МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «ДОНСКОЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

(ДГТУ)

Факультет	Энергетика и нефтегазопромышленность
Кафедра	АММ НГК
Направление	15.03.04 Автоматизация технологических процессов и производств (бакалавриат)

КОНСПЕКТ ЛЕКЦИЙ ПО ДИСЦИПЛИНЕ

«Автоматизация производств нефтегазового комплекса»

Содержание

Глава 1. Диагностика систем управления 4
1.1. Общая характеристика дисциплины 4
1.2. Основные понятия технической диагностики (ТД) 4
1.3. Системы тестовой и функциональной диагностики 5
1.4. Краткая классификация ОД и алгоритмов поиска неисправных элементов 6
1.5. Математические модели ОД и их классификация 7
1.6. Построение логических ФДМ 11
1.7. Построение логической ФДМ для непрерывного комбинационного объекта Д 12
1.8. Тестовый контроль дискретных комбинационных схем 19
1.9. Информационный диагностический тест, близкий к минимальному (метод Синдееева) 26
1.10. Метод критерия различимости – комбинационный метод, реализующий безусловный алгоритм диагностики 32
Глава 2. Диагностика микропроцессорных систем (МПС). 38
Общая характеристика средств диагностики 38
2.1. Общая характеристика и методика использования СА 38
2.2. Точность СА 39
2.3. Условия пригодности модулей МПС для испытания методом СА ... 40
2.4. Методика проверки МПС с помощью СА 41
2.5. Логический анализатор (ЛА) 41
2.6. Простейшие контрольно-испытательные приборы 45
2.6.1. Тестеры логических состояний 45
2.7. Стимулирующий генератор 46
2.8. Бесконтактные индикаторы импульсных токов (ИИТ) 47
2.9. Самоконтроль и самодиагностика в МПС 48
2.10. Диагностика одноплатной микро ЭВМ. 50
Последовательность наладки и контроля микро ЭВМ 50
2.11. Проверка линий системной магистрали в пассивном состоянии 51
2.12. Диагностика ПЗУ 53
2.13. Диагностика ОЗУ 53
2.14. Проверка контроллера последовательного канала связи 54
(На примере платы МП с микросхемой типа К1801ВП033) 54
2.15. Проверки, выполняемые с помощью процессора проверяемой платы 58
2.16. Программное обеспечение наладочных стендов 61
Глава 3. Прогнозирование состояния технических систем. 63
3.1. Задачи прогнозирования 63
3.2. Основные методы прогнозирования 66
3.3. Ресурсные испытания 72
Глава 4. Основные сведения об идентификации систем 75
4.1. Модели, типы моделей и их использование 75
4.2. Идентификация как метод построения моделей 80
4.2.1. Математический аппарат теории идентификации 84

4.2.1. Основные подходы построению алгоритмов идентификации. ... 85
4.3. Идентификация структуры модели ... 86
4.4. Методы упрощения моделей ... 90
4.5. Адекватность модели и объекта ... 96
4.6. Характеристика проблем и методов моделирования объектов ... 101
4.7. Этапы и принципы моделирования объектов и систем на ЭВМ ... 106
4.8. Особенности фиксации и обработки результатов моделирования . 113
Глава 5. Идентификация объектов ... 116
5.1. Постановка задачи и описание проблемы идентификации объектов116
5.2. Идентификация статики линейного детерминированного объекта 118
5.3. Идентификация статики нелинейного детерминированного объекта123
5.4. Построение регрессионной модели стохастического объекта ... 131
5.5. Идентификация динамических объектов ... 147
5.6. Идентификация закона распределения случайной величины ... 152
5.7. Методы переходных функций ... 166
5.8. Частотные методы ... 171
5.9. Непараметрическая идентификация линейных стохастических систем. Уравнение Винера ... 173
Тема 6. Построение модели статики сложных многомерных объектов176
6.1. Выбор существенных переменных модели объекта ... 176
6.2. Оптимальное планирование эксперимента ... 182
6.2.1. Нормирование переменных модели ... 184
6.2.2. Планирование полного факторного эксперимента ... 185
6.2.3. Вычисление коэффициентов модели ... 188
6.2.4. Оценка значимости коэффициентов модели. ... 191
6.2.5. Проверка адекватности модели ... 192
6.2.6. Дробный факторный эксперимент ... 194
6.2.7. Учет влияния помех ... 197
6.2.8. Построение модели с квадратичными эффектами ... 199
6.3. Метод группового учета аргументов ... 202
6.4. Определение наилучшей модели ... 210
Приложение 1 ... 214

Глава 1. Диагностика систем управления

1.1. Общая характеристика дисциплины

Диагностика (Д) – инженерная дисциплина, разрабатывающая теорию и методы организации процессов диагностирования, а также принципы построения технических средств и систем диагностирования.

Существует три типа задач по определению состояния объекта диагностики (ОД). К **первому типу** относятся задачи по определению состояния, в котором находится ОД в настоящее время. Ко **второму типу** – задачи по предсказанию состояния, в котором окажется ОД в некоторый будущий момент времени. Это задача технического прогнозирования. К **третьему типу** – задачи по определению состояния, в котором находился ОД в некоторый момент времени в прошлом. Это задачи технической генетики. Они возникают в связи с расследованием причин аварии (черные ящики) подвижных средств.

Техническая диагностика – первый тип задач. Знание состояния ОД в настоящий момент времени является обязательным для всех трёх задач диагностики.

Родилась техническая диагностика (ТД) в 50-х годах в связи с усложнением технических устройств, которое сопровождает непрерывно идущую научно-техническую революцию.

1.2. Основные понятия технической диагностики (ТД)

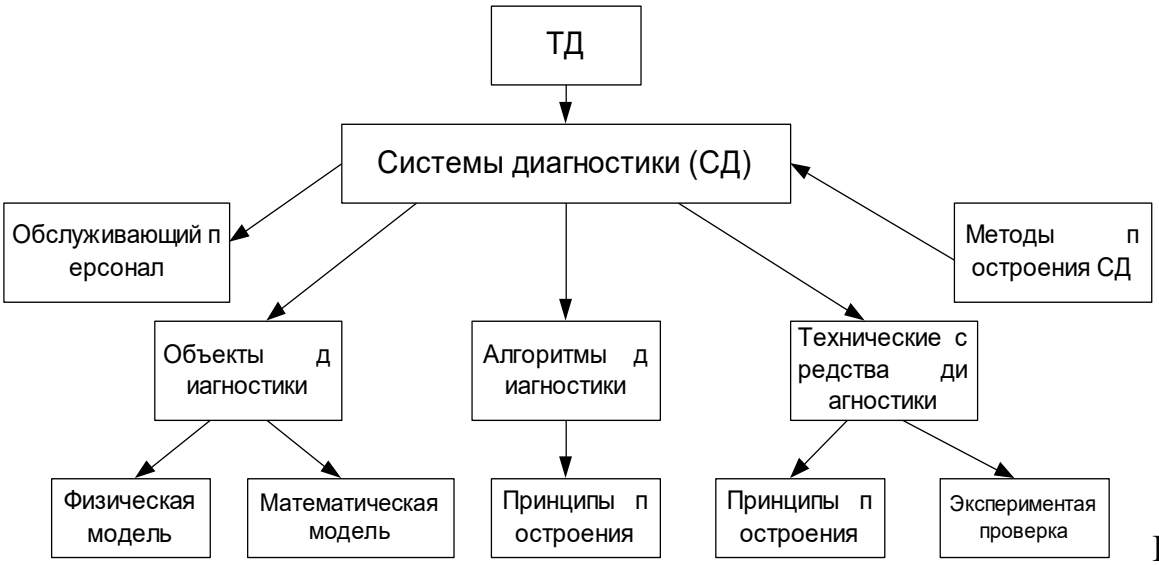


Рис.1

Состояние ОД контролируется с помощью проверок.

Разновидности проверок:

1. часто необходимо выяснить, что ОД исправен, т.е. в нем нет ни одной неисправности. **Проверка исправности** ОД является наиболее полной по объему;

2. на этапе эксплуатации при профилактике ОД необходимо убеждаться в том, что ОД выполняет все функции, предусмотренные его рабочим алгоритмом. Это **проверка работоспособности объекта**. Она является менее полной и может оставлять необнаруженные неисправности, не влияющие на правильную работу ОД;

3. часто осуществляют проверку правильности функционирования ОД в данный момент времени при выполнении одной заданной проверки. Это **проверка правильности функционирования**. Она является еще менее полной.

Степень детализации неисправности, т.е. обнаружение ее с точностью до блока, называют глубиной диагностирования ОД.

Алгоритмы диагностики необходимо применять с глубиной до каскада (усилитель, генератор, электрическая машина и т.д.).

Объекты диагностики, технические средства диагностики и персонал, действующий по определению алгоритмов диагностики, образуют систему диагностирования (СД).

1.3. Системы тестовой и функциональной диагностики

Процессы диагностирования представляют собой многократную дачу на ОД определенных воздействий и анализ реакций на эти воздействия. Воздействия на ОД поступают либо от средств диагностирования, либо являются внешними сигналами, определяемыми рабочим алгоритмом ОД.

Структурная схема тестовой диагностики (ТСТД) представлена на рис. 2.

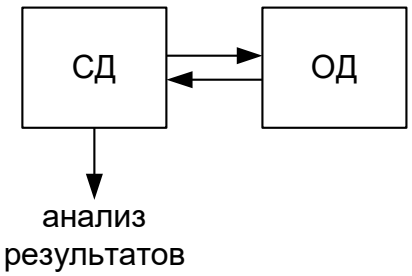


Рис. 2

Такие системы решают задачи проверки исправности, работоспособности и поиска конкретной неисправности. Они используются, когда ОД не применяется по прямому назначению и находится в ремонте. В системах ТСТД тестовые воздействия могут подаваться не только в периоды, когда ОД не

используется по прямому назначению, но также и в процессе функционирования по рабочему алгоритму. В этом случае тестовые сигналы не должны мешать нормальной работе ОД. Это возможно, когда ОД обладает повышенной инерционностью (фоновая программа - диагностическая).

Системы функциональной диагностики представлены на рис. 3.

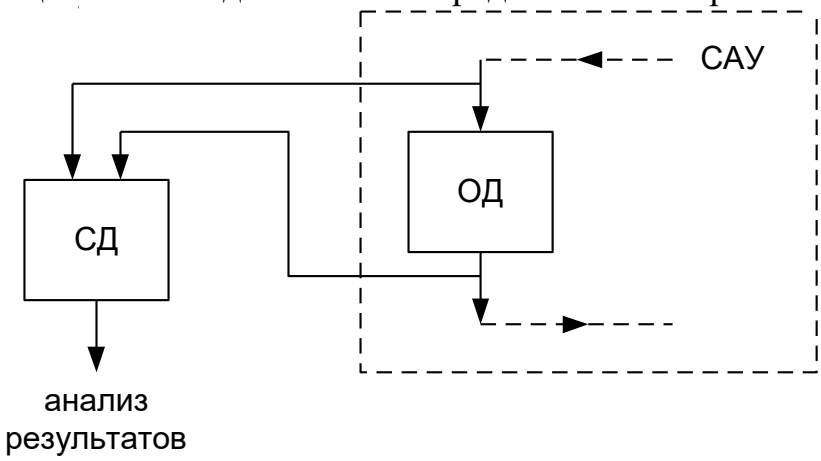


Рис. 3

Для анализа результатов в системах функциональной диагностики необходим эталон (или исправный объект), т.к. надо сравнить отклики с эталона и реального ОД. Для реализации этой системы часто создают упрощенную модель ОД. Например, в системах централизованного контроля температуры, давления, расхода модель ОД представляет собой аппаратуру для задания допустимых значений контролируемых параметров (уставок), а также средства коммутации и подключения этой аппаратуры к устройству сравнения допустимых значений параметров (уставок) с фактическим.

1.4. Краткая классификация ОД и алгоритмов поиска неисправных элементов

По характеру входных и выходных сигналов различают:

1. Непрерывные ОД (эл. приводы, ОУ);
2. Дискретные (САУ с МП – системой управления)
3. Гибридные (ЦАП, АЦП, тиристорные УВ, преобразователи постоянного тока в переменный и наоборот)

Различают ОД комбинационные (или ОД без памяти) и последовательностные.

Аналоговая САУ с несколькими обратными связями является ОД с памятью.

Для обнаружения неисправных функциональных элементов (ФЭ) имеются две группы методов, существенно отличающихся друг от друга: комбинационные и последовательные.

При последовательном методе каждая последующая проверка выбирается в соответствии с принятым алгоритмом поиска в зависимости от результата предыдущей: положительного (1), если выходной сигнал проверяемого ФЭ находится в поле допуска в соответствии с ТУ, или отрицательного (0) в противном случае. Проверки продолжаются до тех пор, пока очередная проверка однозначно не укажет на отказавший ФЭ.

При комбинационном методе обнаружение неисправного ФЭ производится по анализу совокупности (комбинации) наиболее информативных проверок, состав и последовательность которых выбираются заранее в соответствии с принятым безусловным алгоритмом поиска.

Для облегчения анализа совокупности проверок удобно составить двоичные коды неисправностей, каждый из которых соответствует той или иной последовательности «0» и «1». Получающаяся на выходах контролируемых ФЭ разрядность двоичного кода совпадает с числом последовательно выполненных проверок. Критерии отбора проверок в группу наиболее информативных могут быть самыми разными. Так, например, в информационном методе на каждом шаге алгоритма поиска из всех возможных проверок выбирается такая, которая обеспечивает получение наибольшего количества информации относительно предыдущей проверки.

1.5. Математические модели ОД и их классификация

Модели ОД могут быть заданны в явном и неявном виде.

Явная модель – совокупность описания исправного ОД и всех его неисправных состояний.

Неявная модель ОД – описание исправного ОД и некоторых неисправных его состояний.

Как правило, задается математическая модель исправного ОД, по которой можно построить модели его неисправных состояний.

Классификацию диагностических моделей можно выполнить по следующим признакам:

1. По методам представления взаимосвязи состояний ОД, его элементов и параметров различают:

а. Аналитические модели.

Они позволяют решать оптимизационные задачи и получать в явном виде соотношения между состояниями ОД, диагностируемыми параметрами и показателями качества систем управления. Это метод малого параметра и метод функции чувствительности. Для функции чувствительности надо иметь передаточную функцию САУ: $\Phi(p)=F(к_1, к_2, \ldots, к_n, p)$, где $к_1$-$к_n$ – входные (собственные параметры ОД). Под функцией чувствительности понимают

$$S_{k_i} = \frac{\partial \Phi}{\partial \kappa_i} \quad \begin{array}{l} \kappa_i = k_{i0} \\ k_i \neq k_j, \\ k_j = k_{j0}, j = 1 - n \end{array} .$$

Нижний подиндекс 0 указывает на номинальное значение параметра.

(*) Если заменить дифференциальный оператор p на jω, то

$$S_{k_i} = \frac{A + jB}{C + jD}$$

$$\left| S_{k_i} \right| = \sqrt{\frac{A^2 + B^2}{C^2 + D^2}}$$

где $A=f_1(\kappa_1, \ldots, \kappa_n, \omega)$

- - - - - - - - - - - - - -

$D=f_4(\kappa_1, \ldots, \kappa_n, \omega)$

Есть два подхода в использовании функции чувствительности:

2. Частотный подход /начиная с (*)/

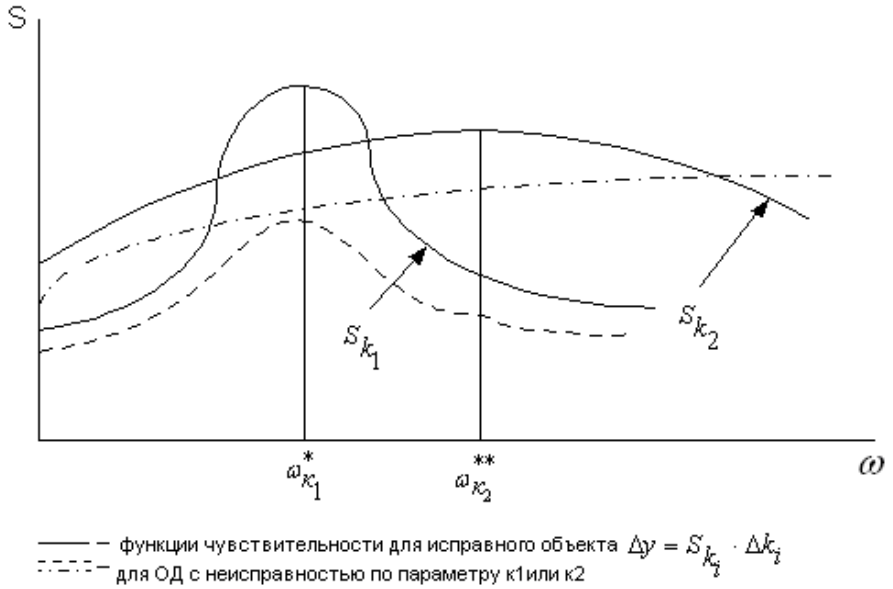


Рис. 4

При изменении параметра схемы κ_i (при ω*=const) следят за тем, чтобы Δy контролируемого ОД было меньше либо равно Δy исправного ОД.

2. Назначение равных классов допусков.

При этом находят связь между относительным изменением выходного параметра y и относительными изменениями входных параметров ОД.

$$\Delta y = S_{x_i} \cdot \Delta x_i \text{, где } S_{x_i} = \left.\frac{\Delta y}{\Delta x_i}\right|_{\substack{x_i = x_{i0} \\ i = 1...n}} . \qquad (1)$$

Выполнив несложные преобразования, получим соотношение, связывающие относительное изменение входного параметра x_i с относительным значением выходного параметра y.

$$\frac{\Delta y}{y_0} = S_{xi} \frac{\Delta x_i}{x_{i0}} \cdot \frac{x_{i0}}{y_0} = A_{xi} \frac{\Delta x_i}{x_{i0}}, \qquad (2)$$

Тогда

$$A_{x_i} = S_{xi} \cdot \frac{x_{i0}}{y_0}, \qquad (3)$$

A_{x_i} - коэффициент влияния i-го параметра.

По (3) можно найти максимальные допуски $\frac{\Delta x_i}{x_{i_0}}$ на параметры схемы x_i, если предварительно вычислить коэффициенты влияния A_{x_i} (отношение $\frac{\Delta y}{y_0}$ задается заказчиком).

В реальных условиях изменяются значения всех параметров схемы (технологический разброс, внешние воздействия и т.д.), поэтому назначают один допуск на все пассивные и активные элементы схемы, используя соотношение по методу равных классов допусков. При этом

$$\frac{\Delta x_1}{x_{10}} = \frac{\Delta x_2}{x_{20}} = ... = \frac{\Delta x_n}{x_{n0}} = \frac{\Delta y}{y_0} \cdot \frac{1}{\sqrt{A_{x_1}^2 + A_{x_2}^2 + ... + A_{x_n}^2}} .$$

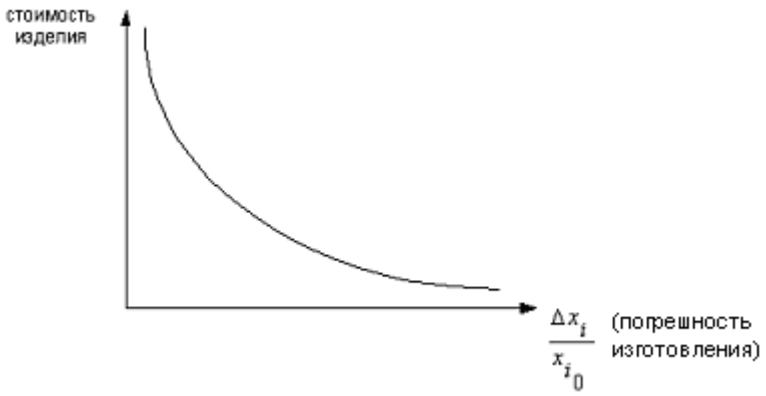


Рис. 5

На рис. 5 представлена зависимость стоимости изделия от точности его изготовления.

Другим видом аналитической модели является корневой годограф, т.е. динамическое распределение корней характеристического уравнения при вариации внутренних параметров схемы. При этом корни характеристического уравнения должны оставаться в левой полуплоскости (комплексной плоскости) UjV и выполнятся условия обеспечения заданного запаса устойчивости как по модулю, так и по фазе.

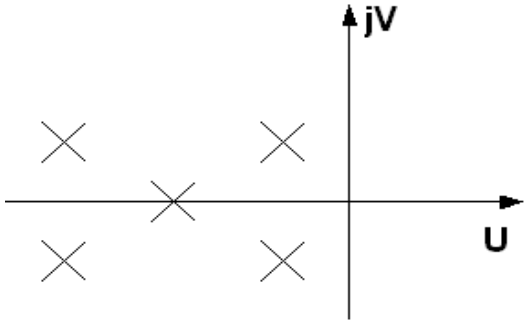


Рис.6

b. Графоаналитические модели.

Это диаграммы прохождения сигналов, протекающих в аналоговых объектах. Они строятся с помощью теории графов и теории множеств.

c. Логические функционально-диагностические модели (ФДМ).

Эти модели построены на основе логического анализа функциональных и принципиальных схем устройств.

d. Информационные модели.

Информационные модели – представляют собой информационное описание систем и объектов диагностики. Методика их построения будет рассмотрена ниже.

Диагностические модели (ДМ) должны дать ответ на вопрос: какой параметр(ы) подлежат контролю в процессе эксплуатации.

ДМ может быть несколько на разных уровнях описания системы автоматики, но главное состоит в том, что ДМ должны упрощать наши представления о сложных зависимостях в системах, а также позволять с минимальными затратами средств и времени решать задачи диагностики.

Одной из наиболее популярных моделей является логическая ФДМ. Для ее построения структура ОД должна быть задана полностью и для нее определены области допустимых значений входных и выходных параметров всех ее блоков. ФДМ представляют в виде графа, вершинами которого является функциональные блоки Q_i, а ребрами – связи между блоками, входными и выходными полюсами схемы. Наиболее рационально искать неисправности последовательно на разных уровнях: система – агрегат – блок – модуль – каскад – деталь. Поэтому строят несколько ФДМ, вначале для устройства в целом с глубиной поиска до неисправного агрегата, затем для каждого блока с глубиной поиска до модуля, а затем до каскада или корпуса.

Для каждой ФДМ разрабатывают свою программу поиска неисправностей, при этом неисправности цепей питания не учитывают, хотя их проверка осуществляется в первую очередь.

Считается, что в ФДМ линии связи между каскадами исправны. Поэтому перед диагностикой осуществляют визуальный контроль трассировки печатной платы и межблочных соединений. В таком случае указанные допущения не являются грубыми.

Модели отдельных каскадов обычно не строят. Для них пригоден способ поэлементной проверки по принципиальной схеме каскада. На это уйдет меньше времени и технических средств.

1.6. Построение логических ФДМ

Если входные и выходные сигналы характеризуются несколькими физическими параметрами (A, f, φ), то каждый из этих параметров должен представляться отдельным входом и выходом на блоке ФДМ. При этом происходит расщепление входов и выходов по числу выделенных признаков сигналов, что значительно усложняет модель схемы.

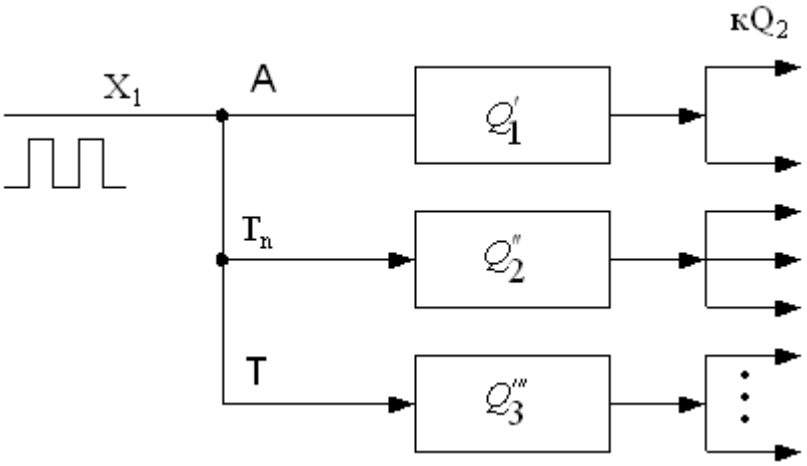


Рис.7

Каждый из блоков ФДМ имеет два состояния S:
«0» -неисправен,
«1» - исправен.

Во всех случаях следует избегать расщепления входов и выходов, т.к. эта операция усложняет схему.

Во многих объектах диагностики это возможно при применении осциллографа, который одновременно контролирует все параметры сигналов.

Входы и выходы блоков ФДМ должны быть контролируемыми. Для этого они должны быть выведены на лицевую панель, заднюю панель, специальный диагностический разъем, КТ или штыри печатной платы.

В случаях без расщепления в ФДМ каждый блок исходной функции схемы заменяется блоком ФДМ, каждый из которых имеет один выход и существенные для данного выхода входы. Блоки ФДМ обозначают $Q_i(i=\overline{1,n})$.

Если расщепление входов не происходит, то ФДМ часто совпадет с исходной функциональной схемой.

Выходной сигнал схемы F можно представить как конъюнкцию входных сигналов $x_i(i=\overline{1,m})$ и выходных сигналов всех блоков схемы $Z_j(j=\overline{1,n})$, где m – число входов схемы, а n – число блоков в схеме.

Условие работоспособности схемы

$$F = x_1 x_2 ... x_m \cdot z_1 ... z_n = 1.$$

Для многорежимных объектов диагностики каждый режим должен рассматриваться отдельно.

1.7. Построение логической ФДМ для непрерывного комбинационного объекта Д

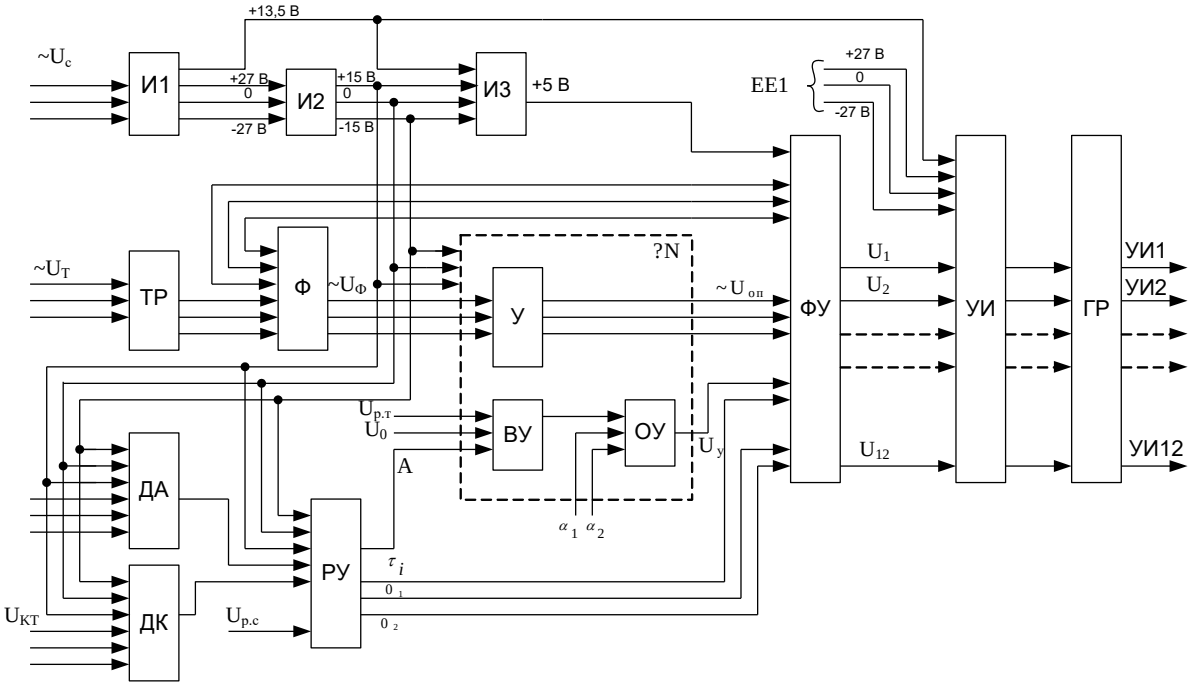


Рис. 8. Функциональная схема СИФУТП КТЭ 100/200

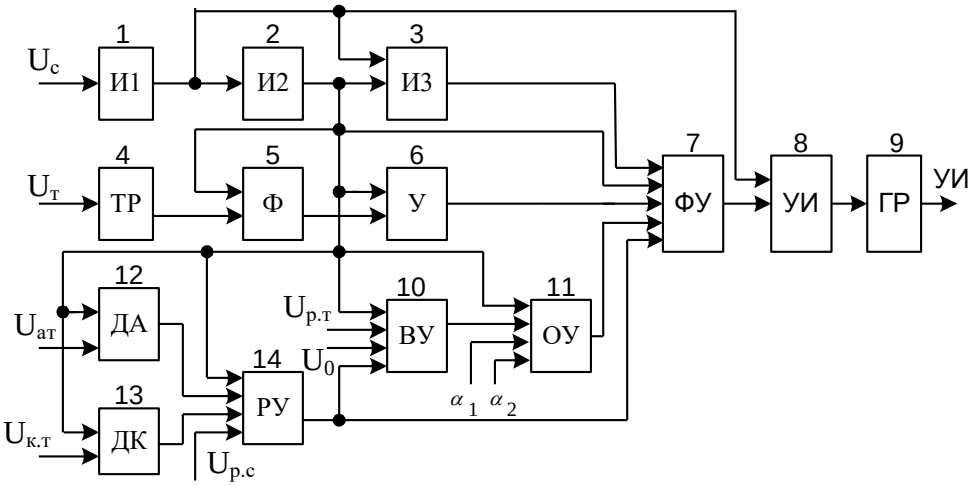


Рис. 9. Функциональная модель СИФУ КТЭ 100/200

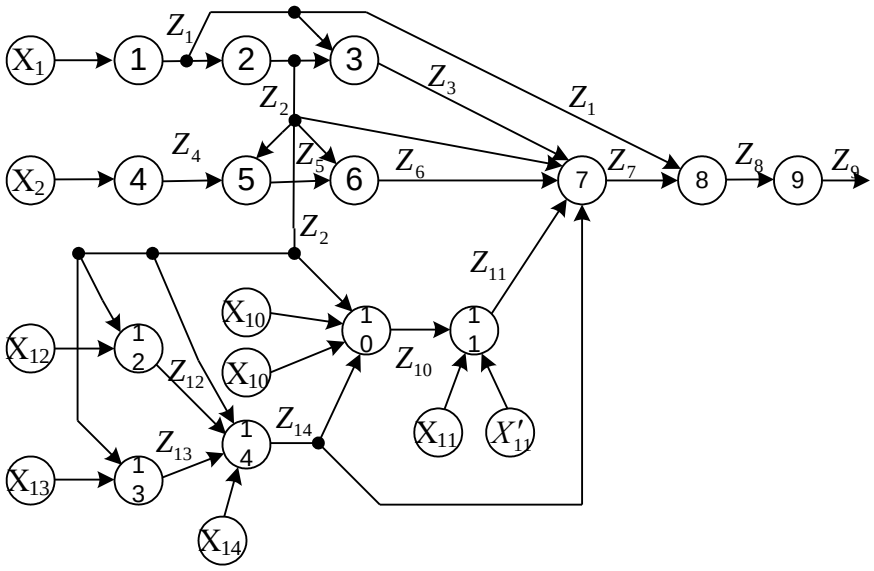


Рис.10. Логическая модель СИФУ КТЭ 100/200

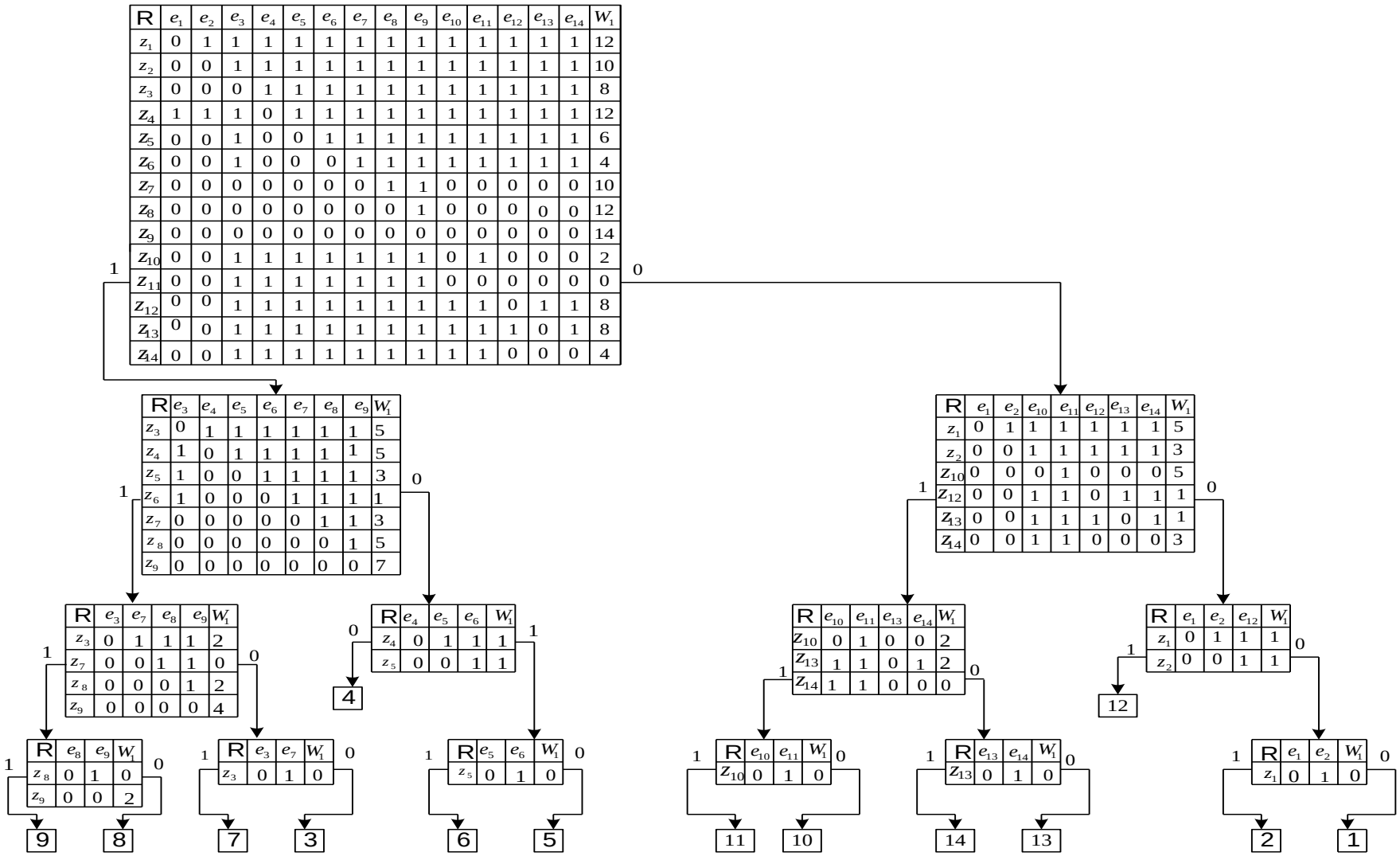


Рис.11. Алгоритм диагностирования по ТФН и функциям

Комбинационной является схема импульсно-фазового управления тиристорным преобразователем электропривода СИФУТП КТЭ 100/220.

Для этого рассмотрим функциональную схему СИФУТП на рис.8;

$И_1$ – источник питания ±27 В, ±13,5 В;

$И_2$ – источник питания ±15 В;

$И_3$ - источник питания ±5 В;

ТР – блок тиристоров для формирования опорных синусных напряжений, синхронных с напряжением питания ТП.

Ф – ячейка фильтров для фильтрации опорных напряжений 50 Гц от высокочастотных составляющих сети.

У – операционные усилители для согласования по уровню опорных напряжений $U_ф$ с входными опорными сигналами.

$И_{оп}$ – ячейки каналов фазового управления ФУ

ВУ – входной усилитель для согласования по мощности сигналов управления ТП со стороны регулятора тока $U_{рт}$ с сигналами управления ячеек СИФУ.

В зависимости от блокирующего сигнала Б на входе ВУ, последний ограничивает уровень своего выходного напряжения.

ОУ – операционный усилитель для ограничения минимального $\alpha_1=30^o$ и максимального $\alpha_2=150^o$ углов управления ТП с RL – нагрузкой

ФУ – ячейка каналов фазового управления для формирования импульсов управления тиристорами ТП.

ЯС – ячейка согласования;

И1 и И2 – сигналы отключения управляющих импульсов на тиристоры анодной или катодной группы.

УИ – ячейка усилителей импульсов.

ГР – ячейка гальванической развязки импульсов управления тиристорами ТП и цепей СИФУ.

ДА, ДК – блоки датчиков состояния тиристоров анодной и катодной группы ТП.

РУ – ячейка раздельного управления комплектами реверсивного ТП.

На базе функциональной схемы СИФУ (рис.8) строится функциональная модель СИФУ (см. рис. 9).

На следующем этапе происходит абстрагирование от конкретного содержания схем, обозначенных вершинами на функциональной модели рис.9.

В вершинах графа логической ФДМ остаются лишь номера ранее выделенных функциональных блоков (узлов).

В результате получится логическая модель СИФУ на рис. 10. Вершины, соответствующие внешним сигналам, обозначены $x_1 \ldots x_m$.

На следующем этапе по логической модели строится таблица функций неисправности ТФН, показанная на рис.11.

ТФН представляет из себя двухмерную матрицу, строками которой является выходы z_i (i=1,…,n) функциональных узлов логической модели, а столбцами – состояния анализируемой схемы e_j (j=1,…,n).

Состояние e_j соответствует наличию одиночной неисправности в j–м функциональном узле схемы.

На пересечении i-й строки и j–го столбца ставится 1, если при j-й неисправности выходной сигнал i-го блока остается в допуске по ТУ, и Ø – в противном случае.

Состояние e_0 соответствует всем исправным блокам схемы (функциональным узлам).

Если Ø или любая неисправность появляются в блоках, охваченных ОС или в звене ОС, то это приводит к появлению Ø во всем контуре, охваченном ОС. Поэтому при диагностике СУ необходимо разрывать цепи ОС, и это надо предусмотреть при проектировании устройства (ввести управляемые разрывы). Это очень легко сделать, обрывая сигналы с датчиков в этом кольце и переключая при диагностике вход с датчика на стандартное напряжение, соответствующее номинальному сигналу с нормирующего усилителя датчика.

После построения ТФН ее можно обработать с помощью различных алгоритмов диагностики.

Пример: для рассматриваемого непрерывного комбинационного ОД., которым является СИФУ, удобно реализовать условный алгоритм диагностики, использующий функцию предпочтения W_i.

В зависимости от поставленной заказчиком задачи и наличия исходных данных, W_i можно рассчитать следующим образом:

1. При равновероятных состояниях отказа функциональных узлов схемы и равных ценах элементарных проверок выходов z_i (в рублях или часах) функция предпочтения рассчитывается по формуле

$$W_i = \left| \sum_i "0" - \sum_i "1" \right|.$$

Первое слагаемое равно сумме нулей в «i» строке. Второе слагаемое равно сумме «единиц» в i-ой строке. Считается функция предпочтения W_i по строкам.

2. При известных (или легко определяемых) вероятностях состояний системы, каждое из которых соответствует отказу одного из функциональных узлов (ФУ).

$$W_i = \left| \sum_i P(S_j)_{"0"} - \sum_i P(S_j)_{"1"} \right|.$$

Известно [10], что интенсивность отказов схемы равна (для экспоненциального закона надежности) сумме интенсивностей отказов $\lambda_i (i = \overline{1,n})$ составляющих ее элементов:

$$\Lambda = \sum_i^n \lambda_i .$$

Вероятность отказа функционального узла (ФУ)

$$Q(t) = 1 - P(t) = 1 - e^{-\Lambda t} \approx \Lambda t ,$$

$$q_i = \lambda_i t$$

$$P(S_j) = P(q_j) = \frac{\lambda_j t}{\Lambda t} = \frac{\lambda_j}{\sum_j^n \lambda_j} .$$

3. Учитывают цены C_j проверок в рублях или часах

$$W_i = \left| \sum_i C_j \cdot P(S_j)_{"0"} - \sum_i C_j \cdot P(S_j)_{"1"} \right| .$$

Первой выбирается проверка, которой соответствует минимальное значение функции W_i. При контроле любого выхода возможен диагностический результат «Ø» или «1». Чтобы выбрать, какой из выходов должен контролироваться вторым, исходная ТФН должна быть разделена на 2 подТФН.

В первую подТФН войдут состояния, которым соответствуют «Ø» в выбранной строке (с минимумом W_i), а в другую – состояния, которым соответствует «1» в выбранной строке. И снова начинается поиск в этих двух под ТФН по минимуму W_i. (см. рис. 11).

Первым будет контролироваться выход Z_{11}, т.к. ему в исходной ТФН соответствует минимальное значение (нулевое) функции W_i.

Согласно изложенным правилам, второй проверкой при положительном результате (R=1) первой проверки стала Z_6, а при отрицательном результате первой проверки - Z_{12}. Далее все происходит по аналогии с изложенным. Деление ТФН на каждом шаге поиска одиночных неисправностей происходит до тех пор, пока очередная под ТФН не превратится в матрицу-строку, что позволяет однозначно указать на неисправный элемент.

На документе, передаваемом пользователю разработчиком, ТФН и все под ТФН не показываются, а приводится только граф, корневая и промежуточные вершины которого соответствуют проводимым

проверкам, а висячие вершины графа однозначно указывают на те или иные одиночные неисправности в диагностируемом объекте.

В графе проверки, передаваемым пользователю, нет смысла приводить исходную ТФН и другие подТФН, а сразу в вершинах бинарного дерева указывают номер следующей проверки. Часть бинарного дерева, соответствующая положительному результату (R=1) первой проверки Z_{11}, показана на рис. 12. Первую диагностическую проверку (в нашем случае 11) называют корневой, а конечные проверки по тому или иному пути – висячими. Висячие вершины графа возможных проверок обозначаются двойным кружком и указывают на номер того или иного дефекта в контролируемой схеме.

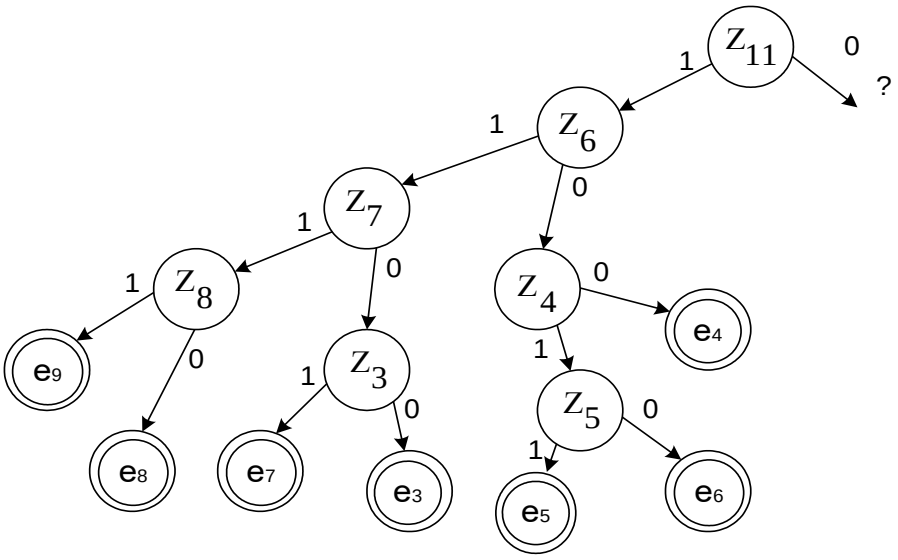


Рис.12. Бинарное дерево для положительного диагностического результата первой проверки

По существу условный алгоритм поиска неисправностей (дефектов) в схеме представляет собой не что иное, как математизированное описание широко известного в электронике метода средней точки, который предполагает контроль в «точках», делящих схему на правомерно равноценные, в смысле надежности, части.

1.8. Тестовый контроль дискретных комбинационных схем

Тестовый контроль предназначен для определения характера и места нахождения неисправности. При этом имеются в виду неисправности типа разрыв цепи и короткое замыкание.

Рассмотрим принцип тестового контроля применительно к комбинационным схемам. Будем считать, что в рассматриваемой схеме

одновременно может быть только одна неисправность – разрыв, эквивалентный постоянному логическому нулю, или короткое замыкание, эквивалентное постоянной логической единице .

Пусть мы имеем логическую схему S0 в исправном состоянии. Она реализует логическую функцию $f_0(x_1,x_2, \dots ,x_n)$, называемую функцией исправной схемы.

При наличии i-й неисправности схема S0 переходит в схему Si, реализующую новую функцию $f_i(x_1,x_2, \dots ,x_n)$, называемую функцией неисправности.

Рассмотрим комбинационную схему, представленную на рисунке 1, и для неё запишем все функции неисправности.

Функция исправной схемы выглядит следующим образом :

$$f_0 = (x_1 \& x_2 \cup \bar{x}_3) \& (\bar{x}_1 \cup \bar{x}_4)$$

Для неисправностей типа “короткое замыкание” функции неисправностей имеют вид:

$$f'_{x_1} = (1 \& x_2 \cup \bar{x}_3) \& (\bar{x}_1 \cup \bar{x}_4) \quad \text{при} \quad x_1 = 1;$$

$$f'_{x_2} = (x_1 \& 1 \cup \bar{x}_3) \& (\bar{x}_1 \cup \bar{x}_4) \quad \text{при} \quad x_2 = 1;$$

$$f'_{\bar{x}_3} = (x_1 \& x_2 + 1) \& (\bar{x}_1 + \bar{x}_4) \quad \text{при} \quad \bar{x}_3 = 1;$$

$$f'_{\bar{x}_1} = (x_1 \& x_2 \cup \bar{x}_3) \& (1 \cup \bar{x}_4) \quad \text{при} \quad \bar{x}_1 = 1;$$

$$f'_{\bar{x}_4} = (x_1 \& x_2 \cup \bar{x}_3) \& (\bar{x}_1 \cup 1) \quad \text{при} \quad \bar{x}_4 = 1.$$

Для неисправности типа “разрыв” функции неисправностей имеют вид:

$$f_{x_1} = (0 \& x_2 \cup \bar{x}_3) \& (\bar{x}_1 \cup \bar{x}_4) \quad \text{при} \quad x_1 = 0;$$

$$f_{x_2} = (x_1 \& 0 \cup \bar{x}_3) \& (\bar{x}_1 \cup \bar{x}_4) \quad \text{при} \quad x_2 = 0;$$

$$f_{\bar{x}_3} = (x_1 \& x_2 \cup 0) \& (\bar{x}_1 \cup \bar{x}_4) \quad \text{при} \quad \bar{x}_3 = 0;$$

$$f_{\bar{x}_1} = (x_1 \& x_2 \cup \bar{x}_3) \& (0 \cup \bar{x}_4) \quad \text{при} \quad \bar{x}_1 = 0;$$

$$f_{\bar{x}_4} = (x_1 \& x_2 \cup \bar{x}_3) \& (\bar{x}_1 \cup 0) \quad \text{при} \quad \bar{x}_4 = 0.$$

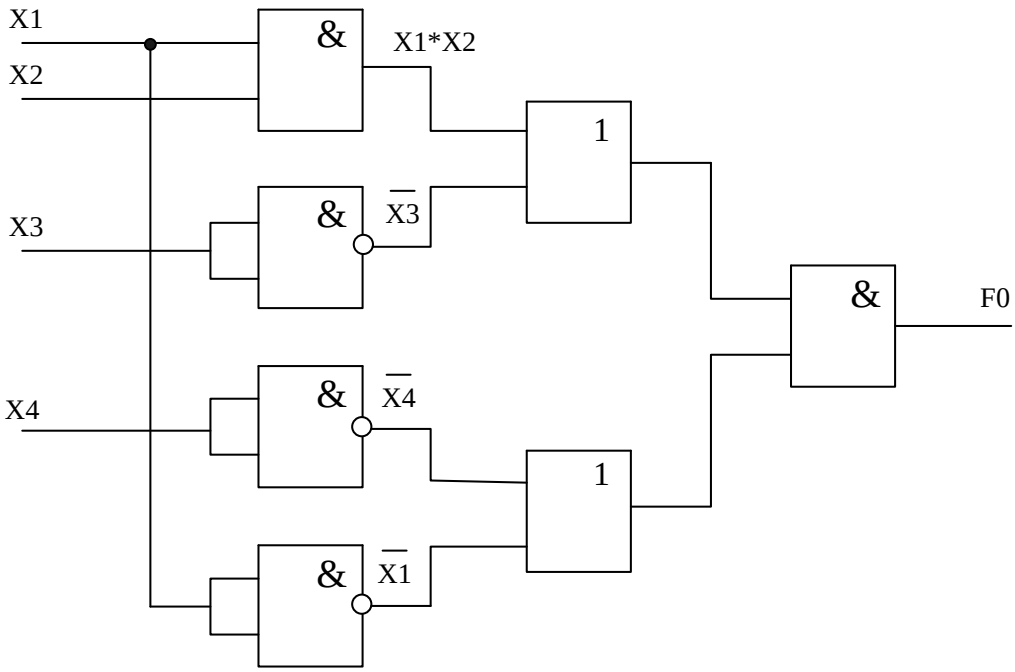


Рис.13. Комбинационная схема №1

По выражениям исправной и неисправной схем строится таблица функций неисправностей, в которую записывают все наборы значений аргументов и значения функций исправной схемы, вычисленные на этих наборах. Для рассматриваемого случая наборы функций неисправностей представлены в таблице 1. В верхней части таблице выделены наборы, на которых f_0=1 и записаны функции неисправностей при разрывах, отличающиеся от значений функции f_0. В нижней части таблицы выделены наборы, на которых f_0=0, и записаны значения функций неисправностей при “коротком замыкании”, отличающийся от функции f_0.

По таблице функций неисправностей можно осуществить поиск неисправностей. Пусть при реализации всех значений наборов аргументов из верхней части таблицы на выходе схемы получен, например, сигнал логического“0” только на наборах е1 и е5. Из таблицы видно, что реализуется схемой в этом случае только функция $f_{\bar{x}_1}$. А это значит, что имеет место разрыв контакта $\bar{x}_1 = 0$.

Такой способ поиска неисправностей предполагает реализацию всех наборов значений аргументов.

Таблица 1

	Наборы значений аргументов				f_0	Значение функции неисправностей при «к.з.»					Значение функции неисправностей при «разрыве.»				
	X_1	X_2	X_3	X_4		f_{x1}	f_{x2}	$f_{\bar{x}3}$	$f_{\bar{x}1}$	$f_{\bar{x}4}$	f_{x1}	f_{x2}	$f_{\bar{x}3}$	$f_{\bar{x}1}$	$f_{\bar{x}4}$
e_0	0	0	0	0	1								0	1	
e_1	0	0	0	1	1								0	0	
e_4	0	1	0	0	1										
e_5	0	1	0	1	1										
e_8	1	0	0	0	1								0	1	0
e_{12}	1	1	0	0	1										0
e_{14}	1	1	1	0	1						0	0			0
e_2	0	0	1	0	0			1							
e_3	0	0	1	1	0			1							
e_6	0	1	1	0	0	1		1							
e_7	0	1	1	1	0	1		1							
e_9	1	0	0	1	0										
e_{10}	1	0	1	0	0		1	1							
e_{11}	1	0	1	1	0										
e_{13}	1	1	0	1	0				1	1					
e_{15}	1	1	1	1	0				1	1					

Рассмотрим методику отыскания неисправностей путём использования лишь некоторого количества наборов, а не всех полностью.

Пусть мы имеем М-множество функций неисправностей, включающее и функцию исправной схемы, заданных на множестве Е наборов значений аргументов. Пусть далее все функции из множества М попарно различимы и имеется множество Р пар функций неисправностей вида: f_i', f_i и множество Р всех возможных пар функций неисправностей за исключением пар f_i, f_j. Тогда **диагностическим тестом** (относительно Е, М, Р) называется множество наборов аргументов $T_g \in E$, на котором все функции неисправностей (f_i, f_j) попарно различимы.

Таким образом, для поиска неисправностей достаточно проверить схему не на всех наборах значений аргументов, а только на тех, которые входят в соответствующий тест. Как правило T_g <E, вследствие чего проверка оказывается более короткой.

Как определить наборы, которые входят в тест T_g?

В общем случае функции f_i и f_j могут различаться не на одном наборе, а на некотором множестве наборов. Поэтому для каждой пары функций неисправностей можно составить функцию таких наборов.

$$e_1 \cup e_2 \cup \ldots e_k = \cup e_m \ . \qquad (4)$$

Для того, чтобы все пары функций неисправностей были одновременно различимы, необходимо образовать конъюнкцию полученных дизъюнкций.

$$\& \cup e_m \qquad (5)$$

Затем (5) преобразуется в дизъюнкцию конъюнкций $\cup \&$ по правилам алгебры логики в предположении, что каждый входной набор отождествляется с логической переменной.

Каждое слагаемое (конъюнкция) полученного выражения будет представлять собой диагностический тест.

Выражение (4) для каждой пары функций неисправностей составляется по таблице функций неисправностей. С этой целью сравниваются значения функций на каждом из наборов значений аргументов и выписываются дизъюнкции наборов всех пар функций неисправностей, то есть составляются выражения (4). Затем составляется конъюнкция дизъюнкций (5), после чего полученные выражения преобразуются в дизъюнкцию конъюнкций.

Каждая конъюнкция наборов будет диагностическим тестом.

Рассмотрим алгоритм получения тестов:

В таблице функций неисправностей вычёркиваются строки, содержащие значения функций либо только с нулями, либо только с единицами;

Из группы неразличимых строк остаётся одна, остальные вычёркиваются;

Для каждой пары функций неисправностей составляется дизъюнкция наборов значений аргументов, при которых эти функции отличаются друг от друга;

Из полученных выражений (4) составляются выражения (5), для чего в начале выписываются одночленные дизъюнкции. Многочленные дизъюнкции, в которые входит набор из числа выписанных одночленных дизъюнкций, вычёркиваются на основании операций поглощения. Из остальных выражений выписываются двучленные дизъюнкции и вычёркиваются те многочлены, которые содержат любую двучленную дизъюнкцию и т.д.;

Выражение, полученное в пункте 4, преобразуется в дизъюнкцию конъюнкций.

Каждая конъюнкция при этом будет диагностическим тестом.

Построим диагностический тест для поиска неисправностей типа “разрыв”в схеме на рисунке 1.

Для построения теста используется верхняя часть таблицы 1.

Произведем поиск неисправности по алгоритму, изложенному выше:

- пункт 1 алгоритма к данному примеру неприменим, так как среди значений неисправностей при разрыве нет строк только с одними единицами;
- по пункту 2 алгоритма вычёркиваем строки e_4 и e_5;
- по пункту 3 составляем дизъюнкцию наборов для каждой пары функций, на которых они отличаются друг от друга.

$\cup\ f_0 f_{x1} = e_{14};$

$\cup\ f_0 f_{x2} = e_{14};$

$\cup\ f_0 f_{\bar{x}3} = e_0 V e_1 V e_8;$

$\cup\ f_0 f_{\bar{x}1} = e_1;$

$\cup\ f_0 f_{\bar{x}4} = e_8 V e_{12} V e_{14};$

$\cup\ f_{x1} f_{\bar{x}3} = e_0 V e_1 V e_8 V e_{14};$

$\cup\ f_{x1} f_{\bar{x}1} = e_1 V e_{14};$

$\cup\ f_{x1} f_{\bar{x}4} = e_8 V e_{12};$

$\cup\ f_{\bar{x}3} f_{\bar{x}1} = e_0 V e_8;$

$\cup\ f_{\bar{x}3} f_{\bar{x}4} = e_0 V e_1 V e_{12} V e_{14};$

$\cup\ f_{\bar{x}1} f_{\bar{x}4} = e_1 V e_8 V e_{12} V e_{14}.$

Просмотрев полученные дизъюнкции, видим, что одночленные дизъюнкции e_1 и e_{14} поглощают все остальные за исключением $e_8 \cup e_{12}$ и $e_0 \cup e_3$.

Тогда конъюнкция дизъюнкций будет иметь вид:

$$\& \cup = e_1 \& e_{14} \& (e_8 \cup e_{12}) \& (e_0 \cup e_3) \ . \qquad (6)$$

По пункту 5 преобразуем выражение (6), используя распределительный закон:

$$\cup \& = e_0 \& e_1 \& e_8 \& e_{14} \cup e_1 \& e_{12} \& e_{14} \& e_3 \cup \ldots \qquad (7)$$

Таким образом, получаем равноценные тесты (конъюнкции), каждый из которых состоит из четырёх наборов:

$$T_1 = e_0 \& e_1 \& e_8 \& e_{14}\ ,\ \ T_2 = e_1 \& e_{12} \& e_{14} \& e_3 \ldots$$

Реализация любого из них достаточна для поиска неисправностей типа “разрыв”.

Построение тестов для “к.з.” аналогично и осуществляется по нижней части таблицы 1.

Поиск неисправности осуществляется путем подачи в произвольно выбранной последовательности наборов теста и фиксации состояний входа. Для этого необходимо закодировать неисправности. Принцип кодирования неисправностей заключается в следующим. Для каждого набора теста по таблице функции неисправности записывается состояние выхода схемы, соответствующее конкретной неисправности. При этом учитывается выбранная последовательности записи наборов в тесте. В результате получаем двоичные числа (коды), каждое из которых описывает поведение схемы при конкретной неисправности.

В ходе “прогонки” теста фиксируются действительные значения выходов схемы, и после реализации всех тестовых наборов производится последовательное сравнение полученного двоичного числа с кодом каждой неисправности, по которому она и определяется.

Для поиска неисправности типа: ”разрыв в схеме ” на рис.13 возьмём тест T_1 и произведем кодирование неисправностей.

Для наглядности тест представим в виде графа, вершинами которого будут являться наборы тестов, а ребра обозначают значения выходов схемы и функции, которые реализуются при данных значениях.

Таблица функций неисправностей

	Наборы значений аргументов				f_0	Значение функции неисправностей при «к.з.»					Значение функции неисправностей при «разрыве.»				
	X_1	X_2	X_3	X_4		f_{x1}	f_{x2}	$f_{\bar{x}3}$	$f_{\bar{x}1}$	$f_{\bar{x}4}$	f_{x1}	f_{x2}	$f_{\bar{x}3}$	$f_{\bar{x}1}$	$f_{\bar{x}4}$
e_0	0	0	0	0	1								0		
e_1	0	0	0	1	1								0	0	
e_4	0	1	0	0	1								0		
e_5	0	1	0	1	1								0	0	
e_8	1	0	0	0	1								0		0
e_{12}	1	1	0	0	1										0
e_{14}	1	1	1	0	1						0	0			0
e_2	0	0	1	0	0			1							
e_3	0	0	1	1	0			1							
e_6	0	1	1	0	0	1		1							
e_7	0	1	1	1	0	1		1							
e_9	1	0	0	1	0										
e_{10}	1	0	1	0	0		1	1							
e_{11}	1	0	1	1	0										
e_{13}	1	1	0	1	0				1	1					
e_{15}	1	1	1	1	0				1	1					

Граф тестовых проверок на обрыв (лог. 0 – const):

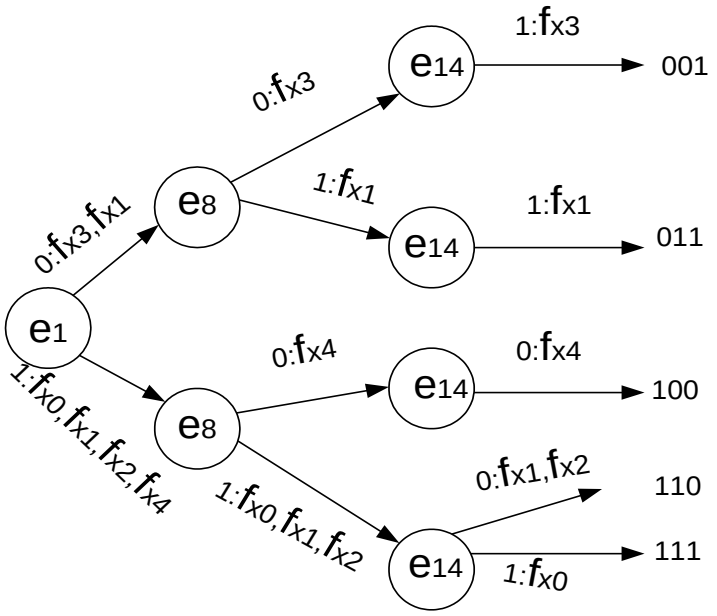


Рис.14

1.9. Информационный диагностический тест, близкий к минимальному (метод Синдеева)

Этот метод реализует безусловный алгоритм диагностики. Основой этого алгоритма является ТФН, у которой столбцы соответствуют всем возможным состояниям, а строки – всем возможным проверкам (проверка π_i означает контроль выхода z_i). Предполагаем, что все n состояний системы, состоящей из n блоков, равновероятны и сумма вероятностей состояний их отказа равна 1, т.е. проверки, как случайные события, образуют полную группу событий:

$$P_1=P_2=\ldots=P_n=\frac{1}{n}. \qquad (8)$$

Тогда с точки зрения теории информации неопределённость (энтропия) H, создаваемая такой схемой для пользователя, в общем случае определяется с помощью формулы Шеннона:

$$H=-\sum_{i=1}^{n}P_i\cdot\log_2 P_i\ ,$$

где P_i – вероятность i – го события (вероятность отказа i – го блока системы).

Для рассматриваемого случая из (8) получим формулу Хартли:

$$H=-\frac{1}{n}\sum_{i=1}^{n}\log_2\frac{1}{n}=-\frac{n}{n}\cdot\log_2 n^{-1}=\log_2 n, \qquad (9)$$

Для определения состояния схемы необходимо провести эксперимент, состоящий в последовательном выборе не более m наиболее информативных проверок (m<n).

Каждая k – ая проверка π_k несет некоторое количество информации I относительно начального состояния (начальной энтропии H_0) рассматриваемой схемы (системы).

$$I=H_0-H(\pi_k)=\Delta H, \qquad (10)$$

где $H(\pi_k)$ – средняя условная энтропия состояния схемы после проведения проверки π_k, $k=\overline{1,n}$.

Т.к. при проведении проверки π_k имеется только два возможных исхода (положительный π_k и отрицательный $\overline{\pi_k}$), т.е. $\pi_k=1$ или $\overline{\pi_k}=0$ с вероятностями соответственно $P(\pi_k)$ и $P(\overline{\pi_k})$, то средняя условная энтропия $H(\pi_k)$ равна

$$H(\pi_k)= P(\pi_k)H(\pi_k)+P(\overline{\pi_k})H(\overline{\pi_k}), \qquad (11)$$

где $H(\overline{\pi_k})$ и $H(\pi_k)$ – энтропия состояний схемы после выполнения проверки π_k соответственно для отрицательного и положительного её исходов.

$$P(\pi_k)=\frac{l}{n}, \qquad (12)$$

$$P(\overline{\pi_k})=\frac{n-l}{n},\ k=1\ldots n, \qquad (13)$$

где l - число единиц в рассматриваемой $k^{\underline{ой}}$ строке ТФН.

Тогда подставляя формулы (12) и (13) в формулу (11), а затем (11) в (10) получим с учетом (9):

$$I=\log_2 n-[\frac{l}{n}\log_2 l+\frac{n-l}{n}\log_2(n-l)]. \qquad (14)$$

Информация по формуле (14) вычисляется для каждой строки ТФН. Первой для теста выбирается проверка π_k, которая приносит максимум информации. Если таких проверок несколько, то выбирается любая из них, что возможно, когда информация вычисляется для равновероятных событий (отказов блоков системы) по формуле Хартли.

Если вероятности состояний $P(S_j)$ блоков устройства не одинаковы, то энтропия (неопределенность) в оценке состояния объекта диагностики (ОД) вычисляется по формуле Шеннона (1). В этом случае вероятность отказа $P(S_j)$ j – го блока системы можно оценить по формуле

$$P_j= P(S_j)=\frac{\Lambda_j}{\sum_{j=1}^{n}\Lambda_j},$$

где Λ_j - интенсивность отказов j – го блока, $час^{-1}$;

$\sum_{i=1}^{n}\Lambda_j$ - интенсивность отказов системы, равное сумме интенсивностей отказов составляющих её блоков.

Строка ТФН, соответствующая лучшей по условию (3) проверке, перемещается на место первой строки ТФН и делит последнюю на две, в общем случае, неравные части, в одну из которых входят столбцы состояний S_j, которым соответствуют “0” в выбранной (лучшей на первом шаге) строке, а в другую – столбцы состояний, которым соответствуют “1” в выбранной строке. Выбранная на первом шаге лучшая по информативности строка больше не участвует в отборах на втором и последующих шагах формирования тестовых проверок.

Успешной (положительной) считается проверка, при которой выход контролируемого блока системы соответствует техническим условиям (ТУ) предприятия-изготовителя. Результаты успешной проверки обозначаются как диагностическая 1. В противном случае проверка считается неуспешной (отрицательной) и обозначается как диагностический 0 в ТФН и других документах.

Выбор второй наиболее информативной проверки проводится одновременно по двум полученным подТФН.

Второй выбирается проверка π_i, которая обладает наибольшей условной информацией $I(\pi_k/\pi_i)$ относительно состояния, характеризуемого энтропией $H(\pi_k)$ после первой выбранной проверки.

$$I(\pi_k/\pi_i)=H(\pi_k) - H(\pi_k/\pi_i) \rightarrow \max \qquad (15)$$

Средняя условная энтропия схемы после 2$^{\underline{ой}}$ проверки.

$$H(\pi_i/\pi_k)= P(\pi_i/\pi_k)H(\pi_i/\pi_k)+P(\overline{\pi_i}/\pi_k)H(\overline{\pi_i}/\pi_k)+$$
$$+P(\pi_k/\overline{\pi_k})H(\pi_i/\overline{\pi_k})+P(\overline{\pi_i}/\overline{\pi_k})H(\overline{\pi_i}/\overline{\pi_k}),$$

(16)

где $P(\pi_i/\pi_k)=\frac{l}{n}\cdot\frac{l_1}{l}=\frac{l_1}{n}$ - вероятность второй успешной проверки после успешной первой. (17)

$P(\overline{\pi_i}/\pi_k)=\frac{l}{n}\cdot\frac{l-l_1}{l}=\frac{l-l_1}{n}$ - вероятность второй неуспешной проверки после успешной первой. (18)

$P(\pi_i/\overline{\pi_k})=\frac{n-l}{n}\cdot\frac{l_2}{n-l}=\frac{l_2}{n}$ - вероятность успешной второй после неуспешной первой проверки.

(19)

$P(\overline{\pi_i}/\overline{\pi_k})=\frac{n-l}{n}\cdot\frac{n-l-l_2}{n-l}=\frac{n-l-l_2}{n}$ - вероятность неуспешной второй проверки после неуспешной первой. (20)

l_1 и l_2 – число единиц в i-ой строке подТФН №2.1 и подТФН №2.2 соответственно, первая из которых соответствует l-единицам, а вторая – (n-l) – нулям в k$^{\underline{ой}}$ – строке исходной ТФН. Из формулы (9) следует

$$H(\pi_i/\pi_k)=\log_2 l_1,\ H(\overline{\pi_i}/\pi_k)=\log_2(l_1-l_2),\ H(\pi_k/\overline{\pi_i})=\log_2 l_2,$$
$$H(\overline{\pi_i}/\overline{\pi_i})=\log_2(n-l_1-l_2).$$

Тогда средняя условная энтропия после 2$^{\underline{го}}$ шага

$$\tilde{H}\,(\pi_i/\pi_k)=\frac{l_1}{n}\cdot\log_2 l_1+\frac{l-l_1}{n}\cdot\log_2(l-l_1)+\frac{l_2}{n}\cdot\log_2 l_2+\frac{n-l-l_2}{n}\cdot\log_2(n-l-l_2)\,.$$

(21)

Информация после $2^{\underline{го}}$ шага для i-той строки будет равна разности выражения в квадратных скобках в уравнении (14) и выражения (21).

