Актуальные модели, методы и технологии оценивания параметров при обработке данных

С.Г. Валеев
Ульяновский государственный технический университет
Кафедра прикладной математики и информатики
sgv@ulstu.ru
http://pmi.ulstu.ru

Введение

Трудно преувеличить значение этапа оценивания параметров, особенно если они являются поправками к константам астрометрических и небесно-механических теорий. Зачастую именно этим завершаются астрономические исследования, связанные с высокоточными наблюдениями и математической обработкой данных. На сегодняшний день можно констатировать, что в ряде случаев математический аппарат по точности уступает прецизионным измерениям.

Другая проблема (она все время будет преследовать нас!) – это проблема размерности моделей обработки данных, недоступной современным компьютерам.

В докладе рассматриваются в конечном итоге информационно-математические технологии в определенной степени, решающие проблемы повышения точности при оценивании параметров, и сокращения размерности используемых моделей обработки данных.

1. Постулирование математической модели обработки данных

1.1. Задача восстановления зависимости

$$MY = \eta(X, \beta) \qquad (1)$$

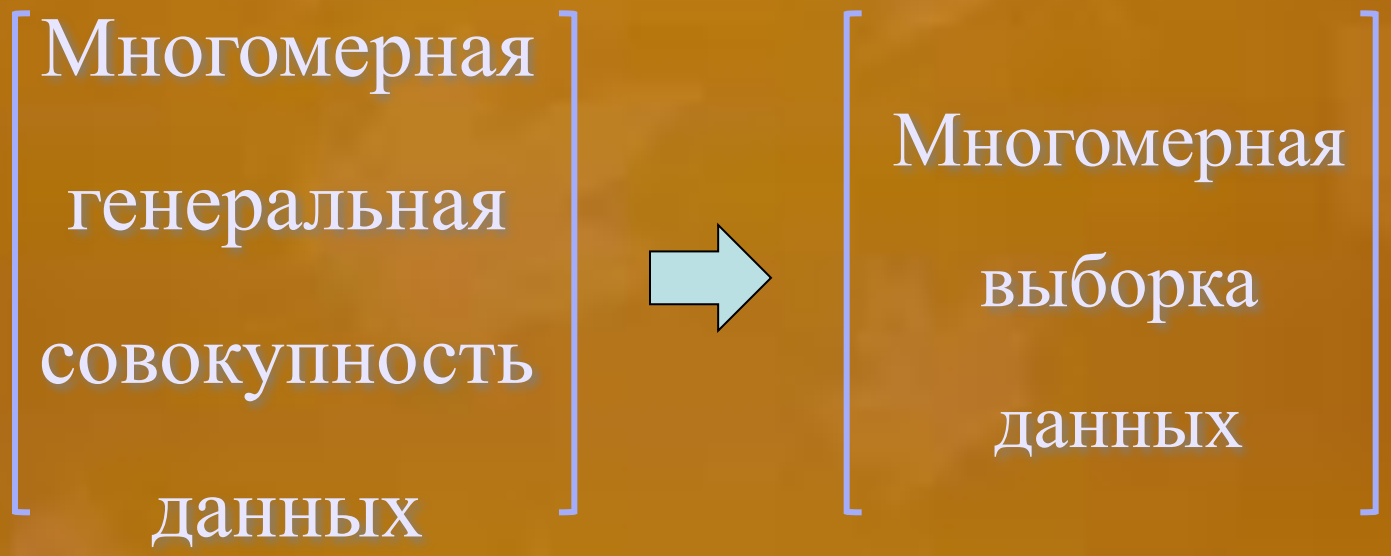
Y – зависимая переменная

X – $(x_0, x_1, \ldots, x_{p-1})^T$ – вектор независимых переменных

$\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})^T$ – вектор неизвестных параметров

1. Постулирование математической модели обработки данных

1.1. Задача восстановления зависимости



Статистическая модель обработки данных

1. Постулирование математической модели обработки данных

1.2. Классификация моделей

По назначению

- параметрические
- предсказательные
- комплексные

По объекту

- стационарные (в пространстве факторов x_i)
- динамические (временные ряды - ВР)
- стационарно-динамические

