

ПАРАМЕТРИЧЕСКАЯ ИДЕНТИФИКАЦИЯ

ФН-12. Магистры - 3 семестр

Лекция 4. СТАТИСТИЧЕСКИЕ МОДЕЛИ.

Линейные относительно параметров статистические модели.

Пусть имеется набор N измерений неизвестной функции $f(x)$ в точках x_i (узлы) со случайными ошибками (шумами) ε_i :

$$y_i = f(t_i) + \varepsilon_i, \quad i = \overline{1, N} \quad (4.1)$$

Требуется по набору измерений $\{(t_i, y_i), i = \overline{1, N}\}$ оценить функцию $f(t)$.
Линейная относительно неизвестных параметров модель имеет вид

$$f(t) = f(t, \theta) = \sum_{j=1}^r \theta_j \psi_j(t), \quad (4.2)$$

где $(\theta_1, \dots, \theta_r) = \Theta^T$, $r \ll N$ – вектор неизвестных параметров, ψ_1, \dots, ψ_r – заданный набор функций (базисные функции).

Обозначим через $\Psi = [\psi_j(t_i)] \in M_{N \times r}$ матрицу – регрессор, элементы $\psi_{i,j}$ – значения базисных функций в момент времени t_i (k -й столбец матрицы – совокупность узловых значений k -й базисной функции).

Пусть Y – вектор-столбец $(y_1, \dots, y_N)^T$ **результатов измерений**. Равенства (3.1), (3.2) запишем в виде

$$Y = \Psi\Theta + \varepsilon, \quad (4.3)$$

где $\varepsilon = (\varepsilon_1, \dots, \varepsilon_N)^T$ – вектор ошибок. Соотношение (3.3) – общая линейная статистическая модель.

Требуется оценить неизвестный вектор параметров Θ^T по заданному вектору Y и матрице Ψ .

Оценка неизвестного вектора параметров Θ^T методом наименьших квадратов (МНК).

Оценку неизвестной функции $f(t)$ запишем в виде $f(t) = f(t, \hat{\Theta})$, вектор $\hat{\Theta}$ найдем из условия минимизации суммы квадратов отклонений измеренных значений y_i от значений функции $f(t)$ в точках измерения:

$$\sum_{i=1}^N \left(y_i - f(t_i, \hat{\Theta}) \right)^2 = \min_{\Theta} \sum_{i=1}^N \left(y_i - f(t_i, \Theta) \right)^2 \quad (4.4)$$

При отсутствии шумов вектор Θ является точным решением системы уравнений

$$\Psi\Theta = Y. \quad (4.5)$$

При наличии шумов решение задачи (3.4) сводится к минимизации квадратичной нормы невязок $Y - \Psi\Theta$. Вектор $\hat{\Theta}$, дающий такой минимум, называют псевдорешением системы (3.5).

Псевдорешение является точным решением вспомогательной системы

$$\Psi^*\Psi\Theta = \Psi^*Y, \quad (4.6)$$

Ψ^* — эрмитово сопряжённая матрица, $\Psi^* = [\psi'_{ij}] \in M_{r \times N}$, $\psi'_{ij} = \psi_{ji}$. Для матриц над полем действительных чисел $\Psi^* = \Psi^T$

Пусть $C = (\Psi^T\Psi)$, $c_{jk} = \sum_{i=1}^N \psi'_{ij}\psi_{ji}$, $B = \Psi^TY$, $b_k = \sum_{i=1}^N \psi'_{ij}y_i$.

Вектор параметров $\Theta = C^{-1}B$.

Столбец $\hat{Y} = \Psi\hat{\Theta}$ называется вектором подгонки, соответствующий вектор невязок обозначается $\hat{E} = Y - \hat{Y}$. Векторы подгонки и невязок ортогональны:

$$(\hat{Y}, \hat{E}) = \sum_{i=1}^N \hat{y}_i \hat{e}_i = 0.$$

Оценки МНК являются несмещенными, если ошибки ε_i – случайные величины, удовлетворяющие условиям

$$M(\varepsilon_i) = 0, \quad M(\varepsilon_i^2) = \sigma^2 > 0, \quad M(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j.$$

При $r < N$ несмещенная оценка дисперсии шумов σ^2 имеет вид:

$$\sigma_N^2 = \frac{1}{N - r} \|E\|^2 = \frac{1}{N - r} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

Полиномиальная регрессия

Пусть в (3.2) $\psi(t) = t^k$, $k = \overline{0, r}$ – полином степени r с неизвестными коэффициентами θ_k . Соотношения (3.1), (3.2) описывают модель полиномиальной регрессии. При $r = 1$ неизвестная функция является линейной.