Вторая лучшая проверка π_i из группы наиболее информативных проверок, записывается второй в исходной ТФН и не участвует в отборе на третьем и последних шагах и т.д.

Выбор проверок продолжается до тех пор, пока средняя угловая энтропия после проверки на каком-то шаге не станет равной нулю. Эта проверка и будет последней в последовательности наиболее информативных, вошедших в тест.

Если отбор проверок для теста заканчивается на m-ом шаге (m<n), то m строк, переставленные вверх по указанным выше правилам в исходной ТФН, автоматически формируют таблицу кодов неисправностей. Причем m – разрядный двоичный код каждой одиночной неисправности читается сверху вниз в каждом столбце ТФН с переставленными вверх наиболее информативными строками (проверками), образующими тестовый набор проверок.

На практике выполняются проверки состояний выходов блоков, вошедших в тестовый набор, в той последовательности, в какой они указаны в тесте. Результат каждой проверки фиксируется в виде диагностической 1 или 0. Комбинация результатов проверок, образующая двоичный код, сравнивается с таблицей кодов неисправностей, которая, как указано выше, автоматически получается из исходной ТФН и состоит из m наиболее информативных строк.

Пример реализации комбинационного метода поиска диагностики неисправностей, реализующего безусловный алгоритм диагностики по методу Синдеева.

Найти информационный тест и таблицу кодов неисправностей для схемы рис.1, в которой имеет место одиночный отказ.

1. Диагностические оценки входных и выходных сигналов:

Z_5=0 – const, x1=1 – const, x2=1 – const.

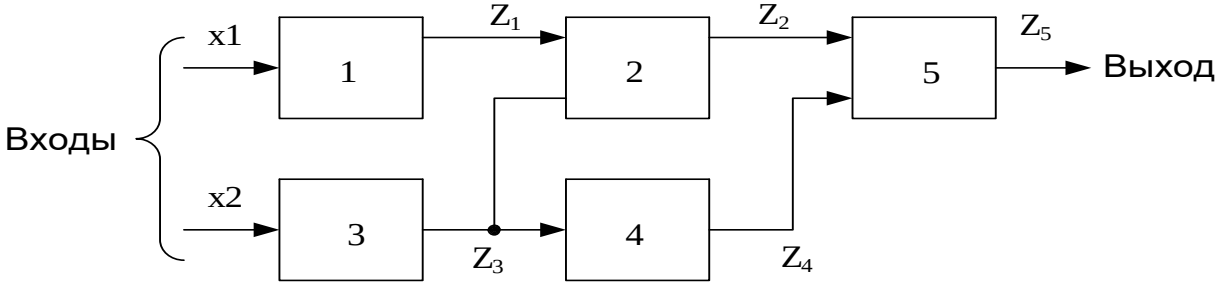


Рис.15

2. Найдем информацию, которую может принести каждая проверка на первом шаге. В этом случае

$$I(\pi_k)=H_0 - \log_2 n,$$

где H_0 – исходная энтропия.

$H_0=\log_2 5=\lg 5/\lg 2=3.332 \lg 5=2.329$бит.

ТФН, соответствующая ФДМ на рис.15 представлена табл. 2.

Таблица 2

ТФН для информационного теста

S_i \ π_i	S1	S2	S3	S4	S5	$I(\pi_k)$, бит	$I(\pi_i/\pi_k)$
П1	0	1	1	1	1	0.729	0.554
П2	0	0	0	1	1	0.975	-
П3	1	1	0	1	1	0.729	0.55
П4	1	1	0	0	1	0.975	0.954

ТФН 2.1 ТФН 2.2

Определим среднюю условную энтропию $H(\pi_2)$ схемы, которую получим в результате проведения проверки π_2. Согласно (4) и (7)

$$H(\pi_k)=\frac{l}{n}\log_2 l+\frac{n-l}{n}\log_2(n-l).$$

Для второй проверки

$$H(\pi_2)=\frac{2}{5}\log_2 2+\frac{5-2}{5}\log_2 3=0.4+0.954=1.354\text{бит}$$

Тогда информация $I(\pi_2)$, полученная в результате проведения проверки π_2, составит:

$I(\pi_2)=H_0-H(\pi_2)$

$I(\pi_2)=2.329-1.354=0.975$ бит.

$$I(\pi_1)=\log_2 5-(\frac{4}{5}\log_2 4+\frac{1}{5}\log_2 1=3.322\lg 5-(0.8\cdot 2+0.2\cdot 0)=0.729\text{бит},$$

$$I(\pi_3)=\log_2 5-(\frac{4}{5}\log_2 4+\frac{1}{5}\log_2 1=0.729\text{бит},$$

$$I(\pi_4)=\log_2 5-(\frac{3}{5}\log_2 3+\frac{2}{5}\log_2 2=0.975\text{бит}.$$

Столбец с вычисленными количествами информации поместим 6-ым в таблицу 1 функций неисправностей (ТФН), соответствующую функционально-диагностической модели (ФДМ) устройства на рис.1. На

пересечении j-го столбца и i-й строки π_i в таблице 1 ставится “1”, если при возникновении одиночной неисправности в j-ом блоке устройства сигнал на выходе i-го блока остается соответствующим техническим условиям (ТУ) его эксплуатации и ставится “0” в противном случае.

Первой в тест выбираем проверку π_2 (контроль выхода блока 2), т.к. она приносит максимум информации и “удачно” делит исходную ТФН на две неравные части ТФН 2.1 и ТФН 2.2, первой из которых соответствуют нули в выбранной строке π_2, а второй - “1”.

3. На втором шаге определим среднюю условную энтропию после второй проверки, которая включает оставшиеся проверки π_1, π_3, π_4. Например,

$$H(\pi_1/\pi_2)=(\frac{2}{5}\log_2 2+\frac{1}{5}\log_2 1)+(\frac{2}{5}\log_2 2+0\cdot\log_2 6)=0.4+0.4=0.8\text{ бит.}$$

Тогда $I(\pi_1/\pi_2)=\acute{I}_{\pi_2}-\acute{I}_{\pi_1/\pi_2}$ =1.354-0.8=0.554 бит.

Аналогично $H(\pi_3/\pi_2)=\frac{2}{5}\log_2 2+0+\frac{2}{5}\log_2 2+\frac{1}{5}\log_2 1$=0.4+0.4=0.8 бит.

$I(\pi_3/\pi_2)$=1.354-0.8=0.554 бит.

Затем $H(\pi_4/\pi_2)=\frac{1}{5}\log_2 1+\frac{1}{5}\log_2 1+\frac{2}{5}\log_2 2\frac{1}{5}\log_2 1$=0.4 бит,

Следовательно $I(\pi_4/\pi_2)$=1.354-0.4=0.954 бит.

Вычисленные количественные оценки информации на втором шаге $I(\pi_i/\pi_k)$ поместим в 7-ом столбце табл. 2.

Анализируя содержимое этого столбца, на втором шаге выбираем 4$^{\underline{ую}}$ проверку π_4 т.к. она несет наибольшее количество информации относительно энтропии, оставшейся после 1$^{\underline{ой}}$, выбранной нами проверки π_2.

4. Перед выбором третьей информативной проверки перенесем на вторую строку исходной ТФН содержимое строки π_4, соответствующей выбранной проверке. Указанная ситуация отображена в табл. 3, где место первой строки занимает информация строки (проверки) π_2, выбранной на первом шаге. Тогда остаются две альтернативные проверки π_1 и π_3. Содержимое строк π_1 и π_3 распределяется в таблице 2 по четырем подТФН, причем ТФН 3.1 и ТФН 3.2 строятся на базе ТФН 2.1, а ТФН 3.3 и ТФН 3.4 - на базе ТФН 2.2.

Таблица 3.

ТФН для выбора 3-ей проверки теста

S_i / π_i	S1	S2	S3	S4	S5	$I(\pi_f/\pi_4/\pi_2)$
П2	0	0	0	1	1	-
П4	1	1	0	0	1	-
П1	0	1	1	1	1	0.4
П3	1	1	0	1	1	0

ТФН 3.1 ТФН 3.2 ТФН 3.3 ТФН 3.4

ТФН 2.1 ТФН 2.2

ТФН 3.2, 3.3 и 3.4 вырождаются в столбцы, что облегчает выбор проверок и указывает на быстрое завершение формирования теста.

Информация, доставляемая третьей проверкой $I(\pi_n/\pi_m/\pi_k)$ определяется по формуле:

$I(\pi_n/\pi_m/\pi_k)=H(\pi_m/\pi_k)-H(\pi_n/\pi_m/\pi_k)$.

Например, если третьей будет проверка π_1, то

$H(\pi_1/\pi_4/\pi_2)=$

$$\frac{1}{5}\log_2 1+\frac{0}{5}\log_2 0+\frac{1}{5}\log_2 1+\frac{0}{5}\log_2 0+\frac{1}{5}\log_2 1+\frac{1}{5}\log_2 1+\frac{1}{5}\log_2 1+\frac{0}{5}\log_2 0=0$$

т.е. $I(\pi_3/\pi_4/\pi_2)$= 0.4-0 = 0.4 бит.

Если третьей будет π_3, то

$H(\pi_3/\pi_4/\pi_2)=$

$$\frac{1}{5}\log_2 1++0+\frac{1}{5}\log_2 1+\frac{0}{5}\log_2 0+\frac{2}{5}\log_2 2+\frac{0}{5}\log_2 0+\frac{0}{5}\log_2 0+\frac{1}{5}\log_2 1=0.4\,\text{бит.}$$

Следовательно, $I(\pi_3/\pi_4/\pi_2)$=0.4-0.4=0 бит, т.е. проверки надо прекратить!

Располагая проверки π_2,π_4,π_1 последовательно друг за другом, получим таблицу кодов неисправностей с трёхразрядными двоичными кодами одиночных неисправностей, которые могут иметь место в устройстве на рис.15. Такое расположение проверок сформировано в табл. 2. Код неисправности 1-го блока (состояние устройства S_1) составляет 010, код состояния S_2 – 011, код состояния S_3 – 001, код состояния S_4 – 101 и код состояния S_5 –111.

Выполняя на практике последовательно проверки π_2,π_4,π_1 (проверяя выходы Z_2,Z_4,Z_1) и фиксируя результаты проверок в виде диагностических “0” или “1”, получают трёхразрядную двоичную комбинацию и идентифицируют неисправность в устройстве, сравнивая полученную комбинацию с таблицей кодов неисправностей.

1.10. Метод критерия различимости – комбинационный метод, реализующий безусловный алгоритм диагностики

1) Постановка задачи

Проблема заключается в том, что разработать набор тестов, который будет иметь дело со всеми неисправностями цифровой системы. Однако, одни методы генерирования тестов вырабатывают их лишь для

однократной специфической неисправности, не пригодны в столь больших схемах. Очевидное решение – применить некоторый метод многократно до полной обработки неисправностей. А затем полученные тесты объединить. Однако набор тестов, полученных таким способом неприемлем из-за емкости потребляемой памяти и времени обработки. Отсюда возникает задача сокращение объема набора теста.

2) **Таблица неисправностей – используется для представления результатов генерирования тестов, является двух мерным массивом, имеющем по одной строке на каждую неисправность и по одному столбцу на каждый тест. Для элемента таблицы** fij **:**

fij **=1, если** i **-я неисправность обнаруживается** j **-м тестом и**

fij **=0 – в противном случае.**

Строка называется отображениями неисправности и является вектор–строками, $fi = (fi1,....., fim)$ - отображение i -й неисправности. Два отображения неисправности fi и fj различимы , если $fik \neq fjk$ для некоторого k, где k - номер теста. Таким образом задача трансформируется в задачу нахождения минимального набора столбцов таблицы так, чтобы каждая строка, соответствующей подтаблицы имела бы по крайней мере одну «1». **Чтобы различать все неисправности необходим минимальный набор столбцов, соответствующий вышеизложенным требованиям и имеющий все различимые строки.**

3) Метод критерия различимости.

Основная идея – в приписывании весов отдельным тестам. Веса отражают относительную способность тестов различать неисправности.

На основе весов отбираются тесты для получения эффективного набора тестов.

Вес W_j теста T_j определяется как и число пар неисправностей, которые он различает:

$$W_j = (n_0^j)(n_1^j),$$

где n_0^j и n_1^j - число соответственно «0» и «1» в j -м столбце таблицы

Метод строится на последовательности итераций. Вычисляются W_j для всех j. Затем в качестве первого выбирается тест с max W_j, так как он различает наибольшее число пар неисправностей. Этот тест делит неисправности на обнаруживаемые и необнаруживаемые. В общем случае, когда выбран i-й тест, неисправности будут разделены на m непересекающихся блоков $b1,.....,bm$. Это означает, что последующие тесты должны выбираться более сложным путем.

Пусть T_j кандидат на $(i+1)$ элемент набора и пусть $(n_0^j)_{bk}$ и $(n_1^j)_{bk}$ - число «0» и «1» в j -м столбце блока bk. Тогда после i-го шага вес для T_j будет равен:

$$W_j^i = \sum_{k=1}^{m} (n_0^j)_{bk} \, (n_1^j)_{bk}$$

Тест, для которого W_j^1-max, выбирается как $(i+1)$ элемент набора тестов. Отбор тестов продолжается до тех пор, пока разделение неисправностей перестанет приносить дальнейшее улучшение, т.е. до тех пор, пока веса всех неиспользованных «тестов» не станет равное «0» (не станет «0» или «1» bk)

Развитие метода

Возможен отбор тестов, которые диагностируют неисправности с заданной степенью диагностической разрешающей способности, которая зависит от объема схемы, содержащейся в каждой сменной ячейке.

Развитие метода – выбор набора тестов, которые идентифицируют неисправность только с точностью до одной БИС – ячейки.

Предположим: неисправности $f1$ и $f2$ связаны с ячейкой $P1$; $f3, f4, f5 - P2$; $f6, f7 - P3$. Ясно, что мы не хотим больше различать пары неисправностей из одной и той же ячейки, вклад таких пар в вес теста нужно вычитать для T_j при условии, что $(n_0)_{bk \cap pt}$ и $(n_1^j)_{bk \cap pt}$ - число «0» и «1» в j-м столбце блока *bk*, связанного с ячейкой *pl*. Тогда выполняется соотношение

$$Wj = \sum_{k=1}^{m} \left[(n_0^j bk) \, (n_1^j) bk - \sum_{l} (n_0^j) bk \, npe (n_1^j) bk \, npe \right],$$

где суммирование производится по всем ячейкам, связанным с блоком bk (l - число сменных ячеек, pl – одна из ячеек, $l = \overline{1, z}$).

Пример: Таблица неисправностей (рис.16) состоит из восьми тестов и двенадцати неисправностей. Объект реализован на пяти сменных ячейках P_1-P_5.

В ячейке P_1 необходимо обнаруживать две неисправности f_1^1 и f_2^1, во второй P_2 и четвертой P_4 – по три неисправности, в третьей P_3 и пятой P_5 ячейках – по две неисправности.

		t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
P_1	f_1^1	1	0	0	1	0	0	0	0
	f_2^1	0	0	0	1	0	0	1	1
P_2	f_1^2	0	0	1	1	0	1	0	0
	f_2^2	0	1	0	1	0	0	1	1
	f_3^2	0	1	0	0	0	0	1	0
P_3	f_1^3	0	0	0	1	1	0	1	0
	f_2^3	1	1	0	0	0	1	0	1
P_4	f_1^4	0	0	0	1	1	0	1	0
	f_2^4	0	1	0	1	1	0	0	0
	f_3^4	0	1	0	0	1	0	0	1
P_5	f_1^5	1	0	1	0	1	1	0	0
	f_2^5	0	1	1	0	1	1	0	0
Вес		24	30	25	30	35	29	29	26

Идентичны

Рис 16. Таблица ТФН №1

Сначала вычисляются веса и получаем, что max весом обладает $T5$.

$W_1^5 = 6 \cdot 6 - 1 \cdot 1 = 35.$

Перестраиваем таблицу 1 в таблицу 2 . Снова вычисляем веса тестов (кроме $T5$). На этот раз выбор пал на $T6$.

		t_5	t_1	t_2	t_3	t_4	t_6	t_7	t_8
P_1	f_1^1	0	1	0	0	1	0	0	0
	f_2^1	0	0	0	0	1	0	1	1
P_2	f_1^2	0	0	1	1	1	1	0	0
	f_2^2	0	0	1	0	1	0	1	1
	f_3^2	0	0	1	0	0	0	1	0
P_3	f_2^3	0	1	1	0	0	1	0	1
P_3	f_1^3	1	0	0	0	1	0	1	0
P_4	f_1^4	1	0	1	0	1	0	0	0
	f_2^4	1	0	1	0	0	0	0	1
	f_3^4	1	1	0	1	0	0	0	0
P_5	f_1^5	1	0	0	1	0	1	0	0
	f_2^5	1	1	1	0	1	1	0	0
			11	13	11	13	$\boxed{14}$	12	9

b1

b2

Рис 17. ТНФ №2

Проиллюстрируем определение веса W_6^2 для теста t_6, лучшего на втром шаге

$$W_6^{21} = \left[(n_0^6)b1\,(n_1^6)b1 - \sum_l (n_0^6)b1 \bigcap pe\,(n_1^6)b1 \bigcap pe \right] +$$

$$+ \left[(n_0^6)b2\,(n_1^6)b2 - \sum_l (n_0^6)b2 \bigcap pe\,(n_1^6)b2 \bigcap pe \right]$$

$$W_6^2 = \{(4\cdot 2)b1 - [(2\cdot 0)b1 \bigcap p1 + (2\cdot 1)b1 \bigcap p2 + (0\cdot 1)b1 \bigcap p3 + (0\cdot 0)\,b1 \bigcap p4 + (0\cdot 0)b1 \bigcap p5]\}$$
$$+ \{(4\cdot 2)b2 - [(0\cdot 0)b2 \bigcap p1 + (0\cdot 0)b2 \bigcap p2 + (1\cdot 0)b2 \bigcap p3 + (0\cdot 2)b2 \bigcap p5]\}$$

$$W_6 = 6 + 8 = 14$$

При следующей перестройке таблицы (ТФН3) блок b4 можно не рассматривать, так как его строки связаны только с P5:

		t_5	t_6	t_1	t_2	t_3	t_4	t_7	t_8	
P_1	f_1^1	0	0	0	0	1	0	0	0	
	f_2^1	0	0	0	0	1	0	1	1	b1
P_2	f_2^2	0	0	1	1	1	1	0	0	
	f_3^2	0	0	1	1	1	0	1	1	
P_2	f_1^2	0	1	1	0	0	0	1	0	b2
P_3	f_2^3	0	1	1	1	0	1	0	1	
P_3	f_1^3	1	0	0	0	1	0	1	0	
P_4	f_1^4	1	0	1	0	1	0	0	0	b3
	f_2^4	1	0	1	1	0	0	0	1	
	f_3^4	1	1	0	1	0	1	0	0	
P_5	f_1^5	1	1	1	0	0	1	0	0	b4
	f_2^5	0	1	1	1	1	1	0	0	
Вес				3	$\boxed{7}$	1	3	5	4	

Рис 18. ТФН №3

Проиллюстрируем определение веса w_2^3 столбца 2, который оказывается лучшим на третьем шаге отбора тестов.

$$W_2^3 = \{2\cdot 2 - 0 - 0\} + \{1\cdot 1 - 0 - 0\} + \{2\cdot 2 - 1\cdot 2\} + \{1\cdot 1 - 1\cdot 1\}$$

$$W_2 = 7$$

Таким образом тест T2 выбираем третьим в группу тестов, имеющих наибольшие веса.

Следующая перестройка ТФН представлена на рис.19..

		t_5	t_6	t_2	t_1	t_3	t_4	t_7	t_8	
P_1	f_1^1	0	1	0	0	1	0	0	0	b1
	f_2^1	0	0	0	0	1	0	1	1	
P_2	f_2^2	0	0	1	1	1	1	0	0	b2
	f_3^2	0	0	1	0	1	0	1	1	
P_2	f_1^2	0	0	1	0	0	0	1	0	b3
P_3	f_2^3	0	1	1	0	0	1	0	1	
P_3	f_1^3	1	0	0	0	1	0	1	0	b4
	f_1^4	1	0	1	0	1	0	0	0	b5
P_4	f_2^4	1	0	1	0	0	0	0	1	b6
	f_3^4	1	1	0	1	0	1	0	0	
P_5	f_1^5	1	0	1	1	0	1	0	0	b7
	f_2^5	0	1	1	0	1	1	0	0	
Вес					0	0	0	0	0	

Рис 19. ТНФ №4

Теперь все тесты имеют веса = 0 – процесс отбора тестов заканчивается, минимальным набором является (Т5, Т6, Т2). Следовательно для идентификации каждой неисправной ячейки требуется только семь комбинаций:
P1=(0,0,0) или (0,1,0); P2=(0,0,1); P3=(0,1,1); P4=(1,0,1) или (1,1,1); P5=(1,1,0)

Метод привлекателен для больших таблиц неисправностей, т.к. для реализации на машине требуется немного памяти, длина наборов тестов равна min (m,n) итераций. Неразличимые итерации автоматически группиризуются в общие группы, а процесс отбора тестов может быть изменен, чтобы учитывался любой вид ограничений, накладываемых ячейками.

Глава 2. Диагностика микропроцессорных систем (МПС). Общая характеристика средств диагностики

Диагностика и контроль МПС осуществляется при комплексной отладке программных и аппаратных средств.

Из аппаратных средств, применяемых при отладке МПС, наиболее мощным является логический анализатор (ЛА), который обеспечивает глубину диагностики с точностью до блока, модуля, платы, а при необходимости – до корпуса. ЛА применяют на этапе разработки, комплексной отладки МПС и анализе причин аварий.

При производственных испытаниях и эксплуатации возникает необходимость обнаруживать неисправность до корпуса. Этой цели служат более простые и дешевые сигнатурные анализаторы (СА).

Наиболее простыми техническими средствами являются тестер логических состояний, стимуляторы логических состояний и бесконтактные индикаторы импульсных токов.

2.1. Общая характеристика и методика использования СА

В эксплуатации, когда мало времени на поиск неисправности, а квалификация обслуживающего персонала невысока, ЛА малопригоден. Поэтому требуется прибор, который позволял бы быстро и точно отыскивать вышедший из строя компонент МПС. Таким прибором является СА. Его принцип действия основан на преобразовании «длинных» последовательностей двоичных сигналов в «короткое» двоичное число («короткое»), называемое сигнатурой. Измеряемые «длинные» двоичные последовательности возбуждаются в контрольных точках МПС под действием тестовой (ПЗУ) или рабочей (ОЗУ) программы. Сигнатуры в контрольных точках измеряются на заведомо работоспособной системе и указываются на принципиальной схеме МПС, подобно тому, как на схемах аналоговых устройств указываются осциллограммы и некоторые параметры непрерывных электрических сигналов.

При поиске неисправности в МПС оператору необходимо установить режим тестовой программы и затем, прослеживая сигнатуры в контрольных точках, найти элемент, у которого входные сигнатуры верны, а выходная – нет. Тогда в этом элементе заключена неисправность (или в его выходной цепи).

Для удобства сигнатура представляется в виде четырех шестнадцатеричных чисел которые получаются с помощью семисегментных светодиодных индикаторов.

2.2. Точность СА

Существует много способов сжатия двоичных последовательностей в сигнатуры:

1) подсчет числа логических переключений;

2) подсчет числа единиц;

3) получение контрольной суммы по одному из модулей

и т.д.

Анализ показал, что наиболее эффективным является способ, основанный на преобразовании с помощью сдвигового регистра с линейными ОС и сумматором по модулю 2.

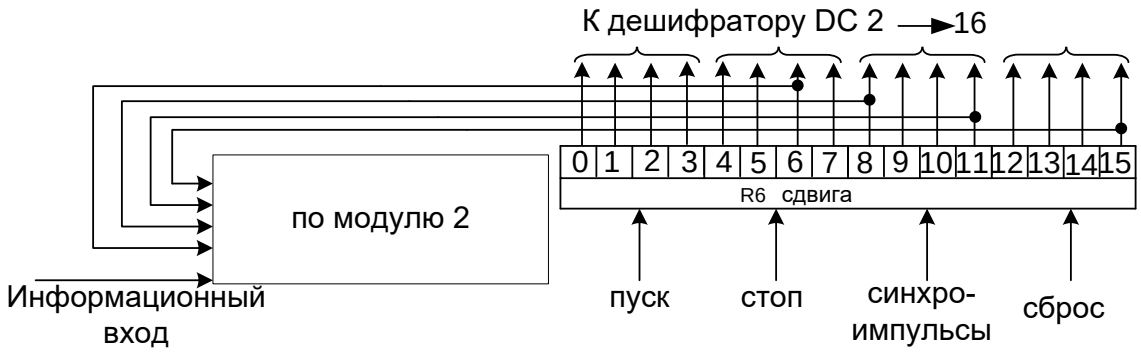


Рис.1

Сдвиговый регистр с сумматором по модулю 2 является линейной цифровой схемой, поэтому для нее справедлив принцип суперпозиции: реакция схемы на сумму двух входящих последовательностей (исходной рабочей и последовательности ошибок) равна сумме реакций на каждое из этих воздействий.

В силу наличия нескольких ОС, сумма никогда не будет равна одному из двух ненулевых слагаемых. Входная последовательность m не должна быть короткой (не менее чем число разрядов регистра сдвига n), т.е.

$$m >> n.$$

Вероятность P обнаружения ошибки в последовательности длинной m при использовании регистра сдвига длиной n равна

$$p = 1 - \frac{2^{m-n} - 1}{2^m - 1},$$

$$\lim_{m \to \infty} P = 1.$$

При длине регистра n=16 и $m > n$ ошибки произвольной кратности обнаруживаются с вероятностью 0,99998 , а однобитовые ошибки - с вероятностью 1.

2.3. Условия пригодности модулей МПС для испытания методом СА

Для реализации СА в схеме должны быть заложены следующие технические средства:

1) Средства, позволяющие разрывать в режиме контроля цепь ОС в МПС.

МПС охвачена ОС по контуру программного управления через ШД, поэтому в одноплатной микро-ЭВМ первое условие реализуется отключением ШД от входа микропроцессора. Это реализуется с помощью разъемов или буферных схем.

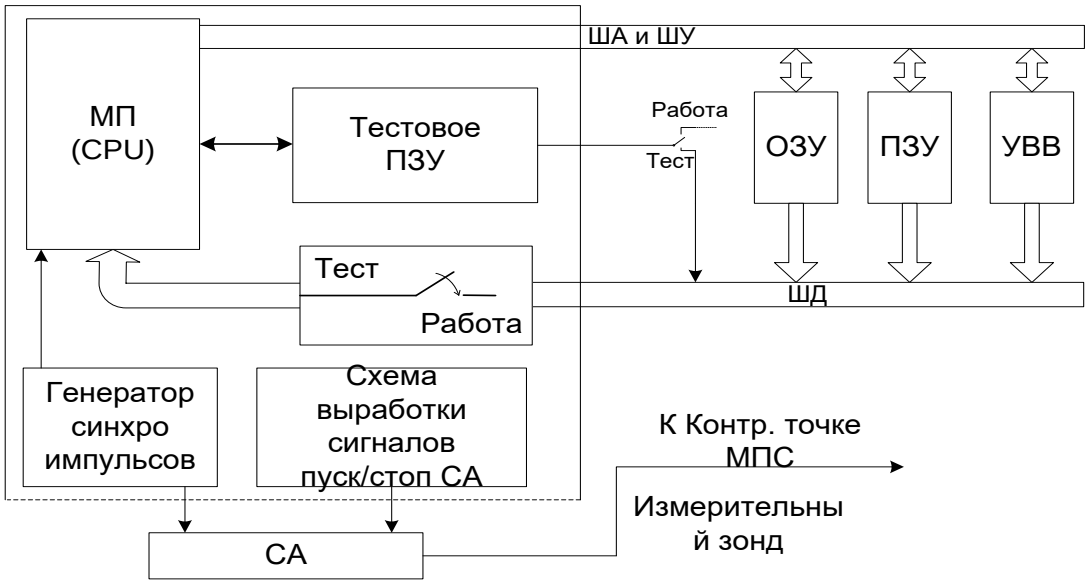


Рис.2. Структурная схема подключения СА к контролируемой МПС

2) Наличие схем, вырабатывающих сигналы «Пуск» и «Стоп», необходимые СА для формирования «Измерительного окна», т.е. интервала времени, в течение которого накапливается сигнатура.

3) Наличие в составе МПС ПЗУ, в котором хранится тестовая программа.

2.4. Методика проверки МПС с помощью СА

Синхрогенератор (на рис. 2), МП, тестовое ПЗУ, схема выработки сигналов пуск - стоп, участок ШД, отключаемый в режиме тестирования от остальной схемы, образуют ядро ЭВМ. В режиме тестирования контролируемые двоичные последовательности в МПС возбуждаются под действием сигналов, вырабатываемых этим ядром, поэтому надо быть уверенным в его работоспособности. Если ядро исправно, то поиск неисправности ведется описанным выше способом - путем переходов от выходов элементов ко входам.

На неработоспособность ядра указывает неверная сигнатура на одной из линий ША или ШУ. В этом случае прежде всего необходимо убедиться в правильной работе синхрогенератора. Если синхронизация МП в норме, то надо убедится в работоспособности схемы пуск - стоп, подключив измерительный зонд СА к линии питания МП +5 вольт, что должно соответствовать подаче на вход СА последовательности из всех единиц. Для локализации неисправности в ядре (МП, тестовом ПЗУ, схеме адресной дешифрации) необходимо иметь возможность отключить выход тестового ПЗУ от МП и подать на вход ШД в месте присоединения ее к МП код команды NOP (00). При этом на адресной шине МП будет происходить последовательный перебор всех адресов $(0-2^n)$, n - число разрядов ША. В этом случае измерительное окно будет равно времени перебора всего поля адресов. В этом режиме надо измерить сигнатуры на ША и на выходах тестового ПЗУ. Неверная сигнатура на ША укажет на неисправность МП, а на выходе тестового ПЗУ - на выход из строя этого элемента.

Указанный режим называют режимом свободной прогонки МП по всему полю адресов. При этом роль схемы выработки сигналов пуск - стоп могут играть сигналы, снимаемые со старшего разряда ША и подаваемые на объединенные входы ПУСК - СТОП многоканального щупа СА.

2.5. Логический анализатор (ЛА)

Напряжения, появляющиеся в разных точках МПС, преобразуются ЛА в четкие последовательности логических уровней. Это преобразование осуществляется с помощью входных компараторов уровня, на которые подаются сигналы с ША, ШД, ШУ. Число входов у ЛА составляет 16, 24 или 32. Входные напряжения сравниваются с пороговым напряжением, которое выбирается как полусумма сигналов, соответствующих логическим нулю и единице.

$$U_{пор} = \frac{U_0 + U_1}{2} = \frac{0{,}4 + 2{,}4}{2} = 1{,}4B$$

ТТЛ:

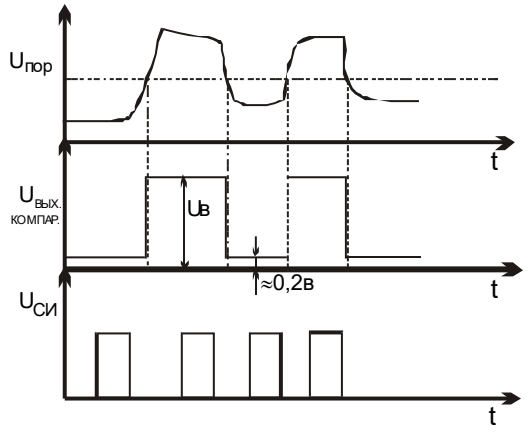


Рис.3

В моменты, соответствующие синхроимпульсам, полученный на выходе компараторов многоразрядный код считывается в буферный регистр (см. структурную схему рис 4.) Из буферного регистра по синхроимпульсам информация передается в ЗУ и одновременно в логический компаратор. ЗУ можно выполнить из нереверсивных сдвиговых регистров, число которых равно числу разрядов буферного регистра. В моменты появления синхроимпульсов, содержимое нереверсивных регистров сдвига перемещается вправо, и в освобождаемый младший разряд каждого регистра принимается содержимое соответствующего разряда буферного регистра. При этом выдвигаемая из регистра сдвига информация теряется. Таким образом, в таком ЗУ удерживаются данные, поступившие из буферного регистра за последние 16 (24, 32) тактов.

Аналогичные процессы можно воспроизвести в ЗУ типа стек, если предусмотреть в нем, например 16 шестнадцатиразрядных ячеек и четырехразрядный счетчик в качестве указателя стека. Синхроимпульсы изменяют на единицу содержимое указателя стека, и по образовавшемуся адресу производится запись информации в стек, которая поступает из буферного регистра.

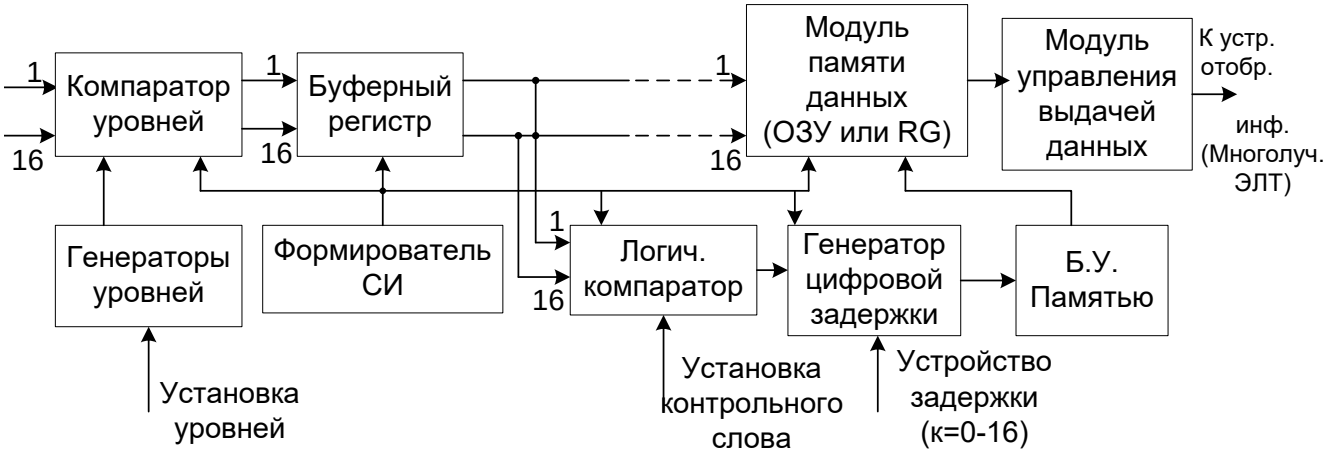


Рис. 4.

Если в процессе отладочных работ обнаруживается выдача МПС неверных результатов, то возникает необходимость выявления участков программы, которые могут быть причиной возникновения неверного результата.

Определение таких точек в программе требует просмотра содержимого шин интерфейса за некоторое число тактов на участке программы, подозреваемом в ошибочном выполнении. Эту информацию можно получить из ЗУ.

Пусть ввод информации в ЗУ прекращается в момент, когда на подключаемой ко входу ЛА шине появляется слово, возникающее при выполнении интересующей нас команды (контрольное слово). Тогда в ЗУ окажутся 16 слов, принятых в течение 16-ти предыдущих тактов. После обнаружения в буферном регистре контрольного слова можно закрыть доступ к ЗУ новых данных с задержкой на некоторое число «к» тактов. При этом в ЗУ окажутся зафиксированными данные, принятые в течение (16-к) предыдущих тактов и «к» последующих тактов.

Выявление момента, когда в буферном регистре появляется контрольное слово, осуществляется логическим компаратором, на входе которого оператор с помощью тумблеров набирает нужное слово в двоичном коде. Генератор задержки по установленному значению «к» воздействует на блок управления памятью, закрывая доступ к ЗУ новых данных на определенном такте. С помощью модуля управления выдачей данных содержимое ЗУ отображается на экране ЭЛТ, встроенной в ЛА. При этом могут использоваться различные формы представления данных, находящихся в ЗУ. Этим данным могут быть приданы следующие формы:

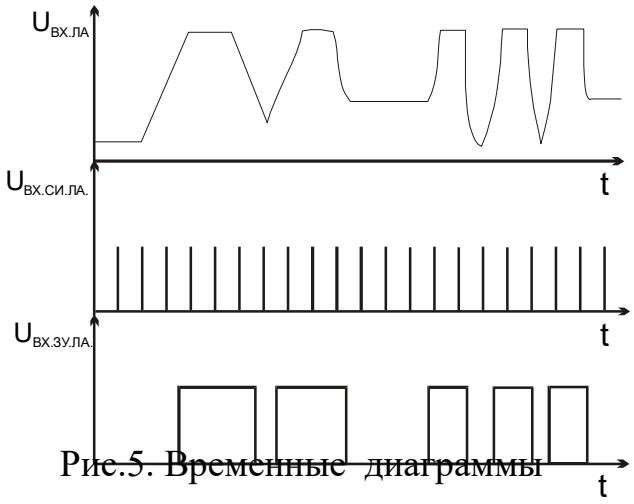


Рис.5. Временные диаграммы

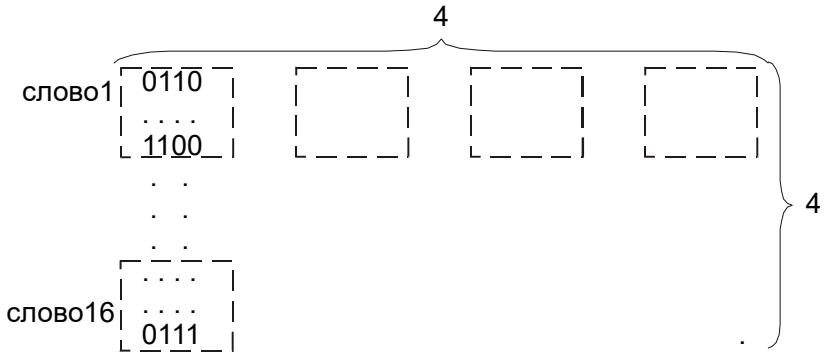
1) временных диаграмм;
2) таблиц данных;
3) карт состояния.

Рассмотрим эти формы.

Первая форма удобна при отладке аппаратных средств, когда необходим анализ временных соотношений между сигналами, возникшими в различных шинах, а также для выявления ложных импульсов и помех. Получение таких диаграмм достигается в результате последовательного циклического считывания содержимого ЗУ и подачей считанных данных на отдельные входы многоканальной ЭЛТ ЛА.

Второй способ реализуется путем установки в каждой входной шине ЛА триггеров, которые сбрасываются в ноль по синхроимпульсу контролируемой МПС и используются для регистрации коротких импульсов помех.

Вторая форма удобна на этапе комплексной отладки МПС. При этом содержимое ЗУ отображается на экране ЭЛТ в виде чисел, представляемых в двоичной или шестнадцатеричной системах счисления. Каждое такое число отображает комбинацию логических уравнений сигналов, поступивших на вход ЛА в последовательные моменты времени.



Третья форма. При ее использовании разряды выводимых из ЗУ данных разбиваются на две группы. Каждая группа разрядов подается на входы отдельного ЦАП. Выходное напряжение одного ЦАП используется в качестве горизонтальной, а другого ЦАП - вертикальной развертки осциллографа. При этом каждому значению выводимых из ЗУ данных, относящихся к одному и тому же моменту времени, соответствует определенное положение светящейся точки на экране ЭЛТ. Последовательности поступающих из ЗУ данных соответствует последовательность точек, образующих на экране фигуру. Такая форма отображения информации в ЗУ удобна для контроля аппаратуры на этапе серийного производства МПС или ее эксплуатации. Если форма фигуры, соответствующая исправной МПС известна, то при

наличии неисправности форма фигуры меняется (луч выходит за контуры эталонной фигуры, нанесённой на экран осциллографа).

Технические характеристики современного ЛА типа В-141 приведены в Приложении 1.

2.6. Простейшие контрольно-испытательные приборы

Используются для поиска простейших неисправностей или для более точной локализации неисправностей, обнаруженных с помощью СА или ЛА. К ним относятся:

1) Тестеры логических состояний (ЛТ);
2) Стимулирующие генераторы логических сигналов;
3) Бесконтактные индикаторы импульсных токов.

Эти приборы при их совместном применении позволяют выявить целый ряд неисправностей в цифровых схемах.

2.6.1. Тестеры логических состояний

Представляют собой малогабаритные и дешевые приборы, позволяющие определять логическое состояние сигналов в контрольной точке. С помощью ЛТ можно определить принадлежит ли уровень контролируемого сигнала к уровню «нулевого», «единичного» или промежуточного состояния. ЛТ выпускают в одноконтактном или многоконтактном исполнениях. Первый ЛТ называют логическим зондом, он снабжен металлическим контактным наконечником, которым оператор касается контролируемой точки (КТ). Источник питания - внешний. Как правило, это блок питания МПС. В большинстве ЛТ индикатор состояний (светодиод) погашен при нулевом логическом уровне, ярко светит при единичном и обеспечивает половинное свечение при промежуточном (висячем) состоянии. Если состояние контрольной точки меняется, то индикатор мигает с частотой 10 Гц. Частота контролируемых сигналов не более 280 МГц, а длительность импульсов $\geq$10 нс.

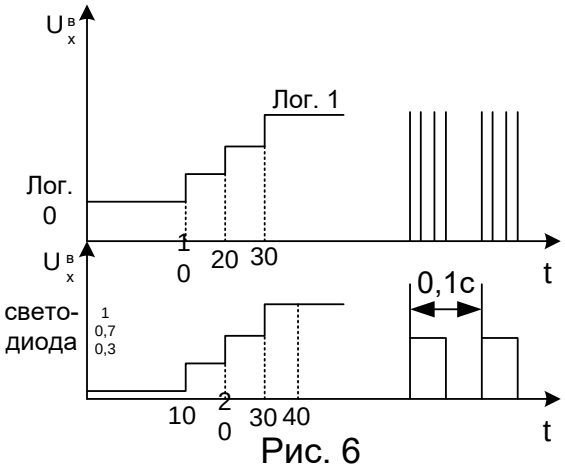


Рис. 6

На корпусе ЛТ имеется переключатель ТТЛ-МОП. В ряде случаев ЛТ снабжен десятичным счетчиком импульсов на двух семисегментных индикаторах. Многоточечный ЛТ или логическая клипса выполнен в виде многоконтактного зажима, который закрепляется сверху интегральной микросхемы. На верхней панели прибора расположены два ряда светодиодных индикаторов, отображающих логическое состояние соответствующих выводов микросхемы (удобна при проверке в пошаговом режиме).

2.7. Стимулирующий генератор

Предназначен для формирования сигналов, амплитуда и длительность которых обеспечивают срабатывание использованных в контролируемом устройстве интегральных микросхем.

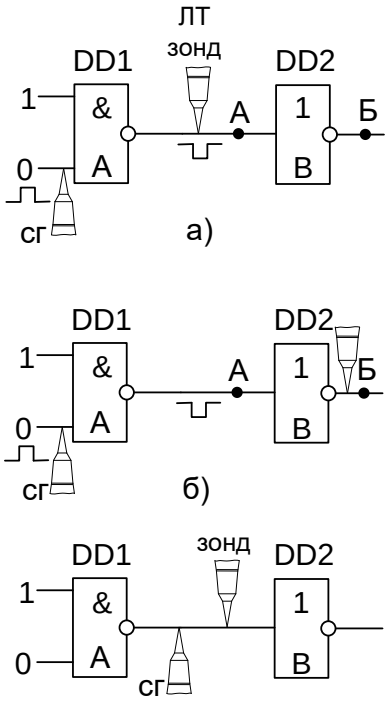


Рис.7. Стимулирующий генератор

Их часто применяют в комплекте с ЛТ и бесконтактными индикаторами импульсных токов.

Конструктивно СГ выполнен аналогично логическому зонду. Он вырабатывает: одиночные импульсы; последовательности импульсов с частотой 1,10,100 Гц; пачки по 10 или 100 импульсов с частотой 1 или 0,5 Гц соответственно.

Примеры совместного применения ЛТ и СГ показаны на рис 7.

Рассмотрим несколько случаев при работе СГ на частоте 10Гц:

1) Индикатор зонда (ЛТ) в точке А мигает, следовательно DD1 исправен (рис.7а);

2) При исправной DD1, индикатор ЛТ при перенесении зонда в точку Б не мигает, следовательно DD2 не исправен (рис.7б);

3) Если при работающем СГ индикатор ЛТ не светится, это говорит о коротком замыкании вывода DD1 на корпус. Если индикатор ЛТ светится и не мигает, это говорит о коротком замыкании вывода на шину +5В.

2.8. Бесконтактные индикаторы импульсных токов (ИИТ)

ИИТ предназначены для контроля токов в защищенных изоляционными покрытиями проводниках печатных плат (ПП). Контроль токов реализуется без разрывов проводников и разрушения покрытий.

ИИТ применяется в комплекте с СГ для поисков КЗ, разрывов в цепи, неисправностей в цепях монтажной логики и в шинах с тремя состояниями.

ИИТ обеспечивают световую индикацию одиночных токовых импульсов длительностью не менее 1 нс и частотой следования до 100 МГц, чувствительность их колеблется от 1 мА до 1 А.

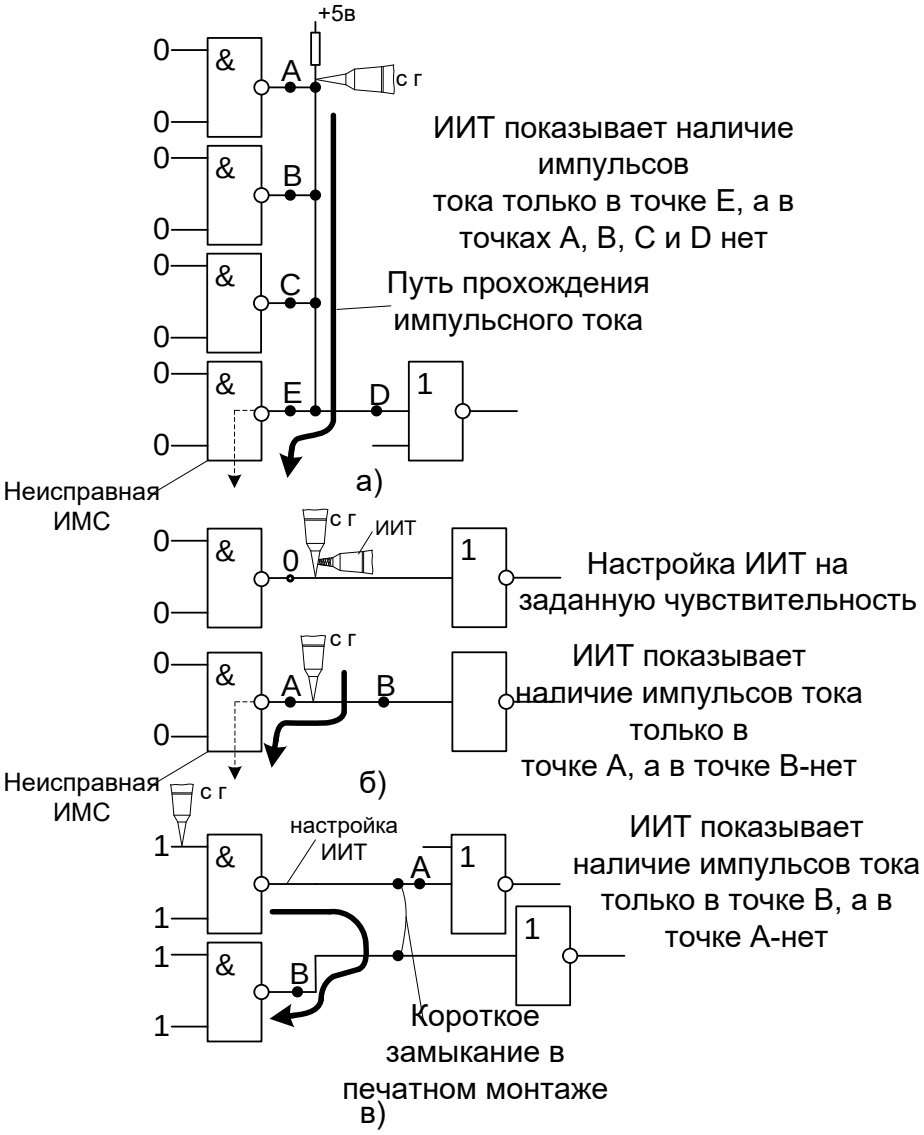


Рис. 8

2.9. Самоконтроль и самодиагностика в МПС

Отметим свойства МПС, позволяющие использовать самоконтроль:

1) Способность МП под действием небольшой тестовой программы генерировать тестовые последовательности очень большого объема;

2) Способность МП к логической обработке информации, что позволяет сравнивать отклики контролируемых устройств с эталонными без применения дополнительной аппаратуры;

3) Программная доступность всех узлов в МПС

Последовательность самодиагностики МПС:

1) программно проверяется МП (вся система команд), счётчик команд, РОН и т.д.;

2) программно проверяется ПЗУ;

3) программно проверяется микросхема ОЗУ;

4) программно проверяются устройства ввода-вывода (микросхемы приёмо-передатчики);

5) программно проверяются внешние устройства (гибкие и жесткие диски);

6) программно проверяются специальные узлы.

Тестовые программы не обязательно находятся в тестовом ПЗУ. Они могут загружаться с одного из дисков.

Для обнаружения неисправностей в ОЗУ используется несколько подходов. Наиболее популярные из них:

1) Тест типа «шахматная доска». При этом тестовая программа заносит в соседние ячейки или все единицы или все нули (рис.2.10а). Каждая клетка выделенной области ОЗУ соответствует одной ячейке. Этот тест позволяет обнаруживать константные неисправности (постоянный «ноль» или постоянная «единица») в одном или нескольких разрядах ячейки, а также паразитные связи между соседними ячейками. После первой проверки делается вторая, перед которой содержимое всех ячеек инвертируется.

	00	01	10	11
00	0	1	0	1
01	1	0	1	0
10	0	1	0	1
11	1	0	1	0

а)

	00	01	10	11
00	0	1	0	1
01	0	1	0	1
10	0	1	0	1
11	0	1	0	1

б)

Рис.2.10

2) Тест с перебором дополнительных адресов. Служит для выявления правильности функционирования внутренних дешифраторов адреса микросхемы ОЗУ. Он начинается записью в память чередующихся столбцов из всех «нулей» или из всех «единиц» (рис.2.10б). Вначале считывается содержимое ячейки с адресом 00-00. Затем содержимое ячейки с обратным адресом, в которой должна находиться информация, противоположная первоначальной. Затем адрес вновь инвертируется и проверяется исходная ячейка. Если все нормально, то происходит переход к следующей ячейке и т.д.

Такой тест представляет серьезное испытание для внутренних дешифраторов адреса микросхемы ОЗУ, а также позволяет выявить все временные задержки, выходящие за допустимые пределы.

3) Тест типа «шахматная линейка». Служит для проверки правильности функционирования разрядовых триггеров внутри самой ячейки. Это обеспечивается записью в ячейку в начале числа 55, а затем АА.

Такая проверка гарантирует от прямого ложного и ложного обратного хода информации за счет паразитных связей между соседними разрядами ячейки.

4) Тест «бегущая единица». При этом записывается единица в один из разрядов контролируемой ячейки памяти. В остальных разрядах записываются «0». После считывания единица переписывается в следующий разряд и т.д. Такая проверка контролирует отсутствие прямого ложного и ложного обратного хода информации через кратное число разрядов ($n \geq 2$) ячейки.

Существует более ста других методов контроля ОЗУ.

2.10. Диагностика одноплатной микро ЭВМ.

Последовательность наладки и контроля микро ЭВМ

Аппаратура контроля:

1. стенды для контроля пустых печатных плат;
2. стенды для контроля комплектующих элементов;
3. стенды для контроля функционально законченных блоков;

3.1. одноплатных устройств;

3.2. блоков питания;

3.3. накопителей на ГМД, ЖМД, клавиатуры и мониторов.

При наладке платы центрального процессора необходимо выполнить следующие действия:

1. Подключить плату к стенду на номинальном напряжении (+5В);

2. Ввести в стендовую микро ЭВМ с одного из накопителей тестовые программы, с помощью которых проверяют:

2.1. Наличие напряжения логических единиц на линиях системной магистрали в пассивном состоянии процессора, эта ситуация имеет место в микро ЭВМ «Электроника» с архитектурой типа DEC. В широко распространенной в настоящее время архитектуре типа IBM PC на всех линиях системной магистрали имеет место высокоимпедансное («висячее») состояние.

При выполнении этой и трёх последующих проверок процессор проверяемой платы поддерживается в пассивном состоянии, а магистраль обслуживается стендовой микро ЭВМ.

2.2. Правильность кодов записанных в ПЗУ при считывании их стендовой микро ЭВМ.

2.3. Работоспособность ОЗУ при взаимодействии со стендовой микро ЭВМ.

2.4. Работоспособность контроллеров параллельного и последовательного каналов связи (LPT и COM порты).

2.5. Работоспособность МП в режимах самоконтроля в реальном темпе времени.

2.6. Работоспособность ОЗУ и ПЗУ при взаимодействии с МП в реальном режиме времени.

2.7. Работоспособность параллельного и последовательного каналов связи при взаимодействии с работающим МП.

2.8. Правильность обращений к системной магистрали со стороны МП.

3. Снизить номинальное напряжение на 10% после чего повторить пункт 2.

4. Повысить номинальное напряжение на 10% после чего повторить пункт 2.

2.11. Проверка линий системной магистрали в пассивном состоянии

При этом на всех линиях магистрали должно присутствовать напряжение логической единицы (около 3В). Это архитектура типа DEC, так как линии выполнены по схеме с открытым коллектором, а элементы передатчики выключены. Чтобы поддержать проверяемую плату ЦП в пассивном состоянии стендовая микро ЭВМ постоянно подает в управляющие линии АСП (аварии сетевого питания) и АИП (аварии источника питания) системной магистрали напряжение низкого уровня. В этом случае МП не проявляет активности на магистрали т.к. с его точки зрения напряжение питания находится за допустимыми пределами, хотя в действительности питание платы в норме.

Стенд в данном случае не проявляет активности, все передающие элементы его выходного порта (за исключением элементов работающих на линии АСП и АИП) выключены. При выполнении дальнейшей проверки линий системной магистрали стендовая микро ЭВМ выдает различные комбинации сигналов на выходы порта и многократно считывает информацию со всех входов порта, подключенных к системной магистрали. Если при многократной проверке выясняется, что на всех входах присутствуют напряжения логической единицы, то проверка считается успешной. Если на одном из входов хотя бы однажды обнаруживается напряжение «0», то стендовая микро ЭВМ выдает сообщение оператору и не переходит к последующим этапам проверок до

тех пор, пока неисправность не будет устранена или с клавиатуры стенда не поступит команда продолжения проверки.

В сообщении на дисплее указывается, на каких линиях системной магистрали присутствовали неисправности. При желании устранить неисправность оператору рекомендуется просмотреть осциллограмму «неисправного» сигнала. Иногда вместо объявленного на экране дисплея нуля можно увидеть на осциллограмме сигнал промежуточного уровня или высокочастотную генерацию. Эту проверку можно усилить введением дополнительного теста, позволяющего обнаружить КЗ между линиями системной магистрали. Дело в том, что даже в маловероятной ситуации, когда все линии короткозамкнуты между собой, кроме линий АСП и АИП, то на них будет напряжение логической единицы. Для обнаружения таких КЗ стендовая микро ЭВМ поочередно выдает через выходы порта напряжение нуля в линии системной магистрали и следит за ответной реакцией проверяемой платы.

Рассмотрим еще один способ проверки системной магистрали и внутренней магистрали, позволяющей обнаружить неисправности без прогонки сложных тестов. Этот способ применим, если МП стоит в сокете (контактирующей колодке). При рассматриваемых проверках, когда МП находится в пассивном состоянии, есть смысл извлечь его из панели. Сокет является дополнительным внешним разъемом платы ЦП, и поэтому к ней можно подключить дополнительный кабель связи с портами стендовой микро ЭВМ. В этом случае стендовая микро ЭВМ получает возможность доступа к внутренней магистрали платы ЦП, что позволяет расширить стендовые проверки цепей, распространив их на внутреннюю магистраль и на единую систему, включающую внешнюю магистраль и адаптер магистрали. На последующих этапах проверок стендовая микро ЭВМ заменяет собой отсутствующий МП.

В панели можно установить пустой корпус микросхемы МП с распаянными на нем перемычками. В этом случае, например, входной сигнал запроса прерывания, поступающий со стенда через системную магистраль на пустой корпус МП с помощью одной из перемычек транслируется на вывод разрешения прерываний и обратно на стендовую микро ЭВМ.

Таким образом, стендовая микро ЭВМ может убедиться в наличии ответного сигнала от МП. Идея таких проверок состоит в том, чтобы из разрозненных цепочек логических элементов связать «гирлянды», концы которых доступны со стороны стенда. Тогда полученные длинные цепочки можно проверять, передовая по ним нули или единицы.

Аналогичные проверки можно выполнить, если есть панель для ПЗУ пользователя. При всей своей простоте это очень «сильные» проверки, т.к. чем проще тест, тем легче найти ошибки (в пассивном состоянии МП).

2.12. Диагностика ПЗУ

МП должен быть по-прежнему отключен. Проверка проводится с помощью стендовой микро ЭВМ, которая последовательно считывает содержимое всех ячеек ПЗУ и вычисляет контрольную сумму , причем перенос из старшего разряда не теряется, а используется в качестве дополнительной единицы, прибавляемой к младшему разряду.

Стендовая микро ЭВМ сравнивает полученную сумму с эталонной.

Виды ошибок:

1) Ошибки из-за нарушения последовательности событий на шинах интерфейса при операциях считывания из ПЗУ. При обнаружении такой ошибки стендовая микро ЭВМ прекращает попытки считывания массива кодов. Оператору на дисплей выдается текстовое сообщение о характере ошибки.

Например, «неправильно формируется сигнал СИП при обращении в ячейку ПЗУ с адресом 140006». При этом стендовая микро ЭВМ **«зацикливается»** на неправильной процедуре обмена таким образом, что с помощью осциллографа можно увидеть реальные временные диаграммы и сравнить их с эталонными;

2) Ошибки, вызывающие несовпадение контрольной суммы с эталоном. В этом случае "под подозрение" попадает внешняя магистраль, адаптер, внутренняя магистраль и микросхема ПЗУ. Чтобы найти ошибку в массиве кодов, стендовая микро ЭВМ считывает с гибкого диска эталонный массив и последовательно, шаг за шагом, сравнивает его с реальным массивом кодов, поступающих из ПЗУ. При первом же несовпадении кодов оператору выдается сообщение (адрес ячейки, эталонное и фактическое содержимое), а стендовая микро ЭВМ "зацикливается" на процедуре считывания неправильного кода, так что регулировщик с помощью осциллографа может быстро найти ошибку.

После окончания проверки ПЗУ появляется уверенность в правильной работе внешней и внутренней шин адресов-данных, в исправности большей части микросхем адаптера магистралей и, конечно, в исправности микросхемы ПЗУ. Отметим, что никаких принципиальных сложностей в наладке не было, да и быть не могло, так как сами методы проверки предельно просты. В то же время достигнутый нами уровень "безошибочности" платы уже намного выше начального.

2.13. Диагностика ОЗУ

Ранее эти вопросы рассматривались. Рассмотренные тесты выявляли тонкие неисправности в ячейках и разрядах ОЗУ. При контроле ЦП на стенде реализуется более простая методика контроля платы.

Вначале в каждую ячейку ОЗУ записывается ее адрес, после этого содержимое ОЗУ считывается и проверяется, затем в каждую ячейку ОЗУ

записывается ее адрес в обратном коде. После этого информация считывается и проверяется.

Процесс чтения информации из ОЗУ полностью совпадает с рассмотренным ранее процессом чтения из ПЗУ. При несовпадении содержимого ячейки с ее адресом стендовая микро ЭВМ зацикливает процесс взаимодействия стенда с неисправной ячейкой ОЗУ, что позволяет с помощью осциллографа на шинах интерфейса определить правильность ее функционирования.

2.14. Проверка контроллера последовательного канала связи

(На примере платы МП с микросхемой типа К1801ВП033)

Контроллер последовательного канала связи (рис. 5.1) состоит из двух независимых секций - передающей и приемной, размещенных в одной и той же микросхеме.

Передающая секция, как и в контроллере АЦПУ, представлена на магистрали регистром состояния *РгС1* и регистром данных *РгД1*. Этим регистрам присвоены адреса, так что программа, выполняемая процессором или стендовой микро ЭВМ, оперирует этими регистрами, как с обычными ячейками памяти. Адреса регистров *РгС1* и *РгД1* не совпадают с адресами других устройств платы центрального процессора.

Обмен информацией между процессором и передающей секцией контроллера последовательного канала связи протекает так же, как и

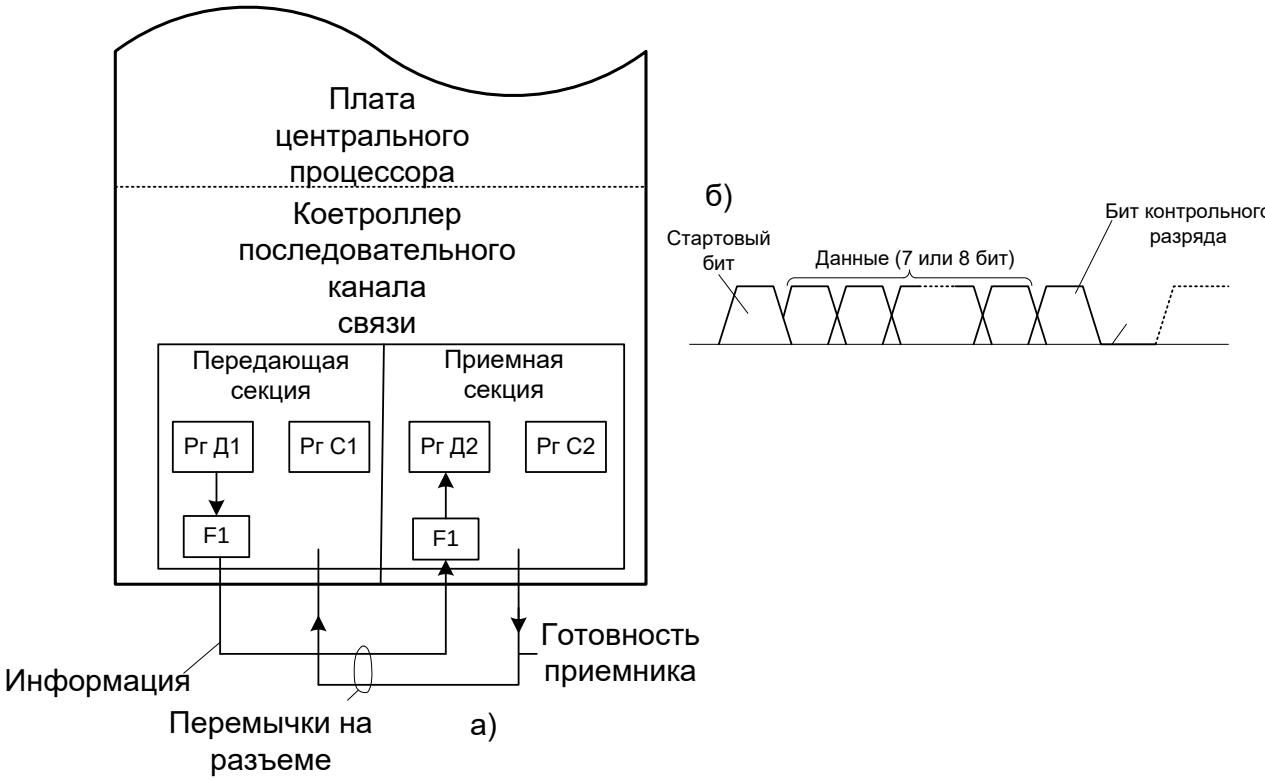


Рис. 9. Схема проверки контроллера последовательного канала связи (а) и формат передаваемой по каналу посылки (б)

при взаимодействии процессора с контроллером АЦПУ. Иными словами, процессор "не знает" о том, как передаваемые байты будут обрабатываться контроллером: пойдут ли они на выход контроллера без изменения формата (как при работе с АЦПУ) или будут преобразованы в стандартные посылки для передачи в последовательный канал связи.

Если в разряде ***Готовность*** регистра *РгС1* присутствует "лог. I", то процессор "знает", что регистр *РгД1* готов принять из магистрали код для передачи в последовательный канал связи. Одновременно с засылкой кода в регистр *РгД1* в разряде *Готовность* регистра *РгС1* автоматически устанавливается сигнал "лог. О". После загрузки регистра *РгД1* преобразователь *F1* параллельного кода в последовательный анализирует состояние линии *Готовность приемника.* Если готовность имеется, то преобразователь *F1* передает в информационную линию посылку, формат которой показан на рис. 9б.

Сначала передается "стартовый" бит, затем 7 или 8 бит данных, затем бит контрольного разряда, дополняющий число единиц в передаваемом коде данных до четного или до нечетного. После окончания передачи посылки, передающая секция контроллера выдерживает паузу определенной минимальной длительности (1; 1,5 или 2 битовых интервала), после чего может последовать новая посылка. Конкретные параметры передаваемой посылки задаются с помощью микропереключателей или перемычек на проверяемой плате центрального процессора. К таким параметрам относятся: число битов данных, наличие

или отсутствие контрольного разряда, при наличии контрольного разряда - вид контроля (по четности или по нечетности), длительность минимальной паузы, длительность битового интервала (скорость передачи) и др. Скорость передачи информации по последовательному каналу связи обычно невелика и составляет, например, 9600 бит/с (учитываются не только "полезные" биты, но и служебные, см. рис. 5.1б, так что фактическая скорость передачи меньшая). При скорости 9600 бит/с длительность битового интервала равна *1/9600 с,* или примерно 104 мкс. (Тем не менее стендовая микро ЭВМ неспособна с нужной точностью воспроизвести диаграмму, показанную на рис. 5.1б, при ее программном моделировании с помощью выходного порта.)

Приемная секция контроллера последовательного канала связи содержит преобразователь *F2* последовательного кода в параллельный и, как обычно, два регистра: регистр данных *РгД2* и регистр состояния *РгС2.*

Принятая из последовательного канала связи информационная посылка (8 бит) проверяется на отсутствие в ней ошибок (подсчетом общей четности или нечетности числа принятых единиц, без учета стартового бита) и записывается в регистр данных *РгД2.* При этом в регистре *РгС2* устанавливается в единичное состояние бит готовности информации, а сигнал *Готовность приемника* снимается для того, чтобы передатчик не послал следующую байтовую посылку, пока не освободится регистр данных приемника.

В момент считывания информации из регистра *РгД2* (во внутреннюю магистраль платы) автоматически устанавливается в нулевое состояние разряд готовности регистра *РгС2.* Одновременно с этим формируется сигнал *Готовность приемника,* который разрешает передатчику выдачу следующей байтовой посылки и т.д.

Контроллер последовательного канала связи, так же как и контроллер АЦПУ, может работать в режиме прерывания процессора.

Проверку контроллера последовательного канала связи с помощью стендовой микро ЭВМ можно осуществлять при выполнении следующих операций.

1. При наличии готовности передающей секции записать код $ООО_8$ в регистр данных *РгД1* как в ячейку памяти, при этом на каждом шаге программы имитации записи проверяют все доступные стенду входные точки с целью выявления различного рода ошибок.

2. Прочитать содержимое регистра *РгС2,* обратившись к нему как к ячейке памяти. Эта операция, как мы знаем, также состоит из целого ряда более мелких операций, каждая из которых должна контролироваться стендовой микро ЭВМ путем анализа содержимого входного порта.