1. Постулирование математической модели обработки данных

1.2. Классификация моделей

По виду

- полиномы (алгебраические, тригонометрические, трансцендентные)
- ортогональные полиномы (полиномы Чебышева, разложения в ряд Фурье)
- разложения в ряд Тейлора (модели дифференциальных поправок)

1. Постулирование математической модели обработки данных

1.2. Классификация моделей

По структуре

- “жестко” фиксированные
- “плавающие”

По количеству “откликов”

- однооткликовые
- многооткликовые (системы одновременных уравнений – СОУ)

2. Оценивание β: МНК и системные подходы

2.1. МНК

Проблемы

- точность расчетов
- скорость расчетов
- машинная память

Вычислительные схемы

- схемы с центрированием данных
- схема Гаусса-Жордана
- схема Холесского
- схемы в искусственном ортогональном базисе
- итерационные схемы

2. Оценивание β: МНК и системные подходы

2.2. Адаптивное регрессионное моделирование

Теорема Гаусса-Маркова

Если будут выполняться предположения:

- $Y = X\beta + \varepsilon$
- X детерминированная матрица, имеющая максимальный ранг ;
- $M(\varepsilon) = 0, \quad D(\varepsilon) = M(\varepsilon^T \varepsilon) = \sigma^2 I_n$

то МНК – оценка

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

является наиболее эффективной (в смысле наименьшей дисперсии) оценкой в классе линейных несмещенных оценок

2. Оценивание β: МНК и системные подходы

2.2. Адаптивное регрессионное моделирование

Запишем модель регрессионного анализа (РА) в матричном виде

$$Y = X\beta + \varepsilon \qquad (2)$$

где Y есть $(n-1)$ вектор (y_i - случайная величина);

X - есть $(n \times p)$ матрица ($x_{10}=x_{20}=\ldots=x_{n0}=1$;

x_{ij} - неслучайная величина, $i=\overline{1,n}$, $j=\overline{0,p-1}$);

β -$(p\times 1)$ вектор (β_j - неслучайная величина);

ε - $(n\times 1)$ вектор (ε_i - случайная величина);

и рассмотрим все предположения (гипотезы) РА – метода наименьших квадратов (МНК) относительно этой модели.

2. Оценивание β: МНК и системные подходы

2.2. Адаптивное регрессионное моделирование

Предположения о выборке

Y является выборочным случайным вектором. Следовательно, для (2) существует генеральная совокупность значений Y объема и $N \to \infty$ выборка объема n, отобранная определенным образом для изучения.

Выборка должна обладать следующими свойствами:

<1.1> - объем наблюдений достаточен,

<1.2> - при организации наблюдений обеспечивается случайный отбор,

<1.3> - ряд наблюдений однороден,

<1.4> - отсутствуют грубые промахи.

2. Оценивание β: МНК и системные подходы

2.2. Адаптивное регрессионное моделирование

Предположения о векторе β.

По оцениваемому параметру β принимаются гипотезы:

<2.1> - адекватная наблюдениям модель (2) линейна по элементам вектора β;

<2.2> - на вектор β не наложено ограничений, т.е. о векторе β нам априори ничего не известно;

<2.3> - вектор β содержит аддитивную постоянную β_0;

<2.4> - элементы β вычислены с пренебрежимо малой компьютерной ошибкой.

2. Оценивание β: МНК и системные подходы

2.2. Адаптивное регрессионное моделирование

Предположения о матрице X

<3.1> - регрессоры $x_0, x_1, \ldots, x_{p-1}$ являются линейно независимыми векторами матрицы X или справедлива запись $rank\ X = p$

<3.2> - элементы матрицы X не являются случайными величинами.

Предположения о векторе ε

<4.1> - ошибки ε_i являются случайными ошибками, аддитивно входящими в модель (2).

<4.2> - ошибки ε_i распределены по нормальному закону.

<4.3> - ошибки ε_i не содержат систематического смещения. При такой гипотезе систематические ошибки, вызванные неучтенными эффектами, войдут в β_0.