Для модели полиномиальной регрессии

$$\psi_{i(k+1)} = t_i^k, \quad c_{j(k+1)} = \sum_{i=1}^N t_i^{k-1+i}, \quad b_{k+q} = \sum_{i=1}^N t_i^k y_i, \quad k = \overline{0, r}.$$

И матрица C невырождена при $r < N$, $t_i \neq t_k$ при $l \neq k$.

В общем случае вектор параметров Θ и полином наилучшего приближения $f(x, \hat{\Theta})$ следует находить составлением и решением системы (3.6).

Полином имеет вид $p(t) = \theta_r t^r + \theta_{r-1} t^{r-1} + \dots + \theta_0$.

С использованием пакета MATLAB можно найти полином $p(t)$ аппроксимирующий $y = f(t)$ и получить коэффициенты полинома.

$p = polyfit(X, Y, r)$ – функция позволяет рассчитать **коэффициенты** $p(t)$ полиномиальной регрессионной модели n -й степени для выборки X , Y методом наименьших квадратов, где X – независимая переменная, Y – зависимая переменная. X , Y – вектора исходных данных одинаковой размерности. Вектор коэффициентов полинома $p \in M_{n+1 \times 1}$, n – степень полинома, элементы вектора расположены по убыванию степени независимой переменной: $p(1)t^n + p(2)t^{n-1} + \dots + p(n)t^1 + p(n+1)$

$y = polyval(p, t)$ – функция позволяет получить **значения полинома** $y = f(t, p)$ для массива t значений аргумента.

$Z = polyval(p, X)$ – функция позволяет получить **вектор подгонки** $\hat{Y} = Z$.

Общий вид задачи оценки неизвестного вектора параметров Θ^T .
Пусть имеется нелинейная функция

$$y = \phi(t) \quad (4.7)$$

Требуется построить модель функции на основе данных измерений.

Пусть измерения производятся в дискретные моменты времени $t_i \in T$,
($i = \overline{1, N}$) в соответствии с условиями опыта

$$z_i = y(t_i) + \varepsilon_i, \quad \varepsilon_i \rightarrow \mathbb{N}(0, \sigma^2) \quad (4.8)$$

где ε_i – ошибки измерений (независимые СВ).

Аппроксимация произвольными функциями.

Рассмотрим модель вида:

$$\tilde{y}(t) = \sum_{j=1}^n a_j \psi_j(t), \quad (4.9)$$

где a_j – неизвестные постоянные коэффициенты, оценку которых необходимо получить; $\psi_j(t)$ – известные функции времени, образующие линейно независимую систему на отрезке T .

Например, $\{\psi_{kj}(t)\}$ — степенные, тригонометрические полиномы, полиномы Лежандра, или другими функции.

Аппроксимирующий полином имеет вид

$$y(t) = \sum_{j=0}^{r-1} a_j \psi_j(t), \quad (4.10)$$

где $y(t)$ – искомая функция, a_j – неизвестные постоянные коэффициенты, $\psi_j(t)$ – известные функции времени, r – количество линейно независимых функций на отрезке T , используемых для аппроксимации, в случае использования полинома – максимальная степень полинома.

Задача аппроксимации функции $y(t)$ заключается в определении неизвестных коэффициентов a_j ($j = \overline{1, n}$) и значения r в (3.10).

Функция правдоподобия выборки $z = (z_1, \dots, z_N)$ имеет вид

$$L(z; a) = (2\pi\sigma_0^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma_0^2}[z - \Psi a]^T[z - \Psi a]\right\}, \quad (4.11)$$

где N – размер выборки, $\Psi = [\psi_j(t_i)] \in M_{N \times r}$ – матрица, составленная из значений функций $\psi_j(t)$ в момент времени t_i .

