^2 \to \min\limits_{\alpha}; \\ \sum\limits_{j=1}^{n} |\alpha_j| \leqslant \varkappa; \end{cases}$$

Лассо приводит к отбору признаков! Почему?

После замены переменных

$$\begin{cases} \alpha_j = \alpha_j^+ - \alpha_j^-; \\ |\alpha_j| = \alpha_j^+ + \alpha_j^-; \end{cases} \quad \alpha_j^+ \geqslant 0; \quad \alpha_j^- \geqslant 0.$$

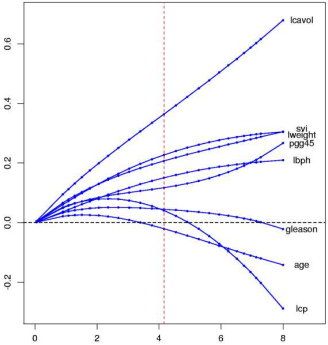
ограничения принимают канонический вид:

$$\sum_{j=1}^{n} \alpha_j^+ + \alpha_j^- \leqslant \varkappa; \quad \alpha_j^+ \geqslant 0; \quad \alpha_j^- \geqslant 0.$$

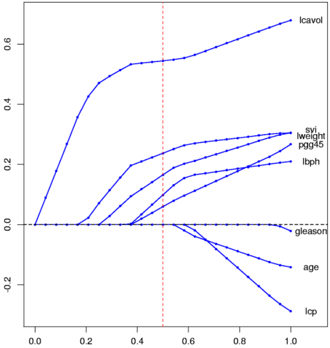
Чем меньше $\varkappa$, тем больше j таких, что $\alpha_j^+ = \alpha_j^- = 0$.

Сравнение гребневой регрессии и Лассо

Зависимость $\{\alpha_j\}$ от σ



Зависимость $\{\alpha_j\}$ от $\varkappa$



Задача диагностики рака (prostate cancer, UCI)

T.Hastie, R.Tibshirani, J.Friedman. The Elements of Statistical Learning. Springer, 2001.

Формула Надарая–Ватсона

Приближение константой $a(x) = \alpha$ в окрестности $x \in X$:

$$Q(\alpha; X^\ell) = \sum_{i=1}^{\ell} w_i(x)\bigl(\alpha - y_i\bigr)^2 \to \min_{\alpha \in \mathbb{R}};$$

где $w_i(x) = K\left(\frac{\rho(x,x_i)}{h}\right)$ — веса объектов x_i относительно x;
$K(r)$ — *ядро*, невозрастающее, ограниченное, гладкое;
h — ширина окна сглаживания.

Формула *ядерного сглаживания* Надарая–Ватсона:

$$a_h(x; X^\ell) = \frac{\sum\limits_{i=1}^{\ell} y_i w_i(x)}{\sum\limits_{i=1}^{\ell} w_i(x)} = \frac{\sum\limits_{i=1}^{\ell} y_i K\left(\frac{\rho(x,x_i)}{h}\right)}{\sum\limits_{i=1}^{\ell} K\left(\frac{\rho(x,x_i)}{h}\right)}.$$

- Ядро $K(r)$
  — существенно влияет на гладкость функции $a_h(x)$,
  — слабо влияет на качество аппроксимации.
- Ширина окна h
  — существенно влияет на качество аппроксимации.
- При неравномерной сетке $\{x_i\}$ — переменная ширина окна:

$$w_i(x) = K\left(\frac{\rho(x, x_i)}{h(x)}\right),$$

  где $h(x) = \rho(x, x^{(k+1)})$, $x^{(k+1)}$ — k-й сосед объекта x.
- Оптимизация ширины окна по скользящему контролю:

$$\mathrm{LOO}(h, X^\ell) = \sum_{i=1}^{\ell} \Bigl(a_h\bigl(x_i; X^\ell \backslash \{x_i\}\bigr) - y_i\Bigr)^2 \to \min_h .$$

Локально взвешенное сглаживание (LOWESS — LOcally WEighted Scatter plot Smoothing)

Основная идея:
чем больше величина ошибки $\varepsilon_i = \bigl|a_h\bigl(x_i; X^\ell\backslash\{x_i\}\bigr) - y_i\bigr|$, тем в большей степени прецедент (x_i, y_i) является выбросом, и тем меньше должен быть его вес $w_i(x)$.