3. Проанализировать биты *Готовность* и *Ошибка* в считанном содержимом регистра состояния *РгС2*. При правильной работе бит ошибки не должен быть установлен, а бит готовности должен присутствовать.

4. Прочитать содержимое регистра данных *РгД2,* обратившись к нему, как к ячейке памяти.

5. Сравнить принятый код с копией переданного и в дальнейшем повторить все указанные действия 256 раз, по числу комбинаций в 8-разрядном коде данных.

После выполнения этих операций имитируются режимы прерывания процессора по сигналам из передающей и приемной секций контроллера. Принцип проверки и здесь остается тем же - стенд выполняет все операции, которые должен был бы выполнить процессор. Для имитации сигнала «*Разрешение прерывания»,* который вырабатывается процессором, необходимо иметь доступ к контактирующей колодке, в которую в дальнейшем вставляется микросхема процессора. Если процессор установлен на плате "жестко", то режим прерывания либо не проверяется на данном этапе, либо проверяется с использованием схем встроенного контроля, которые специально для таких случаев "закладываются" в плату на этапе ее проектирования.

Для контроля целостности линии *Готовность приемника* и связанных с ней схем на одном из этапов проверки нужно эту линию разомкнуть (информация по линии передается в виде тока, а не напряжения), например, с помощью кнопки по указанию стендовой микро ЭВМ. (Это еще один пример того, как микро ЭВМ, выдавая сообщения на экран дисплея, руководит действиями человека, а не наоборот.) Размыкание может быть выполнено и автоматически, с использованием, например, реле, управляемого от выходного порта микро ЭВМ.

Для проверки схем контроля необходим имитатор, вырабатывающий искаженные кодовые посылки с ошибками по четности (нечетности).

Для проверки правильности работы проверяемого устройства при различных положениях микропереключателей, установленных на плате, предусматривается соответствующий тест. При выполнении этого теста стендовая микро ЭВМ выводит на экран дисплея сообщения, которые предписывают оператору перевести тот или иной микропереключатель в нужное положение и подтвердить выполнение этого действия, например, нажатием клавиши «Пробел».

После успешного выполнения всех рассмотренных выше проверок открывается возможность запуска центрального процессора проверяемой платы.

2.15. Проверки, выполняемые с помощью процессора проверяемой платы

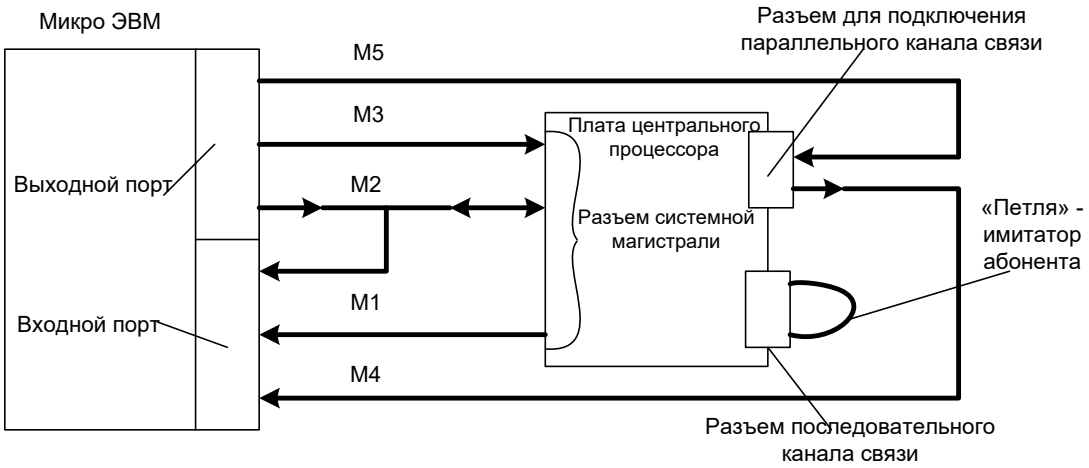


Рис.10. Схема подключения платы ЦП к наладочному стенду

Сначала процессор должен проверить сам себя. Проверка процессора проводится по уже знакомому нам принципу «от простого – к сложному». Вопросы тестирования процессора выходят за рамки этого параграфа, поэтому здесь мы не будем на них останавливаться. То же относится к проверке ОЗУ и контроллеров.

Процессор проверяемой платы не имеет возможности непосредственно загружать тестовые и наладочные программы с НГМД в «свою» память, поэтому загрузка производится с помощью стендовой микро ЭВМ. Сначала она считывает информацию в НГМД в «стендовое» ОЗУ, а затем последовательно, слово за словом, пересылает эту информацию в память проверяемой платы, используя свои порты ввода – вывода. Во время пересылки, процессор проверяемой платы заторможен сигналами на линиях АСП и АИП. После того как информация переписана в память проверяемой платы, стендовая микро ЭВМ снимает напряжение низкого уровня с линий АИП и АСП (см. рис.11) и процессор стартует с определенного адреса, соответствующего началу принятой из стенда тестовой программы.

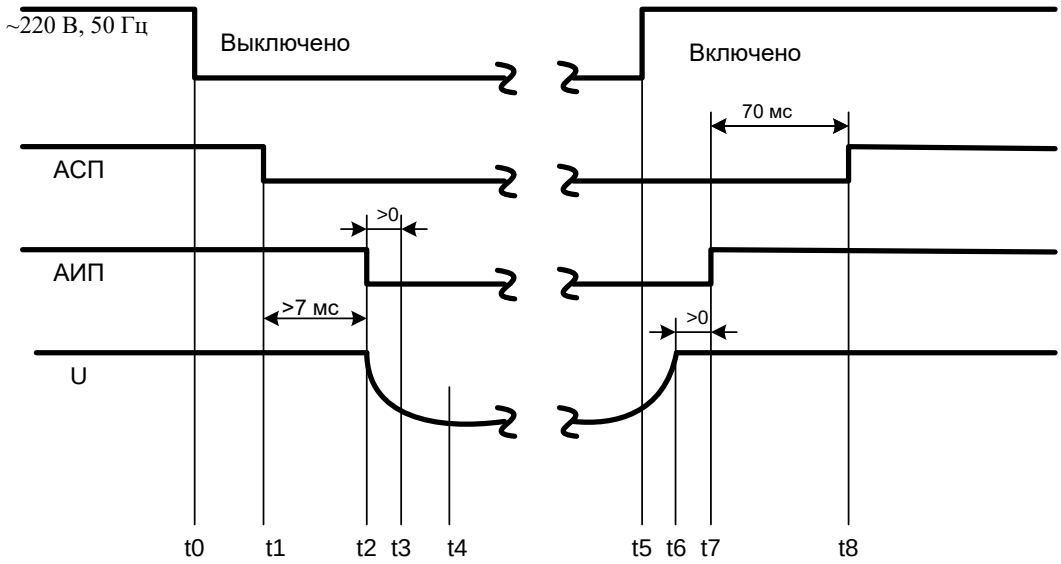


Рис.11. Временная диаграмма сигналов, вырабатываемых блоком питания при включении и при выключении напряжения первичной питающей сети

Проверяемая плата периодически, через канал связи с АЦПУ, сообщает стендовой микро ЭВМ о ходе выполнения тестовых программ. Если проверяемая плата слишком долго не отвечает или отвечает неправильно, то стендовая микро ЭВМ выдает оператору подробное сообщение о том, какая проверка не проходит и что нужно сделать для «зацикливания» участка программы, на котором регистрируется неправильная работа.

Как и в рассмотренных ранее примерах, ошибки можно условно разделить на две группы:

1) ошибки в управляющих цепях, которые полностью или частично лишают процессор возможности выполнять программу;

2) «информационные» ошибки, которые искажают результаты вычислений, но не препятствуют ходу выполнения тестовой программы. Задача состоит в том, чтобы «продемонстрировать» регулировщику те и другие ошибки. Для «демонстрации» ошибок первой группы управление периодически передается той команде (или цепочке команд), которая «не хочет» выполняться. Для периодической передачи управления стендовая микро ЭВМ вырабатывает непрерывную последовательность сигналов начального запуска процессора на линиях АИП и АСП (см. рис.11). Период следования этих сигналов должен быть достаточным для прогона интересующего нас участка программы. Каждый раз при обнаружении условий запуска процессор стартует в определенную ячейку ОЗУ, где его ожидает команда безусловной передачи управления на интересующую нас команду или цепочку команд. Процессор, таким образом, несмотря на возможность выполнения действий в связи с ошибкой, периодически обращается к «ошибочному» месту, и его «поведение» можно легко проконтролировать с помощью осциллографа.

Рассмотрим процесс загрузки тестов и некоторые особенности их выполнения в условиях совместной работы проверяемой платы и наладочного стенда (см. рис.10).

Далее «демонстрацию» ошибок второй группы необходимо «зациклить» проверяемый процессор на интересующей нас команде или группе команд.

На заключительном этапе, после прогона тестов процессора, ПЗУ, ОЗУ и контроллеров, проверяется правильность обращения процессора к внешней системной магистрали (по записи). В качестве внешнего абонента (по отношению к проверяемой плате) выступает стендовая микро ЭВМ – объект чрезвычайно «медленный» с точки зрения процессора. При таких обращениях возникают «безответные» ситуации, и процессор, выждав приблизительно 10мкс, переходит в режим прерывания.

Прерывающая программа вновь предписывает процессору выполнить обращение к стенду через внешнюю системную магистраль, процессор не получает ответа, переходит в режим прерывания и т.д. Таким образом, процессор периодически «приглашает» стендовую микро ЭВМ, но в силу ее инерционности не получает ответа. Стендовая микро ЭВМ, в свою очередь, следит за состоянием линий входного порта (с доступной ей сравнительно низкой частотой) и набирает статические данные о сигналах, сформированных на этих линиях.

Так как «безответные» обращения наиболее продолжительны по времени, вероятность обнаружения стендовой микро ЭВМ значительно выше, чем вероятность обнаружения любого прочего обращения. Поэтому программа, выполняемая стендовой микро ЭВМ, способна выделить в полученном массиве кодов наиболее часто повторяющийся код (периодически выдаваемый процессором проверяемой платы во внешнюю системную магистраль) и сравнить его с эталонным. При обнаружении несовпадения указанных кодов стендовая микро ЭВМ выдает оператору соответствующее сообщении и продолжает выполнение описанного теста, что позволяет обнаружить ошибку с помощью осциллографа.

Итак, включение в работу процессора, установленного на проверяемой плате, позволяет подвести окончательные итоги проверке «внепроцессорной» аппаратуры – теперь она работает на высокой скорости и здесь могут проявиться ошибки, связанные с замедленным включением - выключением логических элементов. Например, при выдаче данных из контроллера последовательного канала связи по внутреннюю магистраль один из разрядов данных может «отстать» от остальных – на 100 или 200 нс. Эту ошибку нельзя зарегистрировать при имитации процессора, так как между получением ответного сигнала от контроллера и моментом считывания данных проходит более 10 мкс. В режиме реального времени данная ошибка приведет к искажению полученной из контроллера информации и, следовательно, будет обнаружена.

После наладки платы центрального процессора с использованием стенда, показанного на рис.10, плата проверяется в рельных условиях в

срставе эталонной микро ЭВМ, так как с помощью наладочного стенда невозможно создать условия, которые точно соответствовали бы реальным. Если налаженная плата все же не работает в составе эталонной микро ЭВМ (такие ситуации чрезвычайно редки, но возможны), то поиск неисправности проводят при работе в реальных условиях, например с помощью логического анализатора. После выбора ошибки анализируют причины, которые препятствовали ее обнаружению аппаратно-программными средствами наладочного стенда, и эти средства дорабатываются.

2.16. Программное обеспечение наладочных стендов

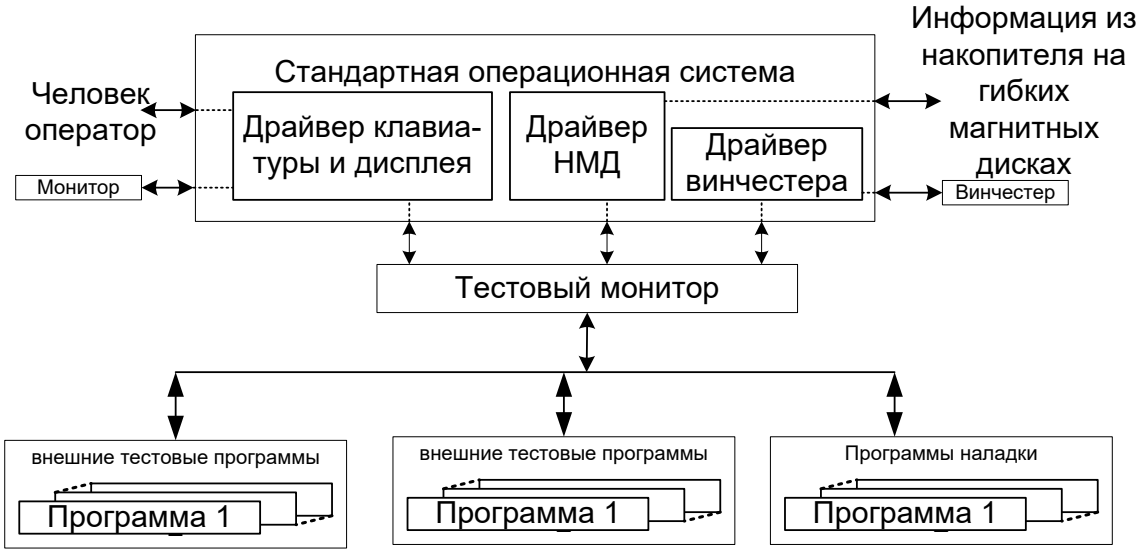


Рис. 12. Структура программного обеспечения стенда

Стенды предназначены для наладки составных частей микро ЭВМ. Принципы построения программного обеспечения стендов во многом схожи. Рассмотрим стенд, построенный на основе универсальной микро ЭВМ с входным и выходными портами и предназначенный для наладки ЦП.

Стандартная операционная система используется:

1. для передачи команд с клавиатуры в тестовый монитор и сообщений от тестового монитора на экран дисплея;
2. для ввода тестовых и наладочных программ из НГМД в ОЗУ стендовой микро ЭВМ по запросам тестового монитора.

Тестовый монитор осуществляет:

- прием команды от человека – оператора;
- анализ команды, выдачу сообщения оператору при обнаружении ошибки в задании команды;

- вызов из НГМД тестовой или наладочной программы, если она еще не размещена в ОЗУ стендовой микро ЭВМ;

- настройка тестовой или наладочной программы на выполнение требуемой функции в соответствии с параметрами, заданными оператором в команде;

- передачи управления внешней тестовой программе;

- пересылка внутренне тестовой программы из ОЗУ стендовой микро ЭВМ в ОЗУ проверяемой платы и слежения за ходом тестирования;

- передачу управления программе наладки;

- последовательную передачу управления внешними тестами программам с выдачей информации о ходе теста на экран дисплея;

- последовательную пересылку внутри тестовых программ из ОЗУ стендовой микро ЭВМ в ОЗУ проверяемой платы ЦП с выполнением этих программ ЦП;

- переход на выполнение тестовой программы к выполнению соответственной наладочной программы при обнаружении ошибки.

Внешние тестовые программы - выполняются процессором стендовой микро ЭВМ. При этом микросхема ЦП заторможена или не установлена на плату.

К внешним тестам относятся:

1. тест проверки линий системной магистрали в пассивном состоянии;
2. тест проверки правильности кодов, записанных в ПЗУ проверяемой платы;
3. тест проверки работоспособности ОЗУ проверяемой платы;
4. тест проверки контроллера АЦПУ проверяемой платы;
5. тест проверки контроллера последовательного КС проверяемой платы.

Внутренние тестовые программы – выполняются процессором проверяемой платы. К внешним тестовым программам относятся:

1. тест проверки ПЗУ (сумма содержимого всех ячеек);
2. тест проверки ОЗУ;
3. тесты проверки контроллера АЦПУ и контроллера последовательного КС;
4. тесты процессора – проверяют правильность выполнения команд из системы команд, работу процессора в режиме прерывания, прямого доступа в память, работу арифметико-логического устройства процессора.

Программы наладки – служат для зацикливания элементарных процедур передачи сигналов по магистралям. Программы наладки позволяют зациклить процедуры:

- запись заданного кода в заданную ячейку памяти или регистр;

- запись (считывания) поочередно информации из ячейки памяти или регистра;

- начальный пуск процессора с определенного адреса, по которому размещена элементарная программа наладки.

Проверяемая плата ЦП является объектом тестирования и выполняет подчиненную роль. Проверяемая плата может отсутствовать, результаты теста при этом будут отрицательными.

Глава 3. Прогнозирование состояния технических систем.

3.1. Задачи прогнозирования

Задача прогнозирования изменения свойств технической системы тесно связана с задачей диагностики текущего состояния этой системы. Принципиальное отличие заключается в том, что при прогнозировании оценка состояния диагностируемой системы производится для будущего момента времени. Эта оценка находится по результатам текущих наблюдений на определенном промежутке времени путем экстраполяции значений параметров на будущий момент времени. Естественно, что достоверность прогноза определяется объемом информации о механизме происходящих в диагностируемой системе процессов, наличием обоснованных физических и математических моделей изменений отдельных параметров, возможностью учета множества действующих факторов. Кроме того, имеется ещё одна сложность при решении задач прогнозирования, которая заключается в индивидуальном характере изменения параметров устройств, связанным с первоначальным разбросом параметров комплектующих элементов, погрешностями сборки и другими факторами.

Поэтому задача прогнозирования качества функционирования технических систем, в общем, сводится к задаче экстраполяции случайного векторного процесса по результатам его наблюдения на некотором интервале времени. При этом прогноз предсказывает будущее значение случайного процесса не точно, а с некоторой вероятностью (точечный прогноз) или предсказывает ту область, в которую с заданной вероятностью попадает будущее значение процесса (интервальный прогноз).

В такой постановке прогнозирование состояния технической системы переходит в текущий диагностический контроль при нулевом времени экстраполяции и, следовательно, **прогнозирование можно рассматривать как более общий случай процедуры диагностирования.**

Итак, множество параметров, характеризующих текущее состояние системы, можно рассматривать в качестве векторного случайного процесса. И именно случайный характер изменения физических параметров технических систем в процессе их эксплуатации делает трудно предсказуемыми будущие значения этих параметров и накладывает естественные ограничения на применение для решения задач прогнозирования так называемых аналитических методов, при которых изменение параметров рассматривается как детерминированный процесс.

Хотя надо заметить, что такой подход из-за его относительной простоты находит достаточно широкое распространение на практике и дает более-менее удовлетворительные результаты при небольших интервалах экстраполяции.

При формальной постановке задачи прогнозирования возможны **два принципиально разных способа** [16].

В первом из них называемым **интегральным,** оператор прогнозирования $\mathbf{G}_1$ определяется из условия экстремума выбранного критерия оптимальности, применяемого ко всему множеству диагностических сигналов **{X(tj)},i** $=\overline{l,n}$**.** Это позволяет получить значение оценки **вектора параметров** в будущий момент времени $(\mathbf{t_n+\tau})$ в виде:

$$\mathbf{U(t_n+\tau)=G_1\{X(t_i)\},i=\overline{1,n},}$$

где $\boldsymbol{\tau}$-время прогнозирования,

$\mathbf{U\ (t_n+\tau)}$ **-** оценка вектора параметров в будущий момент времени.

Во втором способе, называемом **рекуррентным,** оператор прогнозирования $\mathbf{G_2}$ необходимо определить из условия экстремума критерия оптимальности, применение которого осуществляется только к последнему наблюдению $X(t_n)$ и оценке вектора параметров, полученной на предшествующем шаге $\mathbf{t_{n-1}}$:

$$\mathbf{U(t_n+\tau)=G_2\{X(t_n)\},U(t_{n-1}+\tau)\}.}$$

Таким образом, при применении рекуррентного оператора после каждого последующего наблюдения производится прогноз на время $\boldsymbol{\tau}$. Преимущество такого подхода заключается, прежде всего, в минимизации объемов требуемой памяти вычислительной машины, так как для расчетов в этом случае в памяти необходимо хранить только значение оценки вектора параметров $\mathbf{U(t_{n-1}+\tau)}$ и последнего наблюдения $\mathbf{X(t_n)}$**.** Таким образом, если вектор параметров имеет размерность m, то в памяти должно находиться **2m** чисел. При интегральном способе прогнозирования необходимо запоминать m·n чисел, поскольку при обработке используются все предшествующие **n** наблюдений m-мерных векторов. Поэтому **интегральный способ** прогнозирования целесообразно применять в случаях, когда отсутствуют ограничения на время решения задачи прогнозирования и имеется вычислительная машина с достаточным объемом памяти.

Определение операторов прогнозирования $\mathbf{G}_1$ и $\mathbf{G_2}$ возможно традиционным путем минимизации среднего квадрата ошибки D

$$\mathbf{D = \min M\{[U(t_n+\tau)-U^*(t_n+\tau)]^T\,[U(t_n+\tau)-\ U^*(t_n+\tau)]\},}$$

где **M** - оператор математического ожидания,

U*(t$_n$**+τ)** - «истинное» значение вектора параметров в прогнозируемый момент времени. Минимизация величины D осуществляется путем выбора оператора G_1 или G_2 [16].

Таким образом, для решения задачи прогнозирования необходима априорная информация о характеристиках вектора параметров $\mathbf{U_t}$. В процессе функционирования технических систем эти параметры изменяются, причем эти изменения имеют две составляющие, а именно: медленную - дрейф и быструю - скачкообразные отклонения. Под дрейфом параметров понимают совокупность физико-химических процессов коррозии, усталостной деформации, механического износа и т.п., приводящих к необратимым изменениям свойств элементов и конструкции устройства. Причиной скачкообразных отклонений параметров являются обрывы, замыкания, заклинивание и т.п., которые в свою очередь обусловлены дефектами диагностируемого устройства и эксплуатационными перегрузками. Заметим, что дрейф параметров может способствовать возникновению скачков, так как накопление, например, усталостной деформации или процессы коррозии уменьшают пределы прочности конструкции. Из всего вышесказанного можно сделать вывод о том, что **задача прогнозирования состояния технической системы во многом сводится к прогнозированию дрейфа параметров этой системы.**

Экспериментальные данные показывают, что в большинстве случаев дрейф параметров является случайным нестационарным процессом, причем характерной особенностью этого процесса является большое **время корреляции**, превосходящее время ресурса того или иного элемента или устройства в целом [16].

В [16] предлагается в качестве **математической модели дрейфа параметров** использовать полиномиальную зависимость

$$\mathbf{u}_i = \mathbf{\Sigma} \sum_{k=1}^{\lambda} u_{ik}\, \mathbf{t}^{k-1}, \qquad (3.1)$$

где **i** – номер параметра,

λ- порядок полинома,

$\mathbf{U}_{ik}$ - случайные коэффициенты, характеризующие:

$\mathbf{U}_{i1}$ - начальное значение,

$\mathbf{U}_{i2}$ - скорость изменения,

$\mathbf{U}_{i3}$ - ускорение изменения параметра и т.д.

На практике, как правило, **ограничиваются полиномом второго порядка.** В этом случае выражение (3.1) можно записать, опуская номер параметра, в виде

$$\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2\, \mathbf{t} + \mathbf{u}_3\, \mathbf{t}^2 \qquad (3.2)$$

При небольших временах прогноза в ряде случаев можно ограничиться и линейной зависимостью

$$\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2\, \mathbf{t} \qquad \mathbf{(3.3)}$$

Следует заметить, что полное описание моделей дрейфа параметров кроме зависимостей **(3.2)** или **(3.3)** должно включать также **совместный** закон распределения вероятности коэффициентов полинома — случайных величин $\mathbf{U}_к$, который в большинстве случаев аппроксимируется нормальным законом с уровнем значимости 0,05 [16].

Полная модель процесса изменения параметров должна включать и модель, **описывающую скачки параметров.** В качестве последней модели может служить случайный процесс, в котором появление скачков происходит в случайные моменты времени, а значение параметров после скачка является случайной величиной. Обычно ограничиваются

упрощенной моделью, в которой закон распределения вероятности моментов времени появления скачков принимается пуассоновским, а значение параметра после скачка считается известным [16]. В этом случае **математическое ожидание некоторого параметра** записывается в виде

$$\mathbf{u(t)=e^{-\beta\cdot t}(u_1+u_2t+u_3t^2)+(1-e^{-\beta\cdot t})a},$$

где $\mathbf{e}^{-\beta t}$ - вероятность того, что скачка за время t не будет,

$\mathbf{(1-e^{-\beta t})}$ - вероятность хотя бы одного скачка за время t,

$\mathbf{(u_1+u_2\cdot t+u_3\cdot t^2)}$ - значение параметра до скачка,

«а» - значение параметра после скачка.

И ещё одно замечание. Рассмотренные модели не отражают того факта, что реальное измерение параметров сопровождается помехами и ошибками измерения. Для учета в модели этого факта, в предположении, что справедлива гипотеза об аддитивном характере смеси «истинного» значения параметра u(t) и ошибки наблюдения N(t), можно **наблюдаемое значение параметра** представить в виде

$$\mathbf{x(t)=u(t)+N(t).}$$

Рассмотренный подход к прогнозированию параметров реализован в виде алгоритмов вычисления коэффициентов полиномов по наблюдаемым данным. Формулы для оптимальной оценки i-го коэффициента полинома для общего и ряда частных случаев приведены в [16].

3.2. Основные методы прогнозирования

В основу решения задачи прогнозирования могут быть положены два различных подхода. Первый базируется на том, что оценка прогнозируемого явления осуществляется на основе аналогии с достаточно хорошо известными явлениями и процессами, позволяющей построить соответствующую прогностическую модель. Второй подход к получению прогноза предполагает условное продолжение в будущее тенденций и закономерностей изменения состояния диагностируемого объекта в настоящем и прошлом, т.е. является экстраполяционным подходом. Таким образом, методы прогнозирования, реализующие первый подход, предполагают наличие некоторой математической

прогностической модели, например, в виде системы интегро-дифференциальных уравнений, достаточно адекватно описывающей процессы, происходящие в диагностируемом объекте. Однако во многих случаях получение таких адекватных моделей объектов, которые к тому же отражали бы их индивидуальные особенности, является делом весьма трудным, что в значительной мере ограничивает применение на практике подобных методов. Более распространенными методами прогнозирования являются методы, основанные на схеме экстраполяции. Основу этих методов составляет оценка **трендовых тенденций** параметров на базе временных рядов [14].

Генерирование временных рядов значений параметров с целью получения экстраполяционных прогнозных оценок требует **вычисления взвешенных значений предыдущих наблюдений** путем применения различных **способов сглаживания** [23,30]. Одним из таких способов, достаточно широко распространенных, является **метод экспоненциального сглаживания**, фактически являющийся обобщением метода скользящего среднего, также применяемого для операции сглаживания результатов текущих наблюдений.

Экспоненциальным сглаживанием данных называется операция **(7.4).** Выражение **(7.4)** означает, что текущее значение сглаженной величины $\tilde{\mathbf{u}}_t$ равно предыдущему её значению плюс некоторая доля разности между текущим значением результата наблюдения и предыдущим значением сглаженной величины.

$$\tilde{\mathbf{u}}_t = \boldsymbol{\alpha}\, \mathbf{u}_t + (\mathbf{1} - \boldsymbol{\alpha})\, \tilde{\mathbf{u}}_{t-1} = \tilde{\mathbf{u}}_{t-1} + (\mathbf{u}_t - \tilde{\mathbf{u}}_{t-1})\boldsymbol{\alpha} \qquad \textbf{(3.4)}$$

где $\tilde{\mathbf{u}}_t$ - сглаженное значение результата наблюдения, проведенного в момент **t**;

$\mathbf{u_t}$ - текущее значение результата наблюдения за диагностируемой системой в момент **t**;

$\tilde{\mathbf{u}}_{t-1}$ - сглаженное значение результата наблюдения в предыдущий момент **(t - 1)**;

α - постоянная сглаживания (весовой коэффициент), $\mathbf{0 < \alpha < 1}$.

С учетом того, что операция сглаживания проводится одинаково для всех наблюдений, то

$$\tilde{\mathbf{u}}_t = \boldsymbol{\alpha} \sum_{i=0}^{n-1} (1-\alpha)^i\, \mathbf{u}_{t-1} + (\mathbf{1} - \boldsymbol{\alpha})^n\, \mathbf{u}_0,\ \mathbf{n=1,2,3\ldots i=}\overline{0, n-1},$$

где n - число наблюдений;

u_0 - значение начального наблюдения.

Отсюда видно, что в сглаженной оценке $\tilde{u}_t$ наибольший вес имеет текущее значение u_t, а для предыдущих наблюдений веса убывают по геометрической прогрессии.

Нахождение оценок коэффициентов прогнозирующего полинома путем применения операции экспоненциального сглаживания можно показать на простом примере, когда наблюдаемый процесс описывается линейной моделью вида

$$\mathbf{u(t)=a}_0\mathbf{+a}_1\mathbf{t} \qquad (3.5)$$

Применение экспоненциального сглаживания к выражению (3.5) даёт

$$\tilde{\mathbf{u}}\mathbf{(t)=a}_0\mathbf{+a}_1\mathbf{t}+\frac{a_1\cdot(1-\alpha)}{\alpha}\mathbf{,} \qquad (3.6)$$

а повторного экспоненциального сглаживания, называемого «двойным сглаживанием», получим

$$\tilde{\mathbf{u}}^{(2)}\mathbf{(t)=a}_0\mathbf{+a}_1\mathbf{t}+\frac{2a_1\cdot(1-\alpha)}{\alpha} \qquad (3.7)$$

Решая совместно уравнения (3.6) и **(3.7)**, можно найти оценку коэффициента, $\mathbf{a}_1$ в линейной модели процесса

$$\mathbf{a}_1\mathbf{=[}\tilde{\mathbf{u}}_t\mathbf{-}\tilde{\mathbf{u}}_t^{(2)}\mathbf{]}\left[\frac{\alpha}{1-\alpha}\right]$$

а точечный прогноз в точке $\mathbf{(t+\Delta t)}$, будет определяться как

$$\mathbf{u(t+\Delta t)=u}_t\mathbf{+a}_1\Delta\mathbf{t}$$

В [18] приведены рекуррентные соотношения, связывающие текущие значения прогнозируемых коэффициентов линейной модели с их предыдущими значениями и линейным поправочным членом. Используя эти соотношения, прогнозируемое на m измерений возможное значение параметра, обозначенное как $\mathbf{u}_{(n+m)}^{n\cdot p}$**,** можно определить следующим образом:

$$\mathbf{u}_{(n)}^{np}\mathbf{=a}_{0(n)}\mathbf{+a}_{1(n)}\mathbf{m,}$$

где величины $\mathbf{a}_{0(n)}$ и $\mathbf{a}_{1(n)}$ вычисляются по рекуррентным зависимостям:

$$\mathbf{a}_{0(n)}\mathbf{=u}_n\mathbf{+(1-\alpha)}^2\mathbf{[u}_{(n)}^{np}\mathbf{-u}_{(n)}\mathbf{],}$$

$$\mathbf{a}_{1(n)}\mathbf{=a}_{1(n-1)}\mathbf{-\alpha}^2\mathbf{[u}_{(n)}^{np}\mathbf{-u}_{(n)}\mathbf{],}$$

В этих соотношениях $\mathbf{u}_{(n)}$ - значение параметра, вычисленное при **(n – m)** измерении,$\mathbf{u}_{(n)}^{np}$, - прогнозированное на один замер значение параметра, полученное при предыдущем прогнозе, т.е. по **«n-1»** измерениям.

При этом начальные условия, необходимые для вычисления величин $\mathbf{a_0}$ и $\mathbf{a_1}$ задаются в следующем виде:

$$\mathbf{a}_{0(n-k)}\mathbf{=u}_{(n-k)}\mathbf{;}$$

$$\mathbf{a}_{1(n-k)}=\mathbf{0};$$

$$\mathbf{u}_{(n)}^{np}=\mathbf{u}(_{\mathbf{n}}\text{-}_{\mathbf{k}}),$$

где $\mathbf{u}(_{\mathbf{n}}\text{-}_{\mathbf{k}})$ - значение параметра, определенное при ранее проведенном измерении с номером **(n-k).** Величина **k** определяет граничное число значений, используемых при прогнозе.

Важную роль в методе экспоненциального сглаживания играет выбор α. В зависимости от величины **α** прогнозные оценки по-разному учитывают влияние исходного ряда наблюдений. Очевидно, что чем больше о; тем больше вклад последних наблюдений в формирование тренда, а влияние начальных условий быстро убывает. При малых **α** прогнозные оценки учитывают большую часть наблюдений и, естественно, уменьшение влияния более «старой» информации происходит медленно.

Зависимость, определяющая вес наблюдения, сделанного в момент времени **(t-n)**, при применении экспоненциального сглаживания будет

выражаться в виде $\boldsymbol{\alpha(1-\alpha)^n}$, т.е. вес такого наблюдения будет пропорционален $\boldsymbol{\alpha(1-\alpha)^n}$. По данным, приведённым в таблице 3.1, для обычно рекомендуемых значений **α** можно сделать заключение о необходимом числе наблюдений, используемых для прогноза.

Таблица 3.1

α	$\alpha(1-\alpha)^n$					
	n=2	n=4	n=8	n=16	n=32	n=64
0,05	0,04	0,0408	0,0332	0,0220	0,00970	0,00190
0,1	0,081	0,0656	0,0430	0,0185	0,00340	0,00001
0,2	0,128	0,0820	0,0336	0,0056	0,00016	0

Дополнительным фактором, влияющим на ошибку прогноза является **величина интервала упреждения m.** Действительно, с увеличением этого интервала дисперсия возрастает. Для линейной модели прогнозируемого процесса дисперсию прогноза (при действии аддитивной помехи) можно вычислить по формуле [18]:

$$\mathbf{D}_{(u_{t+\Delta t})}=\frac{1-\beta}{(1+\beta)^3}[\mathbf{1+4\beta+5\beta^2+2(1+2\beta-3\beta^2)\Delta t+2(1-\beta^2)/\Delta^2 t]D_n},$$

где $\boldsymbol{\beta = 1 - \alpha}$; $\mathbf{D_n}$ - дисперсия помехи.

Используя оценки дисперсии прогноза как доли дисперсии помехи при различных интервалах упреждения **m**, приведенные в таблице 3.2, можно выбрать возможный интервал упреждения, удовлетворяющий требуемым показателям качества при определенных значениях **α**.

Таблица 3.2

α	Длина интервала упреждения m (число шагов)	Ошибка прогноза
α = 0.05	1 10 20 50	$0{,}07D_n$ $0{,}2\ D_n$ $0{,}53D_n$ $2{,}7\ D_n$
α = 0.1	1 10 20 50	$0{,}15\ D_n$ $0{,}84\ D_n$ $4{,}8\ D_n$ $20{,}6\ D_n$
α = 0.2	1 10 20 50	$0{,}32\ D_n$ $0{,}84\ D_n$ $10{,}8\ D_n$ $62{,}5\ D_n$

Заметим, что все вышеизложенное в основном справедливо при стабильности характера протекания процесса на участках наблюдения и упреждения. Общая рекомендация по уменьшению ошибок прогноза путем увеличения объема исходного количества точек на интервале наблюдения, используемых ля формирования трендовой модели, а также уменьшения интервала упреждения играет особую роль в случае нестационарного характера изменения диагностируемых параметров.

При отсутствии гарантий стационарности выбор величины α позволяет в какой-то степени уменьшить ошибку прогноза. Малое значение се обеспечивает в общем случае более высокую точность оценки коэффициентов при неизменной модели процесса, в то время как увеличение **α** будет способствовать росту скорости реакции прогнозирующего полинома на изменение модели.

Возможным путем уменьшения ошибок прогноза при нестационарном характере изменения прогнозируемого параметра является применение методов адаптации для корректировки значений а. При реализации такого подхода наблюдения за диагностируемой системой для вычисления прогнозных значений контролируемых параметров

используется корректировка величины **a**, производимая на основании анализа ошибки прогноза $\mathbf{e(_{n+m})}$, которую можно представить в виде

$$\mathbf{e(_{n+m})} = \mathbf{u}^{\phi}_{(n+m)} - \mathbf{u}^{np}_{(n)}$$

где $\mathbf{u}^{\phi}_{(n+m)}$ - фактическое значение параметра в момент времени, определяемым (n+m)-m измерением, а $\mathbf{u}^{np}_{(n)}$ - прогнозируемое для этого момента времени значение контролируемого параметра. Для минимизации ошибки прогноза $\mathbf{e(_{n+m})}$ можно использовать любой из

известных способов, например, метод наименьших квадратов. Структурная схема, поясняющая адаптивный метод корректировки коэффициента сглаживания а, приведена на рис. **3.1.**

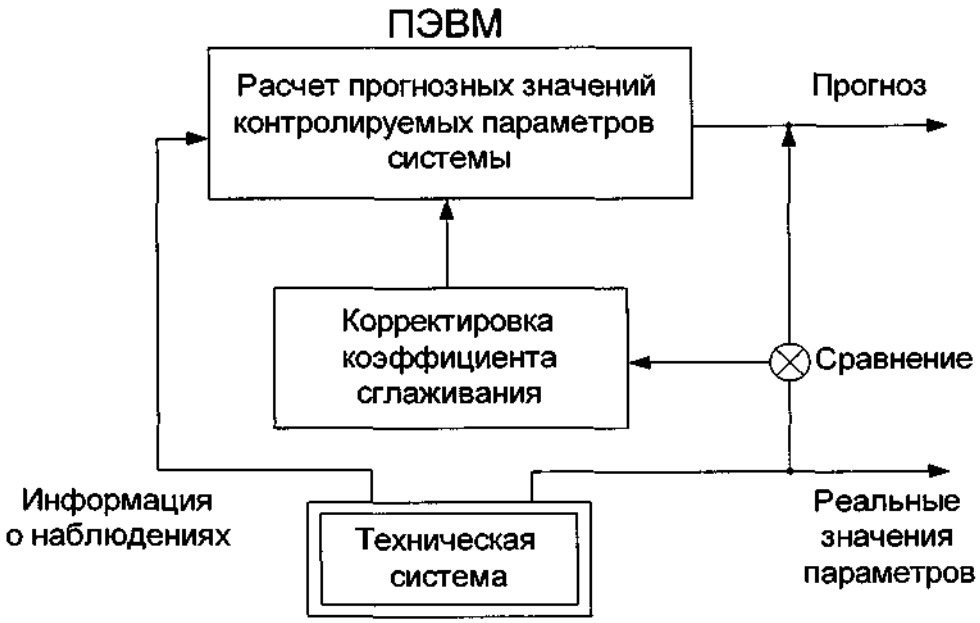


Рис. 3.1.

Использование дополнительной информации о характере изменения прогнозируемого параметра позволяет повысить эффективность прогноза. В частности, информация о системе, полученная в результате испытаний, в процессе ее эксплуатации, наконец, информация об аналогах контролируемой системы позволяет предполагать возможный характер изменений прогнозируемых параметров.

3.3. Ресурсные испытания

Фактором, имеющим исключительно важное значение при эксплуатации технических систем, является их ресурс, оценка которого проводится либо в стендовых условиях (на стадии разработки), либо в эксплуатационных условиях. Последнее, как правило, для систем с длительными сроками активного существования мало подходит, так как информация успевает устареть, теряет ее ценность. В этой связи особое значение приобретают ускоренные методы испытаний.

Методы ускоренных ресурсных испытаний должны обеспечивать связь ускоренных и эксплуатационных режимов работы технических

систем, позволяющих анализировать полученные результаты и переносить их на нормальные эксплуатационные условия. Существующие методы ускоренных испытаний можно разделить на две группы: методы испытаний в форсированном и в нормальном эксплуатационном режиме [13].

Испытания, проводимые в нормальном эксплуатационном режиме, основаны на экстраполяции характеристик случайного процесса появления отказов. Ускорение при этом достигается за счет ограничения времени испытаний до момента появлений отказов только определенной части однотипных технических систем или выявления тенденций и закономерностей изменений контролируемых параметров.

При проведении ускоренных испытаний в форсированном режиме ускорение достигается путем применения более жестких условий функционирования по сравнению с эксплуатационными с целью увеличения скорости протекания всех процессов старения в контролируемой технической системе. Однако следует иметь в виду, что «жесткость» режимов имеет ограничения, превышение которых может приводить к качественным изменениям в испытуемой системе, включая разрушение. Таким образом, главным условием адекватности ускоренных и эксплуатационных испытаний является требование совпадения или незначительного отличия характера и вида, протекающих в испытуемой системе процессов.

Масштабным параметром, связывающим скорости протекания процессов в нормальных условиях и при ускоренных испытаниях, является коэффициент ускорения $\mathbf{K_у}$, под которым понимается следующее соотношение [31]:

$$\mathbf{K}_y = \frac{T_э}{T_y} > \mathbf{1},$$

где **Тэ** - время, затраченное на получение требуемой информации при испытаниях в условиях, аналогичных эксплуатационным;

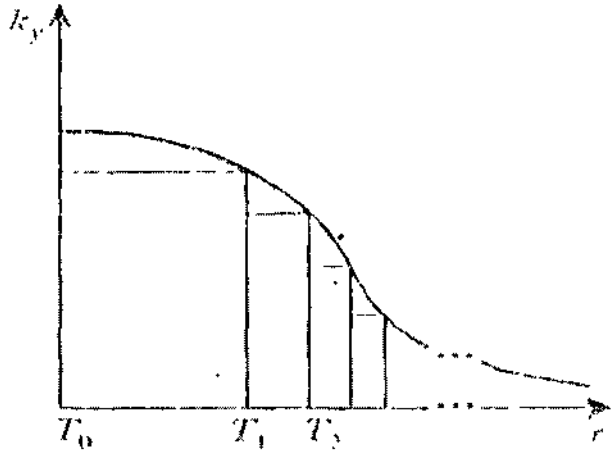


Рис.3.2

Ту - время, в течение которого эта информация получена методом ускоренных испытаний. Следует обратить особое внимание, что коэффициент ускорения в общем случае не может быть постоянным в течение всего интервала испытаний, так как характер протекания процессов, в том числе и их скорость, изменяется в зависимости от времени, прошедшего от начала эксплуатации технической системы.

Для обеспечения эквивалентности процессов, характеризующих изменение состояния технической системы при различных режимах испытаний, коэффициент $\mathbf{K_y}$ приближенно можно определять, используя дискретную аппроксимацию **(рис. 3.2),** при которой весь временной интервал $(\mathbf{T_0}, \mathbf{T_n})$ разбивается на непересекающиеся временные интервалы $(\mathbf{T}_{(i-1)}, \mathbf{T}_i)$, следовательно:

$$(\mathbf{T_0}, \mathbf{T_n}) = \mathop{Y}_{i=1}^{n} (\mathbf{T}_{(i-1)}, \mathbf{T}_i).$$

В этом случае для каждого интервала справедливо равенство [31]:

$$\mathbf{Y}_y = \mathbf{y}_э (\mathbf{K}_{yi}\, \boldsymbol{\tau}), \qquad \textbf{(3.8)}$$

которое, строго говоря, при дискретной аппроксимации будет соблюдаться лишь в узловых точках. В выражении (7.8) $\mathbf{Y_\tau}$ - вектор параметров, характеризующих состояние технической системы; **c(t)** - вектор, характеризующий ускоряющий фактор; **τ** и **t** - время, характеризующее соответственно «медленные» и «быстрые» процессы в испытуемой системе. Выбор вида и параметров ускоряющих воздействий, анализ их влияния на испытуемую систему наиболее удобно производить, если имеются уравнения, описывающие изменения состояния системы во времени.

Ускоряющий фактор может носить детерминистический и стохастический характер, оказывая влияние на протекающие в технических системах процессы путем увеличения интенсивности их изменения. Например, увеличение скоростей движения трущихся элементов технической системы приводит к изменениям интенсивности износа, а также приводит и к увеличению вибраций. В то же время на интенсивность износа трущихся элементов можно повлиять, увеличивая статические нагрузки, но при этом снижается уровень собственных вибраций и, следовательно, фактор, определяющий динамические нагрузки, изменяется в противоположную сторону.

Из приведенных примеров можно видеть, что каждое воздействие, как правило, имеет комплексный характер, изменяющий одновременно интенсивность и «быстрых» (вибрационные колебания), и «медленных» (износ элементов системы) процессов. Теоретическое обоснование ускоренных испытаний, оптимальность их проведения, однозначность анализа полученных результатов возможно, как уже отмечалось выше, при наличии уравнений, описывающих изменение технического состояния. Решить эти задачи оказывается возможным при совместном исследовании этих уравнений при наличии ускоряющего воздействия c(t) и при эксплуатационных условиях **(c(t)=0).**

Однако в практике ресурсных испытаний находят достаточно широкое применение и эмпирические методы определения коэффициента ускорения и других параметров может также осуществляться с использованием моделей технических систем, имитирующих эксплуатационные условия и работу системы при форсирующем воздействии **c(t)** [31]. Но вполне понятно, что такой подход требует создания достаточно точных моделей изменения технического состояния системы.

Глава 4. Основные сведения об идентификации систем

4.1. Модели, типы моделей и их использование

Одним из главных элементов, необходимых для эффективного решения сложных задач, является построение и соответствующее использование модели. Модель - представление объекта или системы в некоторой форме, отличной от формы их реального существования.

Очевидно, что модели могут принимать самую разную форму и записываться с разной степенью математической детализации. Выбор того уровня сложности, который делает модель полезной, определяется планируемым ее использованием.

В повседневной практике при работе с системами пользуются умозрительными (субъективными) моделями, в которых математики нет вообще. Примерами таких моделей могут служить алгоритмы функционирования, правила управления системами и т.д.

Для описания свойств некоторых объектов и систем подходят числовые таблицы и (или) графики. Такие описания обычно называют графическими моделями. Например, линейные системы автоматического управления (САУ) могут быть представлены своими импульсными реакциями, реакциями на единичный скачок или частотными характеристиками. Соответствующие графические представления широко используются при проектировании и исследовании САУ.

В более сложных приложениях используются математические модели, в которых соотношения, описывающие связи между переменными объекта, задаются в виде определенных уравнений. Поэтому такие модели иногда называют аналитическими моделями. Математические модели представляют собой формализованные математические описания, отражающие с требуемой точностью процессы, происходящие в исследуемом объекте. Математические модели могут быть снабжены набором поясняющих прилагательных (линейные, нелинейные, дискретные, непрерывные, детерминированные, стохастические и т.д.) в зависимости от типа исследуемых уравнений.

В процессе машинного моделирования моделью системы является программа для ЭВМ. Программа, которой описывается поведение сложных систем, может представлять собой совокупность взаимодействующих между собой подпрограмм и просмотровых таблиц. Формализация такой совокупности в виде некоторой математической модели может оказаться трудноразрешимой задачей. Такие компьютеризованные представления называют программными (или машинными) моделями. Такие модели в настоящее время играют большую роль в процессе принятия оптимальных решений в сложных системах.

Модели можно классифицировать различными способами. Однако ни один из них не является полностью удовлетворительным, хотя каждый из них служит определенной цели. Укажем некоторые типовые альтернативные группы моделей:
- физические (натурные) и математические (символьные);
- статические и динамические;
- детерминированные и стохастические;
- дискретные и непрерывные;
- линейные и нелинейные;
- сосредоточенные и распределенные;
- стационарные и нестационарные.

Физическими моделями являются модели, в которых свойства реального объекта представляются свойством такого же объекта (макета) или некоторым другим свойством аналогичного по поведению объекта.
К математическим моделям относятся те, в которых для представления процесса используются символы, а не физические устройства.
Математическую модель можно представить в виде множества величин, описывающих процесс функционирования реального объекта:
а) совокупность управляемых входных воздействий на объект

$$u_i \in U, i = \overline{1,n}; \vec{u}(t) = (u_1(t), u_2(t), \ldots, u_n(t));$$

б) совокупность неуправляемых входных воздействий

$$z_i \in Z, i = \overline{1,m}; \vec{z}(t) = (z_1(t), z_2(t), \ldots, z_m(t));$$

в) совокупность внутренних (собственных) параметров объекта

$$x_i \in X, i = \overline{1,r}; \vec{x}(t) = (x_1(t), x_2(t), \ldots, x_r(t));$$

г) совокупность выходных характеристик объекта (переменных состояния)

$$y_i \in Y, i = \overline{1,l}; \vec{y}(t) = (y_1(t), y_2(t), \ldots, u_l(t));$$

Структура моделируемого объекта имеет вид представленный на рис. 4.1

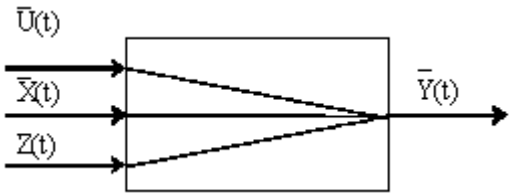


Рис. 4.1

Входные переменные $\vec{u}, \vec{x}, \vec{z}$ являются независимыми (экзогенными), а выходные - зависимыми (эндогенными) переменными.
Процесс функционирования объекта описывается во времени оператором F, который преобразует независимые переменные в зависимые

$$\vec{y}(t) = F(\vec{x}, \vec{u}, \vec{z}, t). \qquad (4.1)$$

Совокупность зависимостей выходных характеристик объекта от времени $y_i(t), i = \overline{1,l}$, называется выходной траекторией $\vec{y}(t)$.

Зависимость (1.1) называется законом функционирования объекта. В общем случае закон функционирования объекта может быть задан в виде функции, функционала, логических условий, в алгоритмической и табличной формах или в виде словесного правила соответствия.

Весьма важным для описания и исследования объекта является **понятие алгоритма функционирования**, под которым понимается метод получения выходных характеристик $\vec{y}(t)$ с учетом входных воздействий $\vec{u}(t), \vec{x}(t), \vec{z}(t)$.

Очевидно, что один и тот же закон функционирования может быть реализован различными способами, т.е. с помощью множества различных алгоритмов функционирования.

Соотношения (1.1) являются математическим описанием поведения объекта моделирования во времени t, т.е. отражают его динамические свойства. Поэтому математические модели такого вида называются ***динамическими***. Они описывают изменения параметров во времени, например:

$$\frac{d\vec{y}(t)}{dt} = f(\vec{u}(t), \vec{x}(t), \vec{z}(t), \vec{y}(t)). \qquad (4.2)$$

Инженеру очень часто приходится сталкиваться с такими моделями при разработке новых технологических процессов, изделий, средств и систем автоматического управления. В сущности, любая задача проектирования, связанная с расчетом потоков энергии или движения тел, в конечном счете сводится к решению дифференциальных уравнений.

Статические модели описывают процессы, не изменяющиеся во времени, т.е. поведение объекта в установившихся режимах

$$\vec{y} = f(\vec{x}, \vec{u}, \vec{z}). \qquad (4.3)$$

Статические модели используют, как правило, при проектной оптимизации объекта.

Обычно динамическая модель задается в виде дифференциальных уравнений, а статическая - в виде алгебраических или трансцендентных.

Модели, у которых существует жесткая связь между переменными, называют ***детерминированными***. Такие модели не содержат случайных факторов и значения выходных переменных однозначно определяются значениями входных переменных.

Стохастическая (вероятностная) модель отражает воздействие случайных факторов. Поэтому между входными и выходными переменными существует не функциональная зависимость (детерминированная модель), а вероятностная. Обычно переменные состояния объекта оцениваются в терминах математического ожидания, а входные воздействия - вероятностными законами распределения.

Непрерывная модель описывает непрерывные изменения переменных объекта в течении определенного промежутка времени, например:

$$\frac{dy(t)}{dt} = f(u(t), x(t), z(t), y(t)), t_j \le t \le t_{j+1}.$$

Дискретная модель описывает зависимость между переменными объекта в дискретные моменты времени, например: где t_j - начало j-ой стадии моделирования объекта; t_{j+1} - ее конец, т.е. состояние объекта в момент времени t_{j+1} определяется по известному его состоянию в момент t_j при условии, что $u(t_j), z(t_j), x(t_j)$ известны и остаются постоянными.

У ***линейной*** модели существует пропорциональная связь между входными и выходными переменными. Модели, не удовлетворяющие этому условию, являются ***нелинейными***.

Динамическая модель, которая описывает изменение переменных объекта только во времени, называется динамической моделью с ***сосредоточенными*** параметрами (искомая величина зависит только от одной переменной).

Эти модели содержат одну или несколько производных от переменных состояния $\vec{y}(t)$ и представляют собой обыкновенные дифференциальные уравнения. Их можно записать в виде:

$$F(t, x(t), u(t), z(t), y(t), y'(t), \ldots, y^{(n)}(t)) = 0. \qquad (4.4)$$

Полная математическая модель наряду с дифференциальным уравнением (1.4) при решении практических задач содержит также некоторые дополнительные условия (например, значения искомых переменных ***y***) в начальный момент времени *t0*, называемыми ***начальными условиями***:

$$y(t_0) = y_0, y'(t_0) = y'_0, \ldots, y^{(n-1)}(t_0) = y_0^{(n-1)}.$$

Во многих практических задачах искомая величина зависит от нескольких переменных. В этом случае математическая модель содержит частные производные и называется ***моделью с распределенными параметрами***.

Если одной из независимых переменных является время t, то такая модель дает описание динамики процесса как во времени, так и в пространстве. Полная математическая модель содержит дифференциальное уравнение в частных производных, начальные условия и граничные условия если математическая модель определена в ограниченном пространстве. Примером такой модели может служить модель теплопроводности или диффузии (параболическое уравнение):

$$\frac{\partial y(t,x)}{\partial t} = a\frac{\partial^2 y(t,x)}{\partial x^2}, \qquad (4.5)$$

где y - параметр состояния (температура или концентрация); t - время; x - пространственная координата (толщина материала); a - константа, при заданных начальных и граничных условиях.

В настоящее время трудно назвать область человеческой деятельности, в которой в той или иной степени не использовались бы модели и методы моделирования. Особенно это относится к сфере управления различными

системами, где основными являются процессы принятия решений на основе получаемой информации.

Идея представления объекта или системы при помощи модели носит столь общий характер, что дать полную классификацию всех функций модели затруднительно. Можно привести, по крайней мере, следующие основания области применения моделей в инженерной практике:

- управление сложными объектами и системами (техническими, экономическими, социальными и т.д.);
- проектирование технических объектов и систем;
- прогнозирование и диагностика с использованием модели объекта;
- создание средств обучения и тренажа;
- постановка численных экспериментов на имитационной модели объекта.

Математическое моделирование является составной частью всех технических и естественно - научных дисциплин. Действительно, основная задача техники заключается в том, чтобы, используя математическую модель, найти хорошее проектно-конструкторское решение, оптимальное управление объектами, наилучшее распределение ресурсов, оптимальный план производства и т.д.

Математические модели являются также мощным инструментальным средством решения задач имитационного моделирования и предсказания (прогнозирования) поведения моделируемых объектов при различных ситуациях, которые часто возникают не только в технике, но и в экономике, экологии, биологии и других областях знания. Модели широко применяются в качестве средств профессиональной подготовки и обучения лиц, которые должны уметь справляться с всевозможными случайностями до возникновения реальной критической ситуации. Широко известны такие применения моделей, как натурные макеты или модели космических летательных аппаратов, используемые для тренировки космонавтов, тренажеры для обучения водителей, деловые игры для обучения персонала, принимающего решения.

Применение моделей позволяет проводить контролируемые эксперименты в ситуациях, когда экспериментирование на реальных объектах практически невозможно или экономически нецелесообразно. При экспериментировании с моделью сложной системы мы часто можем узнать больше о ее внутренних взаимодействующих факторах, чем могли бы узнать, проведя эксперименты с реальной системой. Это становится возможным благодаря наблюдаемости переменных структурных элементов модели, благодаря тому, что мы можем контролировать ее поведение при различных внешних воздействиях, легко изменять ее параметры.

Резюмируя изложенное выше, отметим, что модель может служить для достижения одной из двух основных целей: либо описательной, если модель служит для объяснения и (или) лучшего понимания объекта, либо предписывающей, когда модель позволяет предсказать и (или) воспроизвести характеристики объекта, определяющие его поведение.

4.2. Идентификация как метод построения моделей

Построение моделей опирается в основном на данные наблюдений. Существует два способа (а также их комбинации) формирования математических моделей.

В первом способе исследуемая система расчленяется на такие подсистемы, свойства которых очевидны из ранее накопленного опыта. По существу, это означает, что мы опираемся на известные законы природы и другие надежные соотношения, основанные на ранее проведенных экспериментальных исследованиях. Формальное математическое объединение этих подсистем становится моделью всей системы. Такой подход называется моделированием или аналитическим методом построения моделей. **В его рамках проведение натурного эксперимента не обязательно.** Конкретный вид процедуры моделирования сильно зависит от прикладной задачи и часто определяется традиционными и специфическими средствами из рассматриваемой прикладной области. Основной прием сводится к структуризации процесса в виде блок-схемы, блоки которой состоят из более простых элементов. Процесс восстановления системы по этим простым блокам чаще всего выполняется с помощью ЭВМ и приводит не к математической, **а к машинной модели системы.**

В другом способе построения моделей непосредственно используются экспериментальные данные. В этом случае ведётся регистрация входных и выходных сигналов системы, и модель формируется в результате обработки соответствующих данных. Этот способ называется идентификацией.

Задача идентификации. Задача идентификации формулируется следующим образом: по результатам наблюдений за входными и выходными переменными объекта построить оптимальную в некотором смысле его модель. При этом объект находится в нормальном режиме функционирования (т. е. в обстановке случайных возмущений и помех). Иными словами, если объект описывается некоторым неизвестным оператором F_0, то имея измеренные значения входа и выхода необходимо построить оценку F оператора объекта, оптимальную в смысле некоторого критерия.

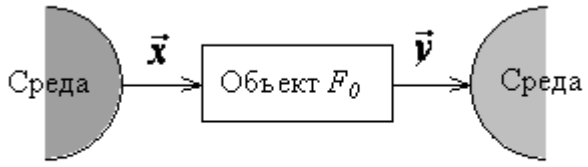


Рис. 1.2

Рис.4.2 иллюстрирует взаимодействие идентифицируемого объекта со средой. Это взаимодействие происходит по каналам $\vec{X}$ и $\vec{Y}$. По каналу $\vec{X}$ среда воздействует на объект, а по каналу $\vec{Y}$ объект

воздействует на среду. Задача идентификации сводится к определению оператора модели F, связывающего вход и выход объекта

$$\vec{Y} = F(\vec{X}) \qquad (4.6)$$

Так как часто отсутствует модель среды, воздействующей на объект, то его вход естественно рассматривать как случайную функцию времени $\vec{X} = \vec{X}(t)$, статистические свойства которой в общем случае неизвестны. Однако известны наблюдения входа и выхода объекта, т.е. реализации функций $\vec{X}(t)$ и $\vec{Y}(t)$. На объект могут воздействовать ненаблюдаемые факторы $\vec{Z}(t)$, которые рассматриваются как случайные помехи.

Пусть $x_1,\dots,x_n$ — наблюдения входа объекта, а $y_1,\dots,y_n$ — соответствующие им наблюдения его выхода в дискретные моменты времени $t_1,t_2,\dots,t_n$. Эти наблюдения связаны неизвестным оператором объекта F_0, т.е. $y_i = F_0(x_i)$ $(i=\overline{1,n})$. Задача идентификации заключается в построении (синтезе) модельного оператора F, т.е. в получении оценки F по наблюдениям x_i, y_i в дискретные моменты времени t_i (в общем случае реализации $x_i(t), y_i(t)$ — непрерывны).

Таким образом, идентификация — это синтез оптимального модельного оператора F исследуемого объекта с использованием результатов наблюдений за его входными и выходными переменными.

Классификация видов идентификации. В соответствии с современной теорией можно предложить следующую классификацию видов идентификации:

1) по конечному результату идентификации:
– структурная;
– параметрическая;
2) по способу изучения объекта идентификации:
– активная;
– пассивная;
3) по типу идентифицируемой модели:
– линейная и нелинейная;
– детерминированная и стохастическая;
– с непрерывным и дискретным временем;
– стационарная и нестационарная;
– одномерная и многомерная;
– статическая и динамическая;
– с сосредоточенными и распределёнными параметрами.

Успех идентификации объекта существенно зависит от соотношения двух факторов: объема априорной информации о структуре объекта и объема измерительной информации. Априорные сведения помогают определить структуру модели, т.е. ее вид (число входов и выходов, характер связи между ними). Эту процедуру называют идентификацией в широком смысле, или структурной идентификацией.

При структурной идентификации объем априорной информации об объекте весьма ограничен. Поэтому необходимо решить следующие задачи :

- выделение объекта из среды;
- задание класса моделей (см. п.3 классификации);
- определение характера связи между входом и выходом модели объекта;
- определение рационального числа информативных переменных (входов и выходов объекта), учитываемых в модели;
- определение возможности представления модели с требуемой точностью в классе линейных операторов и другие.

Структура модели ещё не сама модель, и для определения ее параметров необходимо располагать измерениями. Задачу определения параметров модели по наблюдениям работы объекта при заданной структуре модели называют идентификацией в узком смысле или параметрической **идентификацией.** Например, известна система уравнений, описывающая некоторый объект. Необходимо определить только коэффициенты уравнений.

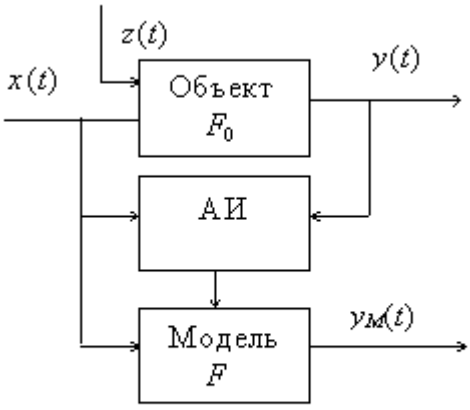


Рис.1.3

При активном способе идентификации реализация входа $\vec{X}(t)$ формируется самим исследователем путем подачи на вход объекта испытательного сигнала желаемой формы (скачкообразного сигнала, импульсного сигнала, сигнала в виде гармонических, прямоугольных, трапецеидальных, треугольных колебаний и др.). Реализацией выхода объекта $\vec{Y}(t)$ является его реакция на испытательный сигнал. При этом в современной теории идентификации широко применяются методы оптимального планирования эксперимента.

При пассивном способе идентификации в качестве реализаций входа $\vec{X}(t)$ и выхода $\vec{Y}(t)$ объекта принимают естественные их изменения в процессе нормального функционирования объекта.

Критерии качества идентификации. Формирование критерия качества, характеризующего адекватность модели реальному объекту, является одним из основных этапов идентификации.

Пусть реальный объект описывается оператором F_0, т.е. $y(t) = F_0(x(t), z(t))$, который нельзя найти точно, но можно сделать его оценку. Применяя некоторый алгоритм идентификации (АИ), необходимо построить модель $y_м(t) = F(x(t))$ с оптимальным оператором F, достаточно близким к F_0.

Но близость операторов непосредственно оценить трудно или просто невозможно, тем более что часто об операторе объекта мало известно. В

связи с этим естественно оценивать близость операторов F_0 и F по их реакциям на одно и то же входное воздействие $x(t)$, т.е. по выходам объекта:

$y(t) = F_0(x(t), z(t))$

и модели $y_M(t) = F(x(t))$, где $z(t)$ — ненаблюдаемое возмущение (рис.4.3)

В общем случае $x(t)$ и $y(t)$ могут быть как детерминированными, так и случайными функциями времени.

Оптимальный оператор F модели ищется в смысле некоторого критерия, связанного с выходами $y(t)$ и $y_M(t)$. С этой целью вводится понятие функции потерь **(функции невязки** $\rho[y(t), y_M(t)]$**)**, которая в любой фиксированный момент времени t зависит от выходов объекта и модели и не зависит от оператора. **Эта скалярная неотрицательная функция векторных аргументов — выходов объекта и модели. В процессе идентификации эта функция минимизируется.**

Для ряда практических задач наиболее естественной, а иногда и единственно возможной является оценка эффективности идентификации по максимально возможному на рабочем отрезке времени $0 \le t \le T < \infty$ отклонению. Тогда проблема идентификации является по существу задачей минимизации максимального отклонения (детерминированный случай):

$$\rho(y, y_M) = \max_{t \in [0,T]} |y(t) - y_M(t)| \to \min \qquad (4.7)$$

При наличии случайных возмущений и шумов, действующих на объект, в качестве критерия выбирается не само отклонение, которое так же является случайным, а его математическое ожидание (стохастический случай) :

$$\rho(y, y_M) = M|y(t) - y_M(t)| \to \min \qquad (4.8)$$

Значение критериев (1.7) и (1.8) едва ли оспоримо, однако в практических расчётах они почти не используются. Это связано с недостаточной разработанностью аналитических приёмов минимизации.

Наиболее часто функция потерь используется в виде квадрата отклонения:

$$\rho(y, y_M) = \max_{t \leftarrow [0,T]} (y(t) - y_M(t))^2 \to \min \quad ; \qquad (4.9)$$

$$\rho(y, y_M) = M[(y(t) - y_M(t))^2 \to \min \quad . \qquad (4.10)$$

В случае наличия непрерывных реализаций входа и выхода объекта в течение интервала времени $[0, T]$:

$$\rho(y, y_M) = \int_0^T [y(t) - y_M(t)]^2 dt \to \min \quad . \qquad (4.11)$$

Если реализации входа $x(t)$ и выхода $y(t)$ объекта получены в дискретные моменты времени $t_i (i = \overline{1,n})$, тогда:

$$\rho(y, y_M) = \sum_{i=1}^{n} \left[y(t_i) - y_M(t_i) \right]^2 \to \min \quad . \qquad (4.12)$$

В отдельных практических задачах автоматического управления в качестве мер сравнения можно принимать различные характеристики (временные, частотные и т.д.) объекта и модели. Критерием идентификации в этом случае является рассогласование этих характеристик. Однако, если модель используется в самонастраивающейся системе автоматического управления, настройка модели по динамическим характеристикам требует наличие измерителей динамических характеристик объекта и модели, что приводит к конструктивному усложнению САУ и уменьшению быстродействия контуров самонастройки. Поэтому приведенные выше критерии, использующие информацию о выходах объекта и модели, более предпочтительны.

4.2.1. Математический аппарат теории идентификации

1. Для идентификации **детерминированных объектов** принимают регулярные функции, связывающие входы и выходы объекта. Это обстоятельство и породило первый подход теории идентификации, который появился в математическом анализе в виде **теории приближения функций многочленами** и ведёт своё начало от трудов Чебышева. Это направление связано с представлением функции в виде разложения по некоторой системе функций (чаще всего по системе полиномов).

Теория приближения имеет две ветви — теорию аппроксимации и теорию интерполяции. Последняя характерна тем, что интерполирующая функция совпадает с исходной в заданном числе точек.

2. Для идентификации **стохастических объектов применяются методы математической статистики**, что даёт начало **теории оценивания**. Основной задачей этой теории является оценка параметров стохастического объекта по наблюдениям в обстановке случайных помех. Другим направлением математической статистики **для целей идентификации статических стохастических объектов стала теория планирования экспериментов**, которая рассматривает активные эксперименты с целью повышения эффективности идентификации.

3. Третьим подходом к решению задач идентификации являются **методы теории систем автоматического управления**. Эта теория породила специальные методы идентификации **динамических объектов управления в режиме нормальной эксплуатации** (т.е. в обстановке случайных возмущений и помех). Именно к этим методам вначале был применён термин “идентификация“.

4.2.1. Основные подходы построению алгоритмов идентификации.

Возможны два подхода к построению алгоритмов идентификации :

1. **В первом подходе** в начале собирается массив данных (реализации входа и выхода объекта на рабочем отрезке времени $[0,T]$) $x_i, y_i, \quad i=\overline{1,n}$ и оценки параметров модели получаются после обработки этого массива (идентификация по массиву или **ретроспективная идентификация**). Место сбора данных и обработки при этом могут быть территориально разобщены.