Оценка вектора a неизвестных коэффициентов по методу максимального правдоподобия равна

$$\hat{a} = C^{-1}\Psi^T z, \quad (4.12)$$

где

$$C = \Psi^T \Psi \in M_{r \times r}. \quad (4.13)$$

Оценка (3.13) является несмещенной, если степень полинома в (3.10) удовлетворяет гипотезе о наличии параболической регрессии n -го порядка.

Степень полинома (3.10) можно установить с помощью величины

$$\frac{1}{\sigma_0^2} \hat{\varepsilon}^T \hat{\varepsilon} = \frac{1}{\sigma_0^2} \sum_{i=1}^N \left(z_i - \sum_{j=1}^r \hat{a}_j \psi_{jt} \right)^2, \quad (4.14)$$

где $\varepsilon \in M_{N \times 1}$ случайный вектор ошибок измерений ε_i ε_i – независимые СВ ($i = \overline{1, N}$).

Величина имеет χ^2 – распределение с $N - r$ степенями свободы. Степень r выбирается таким образом, чтобы доверительный интервал

$$l = \left[\gamma_1 \sqrt{\frac{\hat{\varepsilon}^T \hat{\varepsilon}}{N - r}}, \gamma_2 \sqrt{\frac{\hat{\varepsilon}^T \hat{\varepsilon}}{N - r}} \right] \quad (4.15)$$

с вероятностью p_0 накрывал величину σ_0 .

Коэффициенты γ_1 и γ_2 в выражении (3.15) определяются с помощью таблицы *Границы доверительного интервала* (Линник Ю. В. Метод наименьших квадратов и основы математико-статистической теории обработки наблюдений. Л., Физматгиз, 1962 г., 352 с.)

Нижние γ_1 и верхние γ_2 границы доверительного интервала

$$\gamma_1 s_1 < \sigma < \gamma_2 s_1 \quad \left(s_1 = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \right)^*$$

p \ k	0,99		0,98		0,95		0,90	
	γ_1	γ_2	γ_1	γ_2	γ_1	γ_2	γ_1	γ_2
1	0,356	159	0,388	79,8	0,446	31,9	0,510	15,9
2	0,434	14,1	0,466	9,97	0,521	6,28	0,578	4,40
3	0,483	6,47	0,514	5,11	0,566	3,73	0,620	2,92
4	0,519	4,39	0,549	3,67	0,599	2,87	0,649	2,37
5	0,546	3,48	0,576	3,00	0,624	2,45	0,672	2,090
6	0,569	2,98	0,597	2,62	0,644	2,202	0,690	1,916
7	0,588	2,66	0,616	2,377	0,661	2,035	0,705	1,797
8	0,604	2,440	0,631	2,205	0,675	1,916	0,718	1,711
9	0,618	2,277	0,644	2,076	0,688	1,826	0,729	1,645
10	0,630	2,154	0,656	1,977	0,699	1,755	0,739	1,593
11	0,641	2,056	0,667	1,898	0,708	1,698	0,748	1,550
12	0,651	1,976	0,677	1,833	0,717	1,651	0,755	1,515
13	0,660	1,910	0,685	1,779	0,725	1,611	0,762	1,485
14	0,669	1,854	0,693	1,733	0,732	1,577	0,769	1,460
15	0,676	1,806	0,700	1,694	0,739	1,548	0,775	1,437
16	0,683	1,764	0,707	1,659	0,745	1,522	0,780	1,418
17	0,690	1,727	0,713	1,629	0,750	1,499	0,785	1,400
18	0,696	1,695	0,719	1,602	0,756	1,479	0,790	1,385
19	0,702	1,666	0,725	1,578	0,760	1,460	0,794	1,370
20	0,707	1,640	0,730	1,556	0,765	1,444	0,798	1,358
21	0,712	1,617	0,734	1,536	0,769	1,429	0,802	1,346
22	0,717	1,595	0,739	1,519	0,773	1,416	0,805	1,335
23	0,722	1,576	0,743	1,502	0,777	1,402	0,809	1,326
24	0,726	1,558	0,747	1,487	0,781	1,391	0,812	1,316
25	0,730	1,541	0,751	1,473	0,784	1,380	0,815	1,308
26	0,734	1,526	0,755	1,460	0,788	1,371	0,818	1,300
27	0,737	1,512	0,758	1,448	0,791	1,361	0,820	1,293
28	0,741	1,499	0,762	1,436	0,794	1,352	0,823	1,286
29	0,744	1,487	0,765	1,426	0,796	1,344	0,825	1,279
30	0,748	1,475	0,768	1,417	0,799	1,337	0,828	1,274
40	0,774	1,390	0,792	1,344	0,821	1,279	0,847	1,228
50	0,793	1,336	0,810	1,297	0,837	1,243	0,861	1,199
60	0,808	1,299	0,824	1,265	0,849	1,217	0,871	1,179
70	0,820	1,272	0,835	1,241	0,858	1,198	0,879	1,163
80	0,829	1,250	0,844	1,222	0,866	1,183	0,886	1,151
90	0,838	1,233	0,852	1,207	0,873	1,171	0,892	1,141
100	0,845	1,219	0,858	1,195	0,878	1,161	0,897	1,133
200	0,887	1,15	0,897	1,13	0,912	1,11	0,925	1,09