Эвристика:
домножить веса $w_i(x)$ на коэффициенты $\gamma_i = \tilde{K}(\varepsilon_i)$,
где $\tilde{K}$ — ещё одно ядро, вообще говоря, отличное от $K(r)$.

Рекомендация:
квартическое ядро $\tilde{K}(\varepsilon) = K_Q\bigl(\frac{\varepsilon}{6\,\mathrm{med}\{\varepsilon_i\}}\bigr)$,
где $\mathrm{med}\{\varepsilon_i\}$ — медиана вариационного ряда ошибок.

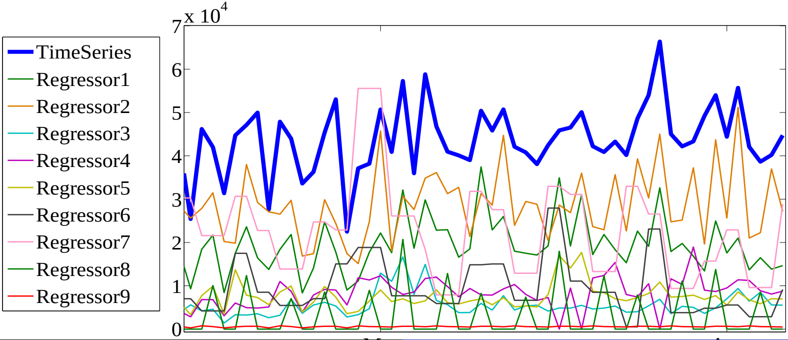
Алгоритм LOWESS

Вход: X^ℓ — обучающая выборка;
Выход: коэффициенты γ_i, $i = 1, \ldots, \ell$;

1: инициализация: $\gamma_i := 1$, $i = 1, \ldots, \ell$;
2: **повторять**
3: **для всех** объектов $i = 1, \ldots, \ell$
4: вычислить оценки скользящего контроля:
$$a_i := a_h\bigl(x_i; X^\ell \backslash \{x_i\}\bigr) = \frac{\sum\limits_{j=1, j\neq i}^{\ell} y_j \gamma_j K\left(\frac{\rho(x_i,x_j)}{h(x_i)}\right)}{\sum\limits_{j=1, j\neq i}^{\ell} \gamma_j K\left(\frac{\rho(x_i,x_j)}{h(x_i)}\right)};$$
5: **для всех** объектов $i = 1, \ldots, \ell$
6: $\gamma_i := \tilde{K}\bigl(|a_i - y_i|\bigr)$;
7: **пока** коэффициенты γ_i не стабилизируются;

Регрессионная модель временного ряда

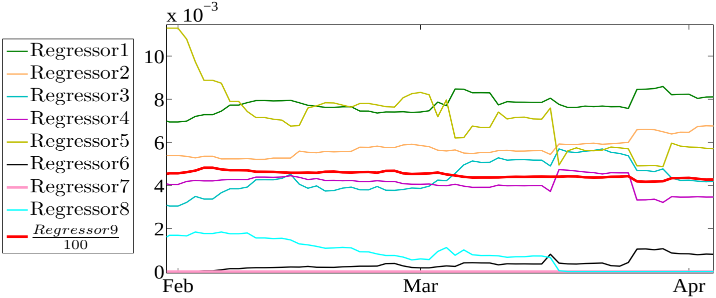
$y_0, y_1, \dots, y_t, \dots$ — временной ряд, $y_t \in \mathbb{R}$, t — срезы моментов времени, $\vec{x}_0, \vec{x}_1, \dots, \vec{x}_t, \dots$ — регрессоры,
$\vec{x}_t = (x_{1,t}, \dots, x_{n,t}) \in \mathbb{R}^n$
$\hat{y}_{t+d}(w) = f_t\left(\vec{x}_{t+d}; \vec{w}_t\right)$ — регрессионная модель временного ряда, где $d = 1, \dots, D$, D — горизонт (отсрочка, delay),
$w_t \in \mathbb{R}^n$ — вектор параметров модели в момент времени t