В другой постановке оценки параметров модели уточняется на каждом шаге поступления новых измерений, т.е. по ходу процесса. Соответствующие алгоритмы и **оценки называются рекуррентными (пошаговыми)**.

В случае, если осуществляется некоторый пошаговый процесс вычисления без поступления новых измерений, такие процедуры будем называть итерационными (их называют также рекурсивными или последовательными приближениями).

Рекуррентные алгоритмы характерны для адаптивных систем управления, когда вследствие дрейфа параметров или недостаточной исходной точности, необходимо уточнение параметров по ходу процесса.

Методы идентификации, реализующие некоторый критерий идентификации, могут быть построены как по рекуррентному, так и по ретроспективному алгоритму.

Алгоритм ретроспективный идентификации с целью оценивания вектора параметров A может быть (для дискретных отсчетов времени $t_i, i=\overline{1,n}$) записан в форме:

$\tilde{A}=\Phi\left(x_0^n, y_0^n\right)$, (1.13)

где (x_0^n, y_0^n) — массив измерений от 0 до n;

$\tilde{A}$ — оценка вектора A, полученная на полной выборке $0,1,\ldots,n$.

Рекуррентный алгоритм записывается в виде

$\tilde{A}_{K+1}=\Phi_1(\tilde{A}_K, x_{K+1}, y_{K+1})$, (1.14)

где $\tilde{A}_K$ — оценка вектора параметров на K-ом шаге ;

x_{K+1}, y_{K+1} — измерения на (K+1)-ом шаге.

4.3. Идентификация структуры модели

Процесс определения структуры оператора модели F составляет задачу структурной идентификации. Если же структура этого оператора определена, то процесс идентификации сводится к определению параметров этой структуры, т.е. к задаче параметрической идентификации, белее простой, чем предыдущая.

Таким образом, идентификация объекта связана прежде всего с предварительным выбором структуры модели. Под структурой модели будем понимать вид оператора F с точностью до его коэффициентов. Заметим, что структура объекта может и не совпадать со структурой модели. Например, стохастические свойства объекта обычно не отражаются в модели (модель выбирается детерминированной). Кроме того модель **может иметь меньше входов и выходов, чем объект**. Это часто делают при малом объеме наблюдений.

Уравнение связи между входными и выходными переменными можно записать в различной форме. Приведем некоторые общие уравнения связи между переменными.

1. Всякий **одномерный статический непрерывный объект** определяется функцией y = F(x). Модель этого объекта можно представить в виде разложения

$$F(x) = \sum_{i=1}^{k} c_i \cdot f_i(x) \tag{4.15}$$

по определенной системе функций $\Phi(x) = \left[f_1(x), f_2(x), \ldots, f_k(x)\right]$.

Здесь **структура модели задается системой функций** $\Phi(x)$ **и числом** k, а ее параметрами являются коэффициенты разложения $c_1, c_2, \ldots, c_k$. Идентификация структуры такого объекта заключается в отыскании удовлетворительной системы функций $\Phi(x)$, а параметрическая идентификация сводится к определению параметров $c_i (i = 1..k)$ при заданной системе функций.

2. Поведение **детерминированного динамического одномерного объекта** удобно описывать оператором

$$F = \frac{B(p)}{A(p)}, \tag{4.16}$$

где $A(p) = 1 + \sum_{i=1}^{m} a_i \cdot p^i;\ B(p) = \sum_{j=0}^{l} b_i \cdot p^j$, $p^r = \frac{d^r}{dt^r}$ - оператор дифференцирования.

Здесь структура оператора модели определяется линейностью оператора и числами m и l. Из физических соображений $l \le m$. Коэффициенты $a_i (i = \overline{1,m})$ и $b_j (j = \overline{0,l})$ являются неструктурными параметрами модели.

Оператор (1.16) эквивалентен обыкновенному дифференциальному уравнению вида

$$y+\sum_{i=1}^{m} a_i \frac{d^i y}{dt^i} = \sum_{j=0}^{l} b_j \frac{d^j x}{dt^j} \qquad (4.17)$$

Непрерывная модель (4.17) преобразуется в форму разностного уравнения, если отсчет времени происходит в дискретные равноотстоящие моменты. При таком отсчете времени непрерывные функции, описывающие поведение переменных, превращаются в решетчатые.

Для дискретного времени в момент времени n разностное уравнение записывается в форме

$$A_1(q^{-1})y[n] = B_1(q^{-1})x[n], \qquad (4.18)$$

$$A_1(q^{-1}) = a_{1m}q^{-m} + a_{1(m-1)}q^{-m+1} + \ldots + 1;$$

$$B_1(q^{-1}) = b_{1l}q^{-l} + b_{1(l-1)}q^{-l+1} + \ldots + b_{10};$$

где $q^{-r}y[n] = y[n-r]$.

При переходе от непрерывной модели (1.17) к дискретной (1.18) порядок уравнения m и l сохраняются, но значения коэффициентов для непрерывной и дискретной форм моделей различны. Если известны коэффициенты в одной из форм и интервал дискретизации, значения коэффициентов в другой форме могут быть вычислены.

В теории автоматического управления широко используются также следующие записи связи между переменными на входе и выходе:

1. **Интеграл свертки (интеграл Дюамеля)**

$$y(t) = \int_{-\infty}^{t} x(\tau)w(t-\tau)d\tau = \int_{-\infty}^{t} w(\tau)x(t-\tau)d\tau, \qquad (4.19)$$

где $x(t), y(t)$ - входная и выходная переменные;

$w(t)$ - импульсная переходная (весовая) функция объекта, т.е. сигнал $y(t)$ на выходе объекта, если входной сигнал $x(t)$ представляет собой дельта-функцию $\delta(t)$.

$$\delta(t) = \begin{cases} 0, t \neq 0, \\ \infty, t = 0. \end{cases}$$

$$\int_{-\infty}^{\infty} \delta(t)dt = 1.$$

Если $x(t) = 0$ при $t < 0$, то уравнение (4.19) можно переписать в форме

$$y(t) = \int_{0}^{t} x(\tau)w(t-\tau)d\tau = \int_{0}^{t} w(\tau)x(t-\tau)d\tau. \qquad (4.20)$$

Для дискретного времени это уравнение записывается в форме

$$y[n] = \sum_{i=0}^{n} x[i]w[n-i] = \sum_{i=0}^{n} w[i]x[n-i]. \qquad (4.21)$$

2. **Запись связи в операторном виде**

$y(p) = w(p)x(p),$

где $x(p), y(p)$ - изображения по Лапласу входного и выходного сигналов;

$w(p)$ - передаточная функция объекта, $w(p) = B(p)/A(p)$.

В более общей форме модель динамики **детерминированного нелинейного непрерывного одномерного объекта** может быть представлена в виде разложения по системе операторов

$$F(x) = \sum_{i=1}^{k} c_i F_i(x),$$

где $F_i(x)(i = \overline{1,k})$ -система нелинейных операторов.

Определение $F_i(x)$ является основной целью структурной идентификации. Отыскание чисел $c_i\,(i = \overline{1,k})$ составляет задачу параметрической идентификации.

Отыскание наиболее подходящей модели из множества моделей-претендентов составляет весьма важную и в тоже время достаточно трудную часть процедуры идентификации. Именно на этом этапе знание формальных свойств моделей необходимо соединить с априорным знанием, инженерным мышлением и интуицией.

При решении задач идентификации, как правило, **существует две возможности выбора структуры модели**. Суть одной из них заключается в том, что после тщательного изучения объекта моделирования на основе законов физики и других достоверных знаний формируется структура модели. Такие **модели** являются **неформальными**, включают неизвестные физические величины, параметрическая идентификация сводится к определению их значений. Другая возможность состоит в том, чтобы без всякого физического обоснования использовать **стандартные формальные модели**. Множество таких моделей, у которых параметры рассматриваются прежде всего как варьируемые средства подстройки моделей к имеющимся данным и не отражают физики процесса называется **черным ящиком**. Множество моделей с настраиваемыми параметрами, допускающими физическую интерпретацию, называются **серыми ящиками.**

Построение по возможности неформальных моделей более предпочтительно, так как они наилучшим образом отражают основные свойства реального объекта. Кроме того, построение моделей на основе анализа физических процессов в объекте позволяет составить уравнение даже на стадии проектирования объекта. Метод предполагает знание тепловых, гидравлических, пневматических, электрических процессов, процессов тепломассопереноса и химических реакций объекта.

Несмотря на большое разнообразие объектов и протекающих в них процессов, можно выделить некоторые общие подходы к составлению моделей объектов. Прежде всего следует расчленить мысленно объект на части и составить уравнение для каждой из них, заменив отброшенные части системы эквивалентными воздействиями.

Рассмотрим пример составления математической модели технологического объекта на основе изучения физических процессов.
Модель динамики резистивной печи: тепловой баланс печи резистивного нагрева складывается из изменения накопленной теплоты Q, теплового потока от нагревателя q_1 и теплового потока теплопотерь q_2 вследствие теплоотдачи и конвекции:

$$\frac{dQ}{dt} = q_1 - q_2 \quad , \qquad (4.24)$$

$$Q = \sum_i c_i m_i \theta \quad , \qquad (4.25)$$

где c_i -удельная теплоемкость материала i - го конструктивного элемента печи;
m_i -масса i -го элемента печи;
θ -температура.

Тепловые потери q_1, q_2 в общем случае определяются формулами:

$$q_1 = \frac{U^2}{R}; \qquad q_2 = K(\theta - \theta_1), \qquad (4.26)$$

где U — напряжение на нагревателе;
R — сопротивление нагревателя;
K — коэффициент теплопередачи;
θ_1 — температура окружающей среды.

Из (1.24), (1.25) и (1.26) получаем уравнение

$$\frac{d\left(\sum_i c_i m_i \theta\right)}{dt} = \frac{U^2}{R} - K\left(\theta - \theta_1\right). \qquad (4.27)$$

Учитывая, что m_i, c_i постоянные величины, из (1.27) находим уравнение

$$\frac{dQ}{dt} = \frac{U^2}{R\sum_i m_i c_i} - \frac{K}{\sum_i m_i c_i}\left(\theta - \theta_1\right) \quad , \qquad (4.28)$$

которое может быть линеаризовано.

Пусть θ_0 -температура заданного режима нагрева, θ_2 -средняя температура окружающей Среды. Тогда

$$\theta = \theta_0 + \Delta\theta; \quad \theta_1 = \theta_2 + \Delta\theta; \quad U = U_0 + \Delta U,$$

где U_0 -постоянное значение напряжения, необходимого для поддержания заданного постоянного значения температуры θ_0.

В установившемся режиме

$$\frac{d\theta}{dt} = 0, \quad \Delta\theta = \Delta\theta_1 = \Delta U = 0.$$

Тогда из (1.28) следует

$$0 = \frac{U_0^2}{R\sum_i m_i c_i} - \frac{K}{\sum_i m_i c_i}(\theta_0 - \theta_2) \quad . \qquad (4.29)$$

Вычитая (1.29) из (1.28), находим

$$\frac{d\Delta\theta}{dt} = \frac{U^2 - U_0^2}{R\sum_i m_i c_i} - \frac{K}{\sum_i m_i c_i}\left[(\theta - \theta_0) - (\theta_1 - \theta_2)\right] \quad .$$

Учитывая, что

$$U^2 - U_0^2 = (U - U_0)\cdot(U + U_0) = \Delta U \cdot (U_0 + \Delta U + U_0) =$$
$$= \Delta U \cdot (2U_0 + \Delta U) = 2U_0 \Delta U + \Delta U^2 ,$$

и пренебрегая величиной ΔU^2, получаем

$$\frac{d\Delta\theta}{dt} = \frac{2U_0}{R\sum_i m_i c_i}\Delta U - \frac{K}{\sum_i m_i c_i}(\Delta\theta - \Delta\theta_1) \quad ;$$

$$\frac{d\Delta\theta}{dt} + \frac{K}{\sum_i m_i c_i}\Delta\theta = \frac{2U_0}{R\sum_i m_i c_i}\Delta U + \frac{K}{\sum_i m_i c_i}\Delta\theta_1 \quad . \qquad (4.30)$$

Обозначим:

$$a_0 = \frac{K}{\sum_i m_i c_i}; \qquad b_0 = \frac{2U_0}{R\sum_i m_i c_i}; \qquad b_1 = \frac{K}{\sum_i m_i c_i}. \qquad (4.31)$$

Тогда получаем следующую модель объекта

$$\frac{d\Delta\theta}{dt} + a_0\Delta\theta = b_0\Delta U + b_1\Delta\theta_1 \quad , \qquad (4.32)$$

где a_0, b_0, b_1-параметры модели, причем, как следует из (4.30), коэффициент b_0 зависит от режима печи (от величины U_0).

При построении модели (4.32) сделано существенное предположение, что температура во всех точках печи одинакова и, следовательно, нагревательная печь является объектом с сосредоточенными параметрами. Определение параметров модели расчетным путем по формулам (4.31) практически не представляется возможным, поскольку характеристики m_i, c_i i-го конструктивного элемента неизвестны, а иногда трудно поддаются к учету. Отметим еще раз, что линеаризованная модель (4.32) справедлива лишь при малых значениях ΔU (когда допустимо пренебрежение величиной ΔU^2). Параметры модели a_0, b_0, b_1 теперь могут быть определены параметрической идентификацией после выбора структуры модели (4.32).

4.4. Методы упрощения моделей

Когда полученная математическая модель является сложной, т.е. неразрешимой, разработчик прибегает к ее упрощению и использованию более глубокой абстракции. В практических задачах исследования процессов функционирования сложных систем часто желателен обратный

процесс — процесс расширения модели. При этом начинают с построения простой модели, а затем усложняют ее. Эволюционный характер процесса конструирования модели упрощает решение поставленной задачи. Сначала решаются более простые задачи с помощью простой модели, а затем ставятся более сложные задачи, что требуют достижения большего соответствия между моделью и реальным объектом, что приводит к усложнению модели.

В обоих случаях возникает необходимость упрощения математических моделей объекта.

Наиболее распространенными являются следующие методы упрощения моделей:

1) расчленение сложной системы на ряд более простых подсистем (декомпозиция);
2) выделение существенных свойств и воздействий и учет остальных в параметрической форме (метод макромоделирования);
3) линеаризация нелинейных процессов в некоторой области изменения переменных;
4) приведение систем с распределенными параметрами к системам с сосредоточенными параметрами (введение более жестких предположений и ограничений);
5) пренебрежение динамическими свойствами процессов.

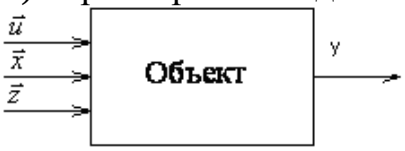


Рис. 4.4

Декомпозиция. В общем случае конечной целью декомпозиции является разбиение пространства переменных объекта (рис.4.4.) $\{y_1, y_2,..., y_n, u_1, u_2,..., u_r, x_1, x_2,..., x_m, z_1, z_2,..., z_l\}$ на q подпространств меньших размерностей, в которых учитывается только связь данного выхода y_i с соответствующими переменными. **Если любой выход имеет связь с остальными выходами, то декомпозиция практически невозможна**. Если общая модель объекта имеет вид неявного выражения достаточно большой размерности

$$F(\vec{y},\vec{u},\vec{x},\vec{z}) = 0 \qquad (4.33)$$

и выходы y_i с объекта не имеют связи между собой, то сложную модель (4.33.) можно представить в виде совокупности эквивалентных ей *n* более **простых частных моделей для каждого из выходов**

$$\vec{y} = \vec{F}(\vec{u},\vec{x},\vec{z}), \qquad (4.34.)$$

где $\vec{F} = (F_1, F_2, \ldots, F_n)$.

Благодаря проведенной декомпозиции системы значительно облегчается задача ее теоретического исследования.

Макромоделирование. При использовании метода макромоделирования в исходном пространстве переменных оставляются (т.е. учитываются) только те из них, которые влияют на выходные переменные наиболее сильно. Остальные неучтенные воздействия могут быть учтены в параметрической форме путем изменения коэффициентов

при учтенных переменных (в случае мультипликативных воздействий) либо путем введения свободных членов (для аддитивных воздействий).

При построении упрощенных моделей с учетом только существенных факторов широко используется **метод адаптивной модели**, т.е. модели, коэффициенты которой подстраиваются таким образом, чтобы некоторая мера расхождения (**невязка) выходов модели и объекта принимала допустимые (минимальные) значения**. Для этого используют **критерии минимизации невязок**. Те переменные, которые стабилизируются и не приводят к изменению выходных переменных, в модели не отражаются. Структура упрощенной модели, называемой макромоделью, может быть трехканальной с каналом управления *u* и каналами контролируемых x и неконтролируемых z воздействий, двухканальной и одноканальной (рис. 4.5.)

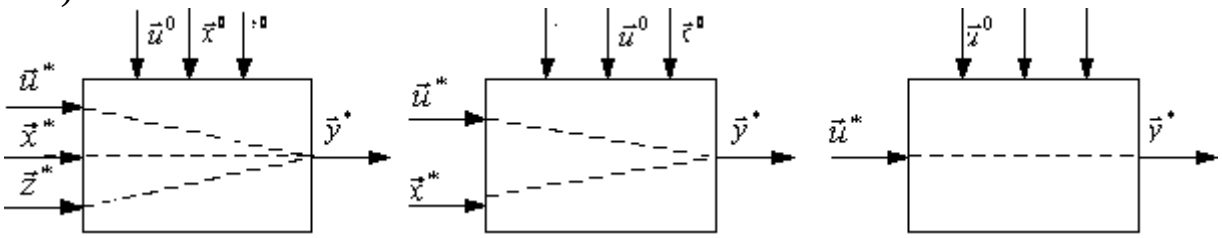


Рис. 4.5

Учет возмущений $\vec{z}$ в двухканальных и $\vec{x}$ и $\vec{z}$ в одноканальных моделях производится параметрически за счет подстройки коэффициентов оставшихся каналов.

Полная математическая модель

$$F(\vec{y},\vec{u},\vec{x},\vec{z}) = 0 . \qquad (4.35)$$

Трехканальная макромодель

$$F(\vec{y}^*,\vec{u}^*,\vec{x}^*,\vec{z}^*,\vec{\alpha}) = 0 . \qquad (4.36)$$

Двухканальная макромодель

$$F(\vec{y}^*,\vec{u}^*,\vec{x}^*,\vec{\alpha}) = 0 . \qquad (4.37)$$

Одноканальная макромодель

$$F(\vec{y}^*,\vec{u}^*,\vec{\alpha}) = 0 , \qquad (4.38)$$

где $\vec{y}^* = (y_1, y_2, \ldots, y_{n1}), \vec{u}^* = (u_1, u_2, \ldots, u_{r1}), \vec{x}^* = (x_1, x_2, \ldots, x_{m1}), \vec{z}^* = (z_1, z_2, \ldots, z_{l1})$ - векторы контролируемых переменных, причем выполняются условия $n_1<n$, $m_1<m$, $r_1<r$, $l_1<l$, свидетельствующие о сокращении числа переменных в макромодели;

$\vec{\alpha} = (\alpha_1, \alpha_2, \ldots)$ - вектор настраиваемых коэффициентов;

$\vec{u}^0, \vec{x}^0, \vec{z}^0$ - векторы неконтролируемых переменных;

Рассмотрим в качестве примера **идею одного из методов адаптации моделей** (компенсационный метод).

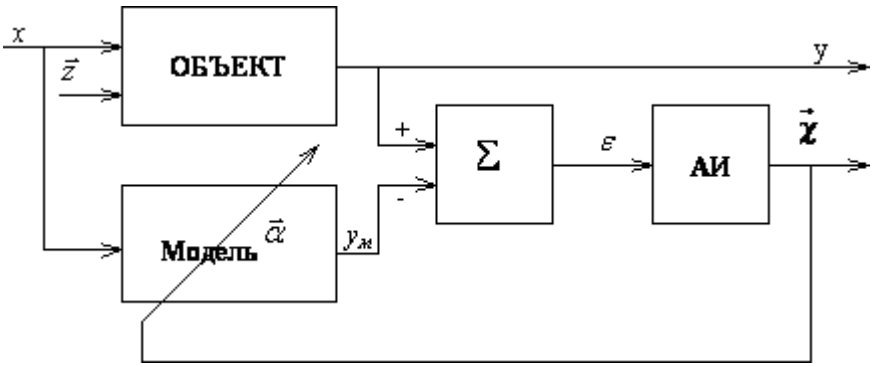


Рис. 4.6

На рис.4.6: e = (y - $y_М$) - сигнал рассогласования выходов объекта и модели;

АИ — алгоритм идентификации.

Алгоритм идентификации позволяет настраивать модель объекта по критерию минимума ошибки ε путем изменения параметров a_1 a_2,... модели (a_i* оптимальное значение параметра a_i).

Приведенная выше схема работает хорошо, если на ее выход подаётся сигнал без помех. При наличии шума на входе ставится задача подавления помех, которую обычно решают с помощью дифференциальных схем, содержащих полосовые фильтры (дифференциальный метод).

Линеаризация. Линеаризация исходной нелинейной модели облегчает решение конкретной задачи исследования. Поэтому для упрощения моделирования и исследования, когда это возможно, желательно заменить нелинейное уравнение приближенным линейным, решение которого с достаточной степенью точности описывает свойство исходной нелинейной системы. **Процесс замены нелинейной модели линейной называется линеаризацией**.

Если дифференциальное уравнение объекта нелинейно из-за нелинейности его статической характеристики, то для линеаризации уравнения необходимо заменить нелинейную статическую характеристику y = F(x) линейной функцией $y = a_0 + a_1x$.

Основное содержание идеи линеаризации состоит в том, что различие в решениях нелинейных уравнений и их линеаризованного представления не столь существенны, чтобы приводить к недопустимым ошибкам в смысле требований к точности решения поставленной задачи.

Для линеаризации нелинейной модели y = F(x) чаще всего применяют общепринятый **метод малых отклонений**. Техника составления линеаризованных уравнений принципиально проста. Математическое обоснование этой процедуры заключается в требованиях к виду нелинейности функции F(x). Для допустимости линеаризации достаточно, что F(x) и $\frac{\partial F}{\partial x}$ существуют и непрерывны в окрестности некоторой точки (x_0, y_0). Тогда линеаризация осуществляется при помощи разложения в ряд Тейлора функции F(x, y) в окрестности точки (x_0, y_0) и отбрасыванием всех нелинейных членов этого ряда

$$y = F(x_0) + \left(\frac{\partial F}{\partial x}\right)_0 (x - x_0)$$

Индекс 0 означает, что производные берутся в точке $x = x_0$, $y = y_0$.

Таким образом, исходная нелинейная модель заменяется линейной моделью вида

$y = a_0 + a_1 x$,

где $a_0 = F(x_0) - \left(\frac{\partial F}{\partial x}\right)_0 x_0$; $a_1 = \left(\frac{\partial F}{\partial x}\right)_0$

Такой способ линеаризации означает замену кривой $y = F(x)$ касательной в точке (x_0, y_0).

В случае многомерной модели, т.е модели вида $y = F(x_1, x_2,..., x_m)$ получим

$$y = F(x_{10}, x_{20}, \ldots, x_{m0}) + \sum_{i=1}^{m}\left(\left(\frac{\partial F}{\partial x}\right)_0 (x_i - x_{i0})\right)$$

или $y = a_0 + \sum_{i=1}^{m} a_i x_i$,

где $a_0 = F(x_{10}, x_{20}, \ldots, x_{m0}) - \sum_{i=1}^{m}\left(\frac{\partial F}{\partial x}\right)_0 x_{i0}$, $a_i = \left(\frac{\partial F}{\partial x}\right)_0$, $i = 1,m$

В этом случае гиперповерхность, описываемая нелинейной функцией $F(\vec{x})$ в пространстве переменных x_1, x_2,..., x_m и y, заменяется гиперплоскостью, касательной к поверхности в точке $(x_{10}, x_{20},..., x_{m0}, y_0)$.

Интуитивно ясно, что линеаризованная модель, полученная при помощи разложения в ряд Тейлора, может оказаться пригодной для описания процессов в нелинейном объекте, не связанных с большими изменениями переменных в окрестности точки $(\vec{x}_0, y_0)$. Ошибка моделирования тем меньше, чем меньше отклонения переменных.

Рассмотрим процесс линеаризации нелинейной модели при помощи разложения в ряд Тейлора на примере объекта, поведение которого описывается нелинейным дифференциальным уравнением общего вида

$$F(y'', y', y, x) = 0. \qquad (4.39)$$

Если x_0, y_0 - некоторое установившееся состояние, то координаты x и y можно записать в виде $x = x_0 + Dx$, $y = y_0 + Dy$,

где Dx и Dy - отклонение координат x и y от установившегося состояния. Уравнение (4.39) в приращениях имеит вид

$$F(Dy'', Dy', y_0 + Dy, x_0 + Dx) = 0. \qquad (4.40)$$

Разложим левую часть уравнения (4.40) в ряд Тейлора относительно точки установившегося состояния $(0, 0, y_0, x_0)$ и пологая Dx, Dy и их производные по времени малыми, отбросим все нелинейные члены этого ряда. При этом получаем уравнение

$$\left(\frac{\partial F}{\partial y''}\right)_0 \Delta y'' + \left(\frac{\partial F}{\partial y'}\right)_0 \Delta y' + \left(\frac{\partial F}{\partial y}\right)_0 \Delta y + \left(\frac{\partial F}{\partial x}\right)_0 \Delta x = 0,$$

которое является линейным дифференциальным с постоянными коэффициентами. Линеаризованное уравнение обычно записывают в следующем виде:

$$a_0Dy'' + a_1Dy' + a_2Dy = b_1Dx. \qquad (4.41)$$

Здесь: $a_0 = \left(\frac{\partial F}{\partial y''}\right)_0; a_1 = \left(\frac{\partial F}{\partial y'}\right)_0; a_0 = \left(\frac{\partial F}{\partial y}\right)_0; b_0 = -\left(\frac{\partial F}{\partial x}\right)_0$.

Необходимым условием линеаризации является разложимость в ряд Тейлора функции F(y”, y’, y, x) в окрестности точки, соответствующей установившемуся состоянию.

Уравнение (4.41) приближенно заменяет уравнение (4.39) лишь в некоторой малой окрестности точки (0, 0, y_0, x_0). Величина этой окрестности зависит от вида функции F(y”, y’, y, x), т.е от величин производных порядка выше первого в рассматриваемой точке. В большинстве случаев при помощи линеаризованной модели (4.41) можно исследовать поведение объекта лишь при малых отклонениях входной и выходной координаты.

Представление исходной нелинейной зависимости $F(\vec{x})$ в виде суммы линейных членов разложения в ряд Тейлора предполагает известным вид этой зависимости. Только в этом случае возможным окажется вычисление величин производных $\frac{\partial F}{\partial \vec{x}}$. Для этого необходимо построение сложной нелинейной модели и подробное выписывание формул в общем виде с целью последующего преобразования, что усложняет процесс идентификации и увеличивают объем вычислительных работ.

Возможна линеаризация нелинейной функции $F(\vec{x})$ при помощи секущей плоскости в многомерном пространстве, описываемой линейным уравнением

$$y = a_0 + a_1x_1 + a_2x_2 +... + a_mx_m,$$

коэффициенты которого определяются методом наименьших квадратов (МНК) так, чтобы получить хорошее приближение исходной и линеаризованной модели в некоторой области возможных изменений переменных.

Применение МНК к вычислению коэффициентов линеаризации существенным образом влияет на процедуру составления линейных уравнений. При таком подходе достаточно задаться структурой линеаризованного уравнения, т.е указать состав переменных, от которых должна зависеть исследуемая величина. Отсюда легко усматривается **преимущество применения МНК для целей линеаризации: далеко не всегда можно дать точный вид функциональной связи между переменными в нелинейном виде. Для статистического подхода этого не требуется,** достаточно знать даже не все, а лишь в основном определяющие переменные для описания свойств реального объекта. Это позволяет отказаться от процедуры построения нелинейных уравнений с последующим их преобразованием, так как можно сразу же отыскивать

линейную связь между требуемыми по условию задачи переменными объекта.

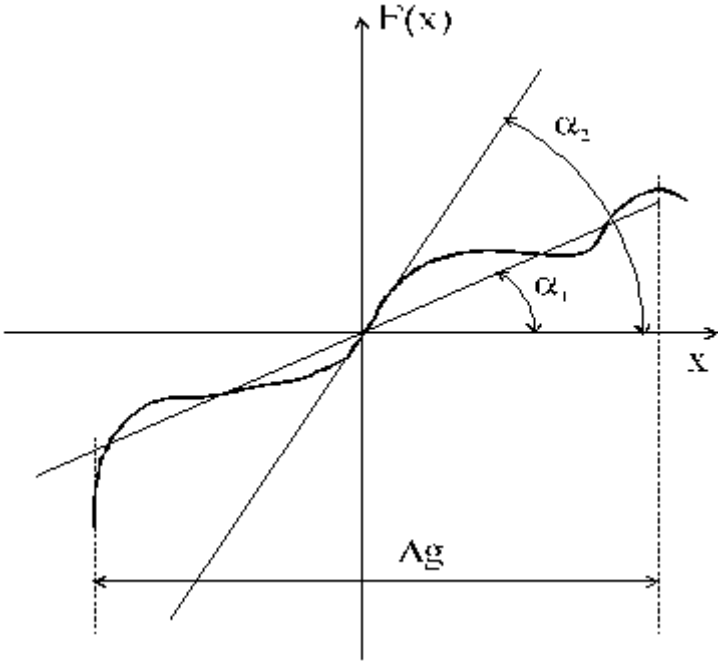


Рис.4.7

На рисунке 4.7 показана геометрическая интерпретация рассматриваемых способов линеаризации, где приняты следующие обозначения: a_1 - угол наклона секущей на интервале ΔD_g (линеаризация МНК); a_2 - угол наклона касательной в точке x = 0 (линеаризация разложением в ряд Тейлора).

Из рис. 4.7 хорошо видно, что при использовании МНК обнаруживается возможность дать оценку допустимого диапазона линеаризации.

Основой для этого служит свойство МНК определять наилучшие значения искомых коэффициентов линейной модели в принятом диапазоне.

4. Упрощение модели с распределенными параметрами. Характеристики состояния объекта могут зависеть не только от времени, но и от пространственных координат. Из множества объектов с распределёнными параметрами можно выделить объекты, параметры которых приводимы к сосредоточенным. **Это такие объекты, для которых достаточно знать значения входных и выходных переменных в конечном числе фиксированных точек пространства**. Например, линейные объекты с распределёнными параметрами структурно могут быть представлены в виде многомерного линейного объекта с сосредоточенными параметрами. Тогда процессы в таких объектах будут описаны совокупностью математических моделей, определяющих изменения **только во времени** исследуемых выходных величин объектов в каждой фиксированной точке пространства.

4.5. Адекватность модели и объекта

Переход от исследования объекта к исследованию модели и подтверждение пригодности модели для решения задач моделирования требует оценки качества полученной модели, т.е. проверку адекватности модели и объекта. Никогда нельзя говорить об абсолютной адекватности, при которой модель по всем свойствам соответствует объекту, так как в зависимости от цели исследования могут строиться различные модели

объекта. Например, при исследовании влияния размещения пассажиров на центровку самолета моделью человека может служить мешок с песком, для конструктора одежды - манекен, а для медико-биологических исследований - животные. Таким образом, всякая модель имеет характер проекции и отражает отдельные свойства объекта. В связи с этим основное подтверждение адекватности модели заключается в том, чтобы убедиться в возможности использования полученной модели для решения той задачи, ради которой эта модель и строилась. Поэтому **адекватность** предполагает воспроизведение моделью с необходимой полнотой всех свойств объекта, существенных для целей данного исследования.

Количественно степень адекватности модели и объекта можно оценить путем сравнения их выходных сигналов при подаче одинаковых входных воздействий на объект и его модель. Такое сравнение предпочтительно на основе новой информации, отличной от того множества данных, которое использовалось в процессе идентификации объекта.

Структурная схема вычисления оценки ошибки модели статического объекта приведена на рис. 4.8.

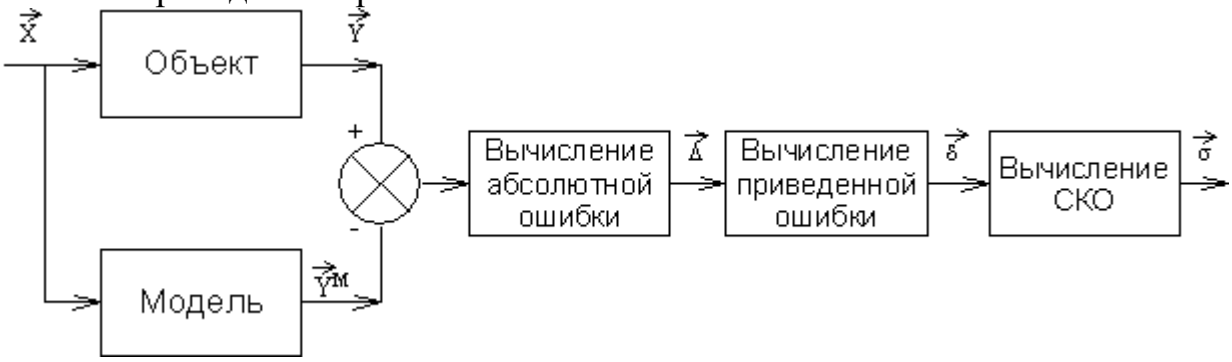


Рис. 4.8.

Пусть проведено l опытов при различных уровнях входных воздействий $\vec{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})$ из области D_x их допустимых значений и получены реализации выходов объекта $\vec{y}_j = (y_{1j}, y_{2j}, \ldots, y_{mj})$ и выходы модели $\vec{y}_j^M = (y_{1j}^M, y_{2j}^M, \ldots, y_{mj}^M)$, $(j = \overline{1,l})$.

Ошибки модели абсолютная $\vec{\Delta} = (\Delta_1, \Delta_2, \ldots \Delta_m)$ и приведенная $\vec{\delta} = (\delta_1, \delta_2, \ldots \delta_m)$ для оценки ее адекватности вычисляются по формулам:

Абсолютная – максимальная ошибка по i-му выходу:

$$\Delta_i = \max_{j=\overline{1,l}} \left| y_{ij} - y_{ij}^M \right|, \quad (i = \overline{1,m});$$

Приведенная ошибка по i-му выходу:

$$\delta_i = \frac{\Delta_i}{\Delta y_i}, \quad (i = \overline{1,m});$$

СКО в серии l опытов:

$$\sigma_i = \sqrt{\frac{1}{l}\sum_{j=1}^{l}(y_{ij} - y_{ij}^{M})^2}, \quad (i = \overline{1,m}),$$

где $\Delta_i, \delta_i, \sigma_i$ - абсолютная, приведенная и среднеквадратичная ошибки модели по i-му выходу $(i = \overline{1,m})$; y_{ij}, y_{ij}^{M} - значение i-го выхода объекта и модели в j-ом опыте $(i = \overline{1,m}, j = \overline{1,l})$; Δy_i - максимальное изменение i-го выхода объекта $(i = \overline{1,m})$ при допустимых значениях входов $x_k\ (k = \overline{1,n})$ из области D_X.

Если величины этих ошибок меньше некоторого заданного положительного числа, то модель адекватна объекту.

Теперь **рассмотрим оценку модели динамического объекта**. Положим, что после идентификации получена модель одномерного объекта в форме линейного дифференциального уравнения вида

$$\sum_{k=0}^{n} a_k \frac{d^k y_M(t)}{dt^k} = \sum_{k=0}^{m} b_k \frac{d^k x_M(t)}{dt^k}, \qquad (4.42)$$

где $x_M(t)$ - входной сигнал модели;

$y_M(t)$ - выходной сигнал модели;

n,m - наивысшие порядки производных $y_M(t)$ и $x_M(t)$ соответственно $(m \le n)$.

Пусть получены реализации входа $x(t)$ и выхода $y(t)$ объекта на отрезке времени $t \in [0,T]$, где T - длина реализации (время наблюдения). Теперь качество модели можно оценить путем сравнения $y_M(t)$ и $y(t)$ либо непосредственно на графике (визуально), либо введя некоторую формальную меру расстояния между этими сигналами.

Выходные сигналы объекта и модели при одном и том же входном сигнале различаются, так как их дифференциальные уравнения и начальные состояния неидентичны. **Для оценки адекватности модели и объекта введем критерий их близости по разности выходных сигналов**, т.е. реакций на один и тот же входной сигнал $x(t)$, например следующего вида

$$I_x = \int_0^T F(y(t) - y_M(t))dt, \qquad (4.43)$$

где $F(\bullet)$ - некоторая выпуклая функция.

В частности

$$F(y(t) - y_M(t)) = (y(t) - y_M(t))^2. \qquad (4.44)$$

В общем случае оценка адекватности проводится для различных форм входного сигнала $x(t)$. Отсюда следует идея необходимости усреднения по входным сигналам и начальным условиям, т.е. введения операции математического ожидания оценки I_x:

$$I = M[I_x] = M[\int_0^T F(y(t) - y_M(t))dt] \quad . \qquad (4.45)$$

Выражение выходного сигнала имеет довольно сложный вид, что затрудняет аналитическое исследование зависимости I от коэффициентов модели, поэтому вводятся и другие критерии. В частности, если **уравнение модели** имеет вид

$$\sum_{k=0}^{n} a_k \frac{d^k y_M(t)}{dt^k} = x_M(t) \quad , \qquad (4.46)$$

то для оценки близости модели и объекта удобным оказывается функционал от разности входных сигналов $(x_M(t) - x(t))$ модели и объекта, обеспечивающих один и тот же выходной сигнал

$$I = M[\int_0^T F(x_M(t) - x(t))dt] \quad , \qquad (4.47)$$

при условии, что $y_m(t) = y(t)$. В этом случае выходной сигнал модели и объекта будем обозначать $y(t)$.

Тогда, подставляя (4.46) в (4.47), имеем

$$I = M[\int_0^T F(\sum_{k=0}^{n} a_k \frac{d^k y(t)}{dt^k} - x(t))dt] \quad , \qquad (4.48)$$

т.е. функционал в явном виде зависит от коэффициентов модели, что удобно для аналитического исследования.

Развивая эту идею, можно формализовать удобный функционал для общего случая модели (1.42)

$$I = M[\int_0^T F(\sum_{k=0}^{n} a_k \frac{d^k y(t)}{dt^k} - \sum_{k=0}^{m} b_k \frac{d^k x(t)}{dt^k})dt] \quad . \qquad (4.49)$$

Выражение

$$\Delta = \sum_{k=0}^{n} a_k \frac{d^k y(t)}{dt^k} - \sum_{k=0}^{m} b_k \frac{d^k x(t)}{dt^k} \quad , \qquad (4.50)$$

называется **обобщенной ошибкой** модели. В качестве функции $F(\bullet)$, как правило, принимают квадрат обобщенной ошибки

$$I = M[\int_0^T \Delta^2(t)dt] \quad . \qquad (4.51)$$

Этот функционал удобен тем, что в явном виде зависит от параметров модели и от доступных измерению входного и выходного сигналов объекта. Однако при вычислении этого функционала возникают определенные трудности, связанные с дифференцированием сигналов $x(t)$ и $y(t)$, а также с необходимостью выполнения операции математического ожидания. Структурная схема вычисления обобщенной ошибки и оценки критерия I представлена на рис. 1.9, где $p = \frac{d}{dt}$ - оператор дифференцирования.

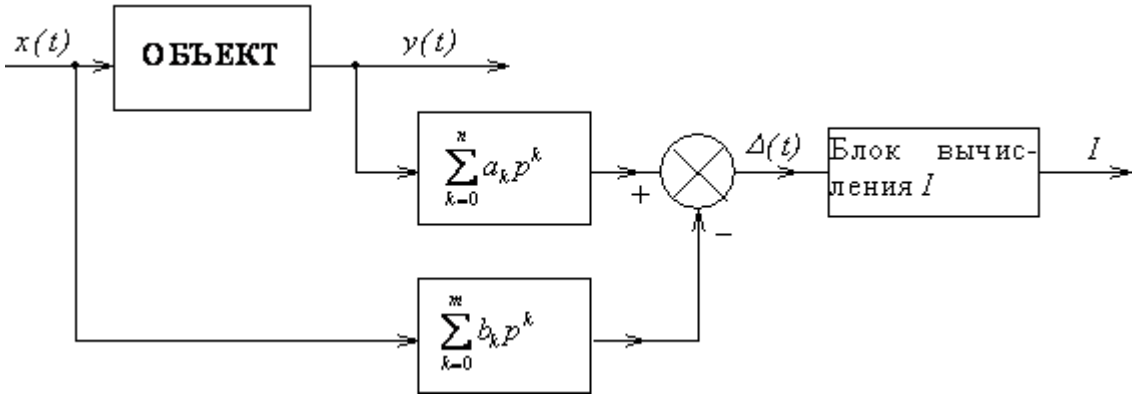


Рис. 4.9

Однако по условиям физической реализуемости можно создавать лишь устройства, порядок числителя которых меньше(или равно) порядка знаменателя, т.е. можно реализовать устройства с операторами

$$\sum_{k=0}^{n} a_k p^k \Big/ D(p) \quad \text{и} \quad \sum_{k=0}^{m} b_k p^k \Big/ D(p),$$

где $D(p)$ - многочлен степени больше или равно n; $m \le n$.

Тогда структурная схема вычисления обобщенной ошибки $\tilde{\Delta}(t)$ и оценки критерия $\tilde{I}$ будет иметь вид, представленной на рис. 4.10.

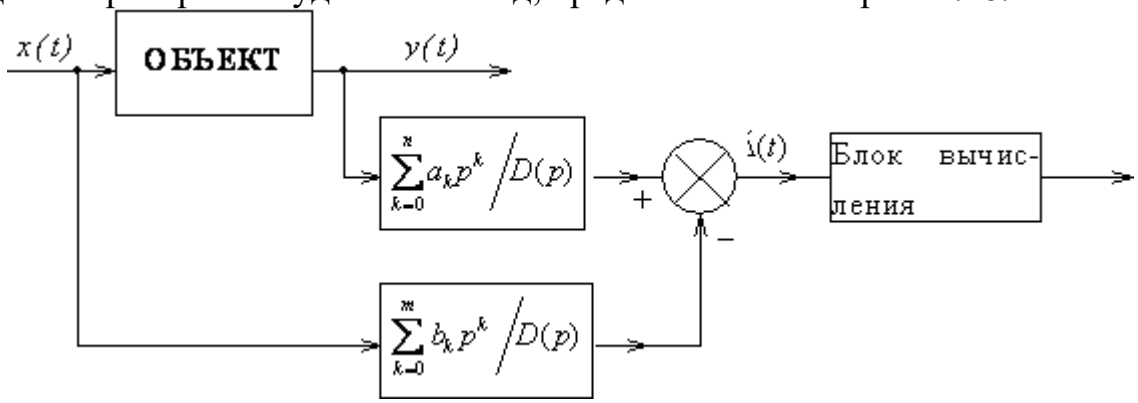


Рис. 4.10

$$\tilde{\Delta}(t) = \Big(\sum_{k=0}^{n} a_k p^k \Big/ D(p)\Big) y(t) - \Big(\sum_{k=0}^{m} b_k p^k \Big/ D(p)\Big) x(t) \quad ; \qquad (4.52)$$

$$\tilde{I} = M\Big[\int_0^T \tilde{\Delta}^2(t)\,dt\Big] \quad . \qquad (4.53)$$

Структурной схеме, изображенной на рис. 4.10, эквивалентна схема, приведенная на рис.4.11.

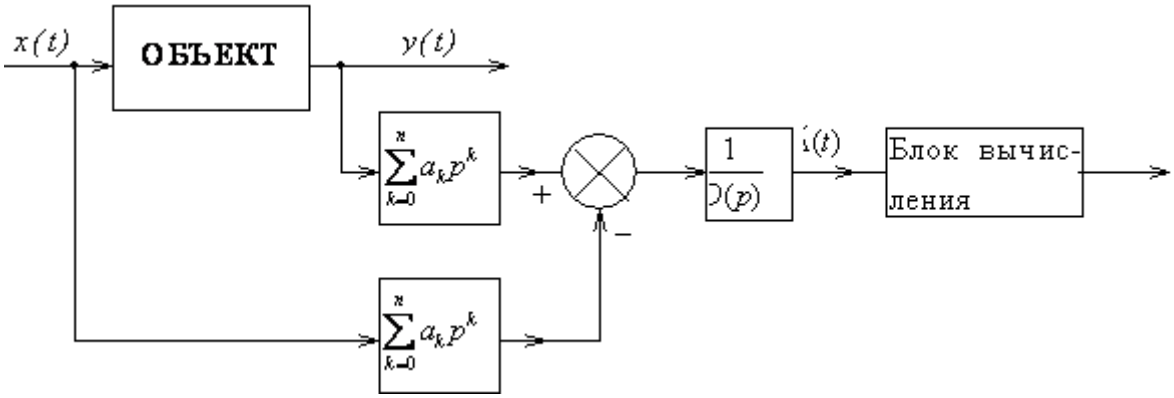


Рис. 4.11

Таким образом, обобщенная ошибка $\tilde{\Delta}(t)$, измеряемая с помощью физически реализуемых устройств, отличается от обобщенной ошибки $\Delta(t)$ тем, что $\tilde{\Delta}(t)$ является результатом преобразования $\Delta(t)$ фильтром с передаточной функцией $1/D(p)$. В силу конечности полосы пропускания этого фильтра происходят искажения сигнала обобщенной ошибки. Эти искажения будут тем меньше, чем больше полоса пропускания фильтра.

Если величина ошибок модели и оценок критериев приближения удовлетворяют требованиям к качеству модели, то модель считается адекватной объекту и может быть использована для решения задач моделирования, оптимизации и управления. В противном случае модель необходимо усовершенствовать путем изменения структуры и введения в нее неучтенных ранее факторов.

4.6. Характеристика проблем и методов моделирования объектов

Под моделированием какого-либо объекта (системы, явления) обычно понимается воспроизведение и исследование другого объекта, подобного оригиналу, в форме, удобной для исследования, и перенос полученных результатов на моделируемый объект. При этом объекты считаются подобными, если характеристики происходящих в них процессов отличаются вполне определенными и постоянными в течение данного процесса коэффициентами. Модель изучаемого объекта при этом может быть иной физической природы, отличной от природы оригинала. Это обстоятельство открывает широкие возможности для применения моделирования в проектировании объектов, обусловленные применением таких эффективных средств, как ЭВМ, динамические стенды, имитаторы, тренажеры.

В основе моделирования лежат информационные процессы, поскольку в процессе реализации модели получают информацию о данном объекте, одновременно в эксперимент с моделью вводится управляющая

информация, далее происходит обработка полученных результатов, т.е. информация лежит в основе всего процесса моделирования.

В настоящее время получили применение системы моделирования, отличающиеся высокой степенью автоматизации процесса моделирования, когда наряду с программными средствами управления вычислительным экспериментом используется возможность диалогового общения исследователя с процессом моделирования. Степень автоматизации процесса моделирования тесно связана с управляемостью модели. В этом смысле наличие многих управляемых параметров и переменных модели в реализованной системе моделирования дает возможность поставить широкий эксперимент и получить обширный спектр результатов.

Характеризуя проблему моделирования в целом, необходимо учитывать, что от постановки задачи моделирования до интерпретации полученных результатов существует большая группа сложных научно-технических проблем. К основным из них можно отнести:
- идентификацию реальных объектов;
- выбор вида моделей;
- получение реализаций переменных объекта;
- построение моделей и их реализацию на ЭВМ;
- взаимодействие исследователя с моделью в ходе вычислительного эксперимента;
- проверку правильности полученных в процессе моделирования результатов и выявление основных закономерностей.

В зависимости от объекта моделирования и вида используемой модели эти проблемы могут иметь разную значимость.

В одних случаях наиболее сложной оказывается идентификация, в других — проблема реализации модели на ЭВМ, особенно в случае имитационного моделирования сложных систем.

Для исследования объектов и систем управления используются различные **методы моделирования**:
- физическое моделирование;
- математическое моделирование на ЭВМ;
- полунатурное моделирование.

Каждый из этих методов имеет свои достоинства и недостатки. Применение того или иного метода определяется в каждом конкретном случае в зависимости от исследуемой системы и условий ее работы. При этом необходимо иметь набор правил и условий, выполнение которых обеспечивает требуемую точность изучения реального объекта по его модели. Эти правила и условия формулируются в теории подобия.

При использовании **физического моделирования** процессы, протекающие в модели и в оригинале, имеют одинаковую иногда и различную физическую природу. Физическая модель может отличаться от оригинала размерами, скоростью протекания и физической природой процессов или материалами, из которых она изготовлена. Примерами такого моделирования может служить изучение аэродинамических свойств

самолета при помощи его макета в аэродинамической трубе, изучение механических колебаний с помощью электрической схемы.

Физическая модель может наиболее полно воспроизводить такие свойства оригинала, которые при теоретическом изучении не могут быть учтены в полной мере. Теоретической базой физического моделирования является теория подобия.

Теория подобия позволяет пересчитать количественные характеристики, полученные при изучении модели, в количественные характеристики оригинала.

При физическом моделировании необходимо для каждого конкретного объекта создать свою модель. Так как изготовление сложных объектов обходится, как правило, очень дорого, требует больших материальных и трудовых затрат, то изменение параметров модели часто нецелесообразно. Поэтому, несмотря на определенные преимущества, физическое моделирование находит ограниченное применение при исследовании сложных объектов и систем управления. Для этой цели широко применяется математическое моделирование.

При **математическом моделировании** исследование объекта осуществляется посредством модели, сформулированной на языке математики и реализованной, как правило, на ЭВМ с использованием тех или иных математических методов.

При моделировании на ЭВМ в качестве объекта моделирования (оригинала) выступает исходные уравнения, представляющие математическую модель реального объекта, а в качестве модели — процессы, протекающие в соответствии с этими уравнениями и воспроизводимые на ЭВМ в виде “машинных решений” путем реализации программ.

Математическое моделирование объектов и систем управления осуществляется на АВМ и ЦВМ, поэтому часто такие способы называют аналоговым и цифровым моделированием.

В математическом моделировании выделяют **имитационное моделирование**, под которым понимается воспроизведение объектов с имитацией случайными величинами и случайными процессами элементов оригинала, которые не удается представить определенными математическими моделями. При этом процесс функционирования объекта (системы) представляется в виде определенного алгоритма, который реализуется на ЭВМ.

На некоторых тактах работы имитационной модели используются параметры, выбираемые исследователем, так называемые управляющие воздействия. Выбор управляющих воздействий осуществляется из некоторого множества и обычно имеет критерий качества этого выбора, т.е. функцию, которую следует оптимизировать. Тогда перед тем как вводить управляющие воздействия в имитационную модель, решается оптимизационная задача по их отысканию, и лишь после этого найденные оптимальные значения вводятся в имитационную модель. В этом случае

имитация позволяет моделировать отклик системы на оптимальные управления ею. Если множество управляющих воздействий не слишком богато, то все они могут быть перепробованы в имитационной системе. Результаты имитационного моделирования в таком случае позволит провести оценку управляющих воздействий — отбросить заведомо плохие, упорядочить и т.п.

Таким образом, имитационное моделирование рассматривают так же, как управляемый эксперимент, производимый на ЭВМ. В таком эксперименте определенные математическими моделями части объекта моделирования взаимодействуют с имитирующими возмущающие воздействия и некоторые части объекта генераторами случайных величин. Это взаимодействие проводится по определенным в эксперименте правилам, а результаты моделирования подвергаются статистической обработке.

Результаты имитационных экспериментов могут влиять на вид модели после завершения серии прогонов ее на ЭВМ (исключение малозначимых параметров и входов и т.д.).

Полунатурное моделирование — это моделирование с реальной аппаратурой, при котором часть системы моделируется, а остальная часть является реальной.

Применение такого метода моделирования становится необходимым в тех случаях, когда не удается описать работу некоторых элементов системы математически. Например, при исследовании радиолокационных (РЛ) САУ летательных аппаратов в условиях помех процесс приема и передачи антенной отраженного сигнала трудно поддается математическому описанию. Поэтому приходится РЛ и другую аппаратуру, не поддающуюся математическому описанию, воспроизводить в натуре, а остальные звенья САУ — математически на ЭВМ. Естественно работа всей замкнутой схемы моделирования должна проходит в натуральном масштабе времени. Такое моделирование системы управления позволяет анализировать как всю систему в целом, так и отдельные ее элементы, исследовать влияния кинематики и динамики элементов реальной аппаратуры на работу САУ, несмотря на то, что их математическое описание отсутствует.

В процессе полунатурного моделирования удачно сочетаются достоинства математического и натурного моделирования и может быть достигнуто оптимальное взаимодействие между вычислительными и натурными экспериментами. В настоящее время методы полунатурного моделирования эффективно применяют при проектировании разнообразных автоматических управляющих систем.

Необходимость применения таких методов возникает, если:

- объект управления находится еще в стадии проектирования, а элементы устройства управления реально существуют;
- объект не может испытываться в лабораторных условиях;

- элементы устройства управления имеют нелинейные характеристики, трения, помехи, которые при составлении их уравнений не учитывались;
- проведение натурных экспериментов по настройке устройства управления на объекте дорого или вообще недопустимо.

Для проведения полунатурного моделирования необходимо иметь сопрягающие устройства для связи модели (ЭВМ) с внешними реальными элементами.

Структура полунатурного моделирования системы управления представлена на рис. 4.12. Система включает также модель генератора возмущающих воздействий для выполнения имитационных экспериментов и устройство сопряжения с объектом (УСО), т.е. преобразующие элементы для связи модели объекта с реальной аппаратурой. Включение в схему моделирования дополнительных преобразующих элементов искажает результаты моделирования реального процесса. Поэтому основные свойства этих элементов должны быть учтены при обработке результатов моделирования или же при организации самого процесса моделирования. Например, если преобразующие элементы инерционны, то можно хотя бы частично скомпенсировать инерционность введением в модель объекта форсирующих звеньев.

Цель моделирования при проектировании САУ — выбор структуры и параметров отдельных устройств автоматической системы, отвечающих требованию выполняемых ею функций и заданным показателям качества ее работы. При выбранной структуре автоматической системы моделированием определяется лишь оптимальные значения параметров устройства управления, в наибольшей степени влияющих на заданные показатели качества.

Моделирование применяется также для исследования спроектированной автоматической системы с целью выяснения ее надежности, устойчивости, чувствительности и поведения при отказе тех или иных элементов, при различных (в том числе аварийных) возмущающих и задающих воздействиях.

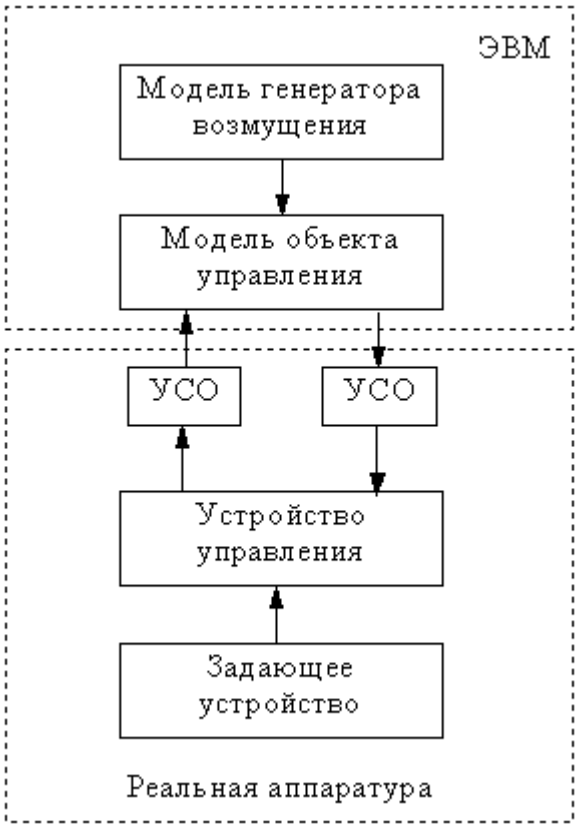


Рис. 4.12

При изучении процессов в реальных системах и их математических моделях различают три типа времени:

1**. Реальное**, во время которого протекают процессы в реальных условиях и системах;

2. **Модельное (системное)**, во время которого протекают процессы в модели системы. Использование модельного времени позволяет анализировать процессы в моделях в ускоренном или замедленном темпе, а также останавливать решение для изучения процессов в какой-то момент времени;

3. **Машинное,** во время которого происходит моделирование системы на ЭВМ, оно регистрируется таймером ЭВМ.

4.7. Этапы и принципы моделирования объектов и систем на ЭВМ

Моделирование с использованием ЭВМ (имитационное моделирование) является эффективным средством исследования сложных объектов и систем.

Разработка имитационной модели реальных объектов на ЭВМ и ее исследование осуществляется поэтапно в определенной последовательности. Традиционно такой процесс включает следующие этапы:

1) определение цели и постановка задачи исследования;

2) формализация объекта (системы) с целью построения его математической модели;

3) построение математической модели объекта;

4) разработка и составление моделирующего алгоритма;

5) построение машинной (имитационной) модели объекта (перевод математической модели объекта и моделирующего алгоритма в моделирующую программу и перенесение ее на машинные носители);

6) планирование имитационных экспериментов;

7) имитационные эксперименты с моделью объекта;

8) обработка и анализ результатов имитационных экспериментов;

9) документирование и практическое использование результатов моделирования.

Структурная схема имитационного моделирования изображена на рис. 4.13. В этой схеме этапы построения модели и проведения имитационных экспериментов разделены. Это обусловлено тем, что машинная модель при планировании и проведении эксперимента рассматривается как черный ящик.

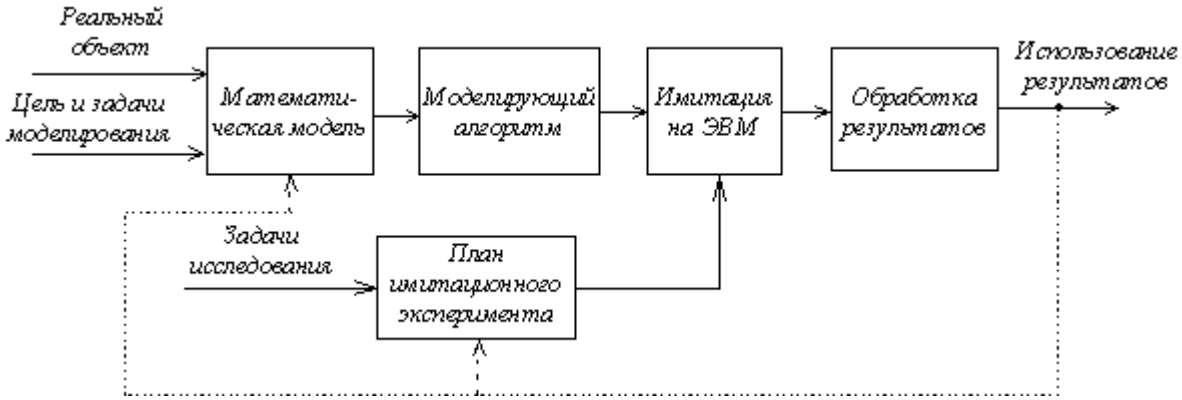


Рис. 4.13

Результаты имитационных экспериментов могут влиять на вид модели объекта и, следовательно, на структуру моделирующего алгоритма. Например, если в процессе эксперимента выяснится, что выходные результаты слабо зависят от того или иного параметра, то это может послужить причиной упрощения модели (исключение данного параметра и соответствующее уменьшение размерности модели) и изменения моделирующего алгоритма.

В зависимости от степени формализации исследуемой системы различают:

- моделирование с применением численных методов;
- вероятностное (статистическое) моделирование с использованием специальных алгоритмических языков программирования.

Моделирование с применением численных методов осуществляют в тех случаях, когда систему удается описать достаточно строгими математическими соотношениями. Его недостаток заключается в том, что структура моделирующего алгоритма и структура информации определяются численным методом, а не структурой и свойствами исследуемого процесса. Изменение требований к режиму моделирования,

корректировка модели объекта, повышение требований к точности результатов могут привести к замене численного метода, а следовательно, к замене алгоритма.

Применение метода статистического моделирования делает моделирующий алгоритм по структуре близким к алгоритму функционирования реального объекта. Изменение условий моделирования не приводит к существенным изменениям алгоритма, который может просто дополняться новыми блоками или проигрываться большее число раз на ЭВМ.

Традиционная схема статистического моделирования включает три блока: блок имитации случайных процессов, действующих на систему; блок программы функционирования системы; блок статистической обработки результатов моделирования.

Статистические имитационные модели являются программным воспроизведением структуры реальной системы и тех элементарных действий, которые выполняют ее отдельные элементы. Недостаток статистического метода — необходимость большого количества машинного времени для его реализации.

Моделирующий алгоритм. Для моделирования процесса на ЭВМ необходимо преобразовать его математическую модель в специальный моделирующий алгоритм. Он представляет собой последовательность некоторых операций, выполняемых ЭВМ.

Запись алгоритма для моделирования сложного процесса сразу в виде программы, как правило, представляет значительные трудности. Кроме того, эта запись оказывается весьма неудобной. В самом деле, нагромождение всевозможных деталей, связанных с организацией вычислительной процедуры, делает ее мало обозримой и затрудняет ориентировку в структуре моделирующего алгоритма. Заметим, что построение моделирующего алгоритма является таким этапом исследования процесса, когда уже решены все принципиальные вопросы создания математического аппарата для исследования. Программирование задачи и решение ее на машине относятся скорее к технической части работы. Поэтому желательно сделать запись алгоритма независимой от типа ЭВМ и языка программирования.

Как правило, моделирующие алгоритмы представляются в графической форме или в виде операторной схемы. Графической формой представления моделирующих алгоритмов являются структурные схемы.

Вся совокупность операторов, составляющих моделирующий алгоритм, может быть разделена на три группы: основные, вспомогательные и служебные операторы. Основные и вспомогательные операторы составляют главную часть моделирующего алгоритма.

Основные операторы реализуют соотношения математической модели, описывающие процессы функционирования реальных элементов системы. В отличие от них вспомогательные операторы не предназначены для имитации реального процесса. Они производят вычисление тех

параметров и характеристик, которые необходимы для работы основных операторов.

Служебные операторы не связаны с соотношениями математической модели, а обеспечивают взаимодействие основных и вспомогательных операторов при моделировании процесса в автоматическом режиме, синхронизацию работы моделирующего алгоритма, фиксацию результатов и их обработку.

Каждый из рассмотренных операторов должен иметь достаточно ясный, наглядный физический смысл, связанный с природой моделируемого объекта. При этом он должен быть представим последовательностью операций.

Операторы обозначают заглавными буквами с индексами, указывающими номер оператора. Применяется сквозная нумерация операторов, входящих в алгоритм.

Для изображения операторных схем алгоритмов удобно пользоваться операторами следующих типов:

1. **Вычислительные операторы** (A_i). Они описывают сколь угодно сложную вычислительную процедуру и должны удовлетворять следующим требованиям:

— управление от операторов после выполнения предусмотренных им операций должно передаваться только одному оператору независимо от результатов;

— все исходные данные. получаемые от других операторов алгоритма, к моменту начала их работы должны быть известны.

2. **Операторы формирования реализаций случайных процессов** (Φ_i). Они предназначены для имитации различных случайных факторов (случайных событий, величин, векторов, функций), сопровождающих исследуемый процесс. Исходным материалом для формирования реализаций случайных объектов на ЭВМ служат случайные числа, равномерно распределенные в интервале (0,1). Преобразование этих чисел в реализации случайных объектов с заданными свойствами осуществляет оператор Φ_i.

3. **Операторы формирования неслучайных величин** (F_i). При моделировании сложных систем приходится формировать не только реализации случайных процессов, но и различные константы и неслучайные функции. Сущность операций при этом принципиально не отличается от сущности обычных вычислительных операций, выполняемых оператором A_i. Однако, операторы A_i, как правило, являются вспомогательными, а F_i — основными. Поэтому выделение операторов F_i в отдельный класс имеет смысл, поскольку они имитируют работу некоторых узлов сложных систем (чаще всего некоторого вычислительного и управляющего устройства системы).

4. **Счетчики** (K_i). Они относятся к классу служебных операторов по назначению и к арифметическим — по выполняемым функциям. Предназначены для подсчета количества различных объектов, обладающих заданными свойствами (количества циклов производственного процесса, количества деталей после определенной стадии обработки, количества бракованных изделий и т.д.). Результаты, выдаваемые счетчиками, часто являются исходными данными для логических служебных операторов, обеспечивающих синхронизацию моделирующего алгоритма. Содержимое счетчиков при моделировании может использоваться для получения вероятностных оценок исследуемых величин.

5. **Логические операторы** (P_i). Предназначены для проверки справедливости заданных условий и выработки признаков, обозначающих результат проверки. После реализации логического оператора управление передается одному из двух операторов алгоритма в зависимости от результата проверки.

6. **Оператор окончания вычислений** (Я). Часто с оператором Я совмещают некоторые другие операции (выдачу результатов, переход к другому варианту задачи и т.д.).

При более детальном рассмотрении операторов моделирующего алгоритма могут быть выделены и другие типы операторов, обладающие некоторыми общими функциональными свойствами.

Раскроем смысловое содержание операторов, используя примеры:

A_{15} — арифметический (вычислительный) оператор № 15;

P_9 — логический оператор № 9;

A_{15}^{10} — оператор A_{15} после выполнения своей операции передает управление оператору с номером 10;

${}^{4}A_{16}^{30}$ — оператор A_{16} получает управление от оператора № 4, выполняет операцию и передает управление оператору № 30;

${}^{5,9}A_{48}$ — оператору A_{48} управление передается от операторов № 5 и № 9;

${}^{5}P_{35\downarrow 12}^{\uparrow 10}$ — оператор P_{35} передает управление оператору № 10, если условие выполнено, или же оператору № 12, если оно не выполнено, получает управление от оператора № 5.

Если данный оператор получает управление от предшествующего оператора, или передает управление оператору, следующему непосредственно за ним, то верхний индекс опускается, например $A_5^6 \equiv A_5$; $A_{20}P_{21\downarrow 30}^{\uparrow 4}$; $A_{10}P_{11\downarrow 4}A_{12}$. Индекс сохраняется, если последующий оператор получает управление не только от предшествующего ему оператора, но и от другого (то же и при передаче управления), например ${}^{4,9}A_{10}$; $A_{10}A_{11}{}^{2,11}P_{12\downarrow 7}A_{13}$; ${}^{2}A_5P_6^{\uparrow 10}F_7$.

В качестве примера построения моделирующего алгоритма рассмотрим простейший вычислительный процесс табулирования функции $y(x)$

$$y = \begin{cases} f_1(x), \ x < 0; \\ f_2(x), \ 0 \le x \le a. \end{cases} \qquad (4.54)$$

Для того, чтобы показать структуру моделирующего алгоритма, соответствующего математической модели (1.54), рассмотрим следующие операторы:

F_1— формирование очередного значения независимой переменной $x_k = x_{k-1} + \Delta x$ (Δx — шаг табуляции), подстановка $x = x_k$;

P_2— проверка условия $x < a$;

P_3— проверка условия $x \ge 0$;

A_4 — вычисление $y = f_1(x)$;

A_5— вычисление $y = f_2(x)$;

F_6— формирование таблицы значений x_k, y_k и переход к следующему шагу вычислений;

$Я_7$— конец вычислений и выдача результатов.

Имея необходимый набор операторов, можно записать операторную схему моделирующего алгоритма в следующем виде:

$^{6}F_1 P_{2\downarrow 7} P_3^{\uparrow 5} A_4^{6} A_5 F_6^{12} Я_7$.

На рис. 4.14 представлена структурная схема этого алгоритма, позволяющая получить наглядное представление о его работе.

На различных этапах моделирования составляются укрупненные и детальные структурные схемы моделирования, а также схемы программ.