Рис. 1. Границы доверительного интервала

Пример.

Аппроксимация вектора параметров полиномом степени r .

Пусть степень полинома $r = 5$. Число измерений $N = 26$, $\sigma_0 = 0.05$.
Аппроксимирующий полином имеет вид

$$y_i = \sum_{j=1}^r a_j t^j, \quad (4.16)$$

Вектор измеряемых параметров $z(t)$ (значения z_i в дискретные моменты времени t_i):

N	z	t
1	-0.4483	0
2	-0.3420	0.0833
3	-0.2779	0.1667
4	-0.1376	0.2500
5	-0.0855	0.3333
6	0.1009	0.4167
7	0.0989	0.5000
8	0.0989	0.5000
9	0.2001	0.5833
10	0.3081	0.6667
11	0.2941	0.7500
12	0.6033	0.8333
13	0.6343	0.9167

N	z	t
14	0.6623	1.0000
15	0.8442	1.0833
16	0.7574	1.1667
17	0.8938	1.2500
18	0.9283	1.3333
19	0.9791	1.4167
20	0.9781	1.5000
21	0.8898	1.5833
22	0.8668	1.6667
23	0.7530	1.7500
24	0.6351	1.8333
25	0.4413	1.9167
26	0.1555	2.0000

Оценки коэффициентов полинома (3.16), найденные по формуле (3.12):

$$\hat{a}_0 = -0.4474 \quad \hat{a}_1 = 1.2195 \quad \hat{a}_2 = -0.4055 \quad \hat{a}_3 = 0.6824 \quad \hat{a}_4 = -0.3732 \\ \hat{a}_5 = 0.0098$$

Вычислим доверительный интервал для величины $\frac{1}{\sigma^2} \hat{\epsilon}^T \hat{\epsilon}$ с использованием (3.15)

Пусть вероятность $p_0 = 0,95$, определим параметр γ_i с $N - r = 100 - 6 = 94$ степенями свободы по таблице (1):

$$\gamma_1(0,95,90) = 0.838, \quad \gamma_1(0,95,100) = 0.845, \quad \text{примем } \gamma_1 = 0.841 \\ \gamma_2(0,95,90) = 1.233, \quad \gamma_2(0,95,100) = 1.219, \quad \text{примем } \gamma_2 = 1,226.$$

Доверительный интервал $[0.0415, \quad l_2 = 0.0605]$ с вероятностью p_0 накрывает величину $\sigma_0 = 0.05$ ($0.0415 < 0.05 < 0.0605$).

Степень полинома выбрана верно.