<4.4> - ошибки ε_i имеют постоянную дисперсию, т.е. наблюдения не коррелированны и при справедливости <4.2> статистически независимы.

2. Оценивание β: МНК и системные подходы

2.2. Адаптивное регрессионное моделирование

Дополнительные предположения о векторе Y

Гипотезы <2> - <4> в совокупности являются одновременно гипотезами о векторе Y. Дополнительно введем следующие два предположения.

<5.1> - метод поиска оптимального набора регрессоров $\{x_j : j = \overline{1, p_1}; p_1 < p\}$ для Y является точным.

Очевидно, при однокритериальном поиске оптимальной модели из 2^{p-1} возможных точным является метод полного перебора.

До сих пор рассматривалась модель для одного отклика Y. Часто возникают ситуации, когда откликов несколько. Это так называемая многооткликовая регрессия, для которой будем считать регрессионные модели независимыми друг от друга. Итак,

<5.2> - для многооткликовой задачи правомерно применение МНК к каждой из регрессий в отдельности.

2. Оценивание β: МНК и системные подходы

2.2. Адаптивное регрессионное моделирование

На практике гипотезы <1> - <5> в той или иной мере не выполняются, поэтому МНК – оценки не обладают требуемыми свойствами наилучших линейных оценок (НЛО): свойствами несмещенности, эффективности, состоятельности. Нарушение каждого из условий РА – МНК приводит к тому или иному виду ошибок.

По степени влияния на окончательные результаты следует в первую очередь обратить внимание на возможные нарушения условий <3.1>,<2.1> и <4.2>.

На начальном этапе адаптивный РМ – подход (АРМ - подход) предусматривает применение линейной по оцениваемым параметрам модели и вычислительной схемы МНК; на последующих этапах: - проверку соблюдения гипотез РА – МНК (диагностики нарушений), ранжирование нарушений по степени

2. Оценивание β: МНК и системные подходы

2.2. Адаптивное регрессионное моделирование

искажения свойств НЛО – оценок или в зависимости от назначения модели (прогноз, описание или описание и прогноз); - последовательную или иную адаптацию к нарушениям путем применения соответствующих вычислительных процедур; - повторные проверки нарушений и ранжирование при необходимости.

Основными элементами РМ или АРМ – подхода, формирующими НЛО – оценки параметров и прогноза, являются выборка, функции, методы оценивания и структурной идентификации, а также вычислительные сценарии адаптации.

2. Оценивание β: МНК и системные подходы

2.3. Динамическое регрессионное моделирование

До сих пор объектом внимания были так называемые статические модели, т.е. модели, не содержащие время в качестве аргумента. Вкратце рассмотрим идею применения (обобщения) РМ-подхода к обработке временных рядов. Задачу оценивания параметров регрессии по временным рядам (ВР) иногда называют задачей динамической регрессии.

Обобщая в целом ситуацию, сложившуюся в практике применения современной методологии обработки временных рядов, можно отметить основные причины неполной адекватности разрабатываемых динамических моделей: использование для оценки адекватности "внутренних" критериев качества, применение вычислительных схем МНК без анализа соблюдения условий схемы Гаусса-Маркова и адаптации к их нарушениям, использование упрощенных схем обработки ВР, одномерность решаемых задач.

2. Оценивание β: МНК и системные подходы

2.3. Динамическое регрессионное моделирование

Пусть (Ω, F, P) - вероятностное пространство, на котором задан стационарный процесс $Y(t)$, наблюдаемый в равноотстоящие моменты времени $t_1, t_2, \ldots, t_N$: $Y(t) = f(t) + \varphi(t) + \psi(t) + \varepsilon(t)$, (3)

где $Y(t_1), Y(t_2), \ldots, Y(t_N)$ - ряд наблюдений случайной функции $\xi(t)$, называемый временным рядом; $f(t)$ - неслучайная (долговременная) функция тренда;

$\varphi(t)$ - неслучайная (сезонная) периодическая функция;

$\psi(t)$ - неслучайная (долговременная, циклическая) функция;

$\varepsilon(t)$ - нерегулярная компонента (случайная величина, ошибка).