Укрупненная схема моделирующего алгоритма задает общий порядок действий при моделировании системы без каких-либо уточняющих деталей, показывает, что необходимо выполнить на очередном шаге моделирования. Детальная схема моделирующего алгоритма содержит уточнения, отсутствующие в обобщенной схеме, и показывает не только, что следует выполнить на очередном шаге моделирования системы, но и как это выполнить.

Структурная схема программы отображает порядок программной реализации моделирующего алгоритма с использованием конкретного математического обеспечения. Схема программы представляет собой интерпретацию логической схемы моделирующего алгоритма разработчиком программы на базе конкретного алгоритмического языка и отражает логику машинной реализации модели с использованием конкретных программно-технических средств моделирования. Схема программы также может быть выполнена как в укрупненной, так и в детальной форме.

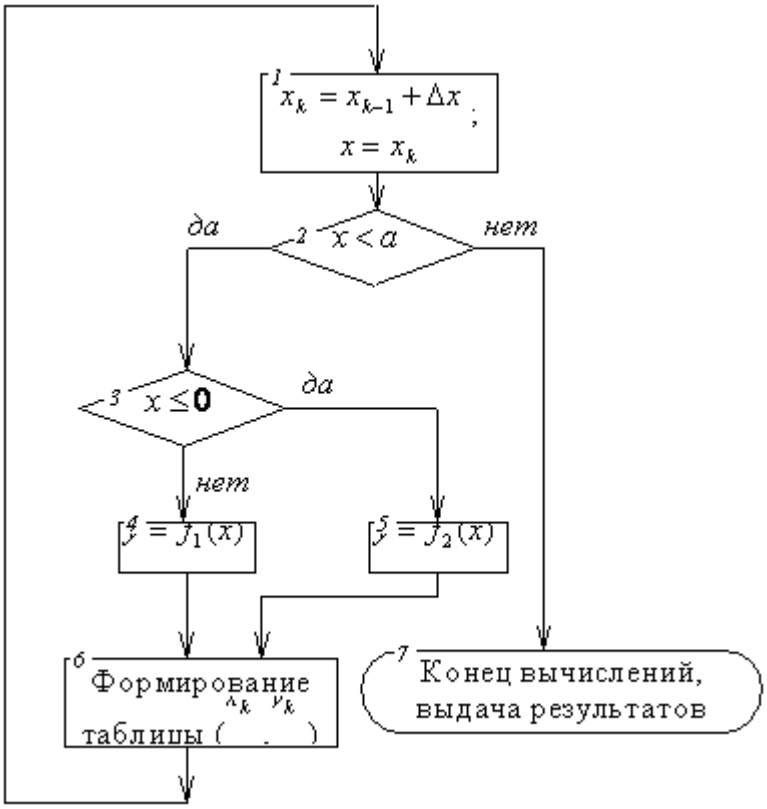


Рис. 4.14

Обычно структурная схема является наиболее удобной формой представления моделирующего алгоритма. В ряде случаев моделирующий алгоритм представляется в форме граф-схемы. В качестве пояснения к граф-схеме алгоритма в тексте дается раскрытие содержания операторов, что позволяет упростить представление алгоритма, но усложняет работу с ним. Для рассмотренного выше примера граф-схема алгоритма имеет вид, представленный на рис. 1.15. Обозначение операторов на граф-схеме соответствует обозначениям для операторной схемы.

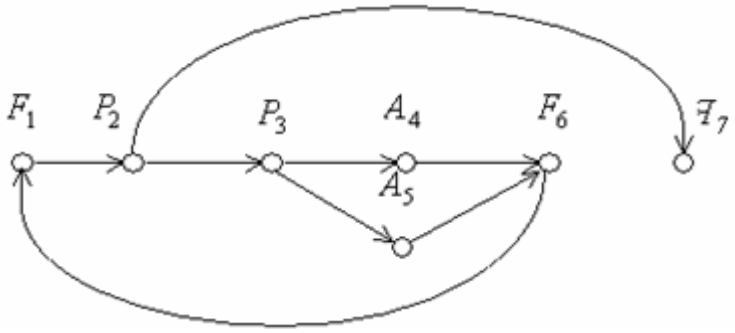


Рис. 4.15.

4.8. Особенности фиксации и обработки результатов моделирования

При проведении имитационного эксперимента необходимо предусмотреть меры по организации эффективной обработки и представления его результатов. В результате эксперимента с моделью системы на ЭВМ вырабатывается информация о состояниях исследуемой системы. Эта информация является исходным материалом для определения оценок искомых характеристик. К качеству оценок, полученных в результате статистической обработки результатов моделирования, предъявляются следующие требования: оценки должны быть несмещенными, состоятельными и эффективными. Возможность получить при моделировании на ЭВМ большие выборки позволяет количественно оценить характеристики системы с учетом этих требований.

При моделировании сложных систем осуществляется многократная реализация на ЭВМ моделирующего алгоритма для получения статистически устойчивых оценок искомых характеристик. При этом накапливается значительный объем статистической информации, для хранения которой требуется большой объем памяти ЭВМ. С целью экономии памяти ЭВМ необходимо производить фиксацию результатов моделирования и их статистическую обработку непосредственно в процессе моделирования с получением текущих оценок для искомых характеристик, используя рекуррентные алгоритмы обработки. то есть без запоминания промежуточных результатов.

Если при моделировании процесса функционирования конкретной системы учитываются случайные факторы, то и среди результатов моделирования присутствуют случайные величины. В такой ситуации в качестве оценок для искомых характеристик рассчитывают средние значения, дисперсии, корреляционные моменты и другие вероятностные характеристики соответствующих случайных величин. При выборе способа фиксации и обработки результатов моделирования необходимо стремиться к использованию минимального объема памяти ЭВМ.

Для иллюстрации рациональной организации вычислительных процессов при моделировании систем на ЭВМ рассмотрим простейшие случаи определения оценок искомых характеристик систем.

Пусть в качестве искомой величины фигурирует вероятность некоторого события А. Предположим, что в результате воспроизведения N реализаций процесса имело место *m* случаев наступления события А. Тогда в качестве оценки $\overline{P}(A)$ для искомой вероятности $P(A)$ можно использовать относительную частоту наступления события А

$$\overline{P}(A) = \frac{m}{N}.$$

Для оценки вероятности достаточно накапливать в памяти ЭВМ лишь числа *m* и N при обработке результатов моделирования.

Аналогично при обработке результатов моделирования можно подойти к оценке вероятностей возможных значений случайной величины, то есть закона распределения. Область возможных значений случайной величины *z* разбивается на *n* интервалов. Накапливается количество попаданий случайной величины в эти интервалы. Оценка для вероятности попадания случайной величины в интервал с номером *k* служит величина

$$P_k = \frac{m_k}{N}.$$

Таким образом, для оценки P_k достаточно фиксировать *n* значений m_k, то есть для обработки результатов моделирования требуется *n* ячеек памяти ЭВМ.

Для оценки среднего значения случайной величины *z* достаточно накапливать сумму возможных значения случайной величины $z_k,\ (k = \overline{1, N})$, которые она принимает при различных реализациях. Тогда среднее значение

$$\bar{z} = \frac{1}{N}\sum_{k=1}^{N} z_k.$$

В качестве оценки дисперсии случайной величины *z* при обработке результатов моделирования можно использовать

$$D_z = \frac{1}{N-1}\sum_{k=1}^{N}(z_k - \bar{z})^2. \qquad (4.55)$$

Непосредственное вычисление дисперсии по этой формуле не рационально, так как среднее значение $\bar{z}$ изменяется в процессе накопления значений z_k. Это приводит к необходимости запоминания всех N значений, что нежелательно. Несложные преобразования (1.55) приводят к более удобной для рациональной организации вычислительного процесса формуле

$$D_z = \frac{1}{N-1}\left[\sum_{k=1}^{N} z_k^2 - \frac{1}{N}\left(\sum_{k=1}^{n} z_k\right)^2\right]. \qquad (4.56)$$

Тогда для вычисления дисперсии достаточно накапливать две суммы: значений z_k и их квадратов z_k^2.

В случае оценки корреляционного момента K_{xy} для случайных величин X и Y с возможными значениями x_k, y_k $(k = \overline{1, N})$ используется выражение

$$K_{xy} = \frac{1}{N-1}\sum_{k=1}^{N}(x_k - \bar{x})(y_k - \bar{y}). \qquad (4.57)$$

Это выражение можно преобразовать к более удобному виду

$$K_{xy} = \frac{1}{N-1}\left(\sum_{k=1}^{N} x_k y_k - \frac{1}{N}\sum_{k=1}^{N} x_k \sum_{k=1}^{N} y_k\right), \qquad (4.58)$$

требующему для вычисления K_{xy} накопления трех сумм.

Если при моделировании системы искомыми характеристиками являются математическое ожидание и корреляционная функция случайного процесса $z(t)$, то для нахождения оценок этих величин интервал моделирования $[0,T]$ разбивают на отрезки с постоянным шагом Δt и накапливают значения процесса $z(t)$ для фиксированных моментов времени $t = t_i = i\Delta t$.

При обработке результатов моделирования оценки для математического ожидания и корреляционной функции можно вычислить по формулам:

$$\bar{z}(t_i) = \frac{1}{N}\sum_{k=1}^{N} z_k(t_i) \quad ; \qquad (4.59)$$

$$R(u,s) = \frac{1}{N-1}\sum_{k=1}^{N}\left[z_k(u) - \bar{z}(u)\right]\left[z_k(s) - \bar{z}_k(s)\right] \quad . \qquad (4.60)$$

где u и s пробегают все значения.

Для уменьшения затрат машинных ресурсов на хранение промежуточных результатов формулу (1.60) целесообразно преобразовать по аналогии с (1.58) и привести к следующему виду:

$$R(u,s) = \frac{1}{N-1}\left[\sum_{k=1}^{N} z_k(u) z_k(s) - \frac{1}{N}\sum_{k=1}^{N} z_k(u)\sum_{k=1}^{N} z_k(s)\right] \quad . \qquad (4.61)$$

Следует отметить особенности фиксации и обработки результатов моделирования, связанные с оценкой характеристик стационарных случайных процессов, обладающих эргодическим свойством. Пусть рассматривается процесс, для которого верны упомянутые предположения. Тогда поступают в соответствии с правилом: “среднее по времени равно среднему по множеству”. Это означает, что для оценки искомых характеристик выбирается одна достаточно продолжительная реализация процесса, для которой целесообразно фиксировать результаты моделирования. Как известно, в рассматриваемом случае математическое ожидание и корреляционная функция процесса выражаются формулами:

$$\bar{z} = \lim_{T\to\infty}\frac{1}{T}\int_0^T z(t)\,dt \quad ; \qquad (4.62)$$

$$R(\tau) = \lim_{T\to\infty}\frac{1}{T-\tau}\int_0^{T-\tau} z(t)z(t+\tau)\,dt - \bar{z}^2 \quad . \qquad (4.63)$$

На практике при моделировании на ЭВМ системы интервал $[0,T]$ оказывается ограниченным и, кроме того, значения $z(t)$ удается определить только для конечного набора моментов времени t_i. Поэтому для получения оценок $\bar{z}$ и $R(\tau)$ можно использовать приближенные формулы:

$$\bar{z} = \frac{\Delta t}{T}\sum_{i=1}^{m} z(t_i) \quad ; \qquad (4.63)$$

$$R(\tau)=\frac{\Delta t}{T-\tau}\sum_{i=1}^{m-l}z(t_i)z(t_i+\tau)-\bar{z}^2 \quad , \qquad (4.64)$$

где $m=\frac{T}{\Delta t}$; $l=\frac{\tau}{\Delta t}$; $\tau=l\Delta t$, $l=0,1,2\ldots$.

Исходя из соотношений (4.63) и (4.64) можно установить целесообразный порядок фиксации и накапливания результатов моделирования.

Мы затронули только простейшие случаи эффективной организации вычислительного процесса для оценок искомых характеристик при моделировании на ЭВМ. Особенности фиксации и статистической обработки результатов моделирования систем на ЭВМ более подробно изложены в источниках [ ] и [ ].

Глава 5. Идентификация объектов

5.1. Постановка задачи и описание проблемы идентификации объектов

Объект идентификации в общем случае представляется в виде многополюсника, изображенного на рис. 5.1, а где $x_1, x_2, \ldots, x_n$ — наблюдаемые входы объекта; $z_1, z_2, \ldots, z_k$ — ненаблюдаемые входы объекта; $y_1, y_2, \ldots, y_m$ — наблюдаемые выходы объекта.

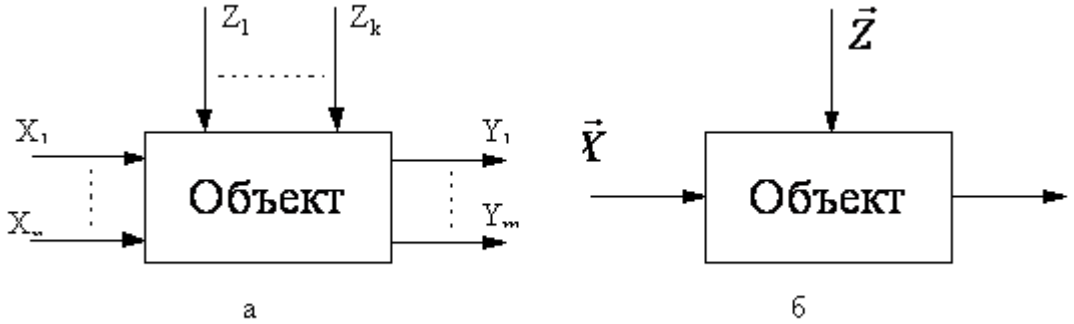


Рис. 5.1

Многомерный объект удобно представлять в векторной форме (рис.2.1,б) где

$\vec{x}=(x_1, x_2, \ldots, x_n)$;

$\vec{z}=(z_1, z_2, \ldots, z_k)$;

$\vec{y}=(y_1, y_2, \ldots, y_m)$.

В общем случае переменные $\vec{x}$, $\vec{y}$, $\vec{z}$ являются случайными функциями времени, $\vec{x}=\vec{x}(t)$, $\vec{y}=\vec{y}(t)$, $\vec{z}=\vec{z}(t)$.

Объект связывает входы $\vec{x}$ и $\vec{z}$ с выходом $\vec{y}$ некоторым априори неизвестным оператором F_0

$$\vec{y} = F_0(\vec{x}, \vec{z}) \quad . \qquad (5.1)$$

Однако идентифицируется не оператор F_0, а оператор модели F, связывающий наблюдаемые входы и выходы:

$$\vec{y} = F(\vec{x}) \qquad (5.2)$$

Ненаблюдаемый вход $\vec{z}$ рассматривается как случайная помеха, затрудняющая определение оператора F.

Задачей идентификации является построение такого оператора модели F, который был бы в определенном смысле близок к оператору объекта F_0, то есть $F \approx F_0$. Однако указанная близость весьма относительна, так как операторы F_0 и F могут иметь разную структуру и разное число входов. Поэтому близость операторов непосредственно оценить трудно или просто невозможно, тем более что часто об операторе F_0 мало что известно.

В связи с этим в теории идентификации близость операторов оценивают по их реакциям на одно и то же входное воздействие $\vec{x}^0$, то есть по выходам объекта

$\vec{y} = F_0(\vec{x}^0, \vec{z})$

и модели

$\vec{y}^M = F_0(\vec{x}^0)$,

где $\vec{y}^M = (y_1^M, y_2^M, \ldots, y_m^M)$ — вектор выхода модели.

Степень близости этих реакций в каждый момент времени в общем случае оценивается функцией невязки ρ. Эта скалярная функция двух векторных аргументов — выходов объекта и модели

$e(t) = \rho(\vec{y}(t), \vec{y}^M(t))$, (2.3)

которая обладает следующими свойствами:

$\rho(\vec{y}(t), \vec{y}^M(t)) \geq 0$ для любых $\vec{y}(t)$, $\vec{y}^M(t)$;

$\rho(\vec{y}(t), \vec{y}^M(t)) = 0$ при $\vec{y}(t) = \vec{y}^M(t)$;

непрерывна и выпукла вниз по обоим аргументам.

Этим требованиям соответствует функция невязки

$$e(t) = \left|\vec{y}(t) - \vec{y}^M(t)\right|^2 = \sum_{i=1}^{m} (y_i(t) - y_i^M(t))^2 \quad , \qquad (5.4)$$

которая чаще всего применяется в задачах идентификации.

Можно ввести единую меру близости выходов объекта и модели на всем интервале наблюдения $[0, T]$, а не только в каждой точке, как (5.3.). Такой мерой в непрерывном случае может быть интеграл

$$E = \int_0^T \rho(\vec{y}(t), \vec{y}^M(t))dt \qquad (5.5)$$

Значение E явно зависит от F:

$$E(F) = \int_0^T \rho(\vec{y}(t), F(\vec{x}(t)))dt \tag{5.6}$$

и задача идентификации заключается в ее минимизации путем соответствующего выбора оператора модели F.

Следовательно, модельный оператор F должен быть таким, чтобы $\vec{y}^M \sim \vec{y}$, то есть выходы модели и объекта при одинаковых входных воздействиях должны быть эквивалентны.

Определение класса оператора, в котором ищется решение задачи идентификации, практически не поддается формализации и нуждается в эвристических решениях.

Для того чтобы начать процедуру идентификации необходимо иметь априорную информацию о структуре модели объекта и достаточном объеме измерительной (апостериорной) информации для определения параметров модели.

Априорная информация, которой необходимо располагать еще до наблюдения входов и выходов объекта, часто имеет качественный характер. Она должна ответить на вопрос, что представляет собой структура модели идентифицируемого объекта. Структура модели определяется в зависимости от основных свойств объекта.

Определение структуры оператора F модели составляет задачу структурной идентификации. Если же структура этого оператора определена и априори известна, то процесс идентификации сводится к определению параметров этой структуры по имеющейся измерительной информации, то есть к решению задач параметрической идентификации.

При решении задач идентификации широко используются процедуры минимизации функции или функционала и задачи оптимального планирования эксперимента.

5.2. Идентификация статики линейного детерминированного объекта

Предполагается, что связь между входными и выходными переменными объекта может быть описана линейным уравнением. При этом выходная переменная изменяется только под воздействием наблюдаемых входных сигналов, а какие-либо ненаблюдаемые помехи, вызывающие дополнительное изменение выходной переменной, отсутствуют или их влиянием можно пренебречь. Следовательно, в схеме на рис. 2.1 предполагается, что $\vec{z} = 0$, т.е. выходная переменная $\vec{y}$ наблюдается без помех. Связь между входом и выходом такого объекта описывается функциональной зависимостью

$$\vec{y} = F(\vec{x})$$

Рассмотрим теперь линейную функцию F и проанализируем специфику ее идентификации.

Модель статики линейного детерминированного объекта с n входами $x_j\,(j=\overline{1,n})$ и m выходами $y_j\,(j=\overline{1,m})$ имеет единственно возможную структуру и описывается системой линейных алгебраических уравнений:

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n + b_1\,; \\ y_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n + b_2\,; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_m = a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n + b_m\,. \end{cases} \qquad (5.7)$$

где идентифицируются mn коэффициентов $a_{ij}\,(i=\overline{1,m};\, j=\overline{1,n})$ и m коэффициентов $b_i\,(i=\overline{1,m})$.

Систему уравнений (2.7) можно записать в компактном виде в векторной форме

$$\vec{y} = A\vec{x} + B\,, \qquad (5.8)$$

где $\vec{x} = (x_1, \ldots x_n)^T$,

$\vec{y} = (y_1, \ldots y_m)^T$,

$B = (b_1, \ldots b_m)^T$.

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix}.$$

Здесь T - знак транспонирования.

Идентификации в данном случае подлежат вектор B и матрица A.

Модель (2.7) можно рассматривать как совокупность моделей с многомерным входом $x_j\,(j=\overline{1,n})$ и одномерным выходом y (m=1).

Поэтому рассмотрим один выход объекта, то есть случай m=1, n>1.

Модель такого объекта в векторной форме имеет вид:

$y = a_0 + \langle A, \vec{x} \rangle$,

где $\langle A, \vec{x} \rangle$ - скалярное произведение векторов $A = (a_1, a_2, \ldots a_n)$ и $\vec{x} = (x_1, \ldots x_n)$:

$\langle A, \vec{x} \rangle = \sum_{i=1}^{n} a_i x_i$.

В скалярной форме модель объекта имеет вид

$$y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n\,. \qquad (5.9)$$

Модель имеет n+1 неизвестных параметров $a_i\,(i=\overline{0,n})$, которые подлежат оценке на основе измерений входов и выхода объекта. Эта информация обычно представляется в виде N соответствующих пар значений $(\vec{x}_i, y_i)$, где $j=\overline{1,n}$, $\vec{x}_j = (x_{1j}, x_{2j}, \ldots x_{nj})$ - j-е состояние входа объекта, а y_j - реакция объекта на этот вход.

Обычным подходом к решению этой задачи является приравнивание выходов объекта и модели в N заданных точках $(\vec{x}_i, y_i)$, в результате чего получают следующую систему уравнений идентификации:

$$a_0 + \sum_{i=1}^{n} a_i x_{ij} = y_j \quad (j = \overline{1, N}). \tag{5.10}$$

Полученные N уравнений с n+1 неизвестными $(N \geq n+1)$ имеют однозначное решение, если матрица

$$A_1 = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{n1} \\ 1 & x_{12} & x_{22} & \dots & x_{n2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1N} & x_{2N} & \dots & x_{nN} \end{bmatrix} \tag{5.11}$$

невырождена, т.е. $\det A_1 \neq 0$ и, следовательно, ранг матрицы равен n+1. Это возможно в том случае, если найдется n+1 линейно независимая строка матрицы (2.11).

Поэтому из N строк следует выбрать n+1 линейно независимых строк $(1, x_{1i}, x_{2i}, \dots x_{ni})$, где $i \in \{1, N\}$.

В этом случае из системы (5.10) будут выделены n+1 линейно независимых уравнений

$$a_0 + \sum_{i=1}^{n} a_i x_{ir} = y_r \quad (r = \overline{1, n+1}), \tag{5.12}$$

совместное решение которых гарантирует определение точных оценок $a_i^* \ (i = \overline{0, n})$ идентифицируемых параметров $a_i \ (i = \overline{0, n})$ объекта, если, разумеется, объект действительно линеен. Покажем это.

Подставим в систему уравнений (5.12) уравнение объекта

$$y_r = a_0^* + \sum_{i=1}^{n} a_i^* x_{ir} = y_r \quad (r = \overline{1, n+1}),$$

где $a_i^* \ (i = \overline{1, n})$ - оценки параметров объекта.

Введем невязки

$$\Delta_i = a_i - a_i^* \ (i = \overline{0, 1}).$$

Тогда система уравнений (5.12) запишется в виде

$$\Delta_0 + \sum_{i=1}^{n} \Delta_i x_{ir} = 0 \ (r = \overline{1, n+1}). \tag{5.13}$$

Для того чтобы решение системы (5.13) было нулевым, необходимо и достаточно, чтобы определитель этой системы не был равен нулю. Легко заметить, что матрица системы (5.13) такая же, что и матрица системы (5.12), содержит n+1 линейно независимых строк матрицы (5.11) и ее определитель не равен нулю.

В результате имеем:

$$\Delta_i = 0 \quad (i = \overline{0, n})$$

и следовательно, решение системы (5.12) гарантирует точную идентификацию параметров объекта, т.е.

$$a_i^* = a_i \quad (i = \overline{0,n}).$$

Однако возможно, что объект не строго линеен и существуют незначительные случайные возмущения. При этом может оказаться, что ранг матрицы (5.11) меньше n+1 и из системы (5.10) невозможно выделить n+1 линейно независимых уравнений (5.12) для определения коэффициентов $a_i \ (i = \overline{0,n})$ объекта. В этом случае возможны следующие подходы к идентификации:

1) повторить измерения входов и выхода объекта в надежде, что первый эксперимент был неудачным, т.е. состояния вектора входа $\vec{x}_j \ (j = \overline{1,N})$ были недостаточно разнообразны. Если и на новом экспериментальном материале не выполнится указанное условие, то можно попытаться изменить структуру модели;

2) понизить число идентифицируемых параметров, т.е. исключить рассмотрение одного из входов, например, того, который мало изменяется. Это означает, что число идентифицируемых параметров стало n (а не n+1). Сказанное следует делать до тех пор, пока ранг матрицы (5.11) не совпадет с ее размерностью. При выполнении этого условия из системы (2.10) всегда можно выделить линейно независимые уравнения (2.12) в количестве равном числу n_1 оставшихся коэффициентов $a_i \ (i = \overline{0,n_1-1})$. Совместно решая их, находят эти коэффициенты;

3) отказаться от метода интерполяции для определения независимых коэффициентов $a_i \ (i = \overline{0,n_1-1})$, который привел к несовместной системе линейных уравнений. **Ввести суммарную невязку выходов модели и объекта.**

$$E = \sum_{j=1}^{N} (a_0 + \sum_{i=1}^{n} a_i x_{ij} - y_j)^2 = E(\vec{a}). \tag{5.14}$$

Величина E характеризует степень несоответствия модели и объекта и зависит от параметров $\vec{a}$ модели.

Задачу оценки параметров $a_i \ (i = \overline{0,n})$ можно теперь представить как задачу минимизации невязки (5.14), например, **методом наименьших квадратов**, т.е. свести к решению системы линейных уравнений:

$$\begin{cases} \dfrac{\partial E}{\partial a_0} = 2\sum\limits_{j=1}^{N} (a_0 + \sum\limits_{i=1}^{n} a_i x_{ij} - y_j) = 0; \\ \dfrac{\partial E}{\partial a_k} = 2\sum\limits_{j=1}^{N} \left[\left(a_0 + \sum\limits_{i=1}^{n} a_i x_{ij} - y_j \right) x_{kj} \right] = 0; \qquad (k = \overline{1,n}). \end{cases} \tag{2.15}$$

Система (2.15) имеет следующий развернутый вид:

$$\begin{cases} Na_0 + a_1 \sum x_{1j} + a_2 \sum x_{2j} + \ldots + a_n \sum x_{nj} = \sum y_j; \\ a_0 \sum x_{1j} + a_1 \sum x_{1j}^2 + a_2 \sum x_{2j} x_{1j} + \ldots + a_n \sum x_{nj} x_{1j} = \sum x_{1j} y_j; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_0 \sum x_{nj} + a_1 \sum x_{1j} x_{nj} + a_2 \sum x_{2j} x_{nj} + \ldots + a_n \sum x_{nj}^2 = \sum x_{nj} y_j \end{cases} \tag{5.16}$$

где суммирование всюду осуществляется по j от 1 до N.

Как видно, эта система линейных алгебраических уравнений относительно искомых параметров $a_i \ (i=\overline{0,n})$. Если ранг матрицы коэффициентов системы (5.16)

$$C=\begin{bmatrix} N & \sum x_{1j} & \dots & \sum x_{nj} \\ \sum x_{1j} & \sum x_{1j}^2 & \dots & \sum x_{nj}x_{1j} \\ \dots & \dots & \dots & \dots \\ \sum x_{nj} & \sum x_{1j}x_{nj} & \dots & \sum x_{nj}^2 \end{bmatrix} \qquad (5.17)$$

равен n+1, то $\det C \neq 0$ и система (2.16) имеет единственное решение $a_0^*, a_1^*, \dots a_n^*$, причем оно доставляет минимум функции $E(\vec{a})$ (2.14), поскольку $E(\vec{a}) \geq 0$.

В некоторых случаях матрица С системы линейных уравнений может оказаться плохообусловленной, т.е. $\det C \approx 0$. В этом случае малые ошибки измерений x_i и погрешности вычислительных процессов приводят к большим погрешностям в определении коэффициентов $a_i \ (i=\overline{0,n})$ модели. Плохая обусловленность матрицы С имеет место в том случае, если некоторые ее строки (или столбцы) почти линейно зависимы. Например, пусть первая и вторая строки матрицы С почти линейно зависимы. Это означает, что

$$\frac{N}{\sum x_{1j}} \approx \frac{\sum x_{1j}}{\sum x_{1j}^2} \approx \dots \approx \frac{x_{nj}}{\sum x_{nj}x_{1j}}.$$

Существует ряд способов определения a_j^* из системы линейных уравнений (2.16), в которой $\det C \approx 0$. Для этого прежде всего стремятся повысить точность вычислений, подвергают С некоторым эквивалентным преобразованиям, изменяют число опытов N и шаг Δx_i дискретности измерения входов $x_i \ (i=\overline{1,n})$.N>n+1.

Идентификация линейных объектов приводит к решению систем линейных уравнений. С этой задачей исследователь часто сталкивается в практике. Это обусловлено, по крайней мере, двумя причинами.

Во-первых, многие задачи линейной оптимизации, идентификации линейных и нелинейных моделей статики, идентификации линейных моделей динамики (дифференциальных уравнений) объекта сводятся к решению систем линейных уравнений.

Во-вторых, большинство нелинейных задач в “малом” линейны, т.е. нелинейные модели в малой окрестности некоторого решения могут быть описаны линейными. Следовательно, первым шагом решения нелинейных задач является исследование линеаризованных моделей, что также связано с решением систем линейных уравнений.

Таким образом, численные методы решения систем линейных уравнений оказываются важным инструментом решения обширного круга научно-технических задач на ЭВМ.

5.3. Идентификация статики нелинейного детерминированного объекта

Предположим, что помехи Z, действующие на выходную переменную объекта, малы или совсем отсутствуют ($\vec{Z}=0$ в схеме на рис. 5.1), т.е. вход $\vec{x}$ и выход $\vec{y}$ объекта являются детерминированными сигналами (без помех), наблюдаемыми точно. Тогда выходной сигнал $\vec{y}$ полностью определяется только наблюдаемым входным сигналом. Поведение моделируемого объекта такого рода в общем случае описывается многомерной функцией многих переменных

$$\vec{y}=F_0(\vec{x}) \tag{5.18}$$

Рассмотрим одномерный случай. Поведение нелинейного одномерного детерминированного объекта описывается нелинейной функцией одной переменной общего вида

$$y=F_0(x) \tag{5.19}$$

Априорные сведения о виде функции F_0 (оператора реального объекта) могут быть различными по объему и содержанию. Трансформация этих сведений в определенную функцию

$$y=F(x,\vec{a}) \tag{5.20}$$

с неизвестными параметрами $\vec{a}=(a_0,\ldots,a_m)$ составляет предмет структурной идентификации. Основные подходы к выбору структуры модели (2.20) опирается на кусочно-линейное и функциональное представление зависимости выхода объекта от его входа, причем частным случаем последнего является разложение в ряд по заданной системе функций.

Относительно функции $F(x,\vec{a})$ возможна различная степень информированности.

1. Вид функции $F(x,\vec{a})$ известен, тогда задача поиска математической модели сводится к определению неизвестных параметров $a_0,a_1,\ldots,a_m$ (параметрическая идентификация).

2. Функция $F(x,\vec{a})$ выбирается из совокупности функций $F_1(x,\vec{a}^{(1)}),F_2(x,\vec{a}^{(2)}),\ldots,F_k(x,\vec{a}^{(k)})$, в которых размерность векторов $\vec{a}^{(1)},\vec{a}^{(2)},\ldots,\vec{a}^{(k)}$ различная. Требуется определить, какая из перечисленных функций является искомой (структурная идентификация) и, соответственно, найти неизвестные параметры $\vec{a}$ (параметрическая идентификация).

Рассмотрим общую для этих случаев задачу параметрической идентификации функциональной модели, когда функция задана в аналитической форме и непрерывно дифференцируема (отметим, что кусочно-линейная модель является частным случаем функциональной).

В этом случае задача идентификации сводится к определению параметров $a_0, a_1, \ldots, a_m$ на основе экспериментальных наблюдений, т.е. апостериорной информации. Для этого достаточно потребовать, чтобы в каждом из зафиксированных состояний модель соответствовала объекту. Наиболее распространенным методом решения такой задачи является интерполирование и приближение.

При интерполировании предполагают, что выходы объекта и модели являются близкими, если они совпадают на заданной системе значений входа $x_1, \ldots, x_n$, т.е.

$$F(x_i, \vec{a}) - y_i = 0; \; i = \overline{1, n}, \qquad (5.21)$$

где $n \geq m+1$.

Выражение (5.21) в общем случае является системой нелинейных уравнений относительно неизвестных параметров. Наличие шумов и погрешностей измерения, как правило, делает систему уравнений (5.21) несовместной и затрудняет ее решение. Поэтому наиболее употребительным является метод приближения.

Для приближения характерна минимизация некоторого функционала, характеризующего различие $F(x, \vec{a})$ и $F_0(x)$ во всем промежутке . В практических приложениях чаще всего используется дискретное квадратичное приближение (метод наименьших квадратов), при котором задача идентификации представляется в виде минимизации функционала

$$E = \sum_{i=1}^{n} \left[F(x_i, \vec{a}) - y_i \right]^2 \to \min_{\vec{a}}, \qquad (5.22)$$

равного сумме квадратов отклонений выходов объекта и модели на заданной системе точек $x_1, x_2, \ldots, x_n$.

Очевидно, что величина E характеризует степень несоответствия модели и объекта и зависит от параметров модели $a_0, a_1, \ldots, a_m$. Эти коэффициенты надо подобрать так, чтобы величина E была наименьшей. Минимум E при правильном выборе структуры модели F равен нулю, т.е. модель может быть адекватной объекту. Если же структура модели отличается от структуры объекта, то минимум E не будет равен нулю, тем не менее задача синтеза модели по-прежнему сводится к задаче минимизации функции невязки E.

Необходимым условием минимума $E(x, a_0, a_1, \ldots, a_n)$ является выполнение равенств

$$\frac{\partial E(x_i, \vec{a})}{\partial a_j} = 0, \; j = \overline{0, m}, \qquad (5.23)$$

или

$$\begin{cases} 2\sum_{i=1}^{n}\left[F(x_i,a_0,\ldots,a_m)-y_i\right]\dfrac{\partial F(x_i,a_0,\ldots,a_m)}{\partial a_0}=0; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ 2\sum_{i=1}^{n}\left[F(x_i,a_0,\ldots,a_m)-y_i\right]\dfrac{\partial F(x_i,a_0,\ldots,a_m)}{\partial a_m}=0. \end{cases} \quad (5.24)$$

После преобразования

$$\begin{cases} \sum_{i=1}^{n}F(x_i,a_0,\ldots,a_m)\dfrac{\partial F(x_i,a_0,\ldots,a_m)}{\partial a_0}-\sum_{i=1}^{n}y_i\dfrac{\partial F(x_i,a_0,\ldots,a_m)}{\partial a_0}=0; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \sum_{i=1}^{n}F(x_i,a_0,\ldots,a_m)\dfrac{\partial F(x_i,a_0,\ldots,a_m)}{\partial a_m}-\sum_{i=1}^{n}y_i\dfrac{\partial F(x_i,a_0,\ldots,a_m)}{\partial a_m}=0. \end{cases} \quad (5.25)$$

Система уравнений (5.25) содержит столько же уравнений, сколько неизвестных коэффициентов $a_0,\ldots,a_m$, и называется системой нормальных уравнений.

Величина $E \geq 0$ при любых $a_0,\ldots,a_m$, следовательно, у нее обязательно должен существовать хотя бы один минимум. Поэтому, если система нормальных уравнений (5.25) имеет единственное решение, то оно и является минимумом для величины E.

Система уравнений (5.25) в общем случае является системой трансцендентных уравнений относительно искомых параметров $\vec{a}$. Поэтому для решения такой системы часто обращаются к поисковым методам минимизации.

Применение поисковых методов для минимизации функции невязки (5.25) открывает большие возможности перед идентификацией. Однако поиска можно избежать, если идентифицируемая модель линейна относительно искомых параметров $\vec{a}$.

Модели такого рода являются частным случаем функциональных моделей и образуются в результате разложения искомой функции по заданной системе функций.

Рассмотрим вкратце такой случай. Пусть модель $F(x,\vec{a})$ линейна относительно неизвестных параметров и имеет вид

$$F(x,\vec{a}) = a_0 f_0(x) + a_1 f_1(x) + \ldots + a_m f_m(x) = \sum_{j=0}^{m} a_j f_j(x), \quad (5.26)$$

где a_j — искомые коэффициенты модели, $f_j(x)$ — известные непрерывно-дифференцируемые функции входной переменной.

Система функций $f_j(x)$ должна быть линейно независимой, т.е. никакая линейная комбинация ее элементов не может быть равна нулю при любых x.

В результате эксперимента можно получить n наблюдений

$$y_i = F_0(x_i); \quad i = \overline{1,n}. \quad (5.27)$$

Для определения неизвестных коэффициентов $\vec{a}$ воспользуемся методом наименьших квадратов. В этом случае функция невязки (5.22) после подстановки в нее выражение (5.26) принимает вид

$$E = \sum_{i=1}^{n}\left[\sum_{j=1}^{m} a_j f_j(x_i) - y_i\right]^2 \to \min_{\vec{a}} . \qquad (5.28)$$

Теперь найдем неизвестные параметры a_j модели из необходимых условий минимума квадратичной функции относительно неизвестных параметров a_j ($j=\overline{1,n}$):

$$\frac{\partial E(x,\vec{a})}{\partial a_j} = \frac{\partial}{\partial a_j}\left\{\sum_{i=1}^{n}\left[\sum_{j=1}^{m} a_j f_j(x_i) - y_i\right]^2\right\} = 0 \quad ; \quad j=\overline{0,m} . \qquad (5.29)$$

Вычислив частные производные и перенеся в правую часть члены, не содержащие коэффициентов a_j, получим систему линейных нормальных уравнений метода наименьших квадратов:

$$\begin{cases} a_0\sum_{i=1}^{n} f_0^2(x_i) + a_1\sum_{i=1}^{n} f_0(x_i)f_1(x_i) + \ldots + a_m\sum_{i=1}^{n} f_0(x_i)f_m(x_i) = \sum_{i=1}^{n} f_0(x_i)y_i; \\ a_0\sum_{i=1}^{n} f_0(x_i)f_1(x_i) + a_1\sum_{i=1}^{n} f_1^2(x_i) + \ldots + a_m\sum_{i=1}^{n} f_1(x_i)f_m(x_i) = \sum_{i=1}^{n} f_1(x_i)y_i; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_0\sum_{i=1}^{n} f_0(x_i)f_m(x_i) + a_1\sum_{i=1}^{n} f_1(x_i)f_m(x_i) + \ldots + a_m\sum_{i=1}^{n} f_m^2(x_i) = \sum_{i=1}^{n} f_m(x_i)y_i, \end{cases} \qquad (5.30)$$

решение которой дает значения идентифицируемых параметров модели.

В векторно-матричной форме система уравнений (5.30) примет вид:

$\Phi^{\mathrm{T}}\Phi\vec{a} = \Phi^{\mathrm{T}}\vec{Y}$, (5.31)

где

$$\Phi = \begin{bmatrix} f_0(x_1)\ f_1(x_1)\ \ldots\ f_m(x_1) \\ f_0(x_2)\ f_1(x_2)\ \ldots\ f_m(x_2) \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ f_0(x_n)\ f_1(x_n)\ \ldots\ f_m(x_n) \end{bmatrix}_{n\times(m+1)}; \quad \vec{a} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix}; \quad \vec{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

Φ — прямоугольная матрица размерности $n\times(m+1)$, задающая значения функций $f_j(x)$ при проведение n наблюдений;

$\vec{a}$ — $m+1$- мерный вектор искомых коэффициентов модели;

$\vec{Y}$ — n- мерный вектор замеров выхода объекта;

T — операция транспонирования матрицы.

Информационная матрица Фишера $\Phi^{\mathrm{T}}\Phi$ является квадратной, положительно-определенной и невырожденной, когда $n \geq m+1$ и хотя бы $m+1$ измерений выхода объекта проведен при различных уровнях входной

переменной x. В этом случае матрица Фишера имеет обратную матрицу $\left(\Phi^{\mathrm{T}}\Phi\right)^{-1}$ и решением системы (5.31) будет вектор

$$\vec{a} = \left(\Phi^{\mathrm{T}}\Phi\right)^{-1}\Phi^{\mathrm{T}}\vec{Y}. \quad (5.32)$$

Систему уравнений (5.32) можно сразу разрешить, если функции $f_0(x), f_1(x), \ldots, f_m(x)$ выбрать в виде ортогональных полиномов, обладающих свойством:

$$\begin{cases} \sum_{i=1}^{n} f_j(x_i) f_l(x_i) = 0; \quad j,l = \overline{0,m}; \quad j \neq l; \\ \sum_{i=1}^{n} \left[f_j(x_i)\right]^2 \neq 0; \quad j = \overline{0,m} \end{cases} \quad (5.33)$$

Последнее означает, что хотя бы в одной из точек x_i, $i = \overline{1,n}$ многочлен $f_j(x) \neq 0$ Такие многочлены называют ортогональными многочленами Чебышева.

При выполнение условий (5.33) в левой части каждого из уравнений системы (5.30) остается только один член (диагональные элементы) и эта система распадается на $m+1$ уравнений с одним неизвестным в каждом. Поэтому сразу можно написать выражение для искомых коэффициентов

$$a_j = \frac{\sum_{i=1}^{n} f_j(x_i) y_i}{\sum_{i=1}^{n} \left[f_j^2(x_i)\right]^2}; \quad j = \overline{0,m}. \quad (5.34)$$

Многочлены Чебышева определяются по формулам:

$$f_0(x) = 1;$$

$$f_1(x) = x - \frac{n+1}{2};$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$f_{m+1}(x) = f_1(x) f_m(x) - \frac{m^2(n^2 - m^2)}{4(4m^2 - 1)} f_{m-1}. \quad (5.35)$$

Из формулы (5.35) следует, что полином $f_{m+1}(x)$ зависит только от объема выборки n. Следовательно, вычисленные по формуле (5.34) коэффициенты a_j не зависят от того, каков будет порядок модели (5.26). Зная многочлены Чебышева $f_{m+1}(x)$ при каждом увеличении степени уравнения статики объекта на единицу используют все ранее найденные коэффициенты a_j $(j = \overline{0,m})$ вычисляют только один коэффициент a_{m+1}. Эта особенность делает полезным использование ортогональных полиномов Чебышева при решении задач идентификации объекта путем последовательного улучшения модели.

Несмотря на преимущества использования полиномов Чебышева, для практических целей, как правило, ограничиваются степенным

разложением функции $F(x,\vec{a})$, т.е. $f_0(x)=1, f_1(x)=x, f_2=x^2, \ldots, f_m(x)=x^m$. Такой подход существенно упрощает программирование решения задач на ЭВМ. В этом случае модель имеет вид:

$$F(x,\vec{a}) = a_0 + a_1 x + \ldots + a_m x^m = \sum_{j=0}^{m} a_j x^j \quad . \qquad (5.36)$$

Подставляя это выражение в функции невязки (5.28), получим

$$E = \sum_{i=1}^{n}\left[\sum_{j=0}^{m} a_j x_i^j - y_i\right]^2 \to \min_{\vec{a}} \quad . \qquad (5.37)$$

Минимизируя эту функцию методом наименьших квадратов, находим следующую систему линейных уравнений:

$$\begin{cases} na_0 + a_1\sum x_i + a_2\sum x_i^2 + \ldots + a_m\sum x_i^m = \sum y_i; \\ a_0\sum x_i + a_1\sum x_i^2 + a_2\sum x_i^3 + \ldots + a_m\sum x_i^{m+1} = \sum x_i y_i; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_0\sum x_i^m + a_1\sum x_i^{m+1} + a_2\sum x_i^{m+2} + \ldots + a_m\sum x_i^{2m} = \sum x_i^m y_i, \end{cases} \qquad (5.38)$$

где всюду суммирование осуществляется по i от 1 до n.

Решая систему линейных уравнений (5.38), находим искомые коэффициенты модели $a_0, a_1, \ldots, a_m$.

Если матрица коэффициентов системы (5.38) неособенная, т.е. ее определитель не равен нулю, то найденное решение единственное, причем оно обеспечивает $\min E$.

Трудности в вычислении коэффициентов $\vec{a}$ из системы нормальных уравнений (5.38) возникает в случае, когда матрица системы плохо обусловлена, тогда приходится привлекать методы регуляризации.

Адекватность модели можно оценить по степени близости выходов объекта и модели на заданной системе точек.

Схема оценки адекватности модели приведена на рис.5.2.

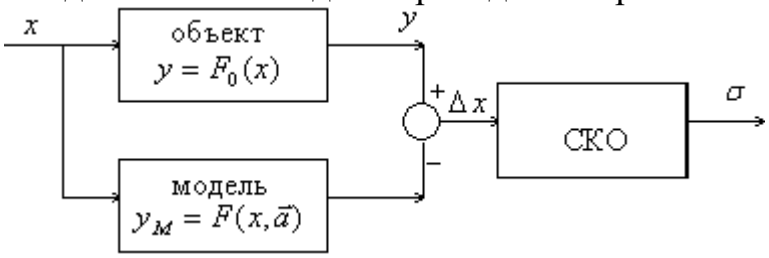


Рис. 5.2

В качестве меры близости выходов используют норму равномерного приближения

$$\Delta = \max\left|y - y_M\right| = \max\left|y - F(x,\vec{a})\right|; \quad x \in \left[x_{\min}, x_{\max}\right] \qquad (5.39)$$

или дискретную норму среднеквадратичного приближения

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[y_i - F(x_i,\vec{a})\right]^2} \quad . \qquad (5.40)$$

Модель считается адекватной, если выполняется условие $\Delta \le \varepsilon_1$, или $\sigma \le \varepsilon_2$,

где $\varepsilon_1, \varepsilon_2$ — заданные положительные числа.

Уменьшение ошибки модели можно достичь путем увеличения степени полинома (5.36). Однако с некоторого момента при повышении степени полинома (5.36) среднеквадратическая ошибка вместо того, чтобы уменьшаться, может увеличиваться из-за искажения результатов измерения выхода неконтролируемыми помехами.

Поэтому в рассмотренной выше процедуре идентификации нелинейной модели неясным остается вопрос выбора числа m — степени модели (5.36).

Обычно его определяют на основе анализа наблюдений выходов объекта и модели.

Если для оценки адекватности выбирается равномерная норма (5.39), то полином, который доставляет

$$\min \Delta = \min_{m} \max_{x_{MIN} \le x \le x_{MAX}} \left| y - \sum_{j=1}^{m} a_j x^j \right|, \qquad (5.41)$$

является оптимальным.

В случае, когда выбирается среднеквадратичная норма, то полином, который доставляет

$$\min \sigma = \min_{m} \left(\frac{1}{n} \sum_{i=1}^{n} \left[y_i - \sum_{j=0}^{m} a_j x_i^j \right] \right)^{1/2}, \qquad (5.42)$$

является полиномом наилучшего среднеквадратичного дискретного приближения.

Следует отметить, что, если результаты измерений выхода объекта в процессе идентификации содержат помехи, то выбор оптимальной величины m из условия (5.41) и (5.42) приведет к неоправданному увеличению степени полинома (5.36) и усложнению модели. В этом случае желательно выбрать m с учетом величины ошибки измерения выхода объекта.

Среднеквадратическую ошибку (СКО) измерения y можно определить по результатам повторных измерений $y_i^0 \quad (i = \overline{1,k})$ выхода объекта при фиксированном значении x^0 входа

$$\sigma_y = \sqrt{\frac{1}{k-1} \sum_{i=1}^{k} \left(y_i^0 - \bar{y}^0 \right)^2}, \qquad (5.43)$$

где $\bar{y}^0 = \frac{1}{k} \sum_{i=1}^{k} y_i^0$.

Среднеквадратическая ошибка модели

$$\sigma_M = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} a_j x^j \right)^2}, \qquad (5.44)$$

при фиксированных значениях $x_i, y_i \quad \left(i = \overline{1,n}\right)$ и $a_j\left(j = \overline{1,m}\right)$ зависит от степени полинома (5.36). Естественно потребовать, чтобы СКО модели была соизмерима с СКО измерения выхода объекта, т.е.

$$\sigma_M \le \sigma_y \qquad (5.45)$$

При выборе m условие (5.45) проверяется в процессе последовательного увеличения степени полинома (5.36), начиная с $m = 1$. При этом для каждого m требуется заново вычислять все коэффициенты $a_j \quad \left(j = \overline{0,m}\right)$. Если же в качестве базисных функций $f_j(x)$ выбрана ортогональная система полиномов Чебышева, то увеличение числа членов рада (5.36) на единицу не требует пересчета найденных ранее коэффициентов. Последовательное увеличение m продолжается до тех пор, пока не будет выполнено условие (5.45).

Рассмотрим другой подход к определению наилучшей степени полинома (5.36). Пусть выход объекта $y = F_0(x)$ измерена достаточно точно и является гладкой и низкочастотной функцией. Если $F_0(x)$ есть полином степени m, то

$$\frac{d^m F_0(x)}{dx^m} = const, \quad \frac{d^{m+1} F_0(x)}{dx^{m+1}} = 0.$$

Используем это свойство полинома для оценки числа m. Производные $F_0(x)$ неизвестны, поэтому заменим их конечными разностями выхода объекта y в точках $y_1 = y_j \quad \left(j = \overline{1,n}\right)$:

$$\Delta^1 y_j = \left(y_{j+1} - y_j\right)/\Delta x, \quad j = \overline{1,n-1};$$

$$\Delta^2 y_j = \left(\Delta^1 y_{j+1} - \Delta^1 y_j\right)/\Delta x, \quad j = \overline{1,n-2};$$

.....

$$\Delta^k y_j = \left(\Delta^{k-1} y_{j+1} - \Delta^{k-1} y_j\right)/\Delta x, \quad j = \overline{1,n-k}; \quad k = \overline{1,n-1},$$

где $\Delta^i y_j$ — конечная разность i -го порядка в точке $y = y_j$;

Δx — шаг изменения x в эксперименте.

Обычно значения $\Delta^k y_j$ флуктуируют, поэтому следует определить среднее значение их модулей

$$\overline{\Delta^k y} = \frac{1}{n-k}\sum_{j=1}^{n-k}\left|\Delta^k y_j\right|, \quad k = \overline{1,n-1}.$$

Если при некотором k все конечные разности порядка k или среднее значение их модулей близко к нулю, то следует принять $m = k - 1$.

5.4. Построение регрессионной модели стохастического объекта

Пусть задан некоторый стохастический объект (рис. 2.6) входная и выходная координаты X и Y которого являются случайными величинами.

Рис. 2.6.

На Y влияет не только входная координата X, но и случайная помеха Z (нестабильность режима работы объекта, стохастические воздействия внешней среды, погрешности измерения Y и т.д.). Поэтому нельзя говорить о функциональной зависимости Y от X. В подобных случаях следует говорить о наличии стохастической (вероятносной) связи между переменными X и Y объекта в статике.

Случайные величины X и Y являются зависимыми, если закон распределения вероятностей одной из них зависит от значений другой, т. е.

$F(Y/X = x) = P$ {Y<y при условии X=x}

Функция $F(Y/X=x)$, или просто $F(Y/X)$, называется условным интегральным законом распределения вероятностей Y при X=x. $\frac{dF(Y/X)}{dx} = f(Y/X)$ называется условной плотностью распределения вероятностей.

Предположим, что можно установить $X=x_1$, тогда поведение случайной величины Y будет полностью характеризоваться условной плотностью распределения вероятностей $f(Y/X_1)$. Обозначим условные числовые характеристики Y - математическое ожидание и дисперсию - через $M[Y/X_1]$ и σ^2_{y/x_1}. Затем зафиксируем $X=x_2$ и найдем $f(Y/X_2)$, $M[Y/X_2]$, σ^2_{y/x_2}, положим $X=x_3$, найдем $f(Y/X_3)$, $M[Y/X_3]$, σ^2_{y/x_3}, и т.д.

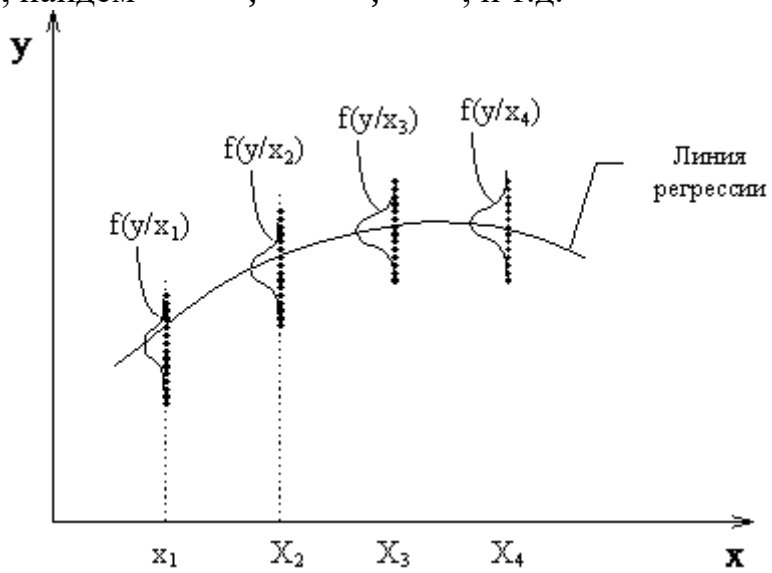


рис. 2.7

Будем считать, что структура $f(y/x)$ не зависит от x, а параметры функции плотности $M[y/x]$ и $\sigma_{y/x}^2$ зависят от того, какое значение x принимает случайная величина X.

Зависимость $M[y/x]$ от x называется регрессионной. Ниже будем говорить только о регрессионной зависимости

$$M[y/x] = f(x). \qquad (5.61)$$

Регрессионная зависимость показывает, как изменяется среднее значение Y при изменении X.

Если соединить плавной линией точки $M[y/x_1]$, $M[y/x_2]$, $M[y/x_3]$ и т.д., то получим линию регрессии. Эта линия есть статическая характеристика объекта.

Уравнением регрессии называют функцию $f(x)$, описывающую линию регрессии. Уравнения регрессии классифицируют на линейные (корреляционные) и нелинейные.

При построении регрессионной модели статики (5.61) объекта широко применяется пассивный метод идентификации. Этот метод применяют при изучении статики объектов, уровень помех которых достаточно велик, а также в тех случаях, когда недопустимо нанесение искусственных возмущений на входе объекта.

Пассивный метод идентификации основан в получении статистической информации об объекте по данным его нормальной эксплуатации. Затем реализации входной x и выходной y величин обрабатываются таким образом, чтобы определить регрессионную модель

$$M[y/x] = f(x, \vec{a}), \qquad (5.62)$$

где $\vec{a}$ - вектор коэффициентов модели.

Определение уравнения регрессии состоит из двух этапов: выбор типа уравнения регрессии и расчет его коэффициентов.

Первый этап обычно осуществляется либо путем эмпирического подбора типа уравнения регрессии по виду корреляционного поля между входными и выходными величинами, либо путем теоретического изучения закономерности физического процесса, отражением которого является стохастическая связь между этими величинами. Иногда оба подхода используются в сочетании друг с другом.

Второй этап - расчет коэффициентов уравнения регрессии - чаще всего выполняется методом наименьших квадратов. При этом минимизация по $\vec{a}$ выражения

$$E = \sum_{i=1}^{n} \left(y_i - M[y/x_i]\right)^2 \qquad (5.63)$$

приводит к оценкам вектора $\vec{a}$. Очевидно, на данной реализации для принятого уравнения модели объекта (5.62) полученные оценки $\vec{a}$ дают

наименьшее СКО реализации, полученной по модели от наблюдаемой реализации $y_i\left(i=\overline{1,n}\right)$ объекта.

Следует отметить, что пассивно - статистический метод имеет целый ряд существенных недостатков по сравнению с активными методами:

1) полученная модель объекта справедлива только в пределах использованного экспериментального статистического материала;

2) трудно разделить эффекты от коррелированности части входных величин многомерного объекта;

3) индивидуальные коэффициенты регрессии не имеют какого-либо физического смысла;

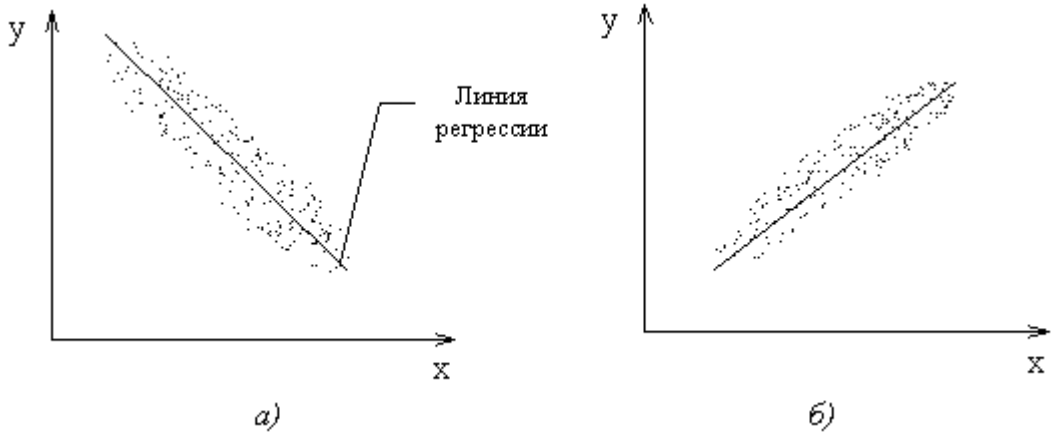
4) не извлекается информация об ошибке опытов;

5) требуется получить большой объем экспериментальных данных и производить трудоемкие вычисления.

Указанные недостатки в значительной степени снижают ценность моделей, полученных пассивно-статистическим методом. К этому методу прибегают только в тех случаях, когда другие методы не могут быть использованы.

Предварительный анализ экспериментального статистического материала составляет основную задачу корреляционного анализа при идентификации стохастического объекта. При этом суть корреляционного анализа сводится к оценке силы стохастической связи между случайными величинами X и Y и установлению вида зависимости между ними в виде уравнения регрессии.

Чтобы предварительно определить наличие корреляционной связи между X и Y, наносят экспериментальные точки $\left(x_i, y_i\right)$ на график и строят корреляционное поле (рис. 2.8).



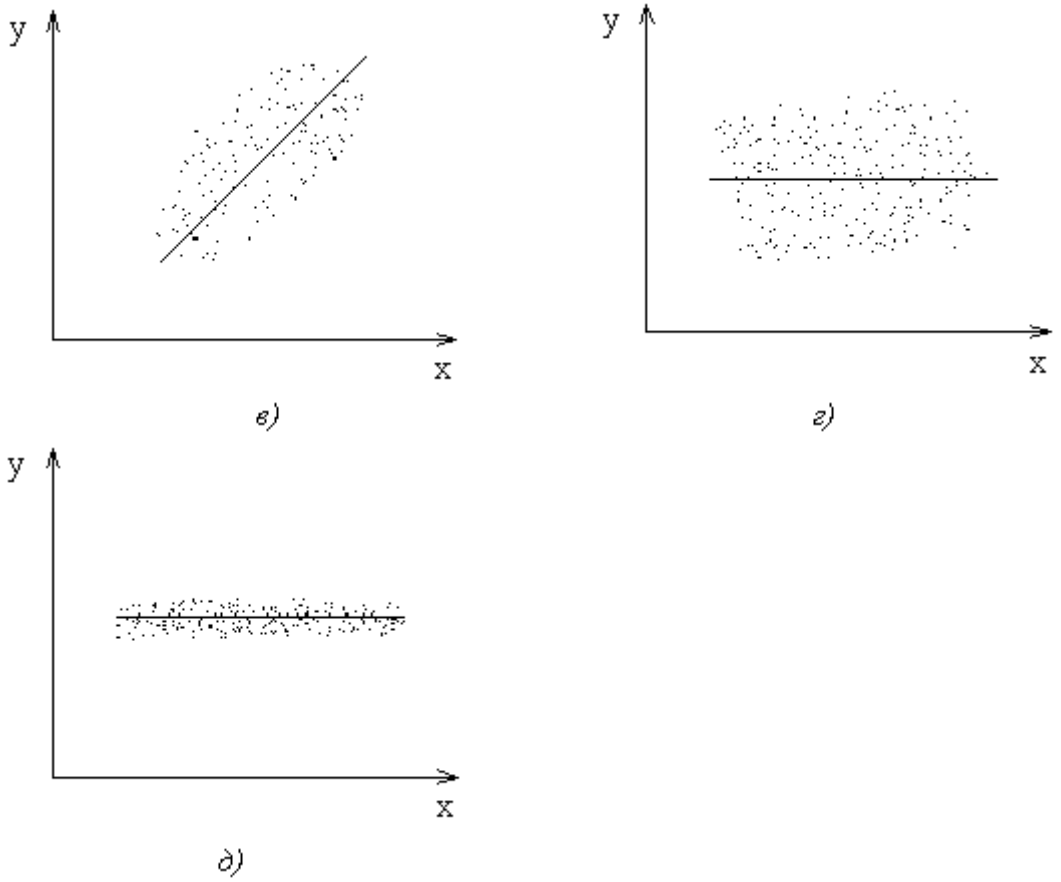


Рис. 2.8

На рис. 2.8 представлены:

а - сильная отрицательная корреляция;

б - сильная положительная корреляция;

в - слабая положительная корреляция;

г, д - отсутствие корреляции.

По тесноте группирования точек вокруг прямой или кривой линии, по наклону линии можно визуально судить о наличии корреляционной связи.

Корреляционное поле характеризует вид связи между X и Y, т.е. наличие линейной или нелинейной зависимости.

Существует три вида корреляции - линейная, нелинейная и множественная корреляция. При линейной корреляции линия регрессии аппроксимируется уравнением прямой, при нелинейной - уравнением кривой. Множественная корреляция определяет связь между многими величинами, и при этом используется уравнение множественной регрессии.

Наиболее распространенной является линейная корреляция.

Понятие корреляции дает возможность судить о том, насколько тесно ложатся экспериментальные точки на аппроксимирующую кривую линии регрессии. Если регрессия определяет предполагаемое соотношение между переменными, то корреляция показывает, насколько хорошо это соотношение отражает действительность.

Задача стохастического анализа объекта ставится таким образом: по данной выборке объема n оценить силу (тесноту) корреляционной связи

между X и Y, найти уравнение приближенной регрессии и оценить допускаемую ошибку.

Оценка тесноты корреляционной связи. В общем виде задача выявления и оценки силы стохастической связи не решена. Существуют показатели, оценивающие те или иные стороны стохастической связи.

Поведение случайной величины Y описывается обычно следующими функциональными и числовыми характеристиками:

- интегральным законом распределения вероятностей $F(Y) = P\{Y < y\}$;
- дифференциальным законом распределения вероятностей $f(y) = \dfrac{dF(Y)}{dy}$ (или плотностью вероятностей);
- математическим ожиданием $M[Y] = \int_{-\infty}^{+\infty} f(y)\, y\, dy$;
- дисперсией $\sigma_y^2 = M\{(Y - M[Y])^2\} = \int_{-\infty}^{+\infty} (Y - M[Y])^2 f(y)\, dy$.

Отметим, что $F(-\infty) = 0$; $F(\infty) = 1$; $F(Y)$ — возрастающая функция. Если F(Y) непрерывна по Y, то $P\{Y = y\} = 0$, однако, $P\{y < Y \le y + dy\} = f(y)dy.$

Кроме того, $\int_{-\infty}^{+\infty} f(y)dy = \int_{-\infty}^{+\infty} dF = 1$.

Поведение системы двух случайных величин X,Y, значения x и y которых изменяются от $-\infty$ до $+\infty$, полностью описывается интегральным законом распределения вероятностей

$$F(x,y) = P\{X < x, Y < y\}$$

или плотностью распределения вероятностей

$$F(x,y) = \frac{\partial^2 F(x,y)}{\partial x\ \partial y},$$

т.е. $F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(x,y)\, dx\ dy. \quad \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\, dx\ dy = 1.$

Отметим, что $F(-\infty,-\infty) = 0$, $F(\infty,\infty) = 1.$ Функция F(x,y) есть неубывающая функция x, y, т.к. с ростом x и y вероятность выполнения двух неравенств X < x, Y < y не может уменьшаться. Далее, $F(x,\infty) = F(x)$, $F(\infty,y) = F(y)$, где F(x) и F(y) - функции распределения X и Y.

По известным f(x,y) всегда можно найти f(x) и f(y).

$$f(x) = \frac{dF(x)}{dx} = \left.\frac{\partial F(x,y)}{\partial x}\right|_{y=\infty} = \frac{dF(x,\infty)}{dx} = \int_{-\infty}^{\infty} f(x,y)dy;$$

$$f(y)=\frac{dF(y)}{dy}=\frac{dF(x,y)}{\partial y}\Big|_{x=\infty}=\frac{dF(\infty,y)}{dy}=\int_{-\infty}^{\infty}f(x,y)dx\ ;$$

Обратная задача - определение f(x,y) по известным f(x) и f(y) в общем случае не имеет решения т.к. эти две функции не характеризуют взаимную зависимость величин X и Y. Но для независимых случайных величин f(x,y) = f(x)f(y).

Основной числовой характеристикой двух случайных величин X и Y является ковариационный (корреляционный) момент

$$K_{xy}=M\left[(X-M[X])(Y-M[Y])\right]=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(x,y)(X-M[X])(Y-M[Y])\,dx\,dy\ .$$

Назовем случайную величину $\dot{X}=X-M[X]$ центрированной, ее значение $\dot{x}$ есть отклонение x от среднего значения M[X]. Тогда

Момент K_{xy} равен среднему значению произведения двух центрированных случайных величин $\dot{X},\dot{Y}$ и характеризует тесноту взаимной связи X и Y.

Действительно, пусть X и Y независимы, тогда f(x,y) = f(x)f(y) и

$$K_{xy}=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(x,y)\dot{x}\dot{y}\,dx\,dy=\int_{-\infty}^{\infty}f(x)\dot{x}\,dx\int_{-\infty}^{\infty}f(y)\dot{y}\,dy\ ;$$

$$\int_{-\infty}^{\infty}f(x)\dot{x}\,dx=\int_{-\infty}^{\infty}f(x)(X-M[X])\,dx=\int_{-\infty}^{\infty}f(x)\,x\,dx-\int_{-\infty}^{\infty}f(x)\,M[X]\,dx=$$

$$=M[X]-M[X]\int_{-\infty}^{\infty}f(x)\,dx=M[X]-M[X]=0\ ;$$

$$K_{xy}=0\ .$$

Обратное утверждение - если $K_{xy}=0$, то X и Y независимы - в общем случае доказать нельзя. Более того, можно найти пример, в котором $K_{xy}=0$, а X и Y - зависимые случайные величины.

Следовательно, величина K_{xy} не является исчерпывающей характеристикой силы связи X и Y. Оказывается, что K_{xy} характеризует только линейную зависимость случайных величин X и Y. Если $K_{xy}=0$, то можно говорить о том, что X и Y линейно независимы (не коррелированны).

Ковариационный момент K_{xy} характеризует разброс случайных величин X и Y относительно их средних. Если $\dot{X}$ мало, то и K_{xy} также будет малым, хотя X и Y в достаточной мере коррелированны между собой.

Размерность K_{xy} не имеет физического смысла, что делает невозможным сравнение ковариационных моментов для двух пар случайных величин X и Y. Поэтому чаще всего используется нормированная величина K_{xy}, т.е. коэффициент корреляции

$$r_{xy} = \frac{K_{xy}}{\sigma_x \sigma_y}, \qquad (5.64)$$

где σ_x^2, σ_y^2 - дисперсии X и Y; $\sigma_x = +\sqrt{\sigma_x^2}$ и $\sigma_y = +\sqrt{\sigma_y^2}$.

Величина r_{xy} безразмерная, кроме того она слабее, чем K_{xy}, зависит от рассеяния X и Y, так как здесь K_{xy} нормируется делением на СКО X и Y от M[X] и M[Y].

Величину r_{xy} можно вычислить по выборке случайных величин X и Y. Допустим, что проведено n испытаний и при каждом отмечались значения случайных величин X, Y. Через $\overline{X}$ и $\overline{Y}$ обозначим их среднее значения:

$$\overline{X} = \widetilde{M}[X] = \frac{1}{n}\sum_{i=1}^{n} x_i \; ; \; \overline{Y} = \widetilde{M}[Y] = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad (5.65)$$

где $\widetilde{M}[X], \widetilde{M}[Y]$ - оценки математического ожидания X и Y.