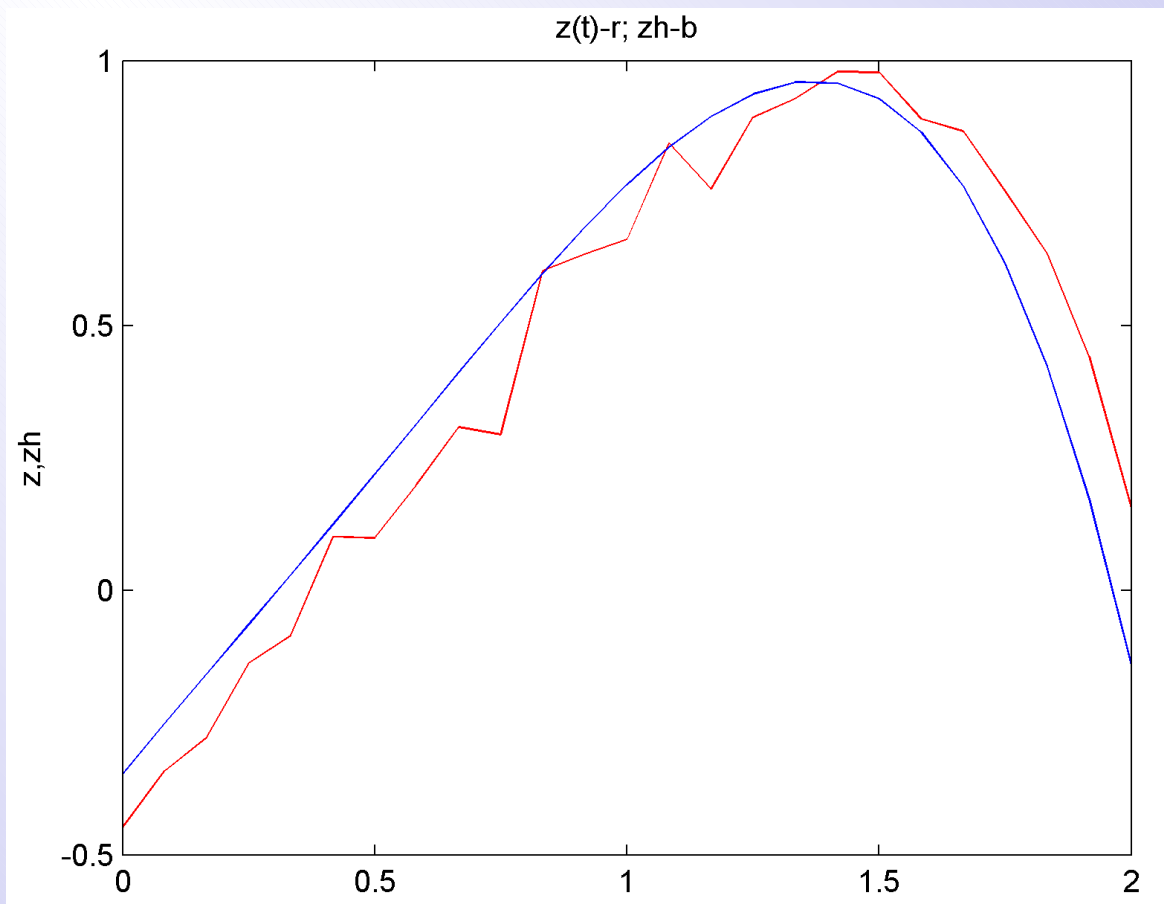


Рис. 2. Измеренные значения. Аппроксимация. N=25

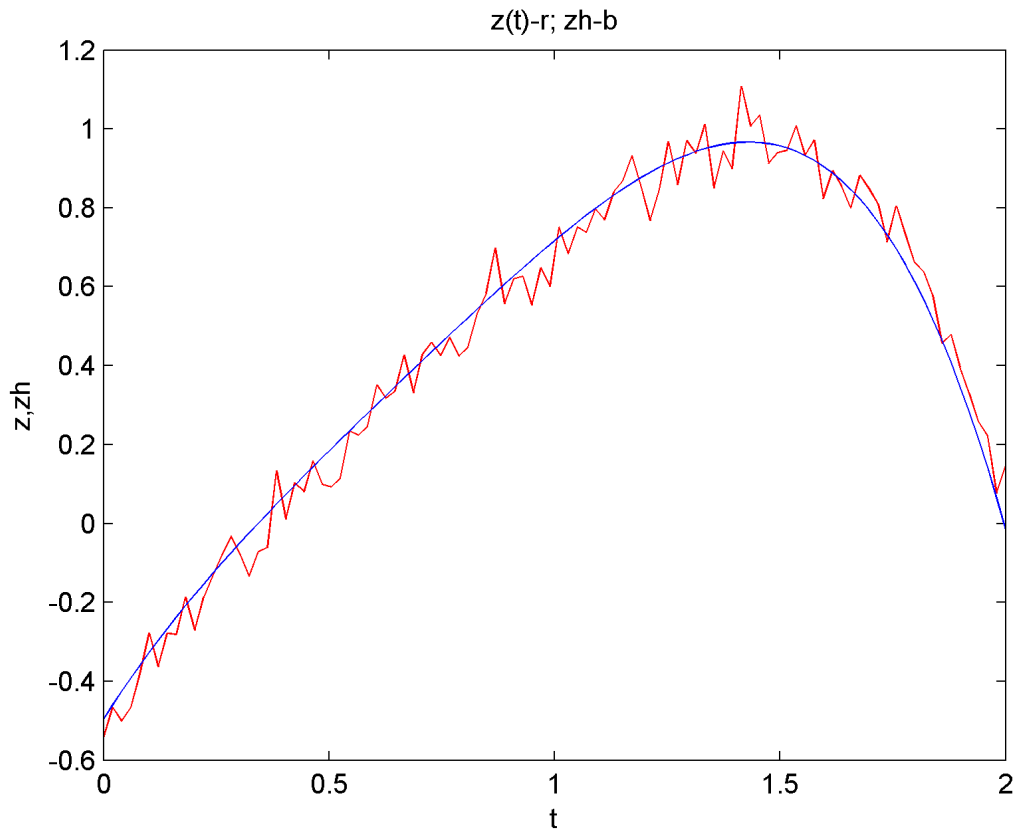


Рис. 3. Измеренные значения. Аппроксимация. N=100

Тригонометрическая регрессия.

Пусть неизвестная функция в (3.2) ищется в виде тригонометрического полинома степени r с неизвестными коэффициентами α_j, β_j , рассматриваемого на $[0, T]$ функциями $\psi_1 = 1, \psi_{2j} = \sin(j\omega t), \psi_{2j+1} = \cos(j\omega t), \omega = 2\pi/T, j = \overline{1, r}$.

Тригонометрический полином степени r имеет вид

$$f(t, \alpha, \beta) = \alpha_0 + \sum_{j=1}^r (\alpha_j \cos(j\omega t) + \beta_j \sin(j\omega t)). \quad (4.17)$$

Запишем модель в комплексной форме

$$f(t, \Theta) = \sum_{j=-r}^r (\theta_j \phi_j(t)), \quad \phi_j(t) = \exp(ij\omega t), \quad (4.18)$$

где $\theta_0 = \alpha_0, \theta_{\pm j} = \alpha_j \pm i\beta_j, j = \overline{1, r}$.

Если $\forall j, j = \overline{1, r}, \alpha_j \in \mathbb{R}, \beta_j \in \mathbb{R}$, то $\alpha_j = \operatorname{Re}\theta_j, \beta_j = \operatorname{Im}\theta_j$