Адаптация весов с регуляризацией

На каждом шаге t веса определяются по МНК и сглаживаются с предыдущими значениями весов:

$$\begin{cases} \sum\limits_{t=0}^{T} \beta^{(T-t)} \Bigl(\sum\limits_{j=1}^{k} w_j x_{j,t} - y_t \Bigr)^2 + \lambda \sum\limits_{j=1}^{k} \bigl(w_j - w_{j,t-1} \bigr)^2 \to \min\limits_{w_j, j=1,\dots,k} \\ \sum_{j=1}^{k} w_j \geqslant 0. \end{cases}$$

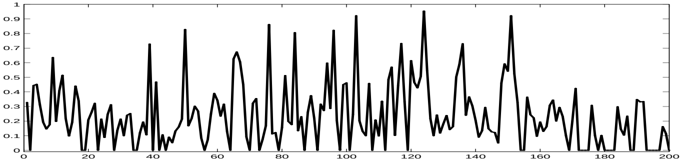


Авторегрессионная модель временного ряда

$y_0, y_1, \dots, y_t, \dots$ — временной ряд, $y_i \in \mathbb{R}$
$\hat{y}_{t+d}(w) = f_t(y_1, \dots, y_t; w)$ — авторегрессионная модель временного ряда,
где $d = 1, \dots, D$, D — горизонт (отсрочка, delay),
w — вектор параметров модели (какого размера?)

Особенности задачи:

- пропуски в данных;
- нестационарность (непостоянство модели);
- модель временного ряда неочевидна;



Экспоненциальное скользящее среднее (ЭСС)

Простейшая регрессионная модель — константа $\hat{y}_{t+1} = c$, наблюдения учитываются с весами, убывающими в прошлое:

$$\sum_{i=0}^{t} \alpha^{t-i}(y_i - c)^2 \to \min_c, \quad \alpha \in (0,1)$$

Аналитическое решение — формула Надарая-Ватсона:

$$c \equiv \hat{y}_{t+1} = \frac{\sum_{i=0}^{t} \alpha^i y_{t-i}}{\sum_{i=0}^{t} \alpha^i}$$

Запишем аналогично $\hat{y}_t$, оценим $\sum\limits_{i=0}^{t} \alpha^i \approx \sum\limits_{i=0}^{\infty} \alpha^i = \frac{1}{1-\alpha}$,

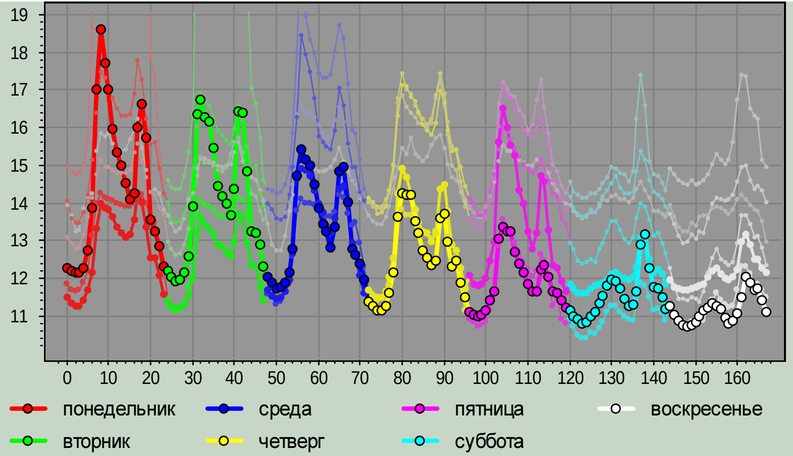
получим

$$\hat{y}_{t+1} := \hat{y}_t + \alpha(y_t - \hat{y}_t) \quad = \quad \alpha y_t + (1-\alpha)\hat{y}_t,$$

$\alpha \in (0,1)$ называется параметром сглаживания.