Тогда выборочный коэффициент корреляции определяется по формуле

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{(n-1)\sigma_x \sigma_y}, \qquad (2.66)$$

где

$$\sigma_x^2 = \frac{1}{(n-1)}\sum_{i=1}^{n} (x_i - \overline{x})^2 \; ; \quad \sigma_x = +\sqrt{\sigma_x^2} \; ; \qquad (5.67)$$

$$\sigma_y^2 = \frac{1}{(n-1)}\sum_{i=1}^{n} (y_i - \overline{y})^2 \; ; \quad \sigma_y = +\sqrt{\sigma_y^2} \; ; \qquad (5.68)$$

σ_x^2, σ_y^2- выборочные дисперсии.

Уменьшение на единицу знаменателя в приведенных выше формулах связано с тем, что величины $\overline{x}$ и $\overline{y}$, относительно которых берутся отклонения, сами зависят от объема выборки.

Для удобства вычислений и повышения точности расчетов преобразуем формулы (5.66), (5.67) и (5.68). Несложные преобразования их приводят к следующим формулам:

$$\sigma_x^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]; \quad \sigma_y^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right]; \qquad (5.69)$$

$$r_{xy} = \frac{1}{(n-1)\sigma_x \sigma_y}\left(\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right). \qquad (5.70)$$

Преимущество формул (5.69), (5.70) в том, что в них нет операций вычитания близких чисел, как в формулах (5.66) - (5.68), приводящих к потере точности вычислений. В формулах (5.69), (5.70) эта операция применяется только один раз.

Коэффициент корреляции не меняется от прибавления к X и Y каких - либо неслучайных слагаемых, от умножения X и Y на положительные числа. Если же одну из величин, не меняя другой, умножить на - 1, то на - 1 умножится и коэффициент корреляции.

Для независимых случайных величин $r_{xy}=0$, но $r_{xy}=0$ может быть и для некоторых зависимых величин, которые при этом называются некоррелированными.

Коэффициент корреляции характеризует не всякую зависимость, а только линейную. Линейная вероятностная зависимость случайных величин X и Y заключается в том, что при возрастании X случайная величина Y имеет тенденцию возрастать (при $r_{xy}>0$) или убывать (при $r_{xy}<0$) по линейному закону. Величина r_{xy} характеризует степень тесноты линейной зависимости $\left(-1\le r_{xy}\le 1\right)$.

При $r_{xy}=\pm 1$ величины X и Y связаны между собой линейной функциональной зависимостью, а при $-1<r_{xy}<1$ связь между этими величинами стохастическая. При $r_{xy}=0$ между X и Y линейной корреляционной связи не существует, однако может существовать нелинейная регрессия.

Линейная регрессионная модель. Пусть с помощью коэффициента корреляции установлено, что зависимость между X и Y существует и она линейная, т.е.

$$M\left[{}^{y}\!/_{x}\right]=a_0+a_1x\ . \qquad (5.71)$$

Надо только определить коэффициенты a_0 и a_1. Определим их методом наименьших квадратов. Построим для этого функцию

$$E=M\left[\left(y-a_0-a_1x\right)^2\right]=E\left(a_0,a_1\right),$$

где M - символ математического ожидания (усреднения на интервале $\left[-\infty,+\infty\right]$).

Для краткости записи обозначим $M\left[X\right]=\overline{X}$, $M\left[Y\right]=\overline{Y}$. Добавим к функции невязки $\left(y-a_0-a_1x\right)$ слагаемое $\overline{Y}+a_1\overline{X}$ и $-\overline{Y}-a_1\overline{X}$ получим

$$E\left(a_0,a_1\right)=M\left[\left(y-a_0-a_1x+\overline{Y}+a_1\overline{X}-\overline{Y}-a_1\overline{X}\right)^2\right]=$$

$$M\left\{\left[\left(Y-\overline{Y}\right)-a_1\left(X-\overline{X}\right)+\left(\overline{Y}-a_0-a_1\overline{X}\right)\right]^2\right\}=M\left[\left(Y-\overline{Y}\right)^2\right]+a_1^2M\left[\left(X-\overline{X}\right)^2\right]+$$

$$+M\left[\left(\overline{Y}-a_0-a_1\overline{X}\right)^2\right]-2a_1M\left[\left(Y-\overline{Y}\right)\left(X-\overline{X}\right)\right]+2M\left[\left(Y-\overline{Y}\right)\left(\overline{Y}-a_0-a_1\overline{X}\right)\right]-$$

$$-2a_1M\left[\left(X-\overline{X}\right)\left(\overline{Y}-a_0-a_1\overline{X}\right)\right]=\sigma_y^2+a_1^2\sigma_x^2+\left(\overline{Y}-a_0-a_1\overline{X}\right)^2-$$

$$-2a_1K_{xy}+2\left(\overline{Y}-a_0-a_1\overline{X}\right)M\left(Y-\overline{Y}\right)-2a_1\left(\overline{Y}-a_0-a_1\overline{X}\right)M\left(X-\overline{X}\right).$$

Так как $M\left[X-\overline{X}\right]=M\left[X\right]-M\left[\overline{X}\right]=\overline{X}-\overline{X}=0$; $M\left[Y-\overline{Y}\right]=0$, получим

$$E\left(a_0,a_1\right)=\sigma_y^2+a_1^2\sigma_x^2+\left(\overline{Y}-a_1\overline{X}-a_0\right)^2-2a_1K_{xy}\ . \qquad (5.72)$$

Из необходимого условия минимума функции (5.72) $\frac{\partial E}{\partial a_0}=0$; $\frac{\partial E}{\partial a_1}=0$ получим следующую систему уравнений для определения a_0,a_1 :

$$\left.\begin{aligned} &\overline{Y} - a_1\overline{X} - a_0 = 0; \\ &a_1\sigma_x^2 + \overline{X}\overline{Y} - a_1\overline{X}^2 - a_0\overline{X} - K_{xy} = 0. \end{aligned}\right\} \qquad (5.73)$$

Первое уравнение системы (5.73) умножим на $\overline{X}$ и вычтем из него второе, получим

$$-a_1\sigma_x^2 + K_{xy} = 0.$$

Отсюда находим

$$a_1 = \frac{K_{xy}}{\sigma_x^2} = r_{xy}\frac{\sigma_y}{\sigma_x}. \qquad (5.74)$$

Из первого уравнения системы (5.73) определяем a_0

$$a_0 = \overline{Y} - r_{xy}\frac{\sigma_y}{\sigma_x}\overline{X}. \qquad (5.75)$$

Линейная регрессионная модель имеет вид

$$M\left[{}^{Y}\!/_{X}\right] = M\left[Y\right] - r_{xy}\frac{\sigma_y}{\sigma_x}M\left[X\right] + r_{xy}\frac{\sigma_y}{\sigma_x}X. \qquad (5.76)$$

Уравнение линейной регрессии, записанное в стандартном виде

$$M\left[{}^{Y}\!/_{X}\right] - M\left[Y\right] = r_{xy}\frac{\sigma_y}{\sigma_x}\left(X - M\left[X\right]\right) \qquad (5.77)$$

наилучшим образом (в смысле величины квадрата ошибок) аппроксимирует зависимость условного среднего Y от X.

Если подставить a_0 и a_1 в (5.72), то получим наименьшее значение суммы квадратов невязок

$$\min E\left(a_0, a_1\right) = \sigma_y^2\left(1 - r_{xy}^2\right). \qquad (5.78)$$

Отсюда следует, что чем ближе r_{xy} к 1, тем лучше качество линейной аппроксимации, тем ближе к линиям регрессии расположены точки $M\left[{}^{Y}\!/_{X}\right]$.

Уравнение (5.77) позволяет прогнозировать среднее значение выхода Y для заданного значения входа X.

Теперь покажем, что $\left|r_{xy}\right| \le 1.$ Если X и Y связаны линейной функциональной связью, то $E\left(a_0, a_1\right) = 0.$ Следовательно, $\sigma_y^2\left(1 - r_{xy}^2\right) = 0.$ Отсюда $\left|r_{xy}\right| = 1.$

Учитывая, что всегда $E \ge 0$, то $\left|r_{xy}\right| \le 1.$

Нелинейная регрессионная модель одномерного объекта. Зададим уравнение регрессии полиномом степени m

$$M\left[{}^{Y}\!/_{X}\right] = a_0 + a_1x + a_2x^2 + \ldots + a_mx^m. \qquad (5.79)$$

Коэффициенты $a_j\left(j = \overline{0, m}\right)$ можно определить методом наименьших квадратов. При этом, как правило, уравнение (5.79) наилучшим образом (в смысле минимума суммы квадратов невязок) аппроксимирует линию регрессии. Коэффициенты $a_j\left(j = \overline{0, m}\right)$ не имеют статистической трактовки.

Для оценки силы связи X и Y в общем случае используется корреляционное отношение

$$\eta_{xy} = \frac{\Delta}{\sigma_y}, \quad (2.80)$$

где $\sigma_y^2 = M\left[\left(Y - M[Y]\right)^2\right]$; $\sigma_y = +\sqrt{\sigma_y^2}$;

$\Delta^2 = M\left[\left(M[Y/X] - M[Y]\right)^2\right]$; $\Delta = +\sqrt{\Delta^2}$,

σ_y^2- дисперсия Y относительно M[Y];

Δ^2- дисперсия линии регрессии $M[Y/X]$, описываемой уравнением (5.79), относительно $M[Y]$. Величина η_{xy} равна отношению СКО $M[Y/X]$ от $M[Y]$ к СКО Y от $M[Y]$ (рис. 2.9). Причем $0 \le \eta_{xy} \le 1$.

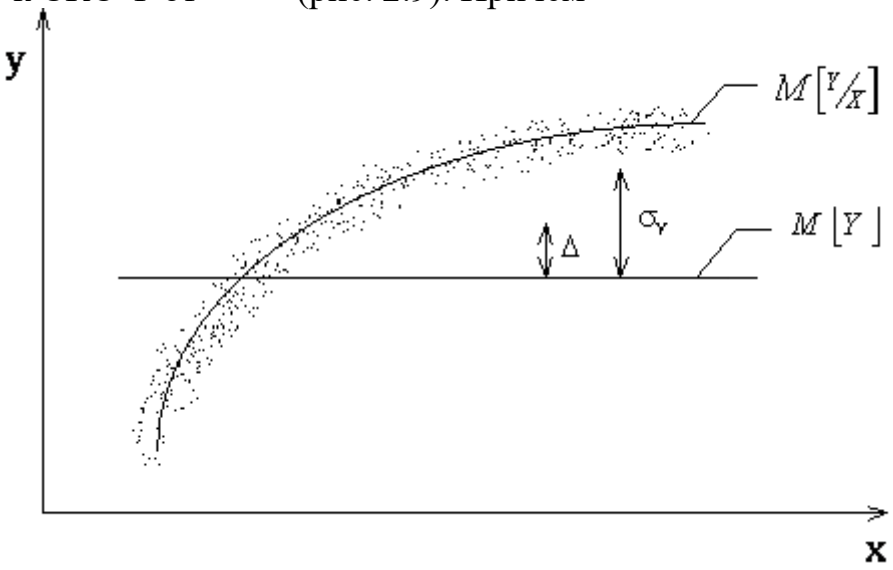


Рис. 2.9

Дисперсию σ_y^2 можно представить суммой

$$\sigma_y^2 = M\left[\left(M[Y/X] - M[Y]\right)^2\right] + M\left[\left(Y - M[Y/X]\right)^2\right] = \Delta^2 + M\left[\left(Y - M[Y/X]\right)^2\right],$$

Последний член характеризует разброс Y относительно $M[Y/X]$. Отсюда находим

$$\Delta^2 = \sigma_y^2 - M\left[\left(Y - M[Y/X]\right)^2\right]. \quad (2.81)$$

Из (2.80) с учетом (2.81) имеем

$$\eta_{xy}^2 = \frac{\Delta^2}{\sigma_y^2} = \frac{\sigma_y^2 - M\left[\left(Y - M[Y/X]\right)^2\right]}{\sigma_y^2};$$

$$\eta_{xy}^2 = 1 - \frac{M\left[\left(Y - M[Y/X]\right)^2\right]}{M\left[\left(Y - M[Y]\right)^2\right]}. \quad (2.82)$$

Если все значения Y лежат на линии регрессии, т.е. абсолютно точно удовлетворяют уравнению (2.79), тогда X и Y связаны функциональной связью и

$$M\left[\left(Y - M\left[Y/X\right]\right)^2\right] = 0\ ; \quad \eta_{xy} = 1\ .$$

Теперь пусть Y не зависит от X, поэтому линия регрессии $M\left[Y/X\right]$ совпадает с $M\left[Y\right]$ и

$$\eta_{xy} = 1 - \frac{M\left[\left(Y - M\left[Y/X\right]\right)^2\right]}{M\left[\left(Y - M\left[Y\right]\right)^2\right]} = 1 - 1 = 0.$$

Чем меньше величина η_{xy}, тем слабее связь между X и Y. Отметим, что при увеличении порядка полинома m, величина $M\left[\left(Y - M\left[Y/X\right]\right)^2\right]$ будет убывать, а $\eta_{xy} \to 1$, хотя, на самом деле, при зтом сила связи между X и Y остается неизменной. Трудность объективного выбора порядка m усложняет и использование корреляционного отношения η_{xy} для решения практических задач.

Множественная линейная регрессия. Множественная регрессия применяется для описания связи входных величин $x_1, x_2, \ldots, x_k$ и выходной случайной величины Y стохастического объекта. Обозначим условное математическое ожидание случайной величины Y при условии, что $X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k$ через $M\left[Y/\vec{X} = \vec{x}\right] = M\left[Y/\vec{x}\right]$, где

$$\vec{X} = \left(X_1, X_2, \ldots, X_k\right)\ ; \quad \vec{x} = \left(x_1, x_2, \ldots, x_k\right).$$

Уравнение линейной множественной регрессии имеет вид

$$M\left[Y/\vec{x}\right] = a_0 + \sum_{j=1}^{k} a_j x_j\ , \tag{5.83}$$

где $a_j\left(j = \overline{1,k}\right)$ находятся методом наименьших квадратов:

$$a_j = r_{x_j y}\frac{\sigma_y}{\sigma_{x_j}}\ , \quad j = \overline{1,k}\ ; \tag{5.84}$$

$$a_0 = M\left[Y\right] - \sum_{j=1}^{k} a_j M\left[X_j\right]; \tag{5.85}$$

$$\sigma^2_{x_j} = M\left[\left(X_j - M\left[X_j\right]\right)^2\right]; \quad K_{x_j y} = M\left[\left(X_j - M\left[X_j\right]\right)\left(Y - M[Y]\right)\right]\ ; \quad j = \overline{1,k}\ ;$$

$$r^2_{x_j} = \frac{K_{x_j y}}{\sigma_{x_j}\sigma_y}, \quad j = \overline{1,k}.$$

Оценки приведенных выше числовых характеристик случайных процессов можно определить по экспериментальным данным. Пусть в результате эксперимента имеем матрицу наблюдений

$$\begin{pmatrix} x_{11} & x_{21} & x_{31} & \mathrm{L} & x_{k1} & y_1 \\ x_{21} & x_{22} & x_{32} & \mathrm{L} & x_{k2} & y_2 \\ \mathrm{L} & \mathrm{L} & \mathrm{L} & \mathrm{L} & \mathrm{L} & \mathrm{L} \\ x_{1n} & x_{2n} & x_{3n} & \mathrm{L} & x_{kn} & y_n \end{pmatrix}, \qquad (5.68)$$

где n - количество опытов; x_{ji} - значение j - го входа в i - том опыте.

Тогда имеем следующие оценки характеристик

$$M\left[X_j\right] = \overline{X}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ji}; \quad M[Y] = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i;$$

$$\sigma_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{Y}\right)^2; \quad \sigma_{x_j}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ji} - \overline{X}_j\right)^2; \quad r_{x_j y} = \frac{\sum_{i=1}^{n}\left(x_{ji} - \overline{X}_j\right)\left(y_i - \overline{Y}\right)}{(n-1)\sigma_{x_j}\sigma_y},$$

где всюду $j = \overline{1,k}$.

Коэффициенты a_j, a_0 получены при условии, что все входные воздействия линейно независимы, т.е. $r_{x_j x_l} \approx 0$, $j = \overline{1,k}$, $l = \overline{1,k}, j \neq l$, где

$$r_{x_j x_l} = \frac{\sum_{i=1}^{n}\left(x_{ji} - \overline{x}_j\right)\left(x_{li} - \overline{x}_l\right)}{(n-1)\sigma_{x_j}\sigma_{x_l}}, \qquad j = \overline{1,k}, l = \overline{1,k}, j \neq l.$$

Рассмотрим общий случай, учитывая возможные связи между факторами (входами).

Перейдем от натурального масштаба к новому, проведя нормировку всех значений случайных величин по формулам:

$$\dot{y}_i = \frac{y_i - \overline{y}}{\sigma_y}; \; \dot{x}_{ji} = \frac{x_{ji} - \overline{x}_j}{\sigma_{x_j}}; \quad j = \overline{1,k}, i = \overline{1,n}. \qquad (5.87)$$

При этом имеем исходный статистический материал в безразмерном масштабе:

$$\begin{pmatrix} \dot{x}_{11} & \dot{x}_{21} & \dot{x}_{31} & \cdots & \dot{x}_{k1} & \dot{y}_1 \\ \dot{x}_{21} & \dot{x}_{22} & \dot{x}_{32} & \cdots & \dot{x}_{k2} & \dot{y}_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \dot{x}_{1n} & \dot{x}_{2n} & \dot{x}_{3n} & \cdots & \dot{x}_{kn} & \dot{y}_n \end{pmatrix}.$$

В новом масштабе имеем:

$$\overline{\dot{x}}_j = 0; \; \overline{\dot{y}} = 0; \; \tilde{\sigma}_{x_j} = 1; \; \tilde{\sigma}_{y_j} = 1, \qquad j = \overline{1,k}.$$

$$r_{x_j y} = \frac{\sum_{i=1}^{n}\left(\dot{x}_{ji} - \overline{\dot{x}}_j\right)\left(\dot{y}_i - \overline{\dot{y}}\right)}{(n-1)\sigma_{x_j}\sigma_y} = \frac{1}{n-1}\sum_{i=1}^{n}\dot{x}_{ji}\dot{y}_i; \qquad j = \overline{1,k}$$

$$r_{x_j x_l} = \frac{1}{n-1}\sum_{i=1}^{n}\dot{x}_j\dot{x}_l; \qquad j,l = \overline{1,k}, \; j \neq l.$$

$$\sigma_{x_j}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\dot{x}_{ji} - \overline{\dot{x}}_j\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\dot{x}_{ji}^2; \qquad j = \overline{1,k}$$

$$\sigma_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\dot{y}_i^2 \quad .$$

Вычисленные по этим формулам выборочные коэффициенты корреляции равны коэффициентам корреляции между переменными, выраженными в натуральном масштабе, т.е. $r_{x_i x_i} = r_{x_i x_i}$; $r_{x_i y} = r_{x_i y}$.

Уравнение регрессии между нормированными переменными не имеет свободного члена, т.к. $\bar{\dot{x}}_j = 0$, $\bar{\dot{y}} = 0$, и имеет вид

$$M\left[\dot{Y}/\vec{\dot{X}}^0\right] = A_1\dot{x}_1 + A_2\dot{x}_2 + \ldots + A_k\dot{x}_k \ . \tag{5.88}$$

Коэффициенты уравнения (5.88) находятся методом наименьших квадратов из условия:

$$E = \sum_{i=1}^{n}\left(\dot{y}_i - M\left[\dot{Y}/\vec{\dot{X}}^0\right]\right)^2 = \sum_{i=1}^{n}\left(\dot{y}_i - A_1\dot{x}_1 - A_2\dot{x}_2 - \ldots - A_k\dot{x}_k\right)^2 \to \min \ .$$

$$\frac{\partial E}{\partial A_j} = 0 \ ; \qquad j = \overline{1,k} \quad ,$$

Которые дают следующую систему уравнений

$$\left.\begin{array}{l} A_1\sum\dot{x}_{1i}^2 + A_2\sum\dot{x}_{2i}\dot{x}_{1i} + A_3\sum\dot{x}_{3i}\dot{x}_{1i} + \ldots + A_k\sum\dot{x}_{ki}\dot{x}_{1i} = \sum\dot{y}\dot{x}_{1i} \ ; \\ A_1\sum\dot{x}_{1i}\dot{x}_{2i} + A_2\sum\dot{x}_{2i}^2 + A_3\sum\dot{x}_{3i}\dot{x}_{2i} + \ldots + A_k\sum\dot{x}_{ki}\dot{x}_{2i} = \sum\dot{y}\dot{x}_{2i} \ ; \\ ---------- \\ A_1\sum\dot{x}_{1i}\dot{x}_{ki} + A_2\sum\dot{x}_{2i}\dot{x}_{ki} + A_3\sum\dot{x}_{3i}\dot{x}_{ki} + \ldots + A_k\sum\dot{x}_{ki}^2 = \sum\dot{y}\dot{x}_{ki} \ , \end{array}\right\} \tag{5.89}$$

где суммирование всюду осуществляется от 1 до n.

Умножим левую и правую части системы уравнений (5.89) на 1/(n-1). В результате при каждом коэффициенте A_j получаются выборочные коэффициенты корреляции $r_{x_i x_i}, r_{x_i y}, r_{x_i x_i} = \sigma_{x_i}^2 = 1$. С учетом этого, а также учитывая $r_{x_i x_i} = r_{x_i x_i}$, получим

$$\left.\begin{array}{l} A_1 + A_2 r_{x_1 x_2} + A_3 r_{x_1 x_3} + \ldots + A_k r_{x_1 x_k} = r_{x_1 y}; \\ A_1 r_{x_2 x_1} + A_2 + A_3 r_{x_2 x_3} + \ldots + A_k r_{x_2 x_k} = r_{x_2 y}; \\ ---------- \\ A_1 r_{x_k x_1} + A_2 r_{x_k x_1} + A_3 r_{x_k x_3} + \ldots + A_k = r_{x_k y}, \end{array}\right\} \tag{5.90}$$

Решая систему линейных уравнений (5.90), находим A_j , $j = \overline{1,k}$. Если все входные воздействия линейно независимы, то $r_{\dot{x}_k \dot{x}_1} = 0$ и имеем: $A_j = r_{\dot{x}_k \dot{y}}$, $j = \overline{1,k}$.

Для изучения тесноты связи между функцией отклика Y и факторами $x_1, x_2, \ldots, x_k$ используют коэффициент множественной корреляции R

$$R = \sqrt{A_1 r_{x_1 y} + A_2 r_{x_2 y} + \ldots + A_k r_{x_k y}} \ . \tag{5.91}$$

Очевидно, что в случае линейно независимых входов

$$R = \sqrt{r_{x_1 y}^2 + r_{x_2 y}^2 + \ldots + r_{x_k y}^2} \ . \tag{5.92}$$

R всегда положителен и принимает значение от 0 до 1. Чем больше R, тем лучше качество предсказаний данной моделью опытных данных.

От уравнения (5.88) можно перейти к уравнению (5.83) с переменными в натуральных масштабах, используя формулы (5.87) нормирования переменных входа X_j:

$$\dot{x}_j = \frac{x_j - \bar{x}_j}{\sigma_{x_j}}, \quad j = \overline{1,k}.$$

Нетрудно получить следующие формулы для пересчета коэффициентов модели:

$$a_j = \frac{\sigma_y}{\sigma_{x_j}}, \quad j = \overline{1,k}; \qquad a_0 = \bar{Y} - \sum_{j=1}^{k} a_j \bar{x}_j.$$

Нелинейная регрессионная модель многомерного объекта. Уравнение нелинейной регрессии стохастического многомерного объекта задается обычно полиномом

$$M\left[Y/\vec{x}\right] = a_0 + \sum_{j=1}^{k} a_j x_j + \sum_{j=1}^{k} b_j x_j^2 + \sum_{l,j=1,j<l}^{k} b_{jl} x_j x_l + \ldots, \tag{5.93}$$

где x_j - входные переменные объекта, $j = \overline{1,k}$.

Коэффициенты уравнения (2.93), как правило, определяются методом наименьших квадратов. Наибольшие трудности вызывает структурная идентификация модели, т.е. выбор порядков полинома по каждому из входов, а также вычисление определителя плохообусловленной матрицы, часто встречающейся при идентификации коэффициентов модели.

Использование метода Брандона для построения уравнения множественной регрессии позволяет избежать эти трудности.

По этому методу уравнение регрессии записывается в виде

$$M\left[Y/\vec{x}\right] = a f_1(x_1) f_2(x_2) \ldots f_k(x_k), \tag{5.94}$$

где $f_j(x_j)$ - неизвестная функция переменной x_j.

Порядок расположения факторов $x_1, x_2, \ldots, x_k$ в уравнении (5.94) не безразличен. Чем сильнее влияние на Y оказывает вход x_j, тем меньше должен быть порядковый номер индекса j. Это необходимо для повышения точности обработки результатов наблюдений.

Вид функции $f_j(x_j)$ выбирается при помощи построения эмпирических линий регрессии. Вначале по точкам выборки $(y_i, \vec{x}_i)$, $\vec{x}_i = (x_{1i}, x_{2i}, \ldots, x_{ki})$, $i = \overline{1,n}$, строятся поле корреляции и эмпирическая линия регрессии $M\left[Y/x_1\right]$ (рис. 2.10).

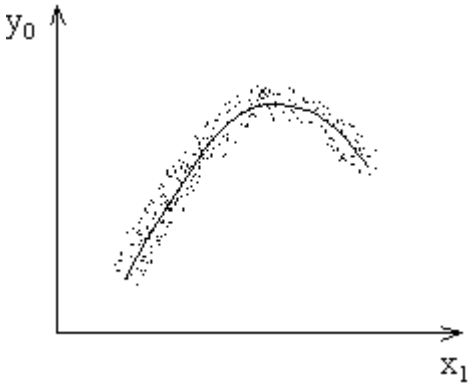


Рис. 2.10

По эмпирической линии регрессии определяется вид зависимости $M\left[Y/x_1\right] = f_1(x_1)$ и методом наименьших квадратов рассчитываются коэффициенты этого уравнения регрессии. Затем составляется выборка новой величины $(y_{1i}, \vec{x}_i)$, где

$$Y_{1i} = \frac{Y_i}{f_1(x_{1i})} \quad , \quad i = \overline{1,n} \tag{5.95}$$

Полученная величина Y_1 не зависит уже от x_1,а определяется только параметрами $x_1, x_2, \ldots, x_k$. Поэтому можно записать

$$M\left[Y_1/\vec{x}\right] = a f_2(x_2) f_3(x_3) \ldots f_k(x_k) \ . \tag{5.96}$$

По точкам новой выборки величин Y_1 и x_2 вновь строятся корреляционное поле и эмпирическая линия регрессии, характеризующая зависимость Y_1 от x_2 (рис. 5.11):

$$M\left[Y/x_2\right] = f_2(x_2) \ . \tag{5.97}$$

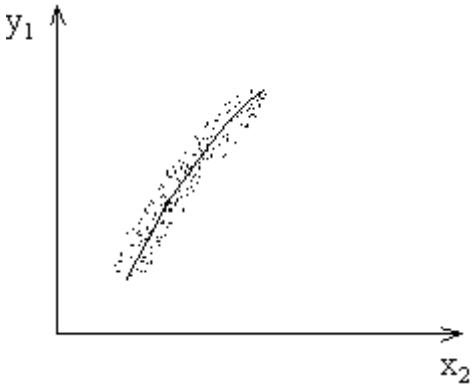


Рис. 2.11

Рассчитываются коэффициенты уравнения (5.97) и вновь составляется выборка новой величины $(y_{2i}, \vec{x}_i)$, где

$$Y_{2i} = \frac{Y_{1i}}{f_2(x_{2i})} = \frac{Y_i}{f_1(x_{1i}) f_2(x_{2i})} \quad , \quad i = \overline{1,n} \ . \tag{5.98}$$

Полученная величина Y_2 не зависит уже от двух факторов x_1 и x_2 и может быть определена из следующего уравнения регрессии:

$$M\left[Y_2/\bar{x}\right] = a f_3(x_3) f_4(x_4) \ldots f_k(x_k) \quad . \tag{5.99}$$

Такая процедура определения функций $f_3(x_3), f_4(x_4)$ и т.д. продолжается до получения выборки величины Y_k:

$$Y_{ki} = \frac{Y_{k-1,i}}{f_k(x_k)} = \frac{Y_{0i}}{f_1(x_{1i}) f_2(x_{2i}) \ldots f_k(x_{ki})} \quad , \; i = \overline{1,n} \; . \tag{5.100}$$

Полученная величина Y_k не зависит от факторов $x_1, x_2, \ldots, x_k$ и определяется коэффициентом исходного уравнения

$$M\left[Y_k/\bar{x}\right] = a = \frac{1}{n}\sum_{i=1}^{n} Y_{ki} \tag{5.101}$$

где n - объем выборки.

Таким образом, окончательно получаем зависимость условного среднего выхода объекта Y от всех факторов (входов) $x_1, x_2, \ldots, x_k$, подставляя в уравнение (5.94) численное значение коэффициента a и найденные выражения функций $f_j(x_j)$, $j = \overline{1,k}$. Таким образом одновременно решаются задачи структурной и параметрической идентификации.

Оценка значимости коэффициентов и адекватности регрессионной модели. После того как уравнение регрессии найдено, необходимо провести статический анализ результатов. Он заключается в проверке значимости коэффициентов модели в сравнении с ошибкой воспроизводимости и оценке адекватности модели эксперименту. Такое исследование называется регрессионным анализом.

Оценку значимости коэффициентов модели выполняют по критерию Стьюдента. Для каждого проверяемого коэффициента a_j вычисляют

$$t_j = \frac{|a_j|}{\sigma_{a_j}} \; , \tag{5.102}$$

где σ_{a_j} - СКО j - го коэффициента.

Как правило, для получения дисперсии воспроизводимости проводится отдельная серия из m повторных опытов при фиксированных значениях входов. Пусть при этом получено y_u^0 , $u = \overline{1,m}$. Тогда дисперсия воспроизводимости определяется по формуле

$$\sigma^2_{\text{воспр}} = \frac{1}{m-1}\sum_{u=1}^{m}\left(y_u^0 - \bar{y}^0\right)^2; \qquad \bar{y}^0 = \frac{1}{m}\sum_{u=1}^{m} y_u^0 \quad . \tag{5.103}$$

Тогда

$$\sigma_{a_j} = \sqrt{\sum_{i=1}^{n}\left(\frac{\partial a_j}{\partial y_i}\right)\sigma^2_{\text{воспр}}} \quad . \tag{5.104}$$

Значения частных производных находят из выражения коэффициентов, найденных, например, методом наименьших квадратов.

Если t_j больше табличного $t(\alpha, f)$ (находят из таблицы распределения Стьюдента) для выбранного уровня значимости $\alpha = 1 - P$ (P - доверительная вероятность) и числа степеней свободы $f = f_{\text{воспр}} = m - 1$, то коэффициент a_j значительно отличается от нуля.

Незначимые коэффициенты исключаются из уравнения регрессии. Оставшиеся коэффициенты пересчитываются заново, поскольку коэффициенты закоррелированы друг с другом.

Адекватность модели проверяется по F - критерию Фишера. Для этого вычисляют дисперсию адекватности

$$\sigma^2_{\text{ад}} = \frac{1}{n-l}\sum_{i=1}^{n}\left(Y_i - M\left[Y/x_i\right]\right)^2, \tag{5.105}$$

где l - число коэффициентов модели, найденных по выборке объема n.

Определяют значение F - критерия $(F \geq 1)$ по формуле

$$F = \begin{cases} \sigma^2_{\text{ад}}/\sigma^2_{\text{воспр}}, & \sigma^2_{\text{ад}} > \sigma^2_{\text{воспр}}; \\ \sigma^2_{\text{воспр}}/\sigma^2_{\text{ад}}, & \sigma^2_{\text{ад}} \leq \sigma^2_{\text{воспр}}. \end{cases} \tag{5.106}$$

5.5. Идентификация динамических объектов

Оператор F0 таких объектов имеет память. В этом случае выход Y в момент времени t отражает не только, точнее не столько состоя ние входа X в этот момент, сколько его значения в предыдущие моменты времени. Объект как бы запоминает возмущение X(t) и реагирует на него лишь через определенный промежуток времени. Поэтому динамические объекты часто называют инерционными.

Здесь будем рассматривать идентификацию непрерывной модели линейного детерминированного стационарного объекта. Эта модель для одномерного случая представляет собой обыкновенное дифференциальное уравнение вида

$$\sum_{i=0}^{p} a_i y^{(i)}(t) = \sum_{j=0}^{l} b_j x^{(j)}(t) \tag{5.107}$$

где $y^{(j)}(t),\ x^{(j)}(t)$ - j-я производная функций y и x, т.е. $y^{(j)} = \dfrac{d^j y}{dt^j},\ x^{(j)} = \dfrac{d^j x}{dt^j},\ a_p = 1,\ l \leq p$.

Модель определяется p + l + 1 параметрами

$$C = (a_0, a_1, \ldots, a_{p-1}, b_0, b_1, \ldots, b_l). \tag{5.108}$$

Структурными параметрами модели (5.107) являются числа p и l, которые должны быть выбраны в процессе структурной идентификации.

Модель (5.107) часто удобно записать в виде системы дифференциальных уравнений первого порядка (в форме Коши), обозначив y1(t) = y(t):

$$\left.\begin{aligned}
&\dot{y}_1(t) = y_2(t);\\
&\dot{y}_2(t) = y_3(t);\\
&\ldots\ldots\\
&\dot{y}_{p-1}(t) = y_p(t);\\
&\dot{y}_p(t) = -\sum_{i=0}^{p-1} a_i y_{i+1}(t) + \sum_{j=0}^{l} b_j x^{(j)}(t).
\end{aligned}\right\} \qquad (5.109)$$

Естественно обобщить эту модель, приведя ее к виду:

$$\left.\begin{aligned}
&\dot{y}_1(t) = \sum_{i=1}^{p} a_{1i} y_i(t) + \sum_{j=0}^{l} b_{1j} x^{(j)}(t);\\
&\ldots\ldots\\
&\dot{y}_p(t) = \sum_{i=1}^{p} a_{pi} y_i(t) + \sum_{j=0}^{l} b_{pj} x^{(j)}(t).
\end{aligned}\right\} \qquad (5.110)$$

Если модель (5.107) сводится к виду (5.110), то последняя в общем виде не сводится к виду (5.107). Поэтому модель (5.110) следует считать наиболее общей формой непрерывной параметрической модели динамического детерминированного линейного стационарного объекта. Эту модель удобно записывать в векторной форме:

$$\vec{Y}'(t) = A\vec{Y}(t) + B\vec{X}'(t), \qquad (5.111)$$

где $\vec{Y}(t) = (y_1(t), y_2(t), \ldots, y_p(t))$ - вектор состояния,

$\vec{X}'(t) = (x(t), x^{(1)}(t), \ldots, x^{(l)}(t))$ - вектор возмущения,

A,B - матрицы коэффициентов модели, причем A = (aqi) - квадратная матрица (p´p), B = (bqj) - прямоугольная матрица (p´(l+1)).

Таким образом, идентифицируемыми параметрами модели (5.111) являются элементы матрицы A и B, т.е. p(p+l+1) параметров модели, которые и образуют две матрицы неизвестных параметров C = (A, B).

Исходной информацией для идентификации в этом случае являются реализации x(t), y(t) в промежутке 0 £ t £ T, характеризующие состояния входа и выхода идентифицируемого объекта

< x(t), y(t) > (0 £ t £ T). (5.112)

Распространим сказанное на многомерный случай, когда число входов n>1, а число выходов m=1, т.е. вектор входа имеет вид:

$\vec{X}(t) = (x_1(t), \ldots, x_n(t))$.

В силу линейности вектор состояния $\vec{Y}$ образуется как сумма векторов:

$$\vec{Y} = \sum_{i=1}^{n} \vec{Y}_i \qquad (5.113)$$

где вектор

$$\vec{Y}'_i(t) = A_i y_i(t) + B_i \vec{X}'_i(t), \qquad (5.114)$$

отражает влияние i-го входа.

Здесь $\vec{X}'_i(t) = (x_i(t), x_i^{(1)}(t), \ldots, x_i^{(l)}(t))$ - вектор возмущения i-го входа; Ai, Bi - матрицы влияния i-го входа.

Теперь, подставляя (5.113) и (5.114) в (5.111), можно записать уравнение для линейной динамической модели со многими входами

$$\vec{Y}'(t) = \sum_{i=1}^{n} A_i y_i(t) + \sum_{i=1}^{n} B_i \vec{X}'_i(t). \qquad (5.115)$$

Эта модель имеет np(p+l+1) идентифицируемых параметров. При этом информация о работе объекта, необходимая для идентификации, имеет вид:

< $\vec{X}(t)$, y(t) > (0 £ t £ T).

Случай n>1 и m>1 сводится к m-кратному повторению предыдущего.

Идентификация параметров модели. Исходной информацией для построения процедуры идентификации является вид идентифицируемой модели (5.107) и наблюдения (5.112). Необходимо определить p+l+1 параметров

$a_i, (i = \overline{0, p-1})$ и $b_j, (j = \overline{0, l})$.

В общем случае, т.е. при произвольных значениях параметров ai и bj, при подстановке реализаций (5.112) в уравнение модели (5.107) равенство в (5.107) выполняться не будет. Естественно подобрать параметры ai и bj такими, чтобы равенство (5.107) было восстановлено или хотя бы правая и левая части выражения (5.107) отличались друг от друга наименьшим образом. Чтобы выполнять это условие, необходимо построить функцию невязки правой и левой частей этого уравнения и минимизировать ее с помощью вариации параметров ai и bj. Сформулируем требования, которым должна удовлетворять эта функция невязки:

1) она не должна принимать отрицательных значений;

2) ее минимум должен соответствовать решению поставленной задачи;

3) этот минимум должен быть равен нулю в случае совпадения структуры модели и объекта.

Этим условиям удовлетворяет невязка в виде среднего квадрата разности правой и левой части уравнения модели (5.107) при подстановке туда функций x(t), y(t) - наблюдений объекта, т.е.

$$E(C) = \int_0^T \left[\sum_{i=0}^{p} a_i y^{(i)}(t) - \sum_{j=0}^{l} b_j x^{(j)}(t) \right]^2 dt. \qquad (5.116)$$

Минимизируя это выражение по параметрам ai и bj, которые образуют вектор (2.108), определим искомые параметры. Такая задача минимизации формулируется в виде

$$E(C) \to \min_C \Rightarrow C^* = (a_0^*, \ldots, a_{p-1}^*, b_0^*, \ldots, b_l^*). \qquad (5.117)$$

Результат минимизации C* и дает значения идентифицируемых параметров, для которых при совпадении структуры объекта и модели E(C*) = 0.

Для решения задачи (5.117) воспользуемся тем, что функция E(C) имеет простую структуру относительно идентифицируемых параметров — это квадратичная форма.

Задача минимизации гладких функций (а квадратичная форма — гладкая функция) сводится к решению системы уравнений, которая образуется в результате приравнивания нулю частных производных:

$$\frac{\partial E(C)}{\partial a_i} = 0, \ (i = \overline{0, p-1});$$

$$\frac{\partial E(C)}{\partial b_j} = 0, \ (j = \overline{0, l}).$$

После очевидных преобразований получим следующую систему линейных алгебраических уравнений:

$$\left.\begin{aligned} &\sum_{i=0}^{p} a_i \phi_{ik} - \sum_{j=0}^{l} b_j \lambda_{jk} = 0, \ (k = \overline{0, p-1}); \\ &\sum_{i=0}^{p} a_i \lambda_{qi} - \sum_{j=0}^{l} b_j \nu_{qj} = 0, \ (q = \overline{0, l}), \end{aligned}\right\} \qquad (5.118)$$

где, напомним, ар = 1 и

$$\left.\begin{aligned} &\phi_{ik} = \int_0^T y^{(i)}(t) y^{(k)}(t)\, dt; \\ &\lambda_{jk} = \int_0^T x^{(j)}(t) y^{(k)}(t)\, dt; \\ &\nu_{iq} = \int_0^T x^{(i)}(t) x^{(q)}(t)\, dt, \end{aligned}\right\} \qquad (5.119)$$

где ϕ_{ik}, λ_{jk}, ν_{iq} - коэффициенты системы уравнений (5.118).

Таким образом, задача идентификации линейной динамической стационарной системы сведена к решению системы линейных алгебраических уравнений, что осуществляется стандартными вычислительными методами.

Однако для вычисления выражений (5.119) необходимо знать производные входного x(t) и выходного y(t) сигналов объекта. Их можно получить двумя способами:

1) численное дифференцирование - связан с локальной (точечной) оценкой производных

$$\left.\begin{aligned} &\dot{z}(t_j) = \frac{1}{\Delta t}\left[z(t_j) - z(t_j - \Delta t)\right]; \\ &\ddot{z}(t_j) = \frac{1}{\Delta t}\left[\dot{z}(t_j) - \dot{z}(t_j - \Delta t)\right]; \\ &\ldots\ldots \\ &z^{(i)}(t_j) = \frac{1}{\Delta t}\left[z^{(i-1)}(t_j) - z^{(t-1)}(t_j - \Delta t)\right], \end{aligned}\right\} \qquad (5.120)$$

где Dt - интервал, характеризующий базу оценок.

Но этот способ обладает двумя недостатками.

Во-первых, точность оценки производных очень быстро падает с ростом порядка производной. Практически удовлетворительно не удается оценить производные выше второго порядка.

Во-вторых, для оценки i-й производной при t= 0 необходимо иметь значения функции при t<0, т.е. z(-Dt), z(-2Dt),...,z(-iDt). Так как этих значений нет в измерениях, то приходится для (5.119) брать пределы интегрирования не [0,T], а [iDt,T], где i - наибольший порядок производной подынтегральной функции.

2) в ряде случаев целесообразно использовать разложение z(t) в ряд по системе заданных в аналитическом виде функций. Тогда полученную аналитическую зависимость, аппроксимирующую z(t), можно дифференцировать нужное число раз.

Все вышеизложенное относилось к простейшей модели (5.107). Если же объект описывается обобщенной моделью (5.111), то функция невязки будет не одна. Количество их будет равно p по числу уравнений (5.110). Для k-ой невязки получаем из k-го уравнения модели (5.110):

$$E_k(a_{k1},\dots,a_{kp},b_{k0},\dots,b_{kl}) = \int_0^T \left[\dot{y}_k(t) - \sum_{i=1}^{p} a_{ki} y_i(t) - \sum_{j=0}^{l} b_{kj} x^{(j)}(t) \right]^2 dt, \qquad (5.121)$$

где $k = \overline{1,p}.$

Дифференцируя эти невязки по коэффициентам a и b и приравнивая производные нулю, получаем систему линейных уравнений, откуда определяются идентифицируемые параметры объекта:

$$\frac{\partial E_k}{\partial a_{ki}} = 0, \ \frac{\partial E_k}{\partial b_{kj}} = 0, \ \left(i = \overline{1,p},\ j = \overline{0,l},\ k = \overline{1,p}\right). \qquad (5.122)$$

Теперь рассмотрим условия применимости рассмотренного метода идентификации. Как было показано, он сводится к решению системы линейных алгебраических уравнений (5.118) или аналогичной ей системы, получающейся из (5.122). Однозначность решения задачи идентификации требует однозначности решения системы (5.118). А это выполняется только в случае, если ее определить

$$\Delta = \begin{vmatrix} \phi_{01}\cdots\phi_{p1} & \lambda_{11}\cdots\lambda_{l1} \\ \cdots & \cdots \\ \phi_{0p-1}\cdots\phi_{pp-1} & \lambda_{1p-1}\cdots\lambda_{lp-1} \\ \lambda_{10}\cdots\lambda_{1p} & \nu_{11}\cdots\nu_{l1} \\ \cdots & \cdots \\ \lambda_{l0}\cdots\lambda_{lp} & \nu_{1l}\cdots\nu_{ll} \end{vmatrix} \neq 0. \qquad (5.123)$$

Это и есть условие однозначности решения задачи идентификации. Если оно не выполняется, то следует либо взять другую реализацию входа и выхода объекта <x(t), y(t)>, либо уменьшить количество

идентифицируемых параметров, т.е. уменьшить числа p и l, тем самым понизить порядок модели.

Обе из указанных мер могут привести к выполнению условия (5.123). Если же, несмотря на предпринятые меры, это условие по-прежнему не выполняется, то следует снова обратиться к одной из них и т.д. до тех пор, пока не выполнится условие (5.123). При этом следует помнить, что если первая мера (новая реализация) не всегда приводит к цели, то вторая (понижение порядка модели) при последовательном применении всегда приводит к выполнению условия (5.123).

Аналогично формулируется условие однозначности для обобщенной модели. Здесь определитель системы (5.122) будет более высокого порядка.

5.6. Идентификация закона распределения случайной величины

Случайная величина. В инженерной практике необходимо исследовать не только детерминированные, но и стохастические процессы. Практически все процессы в технологических и технических объектах выполняются в непрерывно меняющихся непредвиденным образом условиях. В связи с этим приходится анализировать случайные величины в этих объектах.

Под случайной величиной понимают величину, принимающую в результате испытания значение, которое принципиально нельзя предсказать исходя из условий опыта. Случайная величина обладает целым набором допустимых значений, но в результате каждого отдельного опыта принимает лишь какое-то одно из них. Она может принимать различные значения даже при неизменном комплексе основных факторов.

Различают дискретные и непрерывные случайные величины. Возможные значения дискретных случайных величин можно заранее перечислить. Значения непрерывной случайной величины не могут быть заранее перечислены, они непрерывно заполняют некоторый промежуток. Чтобы полностью охарактеризовать случайную величину, необходимо не только указывать, какие значения она может принимать но и как часто.

Законы распределения случайной величины. При решении многих прикладных задач необходимые вероятностные характеристики соответствующих случайных величин неизвестны исследователю и должны определяться по экспериментальным данным. Такое статистическое описание результатов наблюдений, построение и проверка различных математических моделей, использующих понятие вероятности, составляет основное содержание математической статистики.

Фундаментальными понятиями статистической теории являются понятия вероятности и распределения вероятности случайной величины.

Пусть дискретная случайная величина Z может принимать в результате опыта значения $z_1, z_2, \ldots, z_n$. Отношение числа опытов n_i, в результате которых случайная величина Z приняла значение z_i, к общему числу производимых опытов n называется частотой проявления события $Z = z_i$ (частотное определение вероятности). Частота n_i/n сама является случайной величиной и меняется в зависимости от количества произведенных опытов. Но при большом числе опытов она имеет тенденцию стабилизироваться около некоторого значения P_i, называемого вероятностью события $Z = z_i$ (статистическое определение вероятности):

$$P_i = P\{Z = z_i\} = \lim_{n\to\infty} \frac{n_i}{n} \qquad (5.124)$$

Сумма вероятностей всех возможных значений случайной величины равна единице

$$\sum_{i=1}^{n} P_i = 1,$$

так как тот факт, что случайная величина примет в результате опыта одно из своих значений, есть достоверное событие. Эта суммарная вероятность распределена определенным образом между отдельными значениями.

Дискретную случайную величину можно полностью задать вероятностным рядом, указав вероятность P_i для каждого значения z_i:

$$\begin{array}{c} z \\ \cdot\cdot \\ \cdot\cdot \end{array}$$

Всякое соотношение, устанавливающее связь между возможными значениями случайной величины и соответствующими им вероятностями, называется законом распределения. Вероятностный ряд является одним из видов законов распределения случайной величины.

Распределение непрерывной случайной величины нельзя задавать при помощи вероятностей отдельных значений. Число значений так велико, что для большинства из них вероятность принять эти значения равна нулю, т.е. событие может произойти, а вероятность его равна нулю. Для непрерывных случайных величин изучается вероятность того, что в результате опыта значение случайной величины попадет в некоторую заданную область.

Аналитическими выражениями законов распределения случайной величины являются функции распределения вероятностей - интегральная и дифференциальная.

Интегральная функция распределения $F(z)$ случайной величины Z показывает вероятность того, что случайная величина не превышает некоторого заданного или текущего значения z, т.е.

$$F(z) = P\{Z \le z\}. \qquad (5.126)$$

Следовательно, вероятность того, что значение случайной величины Z заключено между z_1и z_2, равна разности значений функции распределения, вычисленных в двух этих точках:

$$P\{z_1 < Z \le z_2\} = F(z_2) - F(z_1). \qquad (5.127)$$

Аналогично,

$$P\{Z > z\} = 1 - F(z). \qquad (5.128)$$

Интегральная функция распределения случайной величины обладает следующими свойствами:

1) $\lim\limits_{z \to -\infty} F(z) = F(-\infty) = 0$;
2) $l\ \underset{z \to \infty}{im} F(z) = F(\infty) = 1$;
3) $F(z) \ge 0$ для всех z;
4) $F(z_2) \ge F(z_1)$ если $z_2 > z_1$.

В виде функции распределения можно задать распределение как непрерывной, так и дискретной случайной величины. Как видно из определения, $F(z)$ есть неубывающая функция z. Ордината кривой $F(z)$, соответствующая точке z_1, представляет собой вероятность того, что случайная величина Z при испытании окажется меньше z_1.

Интегральная функция распределения дискретной случайной величины всегда есть разрывная ступенчатая функция, скачки которой происходят в точках, соответствующих возможным значениям случайной величины, и равны вероятностям этих значений. Сумма всех скачков равна единице.

Возможные виды интегральных функций распределения изображены на рис. 2.12 для непрерывной (а) и дискретной (б) случайной величины:

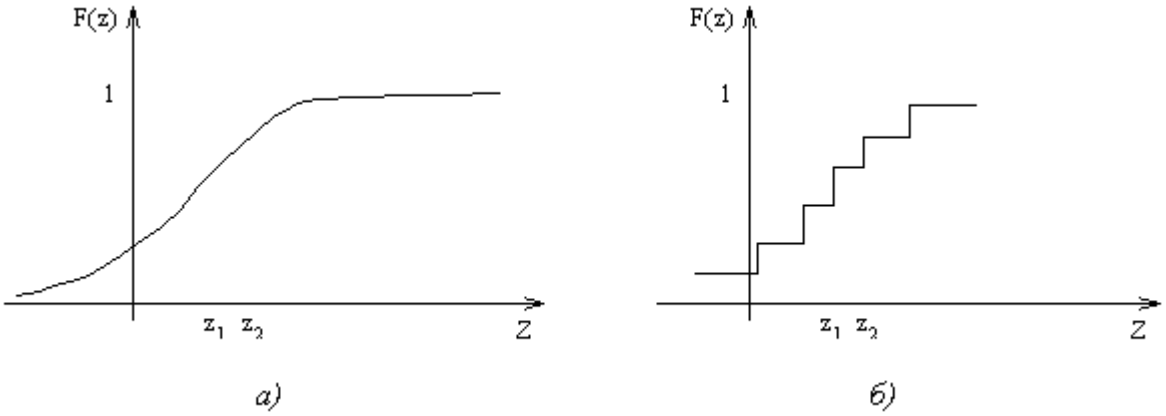


Рис 2.12.

Если функция $F(z)$ дифференцируема для всех значений случайной величины Z, то закон распределения вероятностей может быть выражен в аналитической форме также с помощью дифференциальной функции распределения вероятностей:

$$f(z) = \frac{dF(z)}{dz} = \lim_{\Delta z \to 0} \frac{P\{z < Z \le z + \Delta z\}}{\Delta z} \quad (\Delta z > 0) \qquad (5.129)$$

Таким образом, значение функции $f(z)$ приближено равно отношению вероятности попадания случайной величины в интервал $(z, z+\Delta z)$ к длине Δz этого интервала, когда Δz - бесконечно малая величина. Поэтому функцию $f(z)$ называют также функцией плотности распределения вероятностей (или короче - функцией плотности вероятности).

Отметим основные свойства функции $f(z)$:

1) $f(z) \ge 0$;

2) $\int_{-\infty}^{\infty} f(z)dz = 1$;

3) $\int_{-\infty}^{z} f(x)dx = F(z)$;

4) $\lim_{|z| \to \infty} F(z) = 0$

(x - переменная интегрирования).

С помощью дифференциальной функции распределения вычисляется вероятность нахождения случайной величины в любой области из множества ее возможных значений. Например,

$$P\{Z \le z_1\} = \int_{-\infty}^{z_1} f(z)dz; \qquad (5.130)$$

$$P\{Z > z_2\} = \int_{z_2}^{\infty} f(z)dz; \qquad (5.131)$$

$$P\{z_1 < Z \le z_2\} = \int_{z_1}^{z_2} f(z)dz. \qquad (5.132)$$

Для непрерывной случайной величины вероятность можно определить как относительную долю площади под кривой плотности распределения вероятностей $f(z)$. Так, например, вероятность того, что случайная величина Z примет значение, меньшее z_1, равна относительной доле площади под кривой $f(z)$ слева от точки z_1(рис. 2.13,а); вероятность того, что эта величина Z примет значение, большее z_2, равна относительной доле площади под кривой $f(z)$ справа от точки z_2(рис. 2.13,б); вероятность того что она примет значение, заключенное между значениями z_1и z_2равна относительной доле площади под кривой $f(z)$ между точками z_1и z_2(рис. 2.13,в).

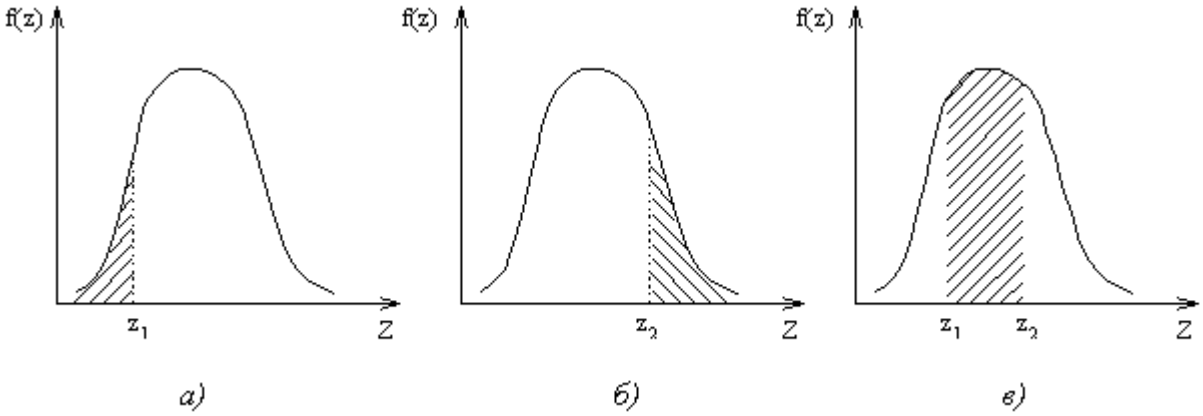


Рис. 2.13.

Теоретическое и эмпирическое распределение. Распределение вероятностей называется теоретическим. Оно может быть установлено для соответствующей случайной величины из анализа генеральной совокупности.

Под генеральной совокупностью понимают все допустимые значения случайной величины. Содержательный смысл этого понятия состоит в том, что предполагается существование некоторых вполне определенных свойств, неслучайных закономерностей присущих данной совокупности, тех свойств, которые и должны быть выяснены исследователем. Фактически эти свойства являются объективным отображением вероятностных свойств изучаемого объекта, которые могут быть охарактеризованы с помощью соответствующих законов распределения вероятностей или связанных с ними числовых параметров. Считается, что указанные свойства не изменяются во времени и присущие генеральной совокупности неслучайные закономерности сохраняют постоянным свой характер, т.е. являются устойчивыми.

На практике исследователь всегда располагает лишь ограниченным числом значений случайной величины, представляющим собой некоторую выборку из генеральной совокупности.

Выборка это конечный набор значений случайной величины, полученный в результате наблюдений. Число элементов выборки называется ее объемом. Если, например, $z_1, z_2, \ldots, z_n$ - наблюдаемые значения случайной величины Z (возможно, и совпадающие), то объем данной выборки равен n.

Выборка называется репрезентативной (представительной), если она достаточно полно характеризует генеральную совокупность. Для обеспечения репрезентативности выборки чаще всего используется случайный выбор элементов (рандомизация). При таком выборе предполагается, что каждая возможная выборка фиксированного объема имеет одну и ту же вероятность выбора, а последовательные наблюдения взаимно независимы.

Смысл статистических методов заключается в том, что по выборке ограниченного объема n, т.е. по некоторой части генеральной

совокупности, высказывать обоснованное суждение о ее свойствах в целом. Подобное суждение может быть получено путем построения эмпирических (выборочных) аналогов вероятностных характеристик исследуемой величины, иначе говоря, путем оценивания параметров (характеристик) генеральной совокупности с помощью некоторых подходящих функций от результатов наблюдений - оценок.

При многократном извлечении выборок одного и того же объема и последующем нахождении множества оценок одного и того же параметра получаются различные числовые значения этих оценок, изменяющиеся от одной выборке к другой случайным образом. Иными словами любая оценка произвольного параметра α есть случайная величина. В этом ее принципиальное отличие от самого оцениваемого параметра α, являющегося неслучайным.

Оценка $\tilde{\alpha}$ параметра α называется состоятельной, если при неограниченном увеличении объема выборки n значение $\tilde{\alpha}$ с полной мерой достоверности (с вероятностью единица) стремится к своему теоретическому значению α. Состоятельность оценки гарантирует исследователю увеличение точности оценивания с ростом n и то, что хотя бы в пределе при $n \to \infty$ он может получить точное значение α.

Оценка $\tilde{\alpha}$ называется несмещенной, если $m_{\tilde{\alpha}} = \alpha$ для любого n ($m_{\tilde{\alpha}}$ - математическое ожидание $\tilde{\alpha}$). Несмещенность означает отсутствие систематической погрешности при оценивании параметра α.

Предположим, что в результате эксперимента получена выборка $z_1, z_2, \ldots, z_n$ значений случайной величины Z. Пусть z - некоторая точка оси z. Обозначим через n_z число (частота) выборочных значений, расположенных левее z на этой же оси. Отношение n_z/n представляет собой относительную частоту наблюдаемых в выборке значений случайной величины Z, меньших z. Эта частота есть функция от z. Обозначим ее $F^*(z)$:

$$F^*(z) = \frac{n_z}{n}. \qquad (5.133)$$

Функция распределения $F^*(z)$, получаемая по выборке, называется эмпирической или выборочной функцией распределения (в отличии от распределения генеральной совокупности, или теоретического распределения). Она устанавливает распределение относительных частот, т.е. определяет для каждого значения z относительную частоту события $Z < z$.

Для каждой выборки эмпирическая функция распределения будет своей, но все эмпирические функции распределения одной и той же случайной величины будут иметь нечто общее, что является информацией о функции распределения этой случайной величины.

Доказано, что с вероятностью 1 при $n \to \infty$ максимальная разность между функциями распределения случайной величины $F^*(z)$ и $F(z)$ стремится к нулю (теорема Гливенко):

$$P\left\{\sup\left|F^*(z) - F(z)\right| \to 0\right\} = 1; \qquad (5.134)$$

Практически это означает, что при достаточно большой выборке функцию распределения генеральной совокупности приблизительно можно заменить функцией распределения выборки.

Эмпирическая функция распределения обладает следующими свойствами:

1) $F^*(z) \in [0;1] \forall z$, т.е. значения эмпирической функции принадлежат отрезку [0;1];

2) $F^*(z_2) \geq F^*(z_1)$,если $z_2 > z_1$, т.е. $F^*(z)$ неубывающая функция;

3) если z_1- наименьшее, а z_2- наибольшее значение случайной величины Z, то $F^*(z) = 0$ при $Z \leq z_1$ и $F^*(z) = 1$ при $Z > z_2$.

График эмпирической функции распределения имеет вид, приведенный на рис. 2.12,б.

Полигоном эмпирического распределения называют ломаную, отрезки которой соединяют точки $(z_1, n_1), (z_2, n_2), \ldots, (z_n, n_n)$, где z_i- элементы выборки, а n_i- соответствующие им частоты, или точки $(z_1, P_1^*), (z_2, P_2^*), \ldots, (z_n, P_n^*)$, где $P_i^* = n_i/n$ - относительные частоты, соответствующие z_i.

При обработке выборок больших объемов весь отрезок, в котором заключены все наблюдаемые значения случайной величины Z, разбивают на ряд интервалов длины h и находят n_i- число элементов выборки, попавших в i -ый интервал.

Гистограммой эмпирического распределения называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат интервалы длины h, а высоты равны отношению n_i/h (плотность частоты). Площадь i -го прямоугольника равна $h\frac{n_i}{h} = n_i$. Площадь гистограммы частот равна сумме всех частот, т.е. объему выборки n.

Если учитываются не абсолютные частоты n_i, а относительные $P_i^* = n_i/n$, тогда гистограмма эмпирического распределения представляет собой ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат интервалы длины h, а высоты равны отношению P_i^*/h (плотность относительной частоты). Площадь i -го прямоугольника равна $h\frac{P_i^*}{h} = P_i^*$. Площадь гистограммы относительных частот равна сумме всех относительных частот, т.е. единице.

На рис 2.14 приведен возможный вид гистограммы эмпирического распределения случайной величины.

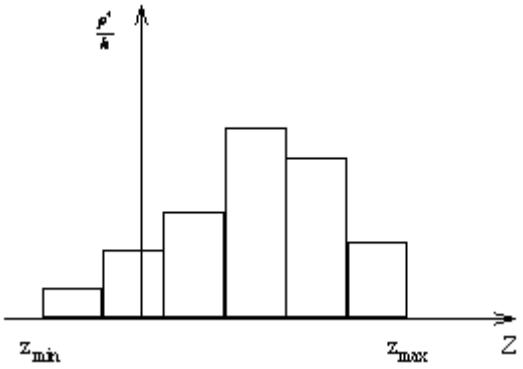


Рис. 2.14

Числовые характеристики случайной величины. Интегральная и дифференциальная функции распределения являются исчерпывающими вероятностными характеристиками случайной величины. Однако в прикладных задачах некоторые основные свойства случайных величин могут быть определены более просто с помощью определенных числовых параметров. Наибольшую роль среди них на практике играют два параметра, характеризующие центр рассеяния (центр распределения) случайной величины и степень ее рассеяния вокруг этого центра. Наиболее распространенной характеристикой центра распределения является математическое ожидание m_z случайной величины z, часто называемое также генеральным средним значением или начальным моментом первого порядка.

Для дискретных случайных величин

$$m_Z = M|Z| = \sum_{i=1}^{N} z_i P_i \quad , \qquad (5.135)$$

где $z_i \left(i = \overline{1, N}\right)$ - элементы генеральной совокупности;

P_i - соответствующие им вероятности.

Для непрерывных случайных величин математическое ожидание выражается интегралом:

$$\mathrm{m}_Z = M[z] = \int_{-\infty}^{\infty} z f(z) dz \quad . \qquad (5.136)$$

Степень рассеяния случайной величины z относительно m_z может быть охарактеризована с помощью генеральной дисперсии σ_Z^2. σ_Z^2 есть математическое ожидание квадрата отклонения случайной величины от ее математического ожидания, т.е.

$$\sigma_Z^2 = M\left[\left(Z - m_Z\right)^2\right]. \qquad (5.137)$$

Для дискретной случайной величины

$$\sigma_Z^2 = \sum_{i=1}^{N} \left(Z_i - m_Z\right)^2 P_i \quad , \qquad (5.138)$$

для непрерывной

$$\sigma_Z^2 = \int_{-\infty}^{\infty} (z - m_Z)^2 f(z)dz \quad . \qquad (5.139)$$

Дисперсия σ_Z^2 называется также вторым центральным моментом случайной величины z (первый центральный момент всегда равен 0).Другие обозначения для дисперсии $D_Z, D[z], s^2$.

Если $f(z)$ все с большей степени концентрируются вблизи m_z,то значения σ_Z^2 уменьшаются. Если же имеются весьма удаленные от m_z значения случайной величины z и для них $f(z)$ не слишком мала, то дисперсия σ_Z^2 увеличивается.

Квадратный корень из дисперсии σ_Z^2 называется средним квадратическим отклонением (или стандартом) σ_Z :

$$\sigma_Z = +\sqrt{\sigma_Z^2} \quad . \qquad (5.140)$$

Определение оценок математического ожидания $\overline{Z}$,дисперсии S_Z^2 и среднего квадратического отклонения S_Z по выборке объема n производят по формулам:

$$\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} z_i \quad ; \qquad (5.141)$$

$$S_Z^2 = \frac{1}{n-1}\sum_{i=1}^{n} (z_i - \overline{Z})^2 \quad ; \qquad (5.142)$$

$$S_Z = +\sqrt{S_Z^2} \quad . \qquad (5.143)$$

Отметим, что множитель $1/(n-1)$ (вместо $1/n$) в формуле (5.142) вводится для получения несмещенной оценки дисперсии.

Оценка результатов наблюдений случайной величины. Рассмотрим некоторые методы статистических оценок. Важнейшая задача математической статистики, решение которой позволило бы, в принципе, решить и все остальные задачи - это нахождение функции распределения наблюдаемой случайной величины.

Пусть имеется экспериментальная выборка большого объема n случайной величины z. Проверяется гипотеза о том, что функция распределения случайной величины есть $P = f(Z)$. Надо установить вид этого теоретического распределения и его адекватность.

Решение поставленной задачи включает следующие этапы:

1) построение эмпирической кривой (гистограммы) распределения случайной величины по заданной выборке и выбор типа теоретического распределения (выдвижение рабочей гипотезы);

2) определение параметров теоретического распределения;

3) проверка адекватности теоретического распределения (подтверждение рабочей гипотезы);

4) уточнение теоретического закона распределения, если первоначальная рабочая гипотеза отвергается.

Построение эмпирической кривой (гистограммы) распределения. Гистограмма $\tilde{f}(z)$ является эмпирическим аналогом функции плотности распределения $f(z)$.

При обработке выборок больших объемов используют метод "сгруппированных данных". Выборка объема n преобразуется в статистический ряд. Для этого весь диапазон изменения случайной величины в выборке $z_{\min} - z_{\max}$ делится на k интервалов. Число интервалов можно выбирать по полуэмпирической формуле (правило Штюргеса):

$$k = 1 + 3,2 \lg n, \qquad (5.144)$$

где найденное значение округляют до ближайшего целого числа.

Длина интервала h равна

$$h = \frac{z_{\max} - z_{\min}}{k}. \qquad (5.145)$$

Подсчитывают число элементов выборки $n_j (j = \overline{1,k})$, попавших в каждый интервал. Величина, равная

$$P_j^* = \frac{n_j}{n} \qquad (5.146)$$

определяет относительную частоту попадания случайной величины в j-ый интервал.

Середина j-го интервала z_j^* определяется формулой

$$z_j^* = \frac{z_{j-1} + z_j}{2}. \qquad (5.147)$$

Полученный статистический ряд можно записывать в виде табл. 2.2.

Таблица 2.2.

Статистический ряд

Номер интер-вала	Интервал	z_j^*	n_j	P_j^*
1	$(z_{\min} = z_0, z_1)$	z_1^*	n_1	P_1^*
2	(z_1, z_2)	z_2^*	n_2	P_2^*
.		.	.	.
.		.	.	.
.		.	.	.
k	$(z_{k-1}, z_{\max} = z_k)$	z_k^*	n_k	P_k^*
Σ			n	1

По данным табл. 2.2. строят гистограмму эмпирического распределения, представляющую собой ступенчатую кривую, значение

которой на j-ом интервале $(z_{j-1}, z_j)(j = 1,2,..k; z_j = z_{j-1} + h)$ постоянно и равно n_j или P_j^*, или с учетом условия $\int_{-\infty}^{\infty} f(z)dz = 1$ равно $\frac{P_j^*}{h}$ (рис. 2.14).