Формулы Эйлера : $\sin x = \frac{e^{ix} - e^{-ix}}{2i} \quad \cos x = \frac{e^{ix} + e^{-ix}}{2}$

Тригонометрическую регрессию применяют к периодическим функциям, заданным на сетке с равноотстоящими узлам: $t_i = (l - 1)T/N$, (значение в узле $t_{n+1} = T$, в силу периодичности, совпадает со значением в начальном узле $t_{n+1} = 0$).

Столбцы матрицы Ψ – ортогональные, т.е. $\forall i, j, i \neq j \quad c_{ij} = \sum_{l=1}^N \psi'_{il} \psi_{lj} = 0$, ψ'_{il} – элемент матрицы Ψ^T , ψ_{lj} – элемент матрицы Ψ , .

Вектор подгонки имеет вид

$$\hat{y}_i = f(t_i, \hat{\Theta}) = \sum_{j=-r}^r \hat{\theta}_j \exp(2\pi i j (l - 1)/N), \quad i = \overline{1, N}. \quad (4.19)$$

Переход (3.20) от вектора Y к вектору B соответствует дискретному преобразованию Фурье (ДПФ) и цифровой фильтрации (обнуление величин b_j при $|j| > r$),

Переход (3.19) от вектора B к вектору Y соответствует обратному дискретному преобразованию Фурье (ОДПФ).