Авторегрессионная модель

Почасовые цены электроэнергии на бирже NordPool, 2000г.



Особенности задачи: три вложенные сезонности, скачки

Линейная модель авторегрессии

В роли признаков — n предыдущих наблюдений ряда:

$$\hat{y}_{t+1}(w) = \sum_{j=1}^{n} w_j y_{t-j+1}, \quad w \in \mathbb{R}^n$$

В роли объектов $\ell = t - n + 1$ моментов в истории ряда:

$$\underset{\ell \times n}{F} = \begin{pmatrix} y_t & y_{t-1} & y_{t-2} & \cdots & y_{t-n+1} \\ y_{t-1} & y_{t-2} & y_{t-3} & \cdots & y_{t-n} \\ y_{t-2} & y_{t-3} & y_{t-4} & \cdots & y_{t-n-1} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ y_n & y_{n-1} & y_{n-2} & \cdots & y_1 \end{pmatrix}, \quad \underset{\ell \times 1}{y} = \begin{pmatrix} y_{t+1} \\ y_t \\ y_{t-1} \\ \ldots \\ y_{n+1} \end{pmatrix}$$

Функционал квадрата ошибки:

$$Q_t(w, X^\ell) = \sum_{i=n+1}^{t+1} \bigl(\hat{y}_i(w) - y_i\bigr)^2 = \|Fw - y\|^2 \to \min_w$$

Адаптивная авторегрессионная модель

Линейная модель авторегрессии (линейный фильтр):

$$\hat{y}_{t+1}(w) = \sum_{j=1}^{n} w_j y_{t-j+1}, \quad w \in \mathbb{R}^n$$

$\varepsilon_t = y_t - \hat{y}_t$ — ошибка прогноза $\hat{y}_t$, сделанного на шаге $t-1$

Метод наименьших квадратов: $\varepsilon_t^2 \to \min\limits_w$.

Один шаг градиентного спуска в каждый момент t:

$$w_j := w_j + h_t \varepsilon_t y_{t-j+1}.$$

Градиентный шаг в методе скорейшего спуска:

$$h_t = \frac{\alpha}{\sum_{j=1}^{n} y_{t-j+1}^2},$$

где α — аналог параметра сглаживания.

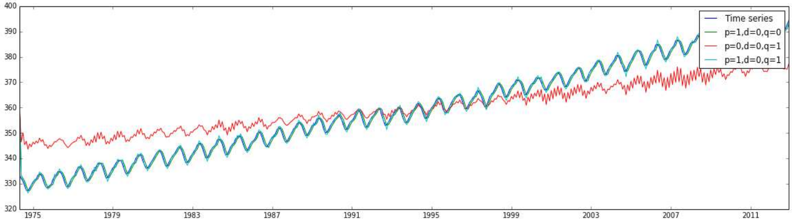
Модели ARMA, ARIMA

ARMA(p,q): $y_1, \dots, y_t$ — стационарный

- $y_t = c + \underbrace{\sum_{i=1}^{p} \alpha_i y_{t-i}}_{AR} + \underbrace{\sum_{j=1}^{q} \beta_j \varepsilon_{t-j}}_{MA} + \varepsilon_t;$

ARIMA(p,q,d): y_t — НЕстационарный

- $\triangle^d = (1 - L_1)^d$, $\triangle^d y_t$ — стационарный
- $\triangle^d y_t = c + \sum_{i=1}^{p} \alpha_i \triangle^d y_{t-i} + \sum_{j=1}^{q} \beta_j \varepsilon_{t-j} + \varepsilon_t;$



Резюме по временным рядам

- Известная модель временного ряда — регрессионный подход (LAWR)
- Модель временного ряда очевидна из стуктуры временного ряда — авторегрессионный подход (ARMA, ARIMA)
- Модель временного ряда неизвестна - простые адаптивные методы (ES)

Резюме в конце лекции

- Многомерная линейная регрессия — сингулярное разложение
- Гребневая регрессия — сингулярное разложение
- Прогнозирование временные ряды требует более сложных моделей
- Для разных классов задач TS Forecasting существенно разные подходы