Кривую эмпирического распределения можно аппроксимировать различной теоретической кривой распределения - законом Пуассона, Вейбула, биномиальным, показательным, логарифмическим, нормальным и др. По виду гистограммы предполагают, что случайная величина Z распределена по тому или иному закону (выдвигают рабочую гипотезу).

Определение параметров теоретического распределения. Окончательное уточнение неизвестной функции распределения сводится к определению некоторых числовых параметров распределения. По выборке могут быть рассчитаны выборочные статистические характеристики (выборочное среднее, дисперсия и т.д.),которые являются оценками соответствующих генеральных параметров. Оценки, получаемые по выборке,сами являются случайными величинами. К оценке обычно предъявляются требования состоятельности и несмещенности.

Для получения оценок используют различные методы. Широко применяется метод максимального правдоподобия. Оценки, получаемые при помощи этого метода отвечают изложенным требованиям.

Сущность метода максимального правдоподобия заключается в нахождении таких оценок неизвестных параметров, для которых функция правдоподобия при случайной выборке объема n будет иметь максимальное значение. Пусть известен общий вид плотности вероятности $f(z,a)$ теоретического распределения; a- неизвестный параметр, входящий в выражение закона распределения. Экспериментально получена выборка значений случайной величины $z_1, z_2, \ldots, z_n$.

Функция

$$L(z,\tilde{a}) = \sum_{i=1}^{n} \ln f(z_i, \tilde{a}) \qquad (5.148)$$

называется функцией правдоподобия, где $\tilde{a}$ - оценка параметра a.

В качестве оценки для a следует взять то значение $\tilde{a}$, для которого функция $L(z,\tilde{a})$ имеет наибольшее возможное значение. Необходимое условие экстремума этой функции дает уравнение

$$\frac{dL(z,\tilde{a})}{d\tilde{a}} = 0, \qquad (5.149)$$

из которого определяют искомое значение $\tilde{a}$.Если максимумов несколько, необходимо выбрать среди них наибольший.

Достаточным условием максимума является выполнение неравенства

$$\frac{d^2 L(z,\tilde{a})}{d\tilde{a}^2} \le 0. \qquad (5.150)$$

В общем случае требуется оценить одновременно несколько параметров распределения. В этом случае формулировка принципа

максимального правдоподобия сохраняется: надо найти такую совокупность допустимых значений параметров $\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_m$ функции распределения $f(z, a_1, a_2, \ldots a_m)$, которая обращает функцию правдоподобия

$$L(z, \tilde{a}_1, \tilde{a}_2, \ldots \tilde{a}_m) = \sum_{i=1}^{n} \ln f(z_i, \tilde{a}_1, \tilde{a}_2, \ldots \tilde{a}_m) \qquad (5.151)$$

в максимум. Необходимое условие экстремума L функции дает система уравнений

$$\frac{\partial L(z, \tilde{a}_1, \ldots \tilde{a}_m)}{\partial \tilde{a}_j} = 0, j = \overline{1, m}, \qquad (5.152)$$

а неотрицательная определенность матрицы

$$\left(\frac{\partial^2 L(z, \tilde{a}_1, \ldots, \tilde{a}_m)}{\partial \tilde{a}_i \partial \tilde{a}_j} \right), i, j = 1, 2, \ldots m, \qquad (5.153)$$

в стационарной точке является достаточным условием того, чтобы этот локальный экстремум был максимумом функции правдоподобия.

Из системы уравнений (5.152) находят оценки $\tilde{a}_j (j = \overline{1, m})$.

Найдем методом максимального правдоподобия оценку для параметра λ показательного распределения с плотностью

$$f(z) = \lambda e^{-\lambda z} \qquad (5.154)$$

по выборке $z_1, z_2, \ldots, z_n$.

Для этого распределения функция правдоподобия имеет вид

$$L = \sum_{i=1}^{n} \ln\left(\tilde{\lambda} e^{-\tilde{\lambda} z_i}\right) = n \ln \tilde{\lambda} - \tilde{\lambda} \sum_{i=1}^{n} z_i. \qquad (5.155)$$

$$\frac{\partial L}{\partial \tilde{\lambda}} = \frac{n}{\tilde{\lambda}} - \sum_{i=1}^{n} z_i = 0;$$

$$\tilde{\lambda} = \frac{n}{\sum_{i=1}^{n} z_i} = \frac{1}{\bar{z}}, \qquad (5.156)$$

где $\bar{z}$ - среднее выборки.

Пусть распределение случайной величины Z подчинено нормальному закону

$$f(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(z-m)^2}{2\sigma^2} \right], \qquad (5.157)$$

где m, σ^2 - параметры распределения.

Функция правдоподобия

$$L\left(z, \tilde{m}, \tilde{\sigma}^2\right) = \sum_{i=1}^{n} \ln\left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(z_i - \tilde{m}\right)^2}{2\tilde{\sigma}^2} \right) \right]. \qquad (5.158)$$

Необходимые условия экстремума L:

$$\left.\begin{aligned}\frac{\partial L(z,\tilde{m},\tilde{\sigma}^2)}{\partial \tilde{m}} &= 0;\\ \frac{\partial L(z,\tilde{m},\tilde{\sigma}^2)}{\partial \tilde{\sigma}^2} &= 0.\end{aligned}\right\} \qquad (5.159)$$

Упуская промежуточные математические операции, из системы (2.159) находим:

$$\tilde{m} = \bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i, \qquad (5.160)$$

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(z_i - \bar{z}\right)^2. \qquad (5.161)$$

Оценка $\tilde{\sigma}^2$ получается несколько смещенной. Для получения несмещенной оценки умножаем $\tilde{\sigma}^2$ на $n/(n-1)$. Несмещенная оценка дисперсии при этом имеет вид

$$\tilde{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(z_i - \bar{z}\right)^2. \qquad (5.162)$$

Уменьшение знаменателя в (5.162) на единицу непосредственно связано с тем, что величина $\bar{z}$ относительно которой берутся отклонения, сама зависит от элементов выборки. Каждая величина, зависящая от элементов выборки и входящая в формулу выборочной дисперсии, называется связью. Знаменатель выборочной дисперсии всегда равен разности между объемом выборки и числом связей ℓ, наложенных на эту выборку. Эта разность

$$f = n - \ell \qquad (5.163)$$

называется числом степеней свободы выборки.

В практических вычислениях для дисперсии $\tilde{\sigma}^2$ удобна формула

$$\tilde{\sigma}^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} z_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} z_i\right)^2\right], \qquad (5.164)$$

легко вытекающая из формулы (5.162). Преимущество формулы (5.164) в том, что в ней нет операций вычитания близких чисел, как в формуле (5.162), что приводит к потере точности.

Как правило, число наблюдений n велико. Поэтому для упрощения расчетов удобно использовать не соотношения (5.160) и (5.162), а менее трудоемкие формулы, позволяющие оценить m и σ^2 по сгруппированным данным:

$$\tilde{m} = \bar{z} = \frac{1}{n}\sum_{j=1}^{k} z_j^{*} n_j = \sum_{j=1}^{k} z_j^{*} P_j^{*}, \qquad (5.165)$$

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{k} z_j^{*2} n_j - \bar{z}^2 = \sum_{j=1}^{k} z_j^{*2} P_j^{*} - \bar{z}^2. \qquad (5.166)$$

Проверка адекватности теоретического распределения. Проверку рабочей гипотезы о законе распределения случайной величины по выборке проводят при помощи критериев согласия. Критерием согласия называется

критерий гипотезы о том, что генеральная совокупность имеет распределение предполагаемого типа (например, нормальное распределение).Критерии согласия позволяют определить вероятность того, что при гипотетическом законе распределения наблюдающееся в рассматриваемой выборке отклонение вызывается случайными причинами, а не ошибкой в гипотезе. Если эта вероятность велика, то отклонение от гипотетического закона распределения следует признать случайным и считать, что гипотеза о предполагаемом законе распределения не опровергается.

Вероятностный характер критериев не позволяет однозначно принять или опровергнуть проверяемую гипотезу. Критерий позволяет утверждать, что гипотеза не противоречит опытным данным, если вероятность наблюдаемого отклонения от гипотетического закона велика, или что гипотеза не согласуется с опытными данными, если эта вероятность мала.

Среди различных критериев согласия наиболее употребителен универсальный критерий согласия χ^2 (Пирсона).

Проверку гипотезы о виде функции распределения с помощью этого критерия производят следующим образом:

1. По выборке строят гистограмму. Если в каком -либо j-м интервале число наблюдений n_j окажется меньше пяти,то его объединяют с соседним интервалом (или интервалами) так, чтобы число наблюдений в таком объединенном интервале оказалось большим или равном пяти (должно быть $n_j \geq 5 \forall j$).

Пусть K_0-окончательное число интервалов группированная после объединения, тогда очевидно, что

$$\sum_{j=1}^{K_0} n_j = n, \quad n_j \geq 5. \qquad (5.167)$$

2. Задаются видом гипотетической функции распределения и для каждого из параметров m этого распределения находят оценки, причем эти оценки можно определить как по исходным, так и по сгруппированным данным.

3. Определяют теоретическую вероятность $P_j^m (j = \overline{1,k})$ попадания в каждый из K_0 интервалов случайной величины с заданным распределением, параметры которого или известны, или оценены в п.2.

P_j^m вычисляют по формуле

$$P_j^m = \int_{z_j^* - h/2}^{z_j^* + h/2} f(z)dz = P\left\{z_j^* - \frac{h}{2} < z \leq z_j^* + \frac{h}{2}\right\}; \qquad (5.168)$$

$n_j^m = nP_j^m$, (2.169)

где $f(z)$ - выбранный теоретический закон распределения;

n_j^m - абсолютная частота попадания величины Z в j-й интервал по теоретическому распределению.

В случае нормального закона распределения

$$P_j^m = P\left\{z_j^* - \frac{h}{2} < z \le z_j^* + \frac{h}{2}\right\} = \frac{1}{\tilde{\sigma}\sqrt{2\pi}} \int_{z_i^*-h/2}^{z_i^*+h/2} \exp\left[-\frac{(z-\tilde{m})^2}{2\tilde{\sigma}^2}\right] dz \quad . \qquad (5.170)$$

Этот интеграл можно вычислить с помощью таблицы значений функции Лапласа $\Phi(x)$:

$$P_j^m = \Phi\left[\frac{\left(z_j^* + \frac{h}{2}\right) - \tilde{m}}{\tilde{\sigma}}\right] - \Phi\left[\frac{\left(z_j^* - \frac{h}{2}\right) - \tilde{m}}{\tilde{\sigma}}\right] = \Phi(\alpha) - \Phi(\beta) \quad . \qquad (5.171)$$

Функция Лапласа $\Phi(x)$- нечетная функция, т.е. $\Phi(-x) = -\Phi(x)$, поэтому таблица значений $\Phi(x)$составлена лишь для $x \ge 0$.

4. Вычисляют значение χ^2критерия Пирсона по формуле

$$\chi^2 = n\sum_{j=1}^{K_0} \frac{\left(P_j^* - P_j^m\right)^2}{P_j^m} = \sum_{j=1}^{K_0} \frac{\left(n_j - nP_j^m\right)^2}{nP_j^m} \quad . \qquad (5.172)$$

Известно, что для данного критерия согласия случайная величина χ^2 при больших nимеет χ^2-распределение (при $nP_j^m \ge 5 \forall j$) с числом степеней свободы

$$f = K_0 - r - 1, \qquad (5.173)$$

где r- число определяемых в п.2. неизвестных параметров гипотетического распределения, а уменьшение числа степеней свободы еще на единицу объясняется наличием линейного соотношения (5.167) между эмпирическими величинами n_jи n, входящими в расчетную формулу (2.172).

5. Задавшись уровнем значимости q(или доверительной вероятностью P, $q = 1 - P$) по таблице χ^2-распределения Пирсона находят критическое значение $\chi^2_{\text{кр}}$при заданном числе степеней свободы f.

6. Сравнивают значения χ^2и $\chi^2_{\text{кр}}$и выносят решение о принятии (если $\chi^2 \le \chi^2_{\text{кр}}$) или отклонении (если $\chi^2 > \chi^2_{\text{кр}}$) рассматриваемой гипотезы о виде функции распределения. Если гипотеза не согласуется с выбранным эмпирическим распределением выдвигается новая рабочая гипотеза с последующей ее проверкой на адекватность.

После того, как определен закон распределения случайной величины, можно говорить о том, что имеется модель изменения этой величины.

5.7. Методы переходных функций

Дифференциальное уравнение и передаточная функция является наиболее общей формой уравнения связи между переменными состояния на входе и выходе **линейной системы**. Экспериментатор наблюдает

только входные и выходные переменные, по которым должен получить уравнение связи.

С помощью эксперимента можно получить только график или таблицу чисел, определяющих значения сигналов на входе и выходе, т.е. получить реализацию некоторого частного дифференциального уравнения при определенном входном сигнале. В дальнейшем, аппроксимировав аналитическим выражением полученные реализации, можно построить дифференциальное уравнение заданной структуры и записать его в одной из форм, приведенных выше.

При эксперименте можно использовать различные виды входных сигналов. Применяя некоторые их виды, можно не только существенно облегчить задачу аппроксимации, но и, что особенно важно, повысить качество (точность) исследования с точки зрения некоторого заданного критерия идентификации. Выбор характера входного сигнала с точки зрения оптимизации качества идентификации. Выбор характера входного сигнала с точки зрения оптимизации качества идентификации исследуется в методических руководствах по оптимизации методов планирования экспериментов [15].

Первые методы идентификации в задачах управления были основаны на использовании ступенчатых, импульсных и гармонических входных сигналов.

Основные виды входных сигналов, используемых для идентификации, приведены в левой колонке рис.2.1.

Наибольшее распространение получили переходные характеристики при ступенчатом воздействии на входе (рис.2.1,а). Реакцию объекта на ступенчатое воздействие часто называют кривой разгона.

Если известна амплитуда ступени на входе d и экспериментально получены ординаты кривой разгона $y(n)$ можно записать, используя (2.6а), выражение дискретного изображения по Лапласу входного и выходного воздействий.

Для дискретного времени, воспользовавшись дискретным преобразованием Лапласа, можно получить выражения (5.6а) и (5.7а). Так, прямое преобразование Лапласа решетчатой функции $x(n)$ определяется соотношением

$$X^*(q) = \sum_{n=0}^{\infty} e^{-qn} x(n), \qquad (5.6а)$$

Как и для непрерывного времени $q = \sigma + j\varpi.$

Соответственно обратное преобразование, определяющее решетчатую функцию по ее изображению, записывается в виде

$$x(n) = \frac{1}{2\pi j} \oint X^*(q) q^{n-1} dq. \qquad (5.7а)$$

Дискретная передаточная функция и соответствующее дискретное уравнение при этом принципиально могут быть найдены.

Для непрерывного времени полученная кривая разгона должна быть аппроксимирована аналитическим выражением.

Обычно для линейных систем такая аппроксимация выполняется с помощью выражения

$$y(t) = \sum_{k=1}^{m} a_k e^{-b_k t}, \qquad (5.8)$$

где b может быть комплексным числом.

Число экспонент в выражении y(t) определяет порядок дифференциального уравнения и передаточной функции объекта идентификации.

Обзор методов аппроксимации экспериментальных кривых аналитическими выражениями приведен в [16], по аналитическим выражениям входа и выхода с помощью уравнений (5.4) – (5.7) или таблиц в [13,14] могут быть найдены изображения этих выражений по Лапласу, передаточные функции и дифференциальные уравнения объекта.

Рассмотрим основные недостатки методов идентификации с помощью переходных функций. Известно, что непериодическая функция x(t), удовлетворяющая условиям Дирихле, может быть представлена в виде интеграла Фурье.

Любой непериодический сигнал *x(t)* может быть представлен в виде бесконечного множества гармонических составляющих, образующих непрерывный спектр.

$$X(j\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt, \qquad (5.9)$$

(5.1) – прямое преобразование Фурье.

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(j\omega) e^{-j\omega t} d\omega, \qquad (5.10)$$

(5.2) – обратное преобразование Фурье

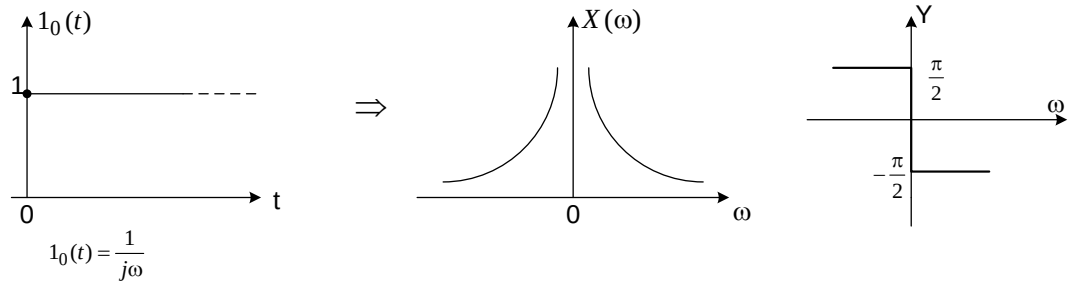


Рис.2.1

Известно, что мощность несинусоидального сигнала, например тока, равна сумме мощностей отдельных гармоник.

Формула Релея дает возможность, зная частотный спектр сигнала, найти его энергию. Из формулы Релея следует, что энергия, передаваемая сигналом, распределяется по частотам пропорционально квадрату

амплитуды частотного спектра. При этом амплитуда плотности спектра *g(ω)* интеграла Фурье пропорциональна энергии сигнала *x(t)* на данной частоте *ω*. Выражение для амплитуды плотности спектра записывается в виде

$$g(\omega) = \left| \frac{1}{2\pi} \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt \right|. \qquad (5.11)$$

Для непрерывной функции

$Y(p) = X(p)W(p)$, где $W(p) = \dfrac{B(p)}{A(p)}$,

$$X(p) = L[x(t)] \int_{-\infty}^{\infty} x(t) e^{-pt} dt, \qquad (5.12)$$

где p=σ+jω, σ – абсцисса абсолютной сходимости, т.е. действительная постоянная, при которой $\int_{-\infty}^{\infty} \left| x(t) e^{-\sigma t} \right| dt < \infty \; j = \sqrt{-1}$,

Обратное преобразование Лапласа для непрерывного сигнала

$$x(t) = \frac{1}{2\pi j} \int_{\sigma - j\omega}^{\sigma + j\omega} X(p) e^{pt} dp. \qquad (5.13)$$

Рис 2.2.а

Здесь g – амплитуда плотности спектра интеграла Фурье,

$2\pi f^* = 1a,\ f = 1{,}6 Гц.$

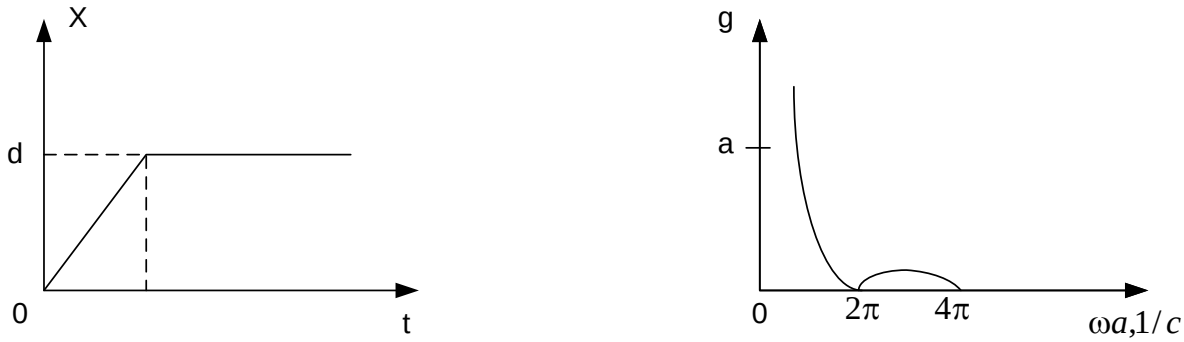


Рис 2.2.б

$\omega a = 2\pi f a.$

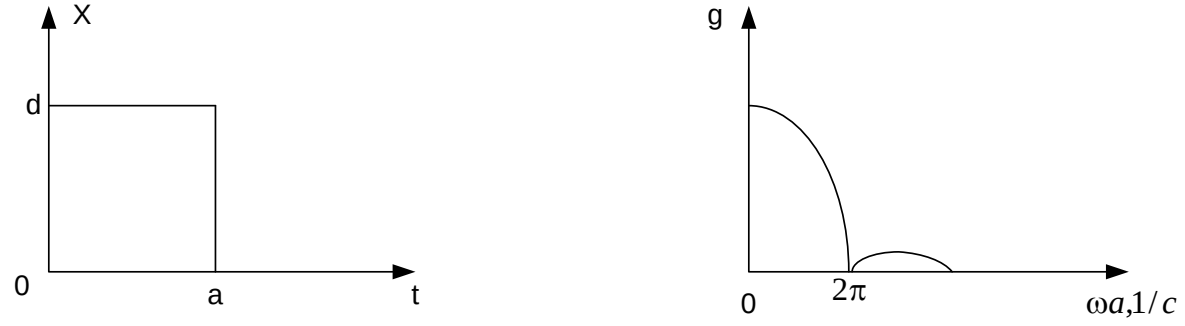


Рис 2.2.в

Если a=1c, то g(ω) практически затухает на f^*=2Гц, т.к. $2\pi f^* a \cong 4\pi.$

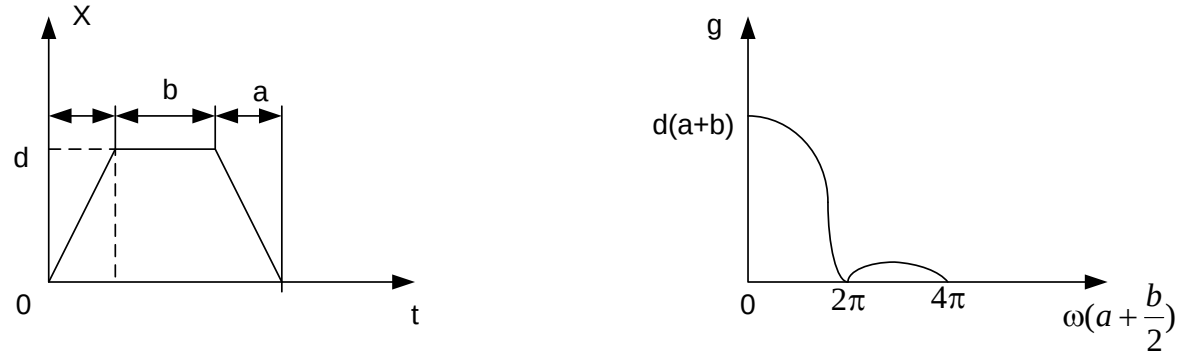


Рис 2.2.г

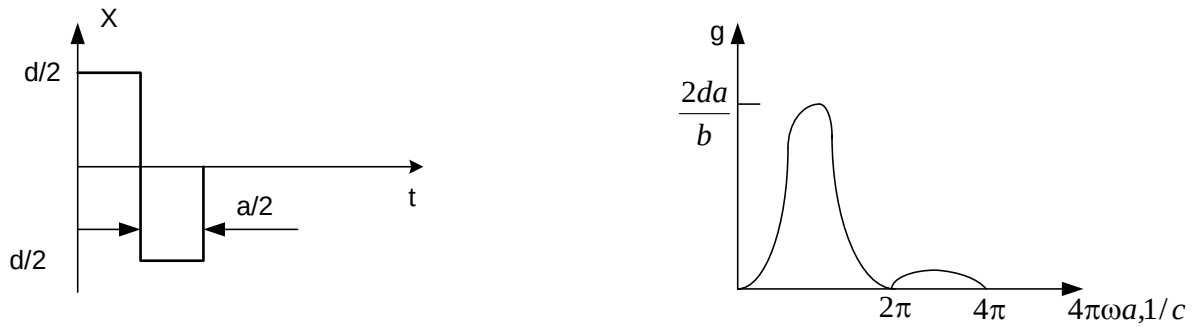


Рис 2.2.д

Рис.2.2 отображает виды переходных характеристик и их амплитудные спектры.

На ри.2.2 в правой колонке приведены зависимости g(ω) для различных видов входных воздействий. Из рисунка видно, что **приведенные виды входных воздействий несут очень мало энергии на высоких частотах**. Кроме того, промышленные объекты обычно являются фильтрами нижних частот, т.е. значительно уменьшают амплитуды на высоких частотах. Поэтому точность определения характеристик объекта в области высоких частот **с помощью переходных функций** весьма низка.

На качестве идентификации с помощью переходных функций весьма неблагоприятно сказываются также случайные помехи, искажающие

реакцию объекта, и неточность аппроксимации объекта линейной моделью.

Частично указанные недостатки устраняются при идентификации объекта частотными методами.

5.8. Частотные методы

Динамические свойства объекта могут быть описаны с помощью частотных характеристик. Частотная характеристика объекта представляет зависимость от частоты в установившемся режиме двух переменных:

1) **Отношения амплитуд** гармонических сигналов на входе и выходе объекта – амплитудно-частотная характеристика (АЧХ) A(ω);

2) сдвига фаз между входными и выходными гармоническими сигналами – фазочастотная характеристика (ФЧХ) φ(ω).

На графиках принято изображать указанные характеристики в следующих формах.

1. В виде совокупности АЧХ и ФЧХ в декартовых координатах в зависимости от частоты. Эта форма графика называется диаграммой Боде.

2. В полярных координатах, где A(ω) является радиусом-вектором, φ(ω) – полярным углом. Частота ω является параметром. Такая кривая с отметками частот носит название **амплитудно-фазовой характеристики (АФХ)** или **диаграммы Найквист**а.

Принято также использовать известную форму записи функций, заданных в полярных координатах, в виде комплексных чисел. Действительная часть W(jω) записывается в виде

$$U(\omega) = A(\omega)\cos\varphi(\omega),$$

$$A(\omega) = \sqrt{U^2(\omega) + V^2(\omega)},$$

$$\varphi(\omega) = arctg\frac{V(\omega)}{U(\omega)}.$$

$V(\omega)$ - мнимая частотная характеристика, $U(\omega)$- вещественная частотная характеристика.

Мнимая – в виде

$$V(\omega) = A(\omega)\sin\varphi(\omega).$$

Амплитудно-фазовая характеристика в символической записи имеет вид:

$$W(j\omega) = U(\omega) + jV\omega.$$

3. Если в диаграммах Боде A(ω) заменены на lgA(ω), то характеристики носят название **логарифмических частотных характеристик.** При этом обычно по оси абсцисс откладывается относительная частота в логарифмическом масштабе в октавах и декадах, а по оси ординат вместо lgA откладывается 20 lgA в децибелах.

Выражения W(jω), A(ω), φ(ω), $U(\omega), V(\omega)$ могут быть легко найдены, если известно дифференциальное уравнение. Для этого достаточно в выражении $W(p) = \dfrac{B(p)}{A(p)}$ подставить jω вместо p. Полученное комплексное число и представляет собой аналитическое выражение АФХ в комплексной форме W(jω). Остальные формы характеристик могут быть легко получены по известным правилам преобразования комплексных чисел. Заданные аналитически или графически частотные характеристики объекта позволяют рассчитать контур стабилизации со стандартным регулятором.

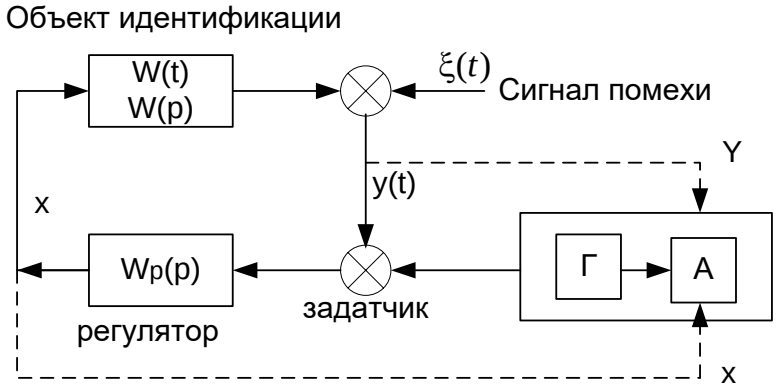


Рис.2.3. Схема определения частотных характеристик в замкнутой системе

Эксперимент, с помощью которого исследуются частотные характеристики, значительно более трудоемок по аппаратуре и времени проведения по сравнению с переходными характеристиками. Для исследования необходима аппаратура, с помощью которой можно подать на вход объекта гармонические сигналы. Выходной сигнал всегда зашумлен и обычно несколько искажен нелинейностями, имеющимися в объекте. Поэтому для определения амплитуды и фазы выходного гармонического сигнала необходимо выделение первой гармонической составляющей вручную или с помощью специальной аппаратуры. Исследование на различных, в том числе низких частотах, требует большого времени.

Поскольку при исследовании частотных характеристик рассматривается вынужденное, а не свободное движение системы, необходимо некоторое время для затухания свободного движения.

На исследование частотных неблагоприятное влияние оказывают тренды, вызывающие «уползание» средней линии выходного сигнала. Для уменьшения ошибок из-за «уползания» средней линии и уменьшения влияния помех определение частотных характеристик производят в замкнутой системе по схеме, приведенной на рис.2.3, где объект с передаточной функцией W(p) охвачен обратной связью с регулятором-стабилизатором, имеющим передаточную функцию $W_p(p)$. Помехи,

приведенные к выходу объекта ξ(t), могут содержать кроме случайных составляющих детерминированные тренды. Но в сигнале y(t) «уползание» будет устранено, а воздействие случайных составляющих ξ(t) значительно ослаблено. Гармонический сигнал теперь будет подаваться с генератора Г не непосредственно на регулирующий орган, а на задатчик регулятора, что, конечно, требует значительно меньшей мощности генератора. На анализатор А, выделяющий первые гармонические составляющие, подаются сигналы *x* и *y*.

Полученные экспериментально частотные характеристики, как упоминалось выше, могут быть использованы для расчета системы управления. Однако более универсальным для различных применений является аналитическое представление.

Разработаны методы аппроксимации частотных характеристик аналитическими выражениями.

5.9. ***Непараметрическая идентификация линейных стохастических систем. Уравнение Винера***

В приведённых в гл. 2 методах предполагалось, что наблюдаемые выходные переменные полностью определяются наблюдаемыми входными воздействиями. В действительности же в подавляющем большинстве случаев выходные переменные определяются также ненаблюдаемыми и неуправляемыми воздействиями (шумами).

Чтобы получить действительные характеристики объекта, нужно исключить из выходного сигнала составляющие, определяемые шумом, оставив только реакцию на входное воздействие, - отфильтровать выходной сигнал. С другой стороны, в условиях нормальной эксплуатации шумы вызывают отклонение выходных переменных от заданных значений. Стремясь поддержать эти переменные ближе к заданным значениям, автоматические устройства или операторы воздействуют на управляющие органы, то есть на выходные переменные. Следовательно, и в режиме нормальной эксплуатации имеют место изменения входных и выходных воздействий, вследствие чего, проведя статистическую обработку записей этих переменных, можно получить ИПФ или частотную характеристику объекта.

Чтобы получить уравнения связи между статистическими характеристиками входа и выхода для стационарных эргодических случайных сигналов, пользуются их корреляционными функциями или спектральными плоскостями. Выражение *взаимной корреляционной функции* сигналов *x(t), y(t)* записывается в виде

$$R_{xy}(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)y(t+\tau)dt. \qquad (5.1)$$

Если в этой формуле участвует одна переменная, т.е. подынтегральное выражение имеет вид $x(t)x(t+\tau)dt$, и полученное выражение называется *автокорреляционной функцией* сигнала.

Структурная схема исследуемого объекта может быть представлена в виде, приведённом на рис. 3.1. Все ненаблюдаемые помехи, воздействующие на различные части объекта, приведены к выходу объекта и представлены в виде аддитивного шума.

Используя уравнение (2.2), можно записать значение выходного сигнала для схемы, приведённой на рис. 3.1, в виде

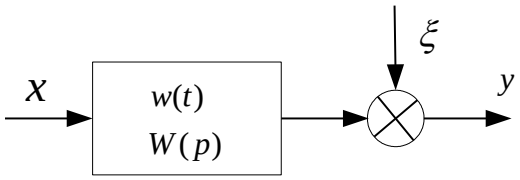


Рис. 3.1. Структурная схема модели Объекта с аддитивным шумом

$$y(t) = \int_{-\infty}^{\infty} w(\tau)x(t-\tau)dt + \xi(t),$$

ω(τ) – импульсная переходная (весовая) функция (ИПФ) объекта, т.е. сигналы y(t) на выходе объекта, если входной сигнал x(t) представляет собой дельта-функцию δ(t) вида

$$\delta(t) = \begin{cases} 0, \ при \ t<0 \ и \ t>0 \\ \infty, \ при \ t=0 \end{cases}$$

$$\int_{-\infty}^{\infty} \delta(t)dt = 1$$

Если входной сигнал x(t)=0 при t<0, то

$$y(t) \equiv \int_0^t x(\tau)\omega(t-\tau)d\tau \equiv \int_0^t \omega(\tau)x(t-\tau)d\tau.$$

Умножив это выражение на $x(t+\tau)$, проинтегрировав левую и правую части по τ_1 в пределах от – T до T и перейдя к пределу при $T\to\infty$, получим выражение

$$R_{xy}(\tau) = \int_0^{\infty} w(t)R_{xx}(t-\tau)dt + R_{x\xi}(\tau).$$

Поскольку $w(t)=0$ при *t<0*, то, если сигналы x и ξ не коррелированы, так что $R_{xy}(\tau)=0$, это выражение преобразуется к виду

$$R_{xy}(\tau) = \int_0^{\infty} w(t) R_{xx}(t-\tau) dt \qquad (5.2)$$

Это интегральное уравнение, справедливое для линейных систем, носит название уравнения Винера-Хопфа.

В дискретной форме уравнение (5.2) записывается в виде

$$R_{xy}(n) = \sum_{\mu=1}^{N} w(\mu) R_{xx}(n-\mu) \qquad (5.2а)$$

Поскольку при эксперименте получаются только оценки корреляционных функций, **значение искомой ИПФ** $w(t)$ так же оказывается приближенным. Структура уравнения такова, что небольшие ошибки в корреляционных функциях приводят к существенным ошибкам при оценке координат ИПФ. Поэтому обычно используют сглаживание решения.

Методы решения уравнения Винера-Хопфа рассмотрены в [4, 11, 18].

Уравнение (5.2) можно решать так же в частотной области. Выполнив преобразование Фурье над уравнением (5.2), получим выражение

$$S_{xy}(p) = W(j\omega) S_{xx}(p) \qquad (5.3)$$

где $S_{xy}(p)$ и $S_{xx}(p)$ - взаимная спектральная плотность сигналов x и y и спектральная плотность сигнала x соответственно (преобразование Фурье от соответствующих корреляционных функций); $W(p)$ - частотная характеристика, $p = j\omega$. Авто и взаимная спектральная плотности

$$S_{xx}(\omega) = \int_{-\infty}^{\infty} R_{xx}(\tau) e^{-j\omega\tau} d\tau$$

$$S_{xy}(\omega) = \int_{-\infty}^{\infty} R_{xy}(\tau) e^{-j\omega\tau} d\tau$$

Непосредственное решение этого уравнения (5.3) относительно неизвестной $W(p)$ может привести к физически нереализуемому результату. Поэтому вначале необходимо провести факторизацию выражений спектральных плоскостей. Физически реализуемое решение находится из выражения

$$W(p) = [S_{xy}(p) / S_{xx}^{-}(p)]_{+} / S_{xx}^{+}(p),$$

где $S_{xx}(p)$ представлено в виде $S_{xx}(p) = S_{xx}^{+}(p) S_{xx}^{-}(p)$, причём $S_{xx}^{+}(p)$ содержит все нули и полюсы в левой полуплоскости, а $S_{xx}^{-}(p)$ - в правой.

Тема 6. Построение модели статики сложных многомерных объектов

6.1. Выбор существенных переменных модели объекта

Для определения структуры модели сложного объекта прежде всего необходимо выяснить, какие входы объекта будут включены в его модель. Для этого прежде всего выявляют всех возможных входов и из них выделяют наиболее существенные. Существенным является тот вход, который наибольшим образом оказывает влияние на выход объекта.

Для этого следует проранжировать выбираемые входы (факторы). В этом случае выбор наиболее существенных факторов очевиден - они должны быть расположены в начале ранжированного ряда.

Поэтому первым этапом задачи идентификации объекта следует считать задачу ранжирования его входов. Так как модели объекта еще нет, то ранжирование можно сделать, например, методом экспертных оценок.

Метод экспертных оценок. Для этого прежде всего определяются все входы объекта. При их выборе принимают во внимание следующее. Входы должны непосредственно воздействовать на объект , не зависеть от других переменных, измеряться и управляться. Исходя из этих соображений составляется ряд входов. Широкому кругу специалистов предлагается расположить эти факторы в порядке убывания степени их влияния на выбранную выходную величину объекта, т.е. проранжировать факторы.

При обработке результатов экспертного опроса весьма эффективным является метод ранговой корреляции. Сущность этого метода заключается в следующем.

Пусть m экспертов ранжируют n факторов. Каждому фактору каждый эксперт присваивает ранг - целое число от 1 до n, приписывая ранг 1 наиболее важному фактору, ранг 2 - следующему по важности и т.д. На основании этих данных составляется матрица рангов

$$A = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ a_{1m} & a_{2m} & \cdots & a_{nm} \end{pmatrix}, \qquad (6.1)$$

где a_{ij} - ранг i-го фактора ($i = \overline{1,n}$), определенного j-м экспертом ($j = \overline{1,m}$), номера строк соответствуют номерам экспертов, а номера столбцов - номерам ранжируемых факторов. Это означает, что j-я строка представляет собой мнение j-го эксперта, а i-й столбец - мнение всех экспертов по поводу i-го фактора.

При назначении рангов экспертами нужно соблюдать следующие условия:

1. Сумма рангов, назначенных всем факторам, должны быть одинакова для каждого эксперта и равна:

$$\sum_{i=1}^{n} i = \frac{n \cdot (n+1)}{2}.$$

Это означает, что сумма элементов любой строки матрицы (6.1) должна быть равна:

$$\sum_{i=1}^{n} a_{ij} = \frac{n \cdot (n+1)}{2}, \quad j = \overline{1,m}.$$

2. Если эксперт какие-то q факторов считает одинаковыми, то приписывает им один ранг. Этот ранг равен среднему из q целых рангов, которые получены при условии, что эксперту удалось их проранжировать. Например, равноценность двух факторов x_3, x_8(q=2), стоящих на шестом месте в ранжированном ряду, приводит к тому, что их ранги равны: (6+7)/2=6.5. Как видно, в этом случае ранги могут быть дробными. Отметим, что дробные ранги кратны 1/2.

На базе данных матрицы (6.1) вычисляем:

$$\sum_i = \sum_{j=1}^{m} a_{ij} \quad , i = \overline{1,n} \quad ; (6.2)$$

$$\overline{\Sigma} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\sum_i\right) \frac{1}{n} \cdot \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij} = \frac{1}{2} \cdot m \cdot (n+1); \qquad (6.3)$$

$$S = \sum_{i=1}^{n} \left[\sum_{j=1}^{m} a_{ij} - \frac{1}{n} \cdot m \cdot (n+1) \right]^2 = \sum_{i=1}^{n} \left(\sum_i - \overline{\sum} \right)^2, \qquad (6.4)$$

где $\sum_i$ - сумма рангов i-го фактора $\left(i = \overline{1,n}\right)$;

$\overline{\sum}$ - среднее значение всех рангов;

S - сумма квадратов отклонений частных рангов от среднего.

Очевидно, максимальное значение S принимает тогда, когда эксперты одинаково проранжируют все факторы и отклонение суммарных рангов от среднего значения будет m(1-n)/2, m(3-n)/2, ... Сумма квадратов членов этого ряда определяется выражением

$$S_m = \frac{1}{12} \cdot m^2 \cdot n \cdot \left(n^2 - 1\right). \qquad (6.5)$$

Степень согласованности мнений экспертов характеризует коэффициент конкордации (согласованности) W

$$W = S / S_m. \qquad (6.6)$$

Коэффициент конкордации меняется от нуля (отсутствие какого-либо согласия в мнениях экспертов) до единицы (полное согласие).

Статистическую значимость коэффициента W определяют в зависимости от числа факторов. При $n \geq 7$ используют χ^2 - критерий Пирсона:

$$\chi_P^2 = m \cdot (n-1) \cdot W. \qquad (6.7)$$

Расчетные значение χ^2 сравнивают с табличным при выбранном уровне значимости α и числе степеней свободы $f = n - 1$. Гипотеза о наличии согласия между специалистами принимается в том случае если $\chi_P^2 \geq \chi_T^2 \cdot \chi_T^2$ определяется из таблицы распределения Пирсона.

В случае $n < 7$ пользуются F-критерием Фишера, расчетное значение которого определяют по формуле:

$$F_P = \frac{0{,}5 \cdot \ln(m-1) \cdot W}{1 - W}. \qquad (6.8)$$

с числом степеней свободы $f_1 = n - 1 - \frac{2}{m}$ и $f_2 = \frac{m-1}{f_1}$.

Мнение специалистов считается согласованным, если расчетное значение критерия Фишера больше или равно табличному $(F_P \geq F_T) \cdot F_T$ находится из таблицы распределения Фишера.

Далее оценивают степень влияния каждого фактора на выбранную выходную величину. Для этого используют ряд показателей, но наиболее часто - сумму рангов данного фактора. Чем меньше эта величина, тем сильнее влияет фактор на выходную величину.

Расположим теперь входы в порядке возрастания суммы их рангов: $x_{l_1}, x_{l_2}, \ldots, x_{\ln}$, где индекс l_i равен номеру фактора с соответствующей суммой рангов (x_{l_1}-фактор с минимальной суммой рангов). Здесь на первом месте стоит самый существенный вход (его номер l_i), а далее следуют остальные в порядке уменьшения их влияния на выход.

Для анализа результатов ранжирования используют диаграмму рангов (рис. 3.1). Наиболее благоприятным видом диаграммы является возрастание рангов по параболе. В этом случае легко разделить факторы на группы и по К-критерию Линка-Уэллеса отсеять несущественные.

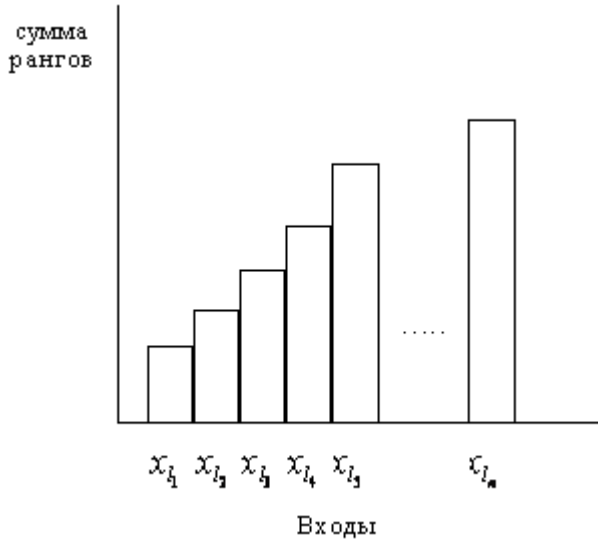


Рис. 3.1

Расчетная величина К-критерия определяется по формуле

$$K_P = \frac{k \cdot \left(\bar{a}_{\max} - \bar{a}_{\min}\right)}{\sum_{i=1}^{k} \bar{a}_{l_i}}, \qquad (6.9)$$

где K - число факторов, включаемых в группу;

$\bar{a}, \bar{a}_{\max}, \bar{a}_{\min}$ - среднее значение суммы рангов фактора (средние ранги), максимальное и минимальное значения $\bar{a}_i$ в группе;

$\bar{a}_{l_i} = m \cdot \sum_{l_i}$

Если значение K_P меньше табличного (табл. 3.1) при выбранном уровне значимости α, числе сравниваемых средних К и числа специалистов *m*, то сравниваем средние ранги $\bar{a}_{l_i} \quad \left(i = \overline{1,k}\right)$ не различаются между собой.

Можно рекомендовать следующий порядок анализа результатов ранжирования. Ориентируясь на вид диаграммы рангов факторы разбивают на отдельные группы. Вычисляют K_P для первой (наиболее значимой) группы факторов. С помощью К-критерия проверяют, различаются ли между собой средние ранги факторов, включенных в эту группу, т.е. образуют ли они в действительности единую группу. Если это подтверждается, тогда объединяют факторы первой и второй (соседней) групп и с помощью К-критерия проверяют эти две группы как единую. Если эти две группы не образуют единую группу, то в первую группу включают не все факторы второй группы, а только один (или несколько) соседний с первой группой фактор и т.д.

Факторы, образующие единую группу включаются в программу исследований. Остальные факторы могут быть отброшены, как не оказывающие существенного влияния на выходную величину объекта.

Метод ранжирования факторов удобно применять при разработке модели объекта, которого еще не существует, т.е. на стадии создания. Если же имеются предварительные экспериментальные данные, полученные на объекте, то отсеивающий эксперимент можно выполнить методом корреляционного анализа.

Корреляционный анализ. Этот метод используют для установления статистической связи между переменными объекта. Суть его заключается в определении коэффициентов парной корреляции $r_{x_i y}$ между каждым входом x_i и выходом y объекта по имеющимся экспериментальным данным.

Коэффициент парной корреляции является мерой тесноты линейной статистической связи между двумя случайными величинами. В общем случае его величина изменяется от 0 до ± 1. Если $r_{x_i y} = 0$, то связь между x_i и y отсутствует или она нелинейна. Если $|r_{x_i y}| = 1$, то связь является линейной, функциональной. При нелинейной зависимости между случайными величинами для анализа меры тесноты их связи используют индекс корреляции.

В процессе отсеивающего эксперимента из рассмотрения исключаются факторы, у которых с выбранной выходной величиной объекта коэффициент корреляции близок к нулю.

Коэффициент парной корреляции между входом x_i и выходом y объекта определяется по формуле

$$r_{x_j y} = \frac{\sum_{i=1}^{N}\left(x_{ji} - \bar{x}_j\right)\cdot\left(y_i - \bar{y}\right)}{(N-1)\cdot\sigma_{x_j}\cdot\sigma_y}, \quad j = \overline{1,n}, \qquad (6.10)$$

где N - число измерений, i - номер опыта, x_{ji} - значение j-го входа в i-м опыте;

$$\sigma_{x_j} = \sqrt{\frac{1}{N-1}\cdot\sum_{i=1}^{N}\left(x_{ji} - \bar{x}_j\right)^2}; \quad \sigma_y = \sqrt{\frac{1}{N-1}\cdot\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2};$$

$$\bar{x}_j = \frac{1}{N}\cdot\sum_{i=1}^{N} x_{ji}; \quad \bar{y} = \frac{1}{N}\cdot\sum_{i=1}^{N} y_i, \quad j = \overline{1,n}.$$

При наличии высокой корреляции у двух входов x_j и x_l

$$r_{x_j x_l} = \frac{\sum_{i=1}^{N}\left(x_{ji} - \bar{x}_j\right)\cdot\left(x_{l_i} - \bar{x}_l\right)}{(N-1)\cdot\sigma_{x_j}\cdot\sigma_{x_l}}, \quad j,l = \overline{1,n}, \quad j \neq l, \qquad (6.11)$$

любой из них также можно исключить из рассмотрения, так как они однозначно влияют на выходную величину объекта. Исключается, как правило, тот фактор, который труднее определить экспериментально.

После расчета коэффициентов парной корреляции необходимо установить их статистическую значимость, т.е. проверить гипотезу об отличии вычисленного значения коэффициента от нуля. Для этого из табл. 3.2 при выбранном уровне значимости α и числе степеней свободы *f=N-2* находят критическое значение коэффициента корреляции r_x. Линейная связь считается статически значимой, если $|r_p| \geq r_x$.

Таблица 3.1.

Критические значения К-критерия при α=0,05

Число экспертов m	Число факторов в группе К										
	2	3	4	5	6	7	8	9	10	11	12
2	3.43	2.35	1.74	1.39	1.15	1.39	0.87	0.77	0.70	0.63	0.58
3	1.90	1.44	1.14	1.94	0.80	0.70	0.62	0.56	0.51	0.47	0.43
4	1.62	1.25	1.01	1.84	0.72	0.63	0.57	0.51	0.47	0.43	0.40
5	1.53	1.19	1.96	1.81	0.70	0.61	0.55	0.50	0.45	0.42	0.39
6	1.50	1.17	1.95	1.80	0.69	0.61	0.55	0.49	0.45	0.42	0.39
7	1.49	1.17	1.95	1.80	0.69	0.61	0.55	0.50	0.45	0.42	0.39
8	1.49	1.18	1.96	1.81	0.70	0.62	0.55	0.50	0.46	0.42	0.39
9	1.50	1.19	1.97	1.82	0.71	0.62	0.56	0.51	0.47	0.43	0.40
10	1.52	1.20	1.98	1.83	0.72	0.63	0.57	0.52	0.47	0.44	0.41
11	1.54	1.22	1.99	1.84	0.73	0.64	0.58	0.52	0.48	0.44	0.41
12	1.56	1.23	1.01	1.85	0.74	0.65	0.58	0.53	0.49	0.45	0.42
13	1.58	1.25	1.02	1.86	0.75	0.66	0.59	0.54	0.49	0.46	0.42
14	1.60	1.26	1.03	1.87	0.76	0.67	0.60	0.55	0.50	0.46	0.43
15	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

.62 .28 .05 .89 .77 68 61 55 51 47 44
20 1 1 1 0 0 0. 0. 0. 0. 0. 0.
.72 .36 .12 .95 .82 73 65 59 54 50 47
50 2 1 1 1 1 0. 0. 0. 0. 0. 0.
.23 .77 .45 .22 .06 94 85 77 71 65 61

Таблица 3.2

Критические значения коэффициента корреляции r_x при α =0.05

f	r_x	f	r_x	f	r_x	f	r_x
1	0.997	9	0.602	17	0.456	45	0.287
2	0.950	10	0.576	18	0.444	50	0.273
3	0.878	11	0.553	19	0.433	60	0.250
4	0.811	12	0.532	20	0.423	70	0.232
5	0.754	13	0.514	25	0.381	80	0.217
6	0.707	14	0.497	30	0.349	90	0.205
7	0.666	15	0.482	35	0.325	100	0.195
8	0.632	16	0.468	40	0.304		

6.2. Оптимальное планирование эксперимента

Главные цели планирования эксперимента заключается в достижении его максимальной точности и минимальной стоимости. Однако достижение максимальной точности требует увеличения числа опытов, что приводит к повышению стоимости эксперимента. Поэтому возникает задача оптимизации в постановке эксперимента.

Методически опыты могут быть поставлены двояко. Традиционный метод постановки опытов состоит в изменении одного какого-либо фактора при сохранении всех других факторов, влияющих на процесс, постоянными. Этот метод постановки опытов известен под названием метода однофакторного эксперимента. При такой методике взаимное влияние факторов учесть невозможно.

Оптимальное планирование эксперимента предполагает одновременное изменение всех факторов, влияющих на процесс, что позволяет сразу установить силу взаимодействия факторов и сократить общее число опытов. Такой метод постановки опытов называется методом

многофакторного планирования эксперимента. В результате исследований получают не сечения статических характеристик объекта, как это имеет место в классическом однофакторном методе, а функциональную зависимость выхода объекта от всех факторов $y = f(x_1, x_2, .., x_k)$, что очень существенно.

Теория планирования эксперимента и ее приложения в настоящее время развились в самостоятельное научное направление. Попытаемся здесь весьма кратко изложить сущность метода планированного факторного эксперимента применительно к построению модели статики объекта исследования.

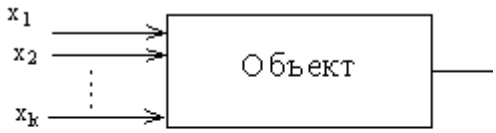


Рис. 3.2.

Сущность метода заключается в следующем. Объект исследования рассматривается как “черный ящик” (рис. 3.2). Его входы называют факторами, а выходы — откликами. Целью исследования является построение модели статики объекта, то есть функцию отклика

$$y = f(x_1, x_2, .., x_k). \qquad (6.12)$$

Координатное пространство с координатами $x_1, x_2, .., x_k$ называют факторным пространством, а геометрическое изображение функции отклика в факторном пространстве — поверхностью отклика.

В процессе эксперимента каждый фактор может принимать одно из нескольких фиксированных значений. Эти значения называются уровнями. Сочетание определенных уровней всех факторов определяет одно из возможных состояний объекта. Множество возможных сочетаний уровней факторов определяет множество состояний данного объекта и, следовательно, число возможных различных опытов, равное

$$N = p^k, \qquad (6.13)$$

где p — число уровней факторов;

k — число факторов.

Для уменьшения размерности факторного пространства и упрощения модели объекта уменьшают число факторов путем отсеивания малосущественных.

Факторы должны быть управляемыми, независимыми и совместимыми. Это означает, что факторы не должны быть функциями других факторов, должна существовать возможность установления фактора на выбранных уровнях независимо от уровня других факторов, а все комбинации уровней факторов должны давать осуществимые и безопасные для объекта режимы.

Перед началом эксперимента необходимо выбрать его план, то есть определить какие сочетания уровней факторов следует реализовать и в каком порядке. Это очень важный этап в рассматриваемом методе, поскольку этим по сути ограничивается класс регрессионных моделей, среди которых отыскивается модель объекта. Следовательно, требуется

задать общий вид отыскиваемой модели: линейная, нелинейная с эффектами взаимодействия, квадратичная и т.п.

Математическая модель чаще всего представляется в виде полинома — отрезка ряда Тейлора, в который разлагается неизвестная функция (6.12):

$$\boldsymbol{y} = \boldsymbol{\alpha}_0 + \sum_{j=1}^{k} \boldsymbol{\alpha}_j \cdot \boldsymbol{x}_j + \sum_{\substack{u,j=1 \\ u<j}}^{k} \boldsymbol{\alpha}_{uj} \cdot \boldsymbol{x}_u \cdot \boldsymbol{x}_j + \sum_{j=1}^{k} \boldsymbol{\alpha}_{jj} \cdot \boldsymbol{x}_j^2 + \ldots \quad , (6.14)$$

где $\alpha_j = \left.\frac{\partial F}{\partial x_j}\right|_0 ; \alpha_{uj} = \left.\frac{\partial^2 F}{\partial x_u \partial x_j}\right|_0 ; \alpha_{jj} = \left.\frac{\partial^2 F}{\partial x_j^2}\right|_0 ; \alpha_0 = F(\vec{x}^0)$;

индекс 0 означает, что производные вычисляются в точке $\vec{x} = \vec{x}^0$.

Поскольку в реальном объекте всегда существуют неуправляемые и неконтролируемые переменные, изменение величины y носит случайный характер. Поэтому при обработке экспериментальных данных получаются так называемые выборочные коэффициенты регрессии $a_0, a_j, a_{uj}, a_{jj}, \ldots$, являющиеся оценками теоретических коэффициентов $\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_j, \boldsymbol{\alpha}_{uj}, \boldsymbol{\alpha}_{jj}, \ldots$. Уравнение регрессии, полученное на основании опыта, запишется следующим образом:

$$\boldsymbol{y} = \boldsymbol{a}_0 + \sum_{j=1}^{k} \boldsymbol{a}_j \cdot \boldsymbol{x}_j + \sum_{\substack{u,j=1 \\ u<j}}^{k} \boldsymbol{a}_{uj} \cdot \boldsymbol{x}_u \cdot \boldsymbol{x}_j + \sum_{j=1}^{k} \boldsymbol{a}_{jj} \cdot \boldsymbol{x}_j^2 + \ldots \quad (6.15)$$

Коэффициент a_0 называют свободным членом уравнения регрессии, коэффициенты a_j — линейными эффектами, $\boldsymbol{a}_{uj}$ — эффектами парного взаимодействия, a_{jj} — квадратичными эффектами.

6.2.1. Нормирование переменных модели

В процессе эксперимента осуществляются различные комбинации уровней факторов. Эти факторы, как правило, имеют различную физическую природу и размерность. Для упрощения записи и обработки результатов уровни факторов нормируются.

Середине области определения выходной переменной объекта соответствует некоторая комбинация уровней факторов. Эти уровни факторов принимаются как основные (исходные, нулевые, начальные).

Построение плана эксперимента сводится к выбору интервала варьирования факторами Δx_j, то есть к выбору экспериментальных точек, симметричных относительно основного уровня.

Чаще всего эксперименты проводятся только на двух уровнях факторов. Такие эксперименты называются экспериментами типа 2^k. При

таком эксперименте уровни факторов представляют собой границы исследуемой области по данному фактору $(x_{j\,\min}, x_{j\,\max})$. Тогда для любого фактора имеем

$$x_j^0 = \frac{x_{j\,\min} + x_{j\,\max}}{2}, \; j = \overline{1,k}; \qquad (6.16)$$

$$\Delta x_j = \frac{x_{j\,\max} - x_{j\,\min}}{2}, \; j = \overline{1,k}. \qquad (6.17)$$

Точка с координатами $(x_1^0, x_2^0, .., x_k^0)$ называется центром плана (основным уровнем); Δx_j — интервал варьирования по оси x_j. Величина Δx_j своя для каждого фактора. Прибавление Δx_j к основному уровню дает верхний, а вычитание — нижний уровень фактора.

От системы координат $x_1, x_2, .., x_k$ перейдем к новой безразмерной системе координат $\hat{x}_1, \hat{x}_2, .., \hat{x}_k$ путем следующего линейного преобразования координат

$$\hat{x}_j = \frac{x_j - x_j^0}{\Delta x_j}, \; j = \overline{1,k}, \qquad (6.18)$$

где $\hat{x}_j$ — нормированное значение j- го фактора;

x_j — натуральное значение j- го фактора $(x_j = x_j^0 \pm \Delta x_j)$;

x_j^0 — натуральное значение основного уровня.

В нормированном виде для всех факторов верхний уровень равен +1, нижний уровень равен $-1 (-1 \le \hat{x}_j \le 1)$, координаты центра плана равны нулю и совпадают с началом координат.

Нормирование переменных существенно упрощает планирование эксперимента, обработку его результатов, построение модели объекта с безразмерными переменными, упрощение модели и проверку ее адекватности. После построения модели с нормированными факторами $y = F(\hat{x}_1, \hat{x}_2, .., \hat{x}_k)$ и оценки ее адекватности от этой модели переходят к модели с натуральными факторами $y = F(x_1, x_2, .., x_k)$, используя нормирующие соотношения (6.18.).

6.2.2. Планирование полного факторного эксперимента

При планировании по схеме полного факторного эксперимента (ПФЭ) реализуются все возможные сочетания уровней факторов. Необходимое количество опытов N при ПФЭ определяется по формуле (6.13).

Если эксперименты проводятся только на двух уровнях факторов и при этом в процессе эксперимента осуществляется все возможные комбинации из k факторов, то постановка опытов по такому плану называется ПФЭ типа 2^k.

ПФЭ типа 2^k широко используется благодаря таким их положительным особенностям, как симметричность относительно центра эксперимента, независимость численных значений коэффициентов уравнения регрессии, полученного экспериментально, от изменения порядка регрессионного полинома, принятого для описания объекта (ортогональность), одинаковая точность модели на равных расстояниях от центра в различных направлениях (ротатабельность). ПФЭ обычно используется при небольших числах факторов $(k \le 5)$.

С помощью ПФЭ типа 2^k может быть получено адекватная модель объекта только в том случае, если в нее не входят степени факторов, то есть этот эксперимент позволяет определить оценки коэффициентов уравнения регрессии только при линейных членах и взаимодействиях:

$$\boldsymbol{y} = \boldsymbol{a}_0 + \sum_{j=1}^{k} \boldsymbol{a}_j \cdot \boldsymbol{x}_j + \sum_{\substack{u,j=1 \\ u<j}}^{k} \boldsymbol{a}_{uj} \cdot \boldsymbol{x}_u \cdot \boldsymbol{x}_j + \sum_{\substack{u,j,l=1 \\ (u<j<l)}}^{k} \boldsymbol{a}_{ujl} \cdot \boldsymbol{x}_u \cdot \boldsymbol{x}_j \cdot \boldsymbol{x}_l + \ldots \qquad (6.19)$$

Взаимодействие факторов имеет место тогда, когда эффект одного фактора зависит от уровня, на котором находится другой фактор, то есть статическая характеристика объекта нелинейна.

Условия эксперимента описывают с помощью специальной таблицы — матрицы планирования, в которой каждая строка соответствует условиям одного опыта. Для проведения опытов исходные данные по всем факторам желательно свести в таблицу (табл. 3.3). Рассмотрим план ПФЭ 2^2. Число опытов $N = 4$.

Матрица плана ПФЭ типа 2^2 приведена в табл. 3.4, в которой введен столбец фиктивной переменной $\hat{x}_0 = +1$ для вычисления свободного члена a_0. Для сокращения записи обычно опускают единицы в матрице планирования. В первом опыте оба фактора находятся на верхнем уровне, во втором фактор x_1 — на нижнем, а фактор x_2 — на верхнем и т.д.

Таблица 3.3				
Исходные данные				
Значения факторов	1	2	....	k
x_j^0				

	Таблица 3.4		
	Матрица плана ПФЭ 2^2		
Номер опыта	$\hat{x}_0$	$\hat{x}_1$	$\hat{x}_2$
1	+1	+1	+1

Δx_j						2	+1	-1	+1
$x_{j\min}$						3	+1	+1	-1
$x_{j\max}$						4	+1	-1	-1

В случае $k > 2$ при построении матрицы плана ПФЭ типа 2^k руководствуются следующим. При любом числе факторов k необходимо дважды повторить матрицу планирования для случая $k-1$ сначала при значении k-го фактора на верхнем уровне, а затем на нижнем. В табл. 3.5 показано последовательное достраивание матрицы ПФЭ при увеличении k от 2 до 4. Для простоты записи плана единицы опущены, а оставлены только знаки.

Таблица 3.5

Матрица плана ПФЭ 22, 23, 24

		План		№ опыта	$\hat{x}_0$	$\hat{x}_1$	$\hat{x}_2$	$\hat{x}_3$	$\hat{x}_4$
				1	+	+	+	+	+
				2	+	—	+	+	+
		2^2		3	+	+	—	+	+
				4	+	—	—	+	+
	3			5	+	+	+	—	+
				6	+	—	+	—	+
				7	+	+	—	—	+
				8	+	—	—	—	+
4				9	+	+	+	+	—
				10	+	—	+	+	—
				11	+	+	—	+	—
				12	+	—	—	+	—
				13	+	+	+	—	—

				14	+	—	+	—	—
				15	+	+	—	—	—
				16	+	—	—	—	—

Матрица плана ПФЭ типа 2^k обладает следующими свойствами:

$$\sum_{i=1}^{N} \hat{x}_{ui} \cdot \hat{x}_{ji} = 0,\ u \neq j,\ u, j = \overline{0,k} \quad ; \qquad (6.20)$$

$$\sum_{i=1}^{N} \hat{x}_{ji} = 0,\ j = \overline{1,k}\ (j \neq 0) \quad ; \qquad (6.21)$$

$$\sum_{i=1}^{N} \hat{x}_{ji}^2 = N,\ j = \overline{0,k} \quad , \qquad (6.22)$$

где $\hat{x}_{ji}$ — нормированное значение j-го фактора в i-ом опыте.

Свойство (3.20) называется свойством ортогональности матрицы планирования. Это свойство резко уменьшает трудности, связанные с расчетом коэффициентов модели.

6.2.3. Вычисление коэффициентов модели

Приведем метод вычисления коэффициентов линейной модели с нормированными факторами

$$y = a_0 + \sum_{j=1}^{k} a_j \cdot \hat{x}_j \qquad (6.23)$$

по результатам ПФЭ типа 2^k.

Используем матричный метод регрессионного анализа, который очень удобен для решения задач на ЭВМ. Для этого введем следующие матрицы:

матрицу независимых переменных

$$X = \begin{pmatrix} \hat{x}_{01} & \hat{x}_{11} & \hat{x}_{21} & \dots & \hat{x}_{k1} \\ \hat{x}_{02} & \hat{x}_{12} & \hat{x}_{22} & \dots & \hat{x}_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{x}_{0N} & \hat{x}_{1N} & \hat{x}_{2N} & \dots & \hat{x}_{kN} \end{pmatrix}, \qquad (6.24)$$

матрицу — столбец откликов (вектор наблюдений) и матрицу — столбец коэффициентов

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \; A = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_k \end{pmatrix}, \qquad (6.25)$$

и матрицу, транспонированную к X

$$X^T = \begin{pmatrix} \hat{x}_{01} & \hat{x}_{02} & \hat{x}_{03} & \dots & \hat{x}_{0N} \\ \hat{x}_{11} & \hat{x}_{12} & \hat{x}_{13} & \dots & \hat{x}_{1N} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{x}_{k1} & \hat{x}_{k2} & \hat{x}_{k3} & \dots & \hat{x}_{kN} \end{pmatrix}. \qquad (6.26)$$

В матричной форме система нормированных уравнений для определения коэффициентов $a_j \; (j = \overline{0,k})$ методом наименьших квадратов запишется следующим образом:

$$X^T X A = X^T Y. \qquad (6.27)$$

Отсюда

$$A = (X^T X)^{-1} X^T Y. \qquad (6.28)$$

Матрица моментов (информационная матрица Фишера) $X^T X$ имеет вид

$$X^T X = \begin{pmatrix} \sum \hat{x}_{0i}^2 & \sum \hat{x}_{0i}\hat{x}_{1i} & \dots & \sum \hat{x}_{0i}\hat{x}_{ki} \\ \sum \hat{x}_{1i}\hat{x}_{0i} & \sum \hat{x}_{1i}^2 & \dots & \sum \hat{x}_{1i}\hat{x}_{ki} \\ \vdots & \vdots & \vdots & \vdots \\ \sum \hat{x}_{ki}\hat{x}_{0i} & \sum \hat{x}_{ki}\hat{x}_{1i} & \dots & \sum \hat{x}_{ki}^2 \end{pmatrix}, \qquad (6.29)$$

где суммирование всюду осуществляется по i от 1 до N.

Учитывая свойства ортогональности (6.20) матрица коэффициентов нормальных уравнений $X^T X$ становится диагональной и ее диагональные элементы равны N

$$X^T X = \begin{pmatrix} N & 0 & 0 & \dots & 0 \\ 0 & N & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & N \end{pmatrix}. \qquad (6.30)$$

Матрица, обратная матрица моментов $(X^T X)^{-1}$, получается равной

$$(X^T X)^{-1} = \begin{pmatrix} 1/N & 0 & 0 & \dots & 0 \\ 0 & 1/N & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1/N \end{pmatrix}. \qquad (6.31)$$

Далее имеем:

$$X^T Y = \begin{pmatrix} \sum_{i=1}^{N} \hat{x}_{0i} y_i \\ \sum_{i=1}^{N} \hat{x}_{1i} y_i \\ \vdots \\ \sum_{i=1}^{N} \hat{x}_{ki} y_i \end{pmatrix}; \qquad (6.32)$$

$$A = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_k \end{pmatrix} = \begin{pmatrix} 1/N & 0 & 0 & \cdots & 0 \\ 0 & 1/N & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1/N \end{pmatrix} \times \begin{pmatrix} \sum_{i=1}^{N} \hat{x}_{0i} y_i \\ \sum_{i=1}^{N} \hat{x}_{1i} y_i \\ \vdots \\ \sum_{i=1}^{N} \hat{x}_{ki} y_i \end{pmatrix} = \begin{pmatrix} \frac{1}{N} \sum_{i=1}^{N} \hat{x}_{0i} y_i \\ \frac{1}{N} \sum_{i=1}^{N} \hat{x}_{1i} y_i \\ \vdots \\ \frac{1}{N} \sum_{i=1}^{N} \hat{x}_{ki} y_i \end{pmatrix}, \qquad (6.33)$$

где суммирование осуществляется по i от 1 до N.