Тогда матрица $C = \Psi^T \Psi$ – диагональная матрица с элементами на диагонали

$$c_{jj} = \sum_{i=1}^N |\psi_{ij}|^2, \quad j = \overline{1, r}, \quad \text{для комплексной формы модели имеем } c_{jj} = N.$$

Решение системы (3.6) имеет вид $\hat{\theta}_j = \frac{b_j}{c_{jj}}, \quad j = \overline{1, r}$, где b_j – элемент матрицы $B = \Psi^T Y$,

$$b_j = \sum_{i=1}^N \psi'_{ij} y_i = \sum_{i=1}^N y_i \exp(-2\pi i j (l - 1)) / N, \quad j = \overline{1, r}. \quad (4.20)$$

$$f(t, \hat{\Theta}) = \sum_{j=-r}^r \hat{\theta}_j \exp(ij\omega t), \quad \hat{\theta}_j = b_j / N$$

Одномерное дискретное прямое и обратное преобразования Фурье

Одномерное дискретное **прямое** преобразования Фурье:

$$X_k = \sum_{n=0}^{N-1} x_n \exp(-i \frac{2\pi nk}{T}), \quad k = \overline{1, N-1}.$$

Одномерное дискретное **обратное** преобразования Фурье:

$$x_k = \frac{1}{N} \sum_{n=0}^{N-1} X_k \exp(i \frac{2\pi nk}{T}), \quad k = \overline{1, N-1}.$$

Здесь N – количество значений сигнала, измеренных за период времени T ; x_k – измеренные значения сигнала (в дискретные моменты времени t_k); X_k – результат ДПФ (N комплексных амплитуд синусоидальных сигналов, составляющих исходный сигнал, $|X_k|$ – амплитуда, $\arg(X_k)$ – фаза k -го синусоидального сигнала); k – индекс частоты.

Частота k -го сигнала равна $\frac{k}{T}$

Реализация ДПФ и ОДПФ пакете MATLAB.

Используют FFT-алгоритм быстрого Фурье-преобразования.

fft (Fast Fourier Transform) Одномерное дискретное прямое преобразования Фурье:

Функция $Y = fft(X)$ вычисляет для массива данных X ДПФ, если массив X двумерный, вычисляется ДПФ каждого столбца.

Функция $Y = fft(X, n)$ вычисляет n -точечное дискретное преобразование Фурье. Если $|X| < n$ – недостающие строки массива X заполняются нулями; если $|X| > n$ – лишние строки удаляются.

ifft (Inverse Fast Fourier Transform) Одномерное дискретное обратное преобразования Фурье:

Функция $X = ifft(Y)$ вычисляет обратное преобразование Фурье для массива Y .

Функция $X = ifft(Y, n)$ вычисляет n -точечное обратное преобразование Фурье для массива Y .

Если $Z = ifft(Y)$, то $b_j = Z_{j+1}$, $b_{-j} = Z_{N-j+1}$, $j = \overline{1, r}$.

При $r \geq N/2$ вектор Y восстанавливается точно ($Y = \hat{Y}$) (соответствует задаче тригонометрической интерполяции).

Функция $f(t)$ может не восстанавливается точно, даже при отсутствии шумов.

Если разложение в ряд Фурье функции имеет вид

$$f(t) = \sum_{j=-\infty}^{\infty} \theta_j \exp(i j \omega t).$$

то коэффициенты Фурье $\hat{\theta}_j$ восстановленной функции имеют вид

$$\hat{\theta}_j = \sum_{j=-\infty}^{\infty} \theta_{j+Nr} \quad (4.21)$$

К коэффициенту $\hat{\theta}_j$ добавляются коэффициенты $\hat{\theta}_{j \pm N}$, $\hat{\theta}_{j \pm 2N}$ и т.д.

Эффект называется наложением частот, связан с дискретизацией значений t . Влияние наложения частот на точность восстановления функции зависит от N и от скорости убывания коэффициентов Фурье $\hat{\theta}_j$ с ростом $|j|$. Скорость убывания коэффициентов Фурье $\hat{\theta}_j$ связана со свойствами гладкости периодического продолжения функции $f(x)$.