Следовательно, любой коэффициент a_j определяется по формуле

$$a_j = \frac{1}{N} \sum_{i=1}^{N} \hat{x}_{ji} y_i \ , \ j = \overline{0,k} \ . \qquad (6.34)$$

Если в рассмотрение ввести нелинейную модель с коэффициентами взаимодействия

$$\boldsymbol{y} = \boldsymbol{a}_0 + \sum_{j=1}^{k} \boldsymbol{a}_j \cdot \boldsymbol{x}_j + \sum_{\substack{u,j=1 \\ u<j}}^{k} \boldsymbol{a}_{uj} \cdot \boldsymbol{x}_u \cdot \boldsymbol{x}_j + \sum_{\substack{u,j,l=1 \\ (u<j<l)}}^{k} \boldsymbol{a}_{ujl} \cdot \boldsymbol{x}_u \cdot \boldsymbol{x}_j \cdot \boldsymbol{x}_l + \ldots \qquad (6.35)$$

то эффекты взаимодействия определяются аналогично линейным эффектам:

$$a_{uj} = \frac{1}{N} \sum_{i=1}^{N} \hat{x}_{ui} \hat{x}_{ji} y_i \ ; \ u, j = \overline{1,k}, \ u < j \ ; \qquad (6.36)$$

$$a_{ujl} = \frac{1}{N} \sum_{i=1}^{N} \hat{x}_{ui} \hat{x}_{ji} \hat{x}_{li} y_i \ ; \ u, j, l = \overline{1,k}, \ u < j < l \ . \qquad (6.37)$$

и т.д.

В связи с тем, что матрица $(X^T X)^{-1}$ для спланированного эксперимента — матрица диагональная коэффициенты уравнения регрессии некоррелированы между собой. При этом выборочные коэффициенты a_j оказываются несмешанными оценками для соответствующих генеральных коэффициентов $\alpha_j (a_j \to \alpha_j)$, то есть величины коэффициентов уравнения регрессии характеризуют вклад каждого фактора в величину y.

Если выход объекта зависит еще и от квадрата факторов

$$y = a_0 + \sum_{j=1}^{k} a_j \cdot x_j + \cdots + \sum_{j=1}^{k} a_{jj} \cdot x_j^2 \quad , \qquad (6.38)$$

то эксперимент типа 2^k не позволяет определить квадратичные эффекты a_{jj}. Действительно, значения $\hat{x}_{ji}^2 \ (j = \overline{1,k})$ не отличаются друг от друга и от вектор столбца $\hat{x}_0$. Поэтому величина a_0 представляет собой смешанную оценку значения свободного члена и вкладов квадратичных членов, то есть

$$a_0 \to \alpha_0 + \sum_{j=1}^{k} \alpha_{jj} \quad ,$$

где a_0 — вычисленный выше коэффициент;

α_0, α_{jj} — неизвестные истинные значения a_0, a_{jj}.

Для получения модели типа (6.38), включающей степени факторов, необходимо проводить эксперимент при $p > 2$.

6.2.4. Оценка значимости коэффициентов модели.

Для упрощения полученной модели проверяют значимость ее коэффициентов. Ее выполняют для каждого коэффициента отдельно. Оценка значимости коэффициентов производится по критерию Стьюдента, либо по доверительному интервалу, который при использовании факторного эксперимента одинаков для всех коэффициентов.

Критерий Стьюдента

$$t_j = \frac{|a_j|}{\sigma_{aj}} \ , j = 0,1,2,... \quad , \qquad (6.39)$$

где a_j — j-й коэффициент уравнения регрессии;

σ_{aj} — среднее квадратичное отклонение коэффициента a_j.

Диагональные элементы ковариационной матрицы $(X^T X)^{-1}$ равны между собой, поэтому все коэффициенты уравнения регрессии определяются с одинаковой точностью:

$$\sigma_{aj} = +\sqrt{\sigma_{aj}^2} = +\sqrt{\frac{\sigma_{\text{воспр}}^2}{N}} \quad , \qquad (6.40)$$

где $\sigma_{\text{воспр}}^2$ — дисперсия воспроизводимости.

Дисперсия воспроизводимости, как правило, определяется с помощью повторных опытов в нулевой точке (центре эксперимента):

$$\sigma_{\text{воспр}}^2 = \frac{S_{\text{воспр}}}{f_{\text{воспр}}} = \frac{1}{m-1} \sum_{i=1}^{m} (y_i^0 - \bar{y}^0)^2 \quad , \qquad (6.41)$$

где m — число повторных опытов;

$f_{\text{воспр}} = m - 1$ — число степеней свободы дисперсии воспроизводимости;

$S_{\text{воспр}}$ — остаточная сумма квадратов воспроизводимости;

$$\bar{y}^0 = \frac{1}{m}\sum_{i=1}^{m} y_i^0 \quad . \qquad (6.42)$$

Таким образом, вычисляют t_i для каждого i-го коэффициента. Для t - критерия имеются специальные таблицы распределения Стьюдента, рассчитанные для различных $f_{\text{воспр}}$, m и доверительных вероятностей P (уровней значимости q).

Коэффициент a_j считается значимым (значимо отличается от нуля) с некоторой доверительной вероятностью, если вычисленный t_i больше табличного значения t_m при такой же доверительной вероятности и соответствующем числе степеней свободы.

Для оценки значимости коэффициентов по доверительному интервалу вычисляют доверительный интервал для j-го коэффициента по формуле

$$\Delta a_j = \pm t_m \cdot \sigma_{aj} = \pm t_m \sqrt{\frac{\sigma^2_{\text{воспр}}}{N}} \quad . \qquad (6.43)$$

Как было отмечено выше, доверительный интервал Δa_j одинаков для всех коэффициентов. Определение значимости коэффициентов уравнения регрессии с помощью соотношения (6.43) удобно тем, что позволяет применить правило: коэффициент значим, если его абсолютная величина больше доверительного интервала.

Незначимые коэффициенты исключаются из модели. При этом если коэффициенты модели некоррелированны между собой (матрица моментов диагональная), то исключение незначимых коэффициентов не скажется на остальных коэффициентах. В обратном случае оставшиеся коэффициенты пересчитываются заново, поскольку коэффициенты закоррелированы друг с другом.

6.2.5. Проверка адекватности модели

Полученную с помощью факторного планированного эксперимента модель объекта необходимо проверить на адекватность. Это осуществляется с помощью критерия Фишера (при наличии степеней свободы):

$$F = \sigma_1^2 / \sigma_2^2 \text{ , (3.44)}$$

где σ_1^2, σ_2^2 — дисперсии двух независимых выборок.

При расчете F — критерия в числитель ставят большую из двух эмпирических дисперсий, поэтому $F \geq 1$. Критерий Фишера показывает значимо ли отличаются друг от друга эти дисперсии.

F — критерий формируют следующим образом:

$$F = \begin{cases} \sigma^2_{\text{ад}} / \sigma^2_{\text{воспр}}, & \text{если } \sigma^2_{\text{ад}} \geq \sigma^2_{\text{воспр}}; \\ \sigma^2_{\text{воспр}} / \sigma^2_{\text{ад}}, & \text{если } \sigma^2_{\text{ад}} < \sigma^2_{\text{воспр}}, \end{cases} \tag{6.45}$$

где $\sigma^2_{\text{ад}}$ — дисперсия адекватности (остаточная дисперсия).

Дисперсия воспроизводимости вычисляется по формуле (6.41), а дисперсия адекватности по формуле

$$\sigma^2_{\text{ад}} = \frac{1}{N-r} \sum_{i=1}^{N} (y_i - y_i^M)^2, \tag{6.46}$$

где r — число значимых коэффициентов модели;

y_i, y_i^M — экспериментальные отклики и отклики, рассчитанные по модели при тех же значениях факторов.

Число степеней свободы для дисперсии адекватности

$$f_{\text{ад}} = N - r. \tag{6.47}$$

Для F-критерия имеются специальные таблицы распределения Фишера, рассчитанные для различных $f_{\text{ад}} = f_1$, $f_{\text{воспр}} = f_2$ и доверительных вероятностей P (уровней значимости q). Если

$$F \leq F_m(f_1, f_2, p), \tag{6.48}$$

то с соответствующей доверительной вероятностью P можно принять, что $\sigma^2_{\text{ад}} = \sigma^2_{\text{воспр}}$ и считать модель в исследованной области адекватной объекту.

Таким образом, с помощью F-критерия проверяется лишь тот факт, что дисперсия адекватности не отличается значимо от дисперсии воспроизводимости, то есть ошибка модели и ошибка связанная с точностью эксперимента, на основе которого построена модель, близки друг другу. Этого достаточно для утверждения, что полученная модель адекватно описывает результаты эксперимента.

Если $F > F_m(f_1, f_2, p)$, то для адекватного описания эксперимента необходимо увеличить порядок модели (например, ввести в модель квадраты факторов). В этом случае эксперимент проводится при числе уровней факторов больше чем 2.

Для случаев, когда повторные опыты ставятся во всех (или в нескольких) точках плана одинаковое или неодинаковое число раз, формулы (3.41) и (3.46) несколько изменяются.

6.2.6. Дробный факторный эксперимент

С увеличением числа факторов количество опытов при ПФЭ растет по показательной функции. В то же время часто заранее известно, что число определяемых коэффициентов в модели объекта намного меньше числа опытов в соответствующем ПФЭ (например, если можно ограничиться линейным приближением модели). В этом случае используя дробный факторный эксперимент (ДФЭ) и получая дробные реплики можно резко сократить число опытов.

ДФЭ позволяет минимизировать число опытов, но при этом оценки коэффициентов получаются смешанными.

Эффективность ДФЭ при определении модели объекта зависит от выбранной системы смешивания. Необходимо выбрать такие дробные реплики, при которых интересующие экспериментатора линейные эффекты смешиваются с теми эффектами взаимодействия, о которых с наибольшей вероятностью можно предположить, что они близки к нулю. Как правило, такое предположение чаще всего справедливо относительно взаимодействий самого высокого порядка.

Для определения системы смешивания оценок используется так называемый определяющий контраст — символическое произведение всех столбцов матрицы планирования равное +1 или -1. Чтобы узнать какие взаимодействия смешаны с данным линейным эффектом (то есть в каких столбцах плана имеются одинаковые элементы) необходимо умножить определяющий контраст на столбец соответствующий данному эффекту.

Положим, что контраст в случае трех факторов равен

$$\hat{x}_1\hat{x}_2\hat{x}_3 = 1. \qquad (6.49)$$

Умножим по очереди определяющий контраст на $\hat{x}_1, \hat{x}_2, \hat{x}_3$, получим

$$\left.\begin{aligned} \hat{x}_1 &= \hat{x}_1^{\,2}\hat{x}_2\hat{x}_3 = \hat{x}_2\hat{x}_3; \\ \hat{x}_2 &= \hat{x}_2^{\,2}\hat{x}_1\hat{x}_3 = \hat{x}_1\hat{x}_3; \\ \hat{x}_3 &= \hat{x}_3^{\,2}\hat{x}_1\hat{x}_2 = \hat{x}_1\hat{x}_2, \end{aligned}\right\} \qquad (6.50)$$

поскольку $\hat{x}_i^{\,2} = 1$.

Полученным соотношениям (6.50) соответствует система смешанных оценок:

$$\left.\begin{aligned} a_1 &\to \alpha_1 + \alpha_{23}; \\ a_2 &\to \alpha_2 + \alpha_{13}; \\ a_3 &\to \alpha_3 + \alpha_{12}. \end{aligned}\right\} \qquad (6.51)$$

При увеличении числа факторов целесообразность использования дробных реплик для получения линейной модели объекта резко возрастает.

Рассмотрим пример ДФЭ. Положим, что интересующая нас модель объекта является функцией трех переменных (факторов), причем заранее известно, что в рабочей области эта модель близка к линейной, то есть

$$y = a_0 + a_1\hat{x}_1 + a_2\hat{x}_2 + a_3\hat{x}_3 \qquad (6.52)$$

При ПФЭ и двух уровнях факторов мы должны осуществить $N = 2^3 = 8$ опытов. Следовательно, для определения четырех коэффициентов $a_j (j = \overline{0,3})$ модели проводится вдвое больше число опытов.

Используя ДФЭ для решения поставленной задачи можно уменьшить число опытов. Чтобы дробная реплика представляла собой ортогональный план, в качестве реплики следует брать ПФЭ для меньшего числа факторов. Число опытов при этом должно быть больше (или равно) числа неизвестных коэффициентов модели. Следовательно, для рассматриваемого примера можно ограничиться четырьмя опытами. При этом ортогональность плана достигается если использовать план ПФЭ типа 2^2, а в качестве плана для $\hat{x}_3$использовать столбец $\hat{x}_1\hat{x}_2$. Матрица такого эксперимента приведена в таблице 3.6.

Таблица 3.6

Полуреплика от ПФЭ типа 23

№ опыта	$\hat{x}_0$	$\hat{x}_1$	$\hat{x}_2$	$\hat{x}_3 = \hat{x}_1 \cdot$	y
1	+	+	+	+	y_1
2	+	—	+	—	y_2
3	+	+	—	—	y_3
4	+	—	—	+	y_4

Такой сокращенный план (половина ПФЭ) называется полурепликой. Пользуясь таким планированием, можно оценить свободный член и три коэффициента модели при линейных членах. Но если модель объекта не строго линейна и коэффициенты при парных взаимодействиях не равны нулю, то найденные коэффициенты будут смешанными в соответствии с системой (3.50), поскольку элементы столбцов $\hat{x}_1\hat{x}_2$, $\hat{x}_1\hat{x}_3$и $\hat{x}_2\hat{x}_3$(эти столбцы в табл. 3.6 не указаны) совпадают с элементами столбцов $\hat{x}_3, \hat{x}_2, \hat{x}_1$ соответственно. Если модель объекта действительно близка к линейной, то со смешанностью коэффициентов вполне можно примириться.

По указанному в табл. 3.6 плану проводится половина ПФЭ типа 2^3и получается полуреплика от 2^3, что принято обозначать как 2^{3-1}. Если знаки одного из элементов столбцов ($\hat{x}_1$, $\hat{x}_2$или $\hat{x}_3$) изменить на противоположные, то получим вторую половину ПФЭ (еще 4 опыта) и, следовательно, вторую полуреплику от 2^3. При осуществлении обоих полуреплик мы провели бы ПФЭ 2^3(8 опытов) и получили бы раздельные оценки коэффициентов как для линейных эффектов, так и для эффектов взаимодействий.

Пусть $k = 7$, $l = 4$, тогда план ДФЭ 2^{7-4}включает матрицу плана ПФЭ 2^3для трех факторов $\hat{x}_1$, $\hat{x}_2$и $\hat{x}_3$, а столбцы плана для остальных факторов представляют собой некоторые произведения факторов $\hat{x}_1$, $\hat{x}_2$и $\hat{x}_3$,

например, $\hat{x}_4 = \hat{x}_1\hat{x}_2$, $\hat{x}_5 = \hat{x}_1\hat{x}_3$, $\hat{x}_6 = \hat{x}_2\hat{x}_3$, $\hat{x}_7 = \hat{x}_1\hat{x}_2\hat{x}_3$. Такие равенства называют генерируемыми соотношениями, их выбор произволен, однако повторение не допускается. При таком планировании все линейные эффекты смешаны с эффектами взаимодействия.

Число опытов в дробной реплике должно удовлетворять неравенству $k+1 \le N < 2^k$, где k — число факторов, для получения несмешанных оценок линейных эффектов.

Если $N = k+1$, то есть число опытов равно числу определяемых коэффициентов линейной модели, дробная реплика представляет собой насыщенный линейный ортогональный план.

Поскольку число опытов в насыщенных планах равно числу определяемых коэффициентов, число степеней свободы дисперсии адекватности $f_{\text{ад}} = 0$. Для поверки адекватности линейной модели, полученной по насыщенному плану, необходим дополнительный эксперимент.

Таким образом оптимальные двухуровневые планы 2^k и 2^{k-l} имеют следующие преимущества:

- планы ортогональны, и поэтому все вычисления просты, все коэффициенты определяются независимо друг от друга;
- каждый коэффициент определяется по результатам всех N опытов;
- для данного числа опытов N эти планы имеют минимальный определитель ковариационной матрицы $(X^T X)^{-1}$; в связи с этим все коэффициенты определяются с одинаковой и минимальной дисперсией, то есть планы обладают свойством Д-оптимальности;
- линейные планы 2^k и 2^{k-l} обладают свойством ротатабельности.

Покажем справедливость свойства ротатабельности в случае линейной модели

$$y = a_0 + \sum_{j=1}^{k} a_j \hat{x}_j \quad . \qquad (6.53)$$

Вследствие отсутствия корреляции между коэффициентами по закону сложения дисперсий для линейной модели (6.53) имеем:

$$\sigma_y^2 = \sigma_{a0}^2 + \hat{x}_1^2\sigma_{a1}^2 + \cdots + \hat{x}_k^2\sigma_{ak}^2 .$$

Так как $\sigma_{aj}^2 = \dfrac{\sigma_{\text{воспр}}^2}{N}, j = \overline{1,k}$, получим

$$\sigma_y^2 = \frac{\sigma_{\text{воспр}}^2}{N}(1 + \hat{x}_1^2 + \hat{x}_2^2 + \ldots + \hat{x}_k^2) = \frac{\sigma_{\text{воспр}}^2}{N}(1+\rho^2) ;$$

$$\rho^2 = \sum_{j=1}^{k} \hat{x}_j^2 ,$$

где ρ — радиус сферы в К-мерном пространстве.

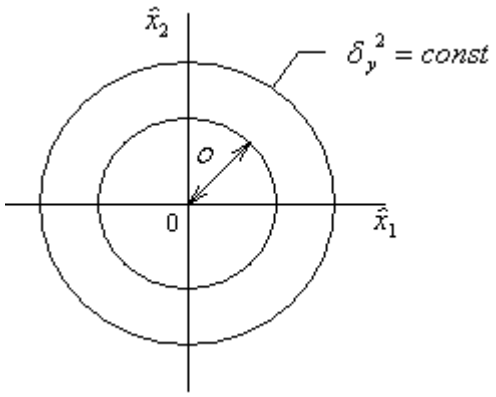


Рис. 3.3.

Дисперсия выхода y растет пропорционально квадрату радиуса сферы ρ^2 и одинаково для всех эквидистантных точек (рис. 3.3).

Планирование обладающее таким свойством называется ротатабельным планированием.

6.2.7. Учет влияния помех

При постановке многофакторного эксперимента на реальном объекте необходимо спланировать его так, чтобы иметь возможность отфильтровать помехи. С этой целью каждый опыт в матрице планирования повторяется в m раз (табл. 3.7). Обычно $m = 3 \div 5$

Таблица 3.7

ПФЭ типа 2k

№ опыта	.	$\hat{x}$	$\hat{x}_2$	.. ...	$\hat{x}$	$\boldsymbol{y}$	$\bar{\boldsymbol{y}}$	$\boldsymbol{\sigma}^2$
1	-	+	+		+	$y_{11}, y_{12}, \dots, y_{1m}$	$\bar{\boldsymbol{y}}_1$	$\boldsymbol{\sigma}^2{}_1$
2	-	—	+		+	$y_{21}, y_{22}, \dots, y_{2m}$	$\bar{\boldsymbol{y}}_2$	$\boldsymbol{\sigma}^2{}_2$
3	-	+	—		+	$y_{31}, y_{32}, \dots, y_{3m}$	$\bar{\boldsymbol{y}}_3$	$\boldsymbol{\sigma}^2{}_3$
.	.	.	.		.	.	.	.
.	.	.	.		.	.	.	.
.	.	.	.		.	.	.	.
N	-	—	—		—	$y_{N1}, y_{N2}, \dots, y_{Nm}$	$\bar{\boldsymbol{y}}_N$	$\boldsymbol{\sigma}^2{}_N$

В каждой строчке матрицы планирования определяются среднее значение и дисперсия выхода объекта по m параллельным опытам:

$$\bar{y}_i = \frac{1}{m}\sum_{j=1}^{m} y_{ij}; \quad \sigma_i^2 = \frac{1}{m-1}\sum_{j=1}^{m}(y_{ij} - \bar{y}_i)^2, \quad i = \overline{1, N}, \qquad (6.54)$$

где y_{ij} — j-ое значение выхода объекта в i-ом опыте.

Проверяется однородность выборочных дисперсий (воспроизводимость эксперимента) по критерию Кохрена. Для этого

определяется отношение максимальной дисперсии к сумме всех дисперсий:

$$G_{\text{рас}} = \frac{\max\limits_i \sigma_i^2}{\sum\limits_{i=1}^{N} \sigma_i^2} \qquad (6.55)$$

Полученное отношение сравнивается с табличным значением (определяется из таблицы распределения Кохрена) G_T при заданной доверительной вероятности P и числе степеней свободы $f_1 = m-1, f_2 = N$.

Если $G_{\text{рас}} < G_m$ дисперсии σ_i^2 однородны (опыты воспроизводимы), в противном случае добиваются достижения однородности дисперсий σ_i^2, как правило, путем увеличения числа повторных опытов m.

Тогда в качестве оценки для дисперсии воспроизводимости

$$\sigma^2_{\text{воспр}} = \sigma_i^2 = \frac{1}{N(m-1)} \sum_{i=1}^{N} \sum_{j=1}^{m} (y_{ij} - \bar{y}_i)^2$$

можно взять среднюю дисперсию

$$\sigma^2_{\text{воспр}} = \frac{1}{N} \sum_{i=1}^{N} \sigma_i^2 \qquad (6.56)$$

с числом степеней свободы $f_{\text{воспр}} = N(m-1)$.

Коэффициенты модели определяются изложенным в разделе 3.2.3 способом. При этом в формулах (6.34), (6.36), (6.37) для вычисления коэффициентов следует вместо y_i принимать значение $\bar{y}_i$. Например, коэффициенты линейных эффектов определяются по формуле

$$a_j = \frac{1}{N} \sum_{i=1}^{N} \hat{x}_{ji} \bar{y}_i, \; j = \overline{0,k}.$$

Упрощение модели путем отсеивания ее незначимых коэффициентов осуществляется по критерию Стьюдента (раздел 6.2.4), а проверка адекватности модели — по критерию Фишера (раздел 6.2.5). Однако, в данном случае необходимо учитывать следующее. Дисперсия воспроизводимости определяется по формуле (6.56). Учитывая, что дисперсия среднего $\bar{y}$, полученного по выборке объема m, в m раз меньше дисперсии единичного измерения y, то есть

$$\sigma^2_{\bar{y}} = \frac{1}{m} \sigma^2_y = \frac{1}{m} \sigma^2_{\text{воспр}}, \qquad (6.57)$$

в рассматриваемом случае дисперсия коэффициентов σ_{aj}^2 определяется следующим образом:

$$\sigma^2_{aj} = \frac{1}{Nm} \sigma^2_{\text{воспр}} \quad \forall j. \qquad (6.58)$$

В матрице планирования каждый опыт повторялся в m раз. Поэтому дисперсию адекватности нужно умножить на m, то есть

$$\sigma^2_{ад} = \frac{m}{N-r}\sum_{i=1}^{N}(\bar{y}_i - y_i^M)^2 \quad . \qquad (6.59)$$

Незначимые коэффициенты отсеиваются из уравнения регрессии. При этом ввиду ортогональности матрицы планирования остальные коэффициенты не приходится пересчитывать. В случае же не подтверждения адекватности модели необходимо увеличить порядок аппроксимирующего полинома.

Необходимость повышения порядка модели может определяться также исходя из следующего соображения. На практике часто оказывается, что линейная модель, адекватно описывающая опытные данные, поставленные в точках плана, неудовлетворительно характеризует внутреннюю часть изученной области факторного пространства. Поэтому имеет смысл поставить опыт в центре плана $(\hat{x}_j = \hat{x}_j^0,\ j = \overline{0,k})$. Среднее значение выхода объекта в центре плана должно быть соизмеримо (в пределах ошибки воспроизводимости) с коэффициентом а0 модели, то есть

$$\left|a_0 - \bar{y}^0\right| < \delta, \qquad (6.60)$$

где $\bar{y}^0 = \frac{1}{m_1}\sum_{i=1}^{m_1} y_i^0$, m_1— число повторных опытов в центре плана, δ— малое число.

В случае нарушения условия (6.60) для математического описания рассматриваемой области факторного пространства потребуется уравнение более высокого порядка.

6.2.8. Построение модели с квадратичными эффектами

Построение модели, содержащую степени факторов, требует иметь такую систему планирования, в которой каждый фактор будет принимать хотя бы три разных уровня. ПФЭ типа 3^kсодержит слишком большое число опытов. При $k > 5$ПФЭ 3^kстановится совершенно неприемлемым для практики.

Сократить число опытов можно, если воспользоваться композиционными или последовательными планами, предложенными Боксом и Уилсоном. Ядро таких планов составляет ПФЭ 2^kпри $k < 5$или полуреплика от него ДФЭ 2^{k-1}при $k \geq 5$. Возможность использования в качестве ядра плана 2^{k-1}при $k \geq 5$обусловлена тем, что уже полуреплика обеспечивает получение несмешанных оценок для линейных эффектов и эффектов парного взаимодействия.

Если линейная модель (а также модель с эффектами взаимодействия) оказалась неадекватной, необходимо:

1) добавить $2k$звездных точек, расположенных на координатных осях факторного пространства. Координаты звездных точек:

$$(\pm\alpha,0,0,\ldots,0);\ (0,\pm\alpha,0,\ldots,0);\ \ldots\ ;(0,0,0,\ldots\pm\alpha),$$

где α — расстояние от центра плана до звездной точки (звездное плечо);

2) увеличить число опытов n_0 в центре плана.

Рассмотрим построение композиционного плана при $k=2$ (рис. 3.4). Точки 1 — 4 образуют ПФЭ 2^2, точки 5 — 8 — звездные точки с координатами $(\pm\alpha,0)$ и $(0,\pm\alpha)$, координаты n_0 опытов в центре плана нулевые — $(0,0)$.

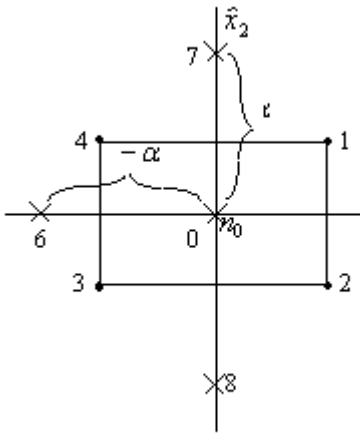


Рис. 3.4

Общее число опытов в матрице композиционного плана второго порядка при k факторах:

$N = 2^k + 2k + n_0$ при $k < 5$;

$N = 2^{k-1} + 2k + n_0$ при $k \geq 5$.

Матрица плана приведена в табл. 3.8. Композиционные планы легко приводятся к ортогональным выбором соответствующего звездного плеча α. На количество опытов в центре плана n_0 при этом не накладывается никаких ограничений (обычно $n_0 = 1$). При этом достаточно определить α из условия:

$\alpha^4 + 2^k \alpha^2 - 2^{k-1}(k + 0{,}5n_0) = 0$ при $k < 5$;

$\alpha^4 + 2^{k-1}\alpha^2 - 2^{k-2}(k + 0{,}5n_0) = 0$ при $k \geq 5$.

Таблица 3.8

Матрица композиционного плана

№ опыта	$\hat{x}_1$	$\hat{x}_2$	$\boldsymbol{y}$
1	+1	+1	$\boldsymbol{y}_1$
2	-1	+1	$\boldsymbol{y}_2$
3	+1	-1	$\boldsymbol{y}_3$
4	-1	-1	$\boldsymbol{y}_4$
5	$+\boldsymbol{\alpha}$	0	$\boldsymbol{y}_5$
6	$-\boldsymbol{\alpha}$	0	$\boldsymbol{y}_6$
7	0	$+\boldsymbol{\alpha}$	$\boldsymbol{y}_7$
8	0	$-\boldsymbol{\alpha}$	$\boldsymbol{y}_8$
9	0	0	$\boldsymbol{y}_9$
.	.	.	.
.	.	.	.
N	0	0	$\boldsymbol{y}_N$

Благодаря ортогональности матрицы планирования все коэффициенты модели определяются независимо друг от друга по формулам:

$$a_0' = \frac{1}{N}\sum y_i;$$

$$a_j = \frac{\sum \hat{x}_{ji} y_i}{\sum \hat{x}_{ji}^2}, \; j = \overline{1,k};$$

$$a_{jl} = \frac{\sum \hat{x}_{ji}\hat{x}_{li} y_i}{\sum \hat{x}^2{}_{ji}\hat{x}^2{}_{jl}}, \; j,l = \overline{1,k}, \; j < l;$$

$$a_{jj} = \frac{\sum\left(\hat{x}_{ji}^2 - \frac{1}{N}\sum \hat{x}_{ji}^2\right) y_i}{\sum\left(\hat{x}_{ji}^2 - \frac{1}{N}\sum \hat{x}_{ji}^2\right)^2}, \; j = \overline{1,k},$$

где суммирование всюду осуществляется по i от 1 до N.

В результате расчетов по ортогональной матрице с преобразованными столбцами для квадратичных эффектов получим уравнение вида

$$y = a_0' + \sum_{j=1}^{k} a_j \hat{x}_j + \sum_{\substack{j,l=1 \\ (j<l)}}^{k} a_{jl}\hat{x}_j\hat{x}_l + \sum_{j=1}^{k} a_{jj}\left(\hat{x}_j^2 - \overline{\hat{x}_j^2}\right); \quad \overline{\hat{x}_j^2} = \frac{1}{N}\sum_{i=1}^{N}\hat{x}_{ji}^2.$$

Чтобы перейти к обычной записи коэффициент a_0 определяют по формуле:

$$a_0 = \frac{1}{N}\sum_{i=1}^{N} y_i - \frac{a_{11}}{N}\sum_{i=1}^{N}\hat{x}_{1i}^2 - \frac{a_{22}}{N}\sum_{i=1}^{N}\hat{x}_{2i}^2 - \ldots - \frac{a_{kk}}{N}\sum_{i=1}^{N}\hat{x}_{ki}^2.$$

В результате получим уравнение вида

$$y = a_0 + \sum_{j=1}^{k} a_j \hat{x}_j + \sum_{\substack{j,l=1 \\ (j<l)}}^{k} a_{jl}\hat{x}_j\hat{x}_l + \sum_{j=1}^{k} a_{jj}\hat{x}_j^2. \qquad (6.61)$$

Коэффициенты модели, получаемые при помощи ортогональных планов второго порядка, определяются с разной точностью. Поэтому для расчета оценки дисперсий коэффициентов a_j используют следующие выражения:

$$\sigma_{a_j}^2 = \frac{\sigma_{\text{воспр}}^2}{\sum_{i=1}^{N}\hat{x}_{ji}^2}, \; j = \overline{1,k};$$

$$\sigma_{a_{jl}}^2 = \frac{\sigma_{\text{воспр}}^2}{\sum_{i=1}^{N}\left(\hat{x}_{ji}\hat{x}l_i\right)^2}, \; j,l = \overline{1,k}, \; l > j;$$

$$\sigma_{a_{jj}}^2 = \frac{\sigma_{\text{воспр}}^2}{\sum_{i=1}^{N}(\hat{x}_{ji}^2 - \overline{\hat{x}_j^2})^2}, \; j = \overline{1,k};$$

$$\sigma_{a0}^2 = \frac{1}{N}\sigma_{\text{воспр}}^2 + \sigma_{a11}^2\left(\overline{\hat{x}_1^2}\right)^2 + \sigma_{a22}^2\left(\overline{\hat{x}_2^2}\right)^2 + \ldots + \sigma_{akk}^2\left(\overline{\hat{x}_k^2}\right)^2 .$$

Дисперсия σ_{a0}^2 получена по закону сложения дисперсий.

После вычисления коэффициентов проверяют значимость коэффициентов по критерию Стьюдента и адекватность модели по критерию Фишера.

Следует отметить, что ортогональные планы второго порядка не обладают свойством ротатабельности, что является их недостатком. Однако простота всех вычислений оставляет за ортогональными планами право на существование.

6.3. Метод группового учета аргументов

При построении математических моделей сложных объектов необходимо учитывать следующие их особенности:

- высокая размерность вектора входных координат $\vec{x} = (x_1, x_2, \ldots, x_k)$;
- наличие большого числа внутренних источников случайных помех, статистические характеристики которых, как правило, неизвестны;
- неизученность механизмов действия различных экономических, социальных и физико-химических явлений и процессов в объектах;
- трудность постановки экспериментов для получения большого числа сигналов $\vec{x}^э, y^э$ (y — одномерный выход объекта).

Указанные особенности затрудняют построение неформальных ММ сложных объектов и вынуждают описывать статические режимы их функционирования уравнениями вида

$$y = f(x) = a_0 + \sum_{j=1}^{k} a_j x_j + \sum_{j=1}^{k} a_{jj} x_j^2 + \sum_{j,\gamma=1}^{k} a_{j\gamma} x_j x_\gamma + \sum_{j,\gamma,\rho=1}^{k} a_{j\gamma\rho} x_j x_\gamma x_\rho + \sum_{j=1}^{k} a_{jjj} x_j^3 + \ldots \qquad (6.62)$$

где $j < \gamma < \rho < \mathrm{K}$;

$\vec{a}$ — вектор неизвестных коэффициентов размерности q;

$\vec{a} = (a_0, a_1, \ldots, a_k, \ldots) \in E^q$.

Если порядок h полинома по всем переменным x_j равен числу k входных координат x_j, то выражение (3.62) называют полным полиномом Колмогорова-Габбора.

При увеличении числа входных координат x_j размерность вектора $\vec{a}$ полного полинома возрастает исключительно быстро. Так, например, при $k = 2$ размерность вектора $\vec{a}$ равна 6, при $k = 4$ она возрастает до 70.

Для определения вектора $\vec{a}$ математической модели сложного объекта по экспериментальным данным $y_i^э; \vec{x}_i^э = (x_{1i}^э, x_{2i}^э, \ldots, x_{ki}^э)$ при $i = \overline{1, n}$ часто применяют МНК.

При этом сталкиваются с двумя затруднениями:

1. Вследствие слабой изученности сложных объектов многие входные координаты, включенные в ММ, оказываются сильно коррелированными, что приводит к плохой обусловленности системы нормальных уравнений МНК. Высокая размерность вектора $\vec{a}$ делает невозможным какой-либо предварительный анализ степени обусловленности матрицы системы уравнений. Поэтому исследователь вынужден убеждаться в неустойчивости задачи определения $\vec{a}$ лишь в результате ее многократного решения на ЭВМ.

2. Задача определения вектора $\vec{a}$ заранее не заданной размерности q полинома типа (6.62) является, как правило, неустойчивой. Обоснованный же выбор числа q структуры полинома (6.62) фактически невозможен из-за большой размерности таблицы экспериментальных данных:

$$\left[x_{ji}, y_i^э\right] \quad , \quad i = \overline{1,n} \quad , \quad j = \overline{1,n} \quad ,$$

и высокого уровня помех, налагаемых на выходную координату объекта. Вследствие этого при построении ММ размерность вектора $\vec{a}$ и структура полинома (6.62) задаются в какой-то мере наугад; затем с помощью МНК определяются значения a; вычисляется СКО σ_a аппроксимации полиномом (6.62) тех табличных данных $\left[x_{ji}^э, y_i^э\right]$, которые не были использованы для нахождения $\vec{a}$; величина σ_a сравнивается с допустимой ошибкой Δ ММ. Если $\sigma_a > \Delta$, размерность $\vec{a}$ увеличивают и (или) видоизменяют структуру ММ (6.62), определяют новые значения a (напомним, что ряд (6.62) не является ортогональным). Затем снова вычисляют величину σ_a, сравнивают ее с Δ и повторяют описанную процедуру до тех пор, пока при некоторой структуре полинома (6.62) и размерности $\vec{a}$ не выполнится условие $\sigma_a < \Delta$. Подобный процесс построения ММ связан с исключительно большим объемом вычислений, так как структура ММ (6.62) и размерность $\vec{a}$ заранее не заданы и должны быть определены в процессе решения задачи.

Указанные трудности вычислительного характера могут быть устранены или уменьшены, если для построения ММ сложного объекта применить метод группового учета аргументов (МГУА).

Согласно этому методу решение трудоемкой исходной задачи (как правило, неустойчивой) аппроксимации таблично заданной функции k переменных $\left[x_{ji}, y_i^э\right]$, $i = \overline{1,n}$, $j = \overline{1,k}$ полиномом (6.62), структура которого и размерность $q < nk$ заранее не заданы, заменяется многостадийным процессом решения большого числа сравнительно простых задач аппроксимации табличных данных функциями двух переменных — полиномами заданной структуры с невысокой размерностью вектора $\vec{a}$.

На каждой из стадий этого процесса производится отбор “наилучших”, в некотором смысле, полиномов, которые используются на следующей стадии в качестве фиктивных аргументов новых полиномов. Подобная процедура построения сложной функции (полинома от

полиномов) продолжается до тех пор, пока на какой-либо стадии не будет достигнута приемлемая точность описания свойств объекта (табличных данных) некоторой математической зависимостью $y = f(\vec{x}, \vec{a})$, быть может и меньшего, чем k, числа аргументов x_j.

Рассмотрим подробно первые две стадии этой процедуры.

Пусть задана совокупность (таблица) экспериментальных данных $[\vec{x}_i, y_i^э]$ при $i = \overline{1,n}$, которую требуется аппроксимировать с допустимой средней квадратической погрешностью Δ некоторой функцией, число переменных x_j которой меньше или равно k.

Разобьем эту совокупность данных на три множества:

обучающее $M_0 = [y_i^э, \vec{x}_i^э]$, $i = \overline{1,n_1}$; $n_1 > 6$;

проверочное $M_n = [y_i^э, \vec{x}_i^э]$, $i = \overline{n_1+1, n_2}$;

контрольное $M_k = [y_i^э, \vec{x}_i^э]$, $i = \overline{n_2+1, n}$.

На первой стадии аппроксимации построим $C_k^2 = \dfrac{k(k-1)}{2}$ полиномов $y_{j\lambda} = f(x_j, x_\lambda)$ двух аргументов x_j и x_λ типа

$$y_{j\lambda} = a_{0j\lambda} + a_{1j\lambda}x_j + a_{2j\lambda}x_\lambda + a_{3j\lambda}x_jx_\lambda + a_{4jj}x_j^2 + a_{5\lambda\lambda}x_\lambda^2 , \quad j, \lambda = \overline{1,k}, \ j \neq \lambda, \ j < \lambda . \quad (6.63)$$

Структура частной модели (6.63) может быть выбрана иная, например, при $a_{1j\lambda} = 0$ или $a_{4jj} = a_{5\lambda\lambda} = 0$ и т.п.

Такие полиномы будем называть частными .

Вектор коэффициентов $\vec{a}_{j\lambda}$ каждого частного полинома определяется МНК по экспериментальным данным $[y_i^э, x_{ji}, x_{\lambda i}] \in M_0$, $i = \overline{1,n_1}$, $n_1 > 6$.

Далее вычислим ординаты каждого частного полином $y_{j\lambda}$ с известными коэффициентами во всех точках $[x_{ji}, x_{\lambda i}] \in M_n$, $i = \overline{n_1+1, n_2}$ проверочного множества M_n; обозначим эти данные через $y_{j\lambda}(i)$. Средняя квадратическая ошибка $\sigma_{j\lambda}$ аппроксимации экспериментальных данных $y_i^э \in M_n$ частным полиномом $y_{j\lambda}$ равна:

$$\sigma_{j\lambda} = \sqrt{\frac{1}{n_2 - n_1}\sum_{i=n_1+1}^{n_2}\left(y^э(i) - y_{j\lambda}(i)\right)^2} , \quad j, \lambda = \overline{1,k} , \quad j \neq \lambda \quad (j < \lambda) . \quad (6.64)$$

Среди всех величин $\sigma_{j\lambda}$ найдем наименьшую и обозначим ее через σ_1^*, $\sigma_1^* = \inf\limits_{i,\lambda} \sigma_{j\lambda}$.

В соответствии с некоторым правилом отбора, в котором используются величины $\sigma_{j\lambda}$, из всех C_n^2 частных полиномов выделим

несколько, например, m_1 “наилучших” полиномов и переобозначим их через $y_1, y_2, \dots, y_\ell, \dots, y_{m_1}$.

Теперь вычислим значения $y_1(i)$ отобранных полиномов во всех точках множеств M_0 и M_k, а также определим СКО δ_1 для всех y_1 на множестве M_k

$$\delta_1 = \sqrt{\frac{1}{n-n_2}\sum_{i=n_2+1}^{n}\left[y_i^э - y_1(i)\right]^2} \quad , \quad 1 = \overline{1, m_1} \quad . \tag{6.65}$$

Наконец найдем величину

$$\delta_1^* = \inf_1 \delta_1 ,$$

которая понадобится на следующих стадиях аппроксимации.

На второй стадии отобранные полиномы $y_1, y_2, \dots, y_\ell, \dots, y_{m_1}$ рассматриваются как аргументы новых частных полиномов

$$Z_{\ell\gamma} = \tilde{a}_{0\ell\gamma} + \tilde{a}_{1\ell\gamma} y_\gamma + \tilde{a}_{2\ell\gamma} y_\gamma + \tilde{a}_{3\ell\gamma} y_\ell y_\gamma + \tilde{a}_{4\ell\ell} y_\ell^2 + \tilde{a}_{5\gamma\gamma} y_\gamma^2 \quad , \tag{6.66}$$

$$1, \gamma = \overline{1, m_1} \quad , \quad 1 \neq \gamma \ \ (1 < \gamma) .$$

Количество этих частных полиномов равно числу сочетаний из m_1 по два

$$C_{m_1}^2 = \frac{m_1(m_1-1)}{2} .$$

Коэффициенты $\tilde{a}_{0\ell\gamma}, \tilde{a}_{1\ell\gamma}, \dots, \tilde{a}_{5\gamma\gamma}$ каждого из этих полиномов определяются методом наименьших квадратов по множеству значений $\left[y_i^э, y_1(i), y_\gamma(i)\right] \in \mathrm{M}_0$ при $1, \gamma = \overline{1, m_1}$, $1 \neq \gamma$; $i = \overline{1, n_1}$, где значения $y_1(i), y_\gamma(i)$ вычисляются во всех точках множества M_0 по модели (6.63).

Определим СКО $\sigma_{1\gamma}$ аппроксимации каждым частным полиномом $z_{1\gamma}$ экспериментальных данных $y_i^э \in \mathrm{M}_n$. Для этого подставим полиномы $y_1, y_2, \dots, y_{m_1}$ в выражение (6.66), вычислим значения $z_{1\gamma}(i)$ в каждой точке $\vec{x}_i^э \in \mathrm{M}_n$, $i = \overline{n_1+1, n_2}$ и, аналогично формуле (6.64), найдем $\sigma_{1\gamma}$:

$$\sigma_{1\gamma} = \sqrt{\frac{1}{n_2-n_1}\sum_{i=n_1+1}^{n_2}\left[y_i^э - z_{1\gamma}(i)\right]^2} \quad ; \quad 1, \gamma = \overline{1, m_1} \quad , \quad 1 \neq \gamma \ \ (\gamma > 1) \quad . \tag{6.67}$$

Среди всех величин $\sigma_{1\gamma}$ найдем наименьшую и обозначим ее через σ_2^*:

$$\sigma_2^* = \inf_{1,\gamma} \sigma_{1\gamma} .$$

Далее, в соответствии с принятым правилом отбора, из всех полиномов $z_{1\gamma}$ выделим множество “наилучших” полиномов и переобозначим их через $z_1, z_2, \dots, z_\mu, \dots, z_{m_2}$, где $m_2 < C_{m_1}^2$.

Каждый полином z_μ может быть функцией четырех независимых переменных x_j, поэтому вероятность аппроксимации им с приемлемой

точностью экспериментальные данные $\left[\vec{x}_i^{\,\text{э}}, y_i^{\text{э}}\right]$ выше, чем на первой стадии. Вследствие этого необходимо выяснить целесообразность перехода к третьей стадии аппроксимации.

Процедура последовательной аппроксимации оканчивается, когда точность приближения независимых экспериментальных данных $y_i^{\text{э}} \in \mathrm{M}_k$ каким-либо “наилучшим” полиномом удовлетворить нашим требованиям. Заметим, что значения $\vec{x}_i^{\,\text{э}}, y_i^{\text{э}}$ из M_0 уже использовались для определения коэффициентов полиномов, а данные из M_n — для отбора “наилучших” полиномов как на первой, так и на второй стадии.

Для проверки точности приближения вычислим сначала значения отобранных полиномов $z_\mu(i)$, $\mu = \overline{1, m_2}$, $i = \overline{n_2, n}$ во всех точках множества M_k, а затем значения $\delta_\mu\left(\mu = \overline{1, m_2}\right)$ на множестве M_k по формуле

$$\delta_\mu = \sqrt{\frac{1}{n - n_2} \sum_{i=n_2+1}^{n} \left[y_i^{\text{э}} - z_\mu(i)\right]^2} \quad , \quad \mu = \overline{1, m_2} \quad . \tag{6.68}$$

Найдем величину

$$\delta_2^* = \inf_\mu \delta_\mu .$$

Если окажется, что $\delta_2^* \le \Delta$, то соответствующий полином z_μ типа (6.66) может рассматриваться как математическая модель сложного объекта. Если же $\delta_2^* > \Delta$, то необходима третья стадия аппроксимации, на которой в качестве фиктивных аргументов будут выступать $z_1, z_2, \dots, z_{m_2}$. Значения этих аргументов предварительно вычисляются во всех точках множества M_0 и сводятся в таблицу $\left[z_\mu(i), y_i\right]$, $i = \overline{1, n_1}$, $\mu = \overline{1, m_2}$.

Во многих случаях допустимая СКО (Δ) ММ бывает известна весьма неточно и фактически не может быть использована для сравнения с δ_2^*. В таких случаях для окончания процесса аппроксимации можно применять величину

$$E = \frac{\left|\delta_1^* - \delta_2^*\right|}{\delta_2^*} .$$

Если E относительно невелика $\left(E \le 0.1 \div 0.15\right)$, то дальнейшая аппроксимация малоэффективна и в качестве ММ объекта можно использовать полином Z_j, для которого величина δ_j будет наименьшей. Если же E относительно велика, то необходимо третья стадия аппроксимации, вся последовательность расчетов на которой такая же, как и на второй стадии.

Приведем еще одно условие окончания процесса аппроксимации, рекомендованное в трудах Ивахненко А.Г.

Практика применения МГУА показывает, что в большинстве случаев с ростом β величина δ^*_β вначале убывает, а затем начинает быстро возрастать. В таких ситуациях необходимо оканчивать процесс последовательной аппроксимации при достижении наименьшего значения СКО δ^*_β. При этом, оказывается, число стадий β обычно не превышает $k-1$.

Анализ особенностей МГУА.

1. На начальных стадиях аппроксимации экспериментальных данных $\left[\vec{x}_i^{\,\prime}, y_i^{\,\prime}\right]$, $i=\overline{1,n_1}$, $\vec{x}_i=\left(x_{1i}, x_{2i}, \ldots, x_{ki}\right)$ отношение числа kn_1 этих данных к числу коэффициентов частных полиномов, определяемых в совокупности, намного больше единицы. Вследствие этого коэффициенты “наилучших” полиномов, приближающих точки $y_i^{\,\prime} \in \mathrm{M}_0$ и, отчасти, $y_i^{\,\prime} \in \mathrm{M}_n$, весьма слабо зависят от помехи, налагаемой на выходную координату объекта, а СКО ММ остается примерно одинаковой как на M_n, так и на M_k. С увеличением номера стадии β возрастает число аргументов полинома и общее число коэффициентов, определяемых по точкам $y_i^{\,\prime} \in \mathrm{M}_0$ и $y_i^{\,\prime} \in \mathrm{M}_n$, что, естественно, приводит к уменьшению величины σ^*_β. Одновременно с ростом β уменьшается отношение числа экспериментальных данных к числу определяемых коэффициентов (это отношение может стать <1). Поэтому получаемая ММ фактически интерполирует $y_i^{\,\prime} \in \mathrm{M}_0$ и, отчасти, $y_i^{\,\prime} \in \mathrm{M}_n$, т.е. весьма точно описывает неточные экспериментальные данные, искаженные случайными помехами . ММ, коэффициенты которой найдены интерполиационным методом по случайным величинам, не может надежно описывать независимые экспериментальные данные $y_i^{\,\prime} \in \mathrm{M}_k$, поэтому во многих случаях величина δ^*_β будет нарастать с увеличением номера стадии β .

Такое явление будем называть неустойчивостью МГУА.

Неустойчивость МГУА наблюдается и при анализе величины σ^*_β, которая часто убывает с ростом β , а затем начинает также возрастать. Величина σ^*_β менее чувствительна к росту β , так как данные $y_i^{\,\prime} \in \mathrm{M}_n$ участвуют в формировании ММ.

Для предотвращения неустойчивости МГУА в работах Ивахенко А.Г. рекомендуется выбрать число стадий β из условия минимума σ^*_β. Но во многих случаях величина δ^*_β более чувствительна к увеличению β и более целесообразно оканчивать процесс аппроксимации, как только δ^*_β достигнет своего минимума.

2. После завершения процесса аппроксимации строится окончательное выражение функции $f(\vec{x},\vec{a})$ путем последовательного исключения промежуточных аргументов. При большом числе стадий структура модели $f(\vec{x},\vec{a})$ может оказаться весьма сложной и неудобной в расчетах. В таких случаях $f(\vec{x},\vec{a})$ задается набором полиномов.

3. Остановимся теперь на правиле отбора “наилучших” полиномов. По существу в МГУА отсутствует формализованная процедура выбора на каждой стадии таких полиномов, на базе которых можно построить функцию $f(\vec{x},\vec{a})$, описывающую экспериментальные данные $\left[\vec{x}_i, y_i^{э}\right]$ с заранее заданной погрешностью. Можно только предполагать, что подобные полиномы находятся среди тех частных полиномов, СКО аппроксимации которых точек $y_i^{э} \in \mathrm{M}_n$ относительно малы. Чрезмерное уменьшение (особенно на начальных стадиях) числа отбираемых полиномов увеличивает вероятность потери настоящих аргументов функции $f(\vec{x},\vec{a})$; чрезмерное увеличение ведет к резкому росту объема вычислений. Поэтому на каждой стадии целесообразно отбирать примерно одно и тоже число m_p полиномов с наименьшими значениями σ. Если при выбранном наугад числе m_p не удается построить желаемую функцию $f(\vec{x},\vec{a})$, то все расчеты повторяются при другом, несколько большем m_p.

4. В методе МГУА не требуется предварительно задавать структуру модели $f(\vec{x},\vec{a})$. Она формируется в процессе многостадийного отбора “наилучших” частных полиномов.

5. Полученная данным методом зависимость $f(\vec{x},\vec{a})$ может быть функцией меньшего чем k числа входных переменных x. При исследовании сложных объектов всегда имеется опасность включения в число входных координат таких факторов, влияние на y которых незначимо или не может быть оценено при данном уровне помех. Устранение из ММ таких незначимых факторов также осуществляется МГУА в процессе многостадийного отбора “наилучших” частных моделей.

6. Число стадий аппроксимации зависит от многочисленных факторов. Важнейшими из них являются число входных координат, допустимая погрешность Δ ММ, структура частных полиномов, число отбираемых полиномов, соотношение чисел $n_1, n_2, n,$ уровень помехи и др.

7. При использовании МГУА на каждой стадии имеем дело с матрицей системы нормальных уравнений МНК порядка, не превышающего 6×6, что, в общем случае, способствует улучшению ее обусловленности. Если же размерность вектора $\vec{a}$ частных полиномов равна 3-4, то появляется возможность сравнительно просто оценивать степень обусловленности матрицы B системы нормальных уравнений МНК по критерию Фоке

$$F = |\det \mathrm{B}| \cdot \|\mathrm{B}\|_{\text{Е}}^{-1} \sqrt{\mathrm{K}_{\text{К}}},$$

где $\|\mathrm{B}\|_{\text{Е}}$ — эвклидова норма матрицы B с элементами $b_{ij}, i, j = \overline{1,k}$;

$\mathrm{K}_{\text{К}}$ — порядок матрицы B;

$$\|\mathrm{B}\|_{\text{Е}} = \sqrt{\sum_{i,j=1}^{k_1} b_{ij}^2}.$$

Частные полиномы, матрицы B которых плохо обусловлены, т.е. величины F малы, не должны включаться в число отбираемых на следующую стадию.

В настоящее время многие вопросы применения и обоснования МГУА еще нуждается в детальной разработке, недостаточен и опыт его применения для решения задач моделирования объектов.

Практическая рекомендация по группировке исходных данных $\left[\vec{x}_i^{\text{э}}, y_i^{\text{э}}\right], i = \overline{1,n}$ **по множествам** $\mathrm{M}_0, \mathrm{M}_n, \mathrm{M}_k$. Определяются среднее арифметическое значение $\bar{y}^{\text{э}}$ по всем данным и отклонения случайных величин $y_i^{\text{э}}$ от $\bar{y}^{\text{э}}$. Ординаты $y_i^{\text{э}}$, обладающие наибольшими отклонениями включают в M_0, наименьшими — в M_k, остальные — в M_n.

Структура алгоритма самоорганизации модели на ЭВМ. Многорядный алгоритм селекции с пороговым самоотбором реализует МГУА. Его можно эффективно применять для большого числа переменных (до 1000), когда таблица исходных экспериментальных данных содержит всего 15-20 точек.

Структурная схема алгоритма показана на рис.3.5

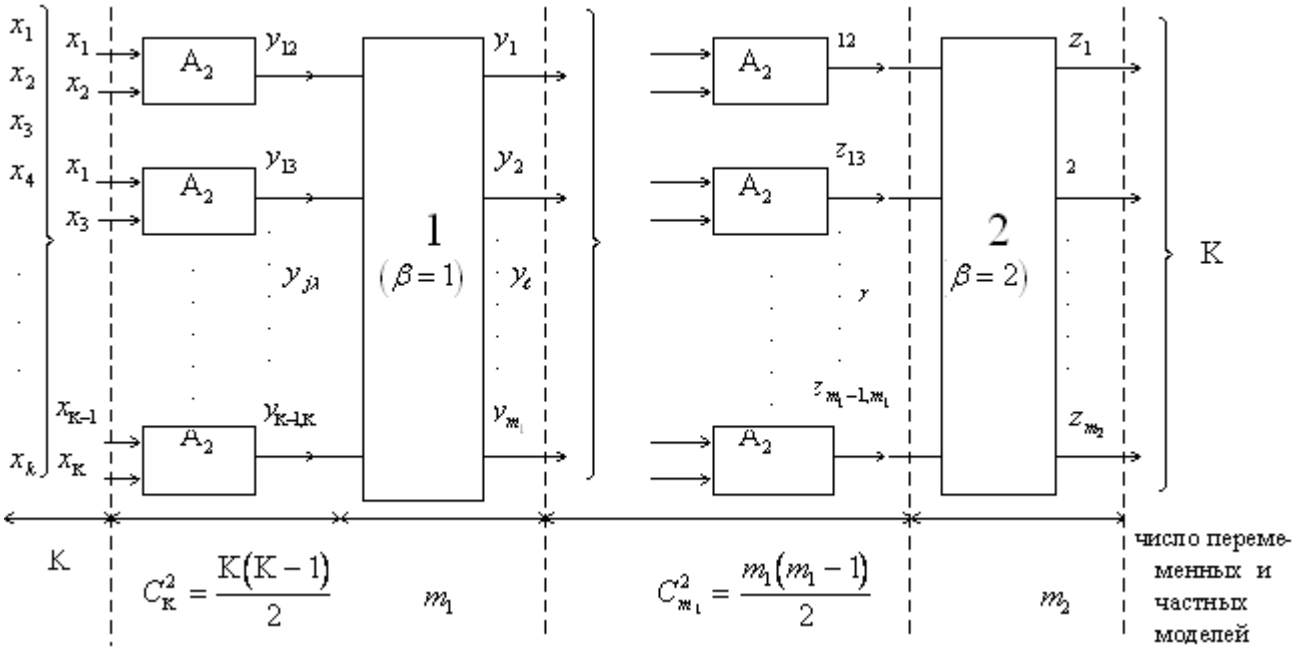


Рис.3.5

Здесь: 1,2 — пороговые отборы $C_{m_\beta}^2 = \frac{m_\beta(m_\beta - 1)}{1 \cdot 2}$ лучших частных моделей $y_1\left(1 = \overline{1, m_1}\right)$ и $z_\mu\left(\mu = \overline{1, m_2}\right)$ по критерию селекции;

β — номер ряда селекции;

A_2 — алгоритм попарной обработки переменных.

В первом ряду селекции образуются всевозможные пары аргументов, и для каждой из них находится частная модель вида

$$y_{j\lambda} = f_i\left(x_j, x_\lambda\right) = a_{0j\lambda} + a_{1j\lambda}x_j + a_{2j\lambda}x_\lambda + a_{3j\lambda}x_j x_\lambda + a_{4jj}x_j^2 + a_{5\lambda\lambda}x_\lambda^2,$$

где $i = \overline{1, d_1};\ d_1 = C_k^2 = \frac{k(k-1)}{2};\ j = \overline{1, k-1};\ \lambda = j+1, j+2, \ldots, k$;

k — число входных переменных первого ряда селекции.

Оценки коэффициентов частных моделей $y_{j\lambda}$ рассчитываются МНК по части опытных данных, называемой обучающей последовательностью. Из всех частных моделей выбирается m_1 лучших моделей $y_1, y_2, \ldots, y_{m_1}$.

Во втором ряду селекции образуются пары выходных переменных, прошедших первый ряд и для каждой из них находятся частные модели второго ряда $z_{\epsilon\gamma}$ и т.д.

Число предложенных алгоритмов МГУА непрерывно возрастает, но все они основаны на указанном выше едином принципе самоорганизации.

6.4. Определение наилучшей модели

Критерий выбора модели. При анализе сложных систем большое практическое значение имеют способы определения наилучшей модели объекта из некоторой их совокупности. Рассмотрим этот вопрос на примере регрессионных моделей.

Цель предварительного выбора наилучшей модели достигается применением в отдельности или в некоторой совокупности следующих критериев:

- наименьшее число коэффициентов модели, совместимое с заданной допустимой погрешностью;
- простота структуры модели, совместимая с допустимой погрешностью;
- разумные физические основания;
- минимальная сумма квадратов отклонений между предсказанными y_i^M и эмпирическими значениями y_i выходной переменной y;
- минимальная оценка дисперсии σ_y^2 выходной переменной.

Анализ погрешностей. Прежде всего, при оценке моделей проводят анализ остатков (погрешностей) в виде разностей $y_i - y_i^M$. В регрессионном анализе используется ряд основных предположений:

- независимость ошибок;
- постоянства их дисперсий;
- нормальный закон распределения ошибок.

Если модель адекватно описывает экспериментальные данные, то погрешности должны не противоречить этим предположениям. Анализ погрешностей - это способ проверки того, что то или иное предположение не нарушено.

Можно выделить следующие наиболее характерные результаты анализа погрешностей:

1. Обнаружение *выбросов*, т.е. резко отличающихся, экстремальных значений y_i от предполагаемого значения y_i^M. В этом случае необходимо повторить эксперимент в точке выброса, предварительно проверив правильность его условий. Если выброс устойчив, то его надо или выбросить и пересчитать коэффициенты модели без него, или исследовать физическую природу его существования.

2. Обнаружение некоторого тренда в остатках, т.е. тенденции в их изменении с течением времени. Например, в остатках наблюдается тенденция к линейному росту. Для улучшения модели вычисляют поправку в виде разности $\frac{y_i - y_i^M}{y_i^M} - \alpha_i$, где $\alpha_i = kN$, $k = tg\alpha$, α - угол наклона усредненной линии остатков к оси абсцисс, N - номер эксперимента.

3. Обнаружение разного сдвига уровня процесса. В этом случае нужно выяснить причину резкого скачка погрешности, а затем разбить выборку на две и для каждого уровня построить модель.

4. Обнаружение изменений в дисперсии ошибки. Анализ, который проводят при построении регрессионных моделей, позволяет найти некоторое среднее значение дисперсии ошибки (например, дисперсию воспроизводимости). Если дисперсии ошибки неоднородны, то найденная средняя дисперсия может неверно описывать часть экспериментальных данных.

5. Исследование остатков для проверки того, описываются ли они нормальным законом распределения. В этом случае можно проверить случайность значений остатков и воспользоваться кривой нормального распределения.

Сравнение двух регрессионных моделей. Рассмотрим некоторые методы сравнения регрессионных моделей, применимые как к линейным, так и нелинейным моделям.

1. Критерий Хоэла. Предполагается, что имеется адекватная модель y_1^M, которая удовлетворительно описывает экспериментальные данные в области ее применимости. С ней конкурирует другая модель y_2^M. Ставится

задача: не следует ли отказаться от модели y_1^M и отдать предпочтение модели y_2^M. Строят тестовую линейную зависимость в виде уравнения в параметрической форме

$$y - y_1^M = \lambda(y_2^M - y_1^M), \qquad (6.69)$$

где □>0. Проверка сводится к оценке в уравнении (6.69) углового коэффициента (параметра) □. Если □ значимо положителен (□□1), от модели y_1^M отказываются в пользу второй модели y_2^M. Действительно, из (6.69) имеем:

$$y - y_1^M = \lambda(y_2^M - y_1^M);$$

$$\lambda(y_2^M - y) = (\lambda - 1)(y_1^M - y).$$

Обозначим $\Delta_1 = y_1^M - y$, $\Delta_2 = y_2^M - y$ - ошибки модели y_1^M и y_2^M.

Тогда $\lambda\Delta_2 = (\lambda - 1)\Delta_1$.

Отсюда при □□1 имеем

$|\Delta_2| < |\Delta_1|$.

Если □ незначимо положителен (□<1), то нельзя определить, какая из моделей лучше. Действительно, □<1 не дает определенного результата, так как возможно выполнение и неравенства $|\Delta_2| < |\Delta_1|$(в пользу модели y_1^M) и противоположного ему неравенства $|\Delta_2| < |\Delta_1|$(в пользу модели y_2^M), т.е. неизвестно условие нарушения неравенств. В этом случае отдается предпочтение модели y_1^M.

Критерий Хоэла называется *несимметричным*, так как он может использоваться только при □>0. Таким образом, подставляя в уравнение (3.69) значения y_i, y_{1i}^M, y_{2i}^M можно дать оценку погрешностей регрессионных моделей y_1^M, y_2^M и выбрать путем оценки параметра □ лучшую из них.

2. Критерий Вильяма и Клута. Для сравнения двух регрессионных моделей, которые, по крайней мере, первоначально представляются равноценными, можно использовать симметричный критерий Вильяма и Клута. Проверка осуществляется путем оценки параметра □ в тестовом уравнении

$$y - \frac{1}{2}(y_1^M + y_2^M) = \lambda(y_2^M - y_1^M), \qquad (6.70)$$

где y принимает эмпирические значения. Предположим, что модель y_1^M удовлетворительна тогда

$$y = \eta_1 + \delta, \qquad (6.71)$$

где □$_1$- теоретическое значение, предсказанное первой моделью в предположении, что ее коэффициенты регрессии вычислены точно, □ - погрешность модели. Подставляя (6.71) в (6.70) получим

$$\eta_1 + \delta - \frac{1}{2}(y_1^M + y_2^M) = \lambda(y_2^M - y_1^M). \qquad (6.72)$$

Вследствие того, что модель y_1^M достаточно адекватна, можно положить $\eta_1 \approx y_1^M$ и преобразовать выражение (6.72) к виду

$$-\frac{1}{2}(y_2^M - y_1^M) + \delta = \lambda(y_2^M - y_1^M).$$

Отсюда при малых ошибках □ следует, что модели y_1^M отдается предпочтение при $\lambda \approx -\frac{1}{2}$.

Теперь предположим, что удовлетворительна модель y_2^M, тогда

$$y = \eta_2 + \delta, \quad (3.73)$$

Подставляя (3.73) в (3.71) и предполагая, что $\eta_2 \approx y_2^M$, получим

$$\frac{1}{2}(y_2^M - y_1^M) + \delta = \lambda(y_2^M - y_1^M).$$

Следовательно, если верна модель y_2^M, то получим оценку $\lambda \approx \frac{1}{2}$.

Таким образом, значимое отрицательное значение параметра $\lambda\left(\lambda \approx -\frac{1}{2}\right)$ указывает на то, что модель y_1^M лучше модели y_2^M. Если □ значимо положителен, то окажется лучше модель y_2^M. Если □ незначительно отличается от нуля, то выбора между двумя моделями сделать нельзя - они равнозначны.

Итак, подставляя в уравнение (6.70) значения y_i, y_{1i}^M, y_{2i}^M взятые из допустимой области применения регрессионных моделей y_1^M, y_2^M можно путем количественной оценки параметра □ выбрать лучшую из этих моделей. критерий Вильяма и Клута называется симметричным, так как он может применяться и при положительных и при отрицательных значениях □.

Приложение 1

Основные технические характеристики логического анализатора В-141:

Каналы:

- Число входных измерительных каналов – 32
- Активные 8-канальные пробники (4 пробника)
- Параметры входов – 100 МОм, 10 пФ
- Полоса пропускания входного тракта – 0 …200 Мгц
- Входной диапазон анализируемых сигналов - ±5 В
- Программная установка уровня дискриминации анализируемых сигналов в пределах всего входного диапазона с дискретностью 150 мВ

Общие:

- Максимальная частота дискретизации сигнала – 400 МГц
- Максимальная глубина памяти – 128 квыб
- Внутреннее (от встроенного генератора) и внешнее (от сигнала любого из входных каналов) тактирование
- Генерация сигналов ТТЛ по 16 каналам с максимальной частотой дискретизации сигнала 100 МГц

Основные рабочие конфигурации анализатора:

- Анализ сигналов от произвольных 8 входных каналов с частотой дискретизации сигнала до 400 МГц и глубиной памяти до 128 Квыб/канал
- Анализ сигналов от произвольных 16 входных каналов с частотой дискретизации сигнала до 200 МГЦ и глубиной памяти до 64 Квыб/канал
- Анализ сигналов от каждого из 32 входных сигналов с глубиной памяти до 32 Квыб/канал и частотой дискретизации сигнала в диапазоне от 1 МГц до 100 МГц

Режимы запуска развертки:

- От произвольного входного каналу о заданному перепаду логических уровней
- По заданной комбинации логических уровней от 2 до 32 произвольных входных каналов
- От произвольного входного канала по заданной последовательности логических уровней, а также с учетом их длительности
- По заданной последовательности комбинаций логических уровней от 1 до 32 произвольных каналов с учетом их длительности

Прочие:

- Аппаратно-логический анализатор/генератор цифровых сигналов В-141 включает в себя корпус УНИПРО с блоком питания и интерфейсом В-101 и измерительный модуль В-141
- Интерфейсы с компьютером:
  - Принтерный порт, протоколы ECP/EPP;
  - Последовательный порт RS-232
- Питание 220 В/50 Гц
- Потребление 15 Вт

- Размеры:
  - Измерительного модуля В-141-190х140х120 мм;
  - Комплекса УНИПРО – 290х260х70мм.
- Масса:
  - Измерительного модуля В-141-0,8 кг
  - Анализатора/генератора В-141 в сборе – 2,3 кг.

Основные программно функциональные характеристики В-141

- Программное обеспечение реализовано для операционных систем Windows95/98/NT/2K/XP
- Маркерные измерения
- Функции поиска заданной последовательности по всему сигналу, либо по его части, а также функции перехода к любой точке сигнала
- Файловые функции обеспечивают создание, открытие, сохранение, копирование и печать данных, комментарии

Состав программного обеспечения В-141:

- Программа генератора сигналов UniLA 1.7
- Драйвер UniLA для Windows95/98/NT/2K/XP
- Документация и библиотека дополнительных функций по драйверу UniLA для использования В-142 в качестве встраиваемого устройства
- Примеры использования драйвера для MS Visual C++, Delphi 4.0.