2. Оценивание β: МНК и системные подходы

2.3. Динамическое регрессионное моделирование

Анализ ВР сводится к выделению регулярных компонент $f(t)$, $\varphi(t)$, $\psi(t)$, если они существуют в реальности, и описанию нерегулярной ее части $\varepsilon(t)$ при условии стационарности ряда. Временной ряд, описанный соотношением (3), может помимо этого либо не содержать вовсе аддитивных членов $f(t)$, $\varphi(t)$, $\psi(t)$, либо включать их в различных комбинациях друг с другом; при этом слагаемое $\varepsilon(t)$ присутствует всегда.

Сравнение методологий обработки ВР и регрессионного моделирования приводит к заключению о том, что системный РМ-подход включает в себя технологию обработки временных рядов. В случае использования последней так же, как и в РМ, постулируется модель обработки данных. Комбинированный характер этой модели, включающей при описании колебательных процессов, возможно, фильтр Калмана, а при представлении ε_t – мартингал или другие описания, тем не менее не помешает формированию банка функций при обработке данных.

2. Оценивание β: МНК и системные подходы

2.3. Динамическое регрессионное моделирование

Наибольшее сходство фиксируется на этапе оценивания параметров модели, точнее, ее регулярной и нерегулярной частей. Отсюда и появление термина “динамическая регрессия” в задачах обработки ВР. Следует заметить, что множество используемых при обработке ВР методов идентификации параметров заметно менее мощно, чем при РМ. С другой стороны, при обработке ВР появляется набор функций, аппроксимирующих случайную составляющую ε_t, чего, естественно, нет при РМ. Этап идентификации структур, основной для РМ, достаточно скромно представлен в схемах обработки ВР. К тому же с этим этапом тесно связана проблема критериев, которая для моделей ВР разрешается путем использования их в режиме прогноза-экстраполяции.

Наконец, помимо сходства отметим принципиальные отличия ВР от случайной выборки, обрабатываемой при РМ-подходе: члены ВР не являются статистически независимыми и одинаково распределенными.

2. Оценивание β: МНК и системные подходы

2.3. Динамическое регрессионное моделирование

Сказанное означает, что не все приемы статистического анализа выборки могут быть распространены на ВР. Однако это не помешает их применять при конструировании регулярных составляющих модели ВР (особенно при анализе регулярных факторов x_t).

Отмеченные ранее трудности построения адекватных моделей ВР могут быть разрешены применением подхода регрессионного моделирования. Учитывая, что постулируемыми моделями могут быть как статические зависимости, так и случайные процессы в виде временных рядов, предлагаемую методологию можно назвать методологией динамического регрессионного моделирования (ДРМ).

Математический инструментарий этого подхода рассматривается при описании соответствующего программного обеспечения.

3. Алгоритмы оценивания параметров и структурно-параметрической идентификации

3.1. Алгоритмы оценивания параметров

Укажем несколько достаточно современных вычислительных схем МНК, полагая, что модель (2) жестко фиксирована по количеству слагаемых.

Представим (2) в виде

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{i,p-1} + \varepsilon_i,\ i = \overline{1,k}, \quad (4)$$

где Y - отклик; $x_j (j = \overline{1, p-1})$ - регрессоры,

$\beta_j (j = \overline{0, p-1})$ - параметры, оцениваемые по одному из рассматриваемых ниже алгоритмов.

3. Алгоритмы оценивания параметров и структурно-параметрической идентификации

3.1. Алгоритмы оценивания параметров

Метод исключения Гаусса

Из существующих схем в первую очередь следует отметить схему Гаусса – Жордана.

Метод треугольного разложения

Для решения задачи МНК можно использовать один из вариантов этого метода, применимый к симметричным матрицам и называемый методом квадратного корня (метод, основанный на разложении Холесского).

Метод Холесского обладает привлекательными свойствами экономичности и устойчивости для задач большой размерности. Также метод весьма устойчив к плохой обусловленности, особенно для задач с положительно определенными матрицами.

3. Алгоритмы оценивания параметров и структурно-параметрической идентификации

3.1. Алгоритмы оценивания параметров

Ортогональное разложение (QR - разложение)

Один из вариантов ортогонального разложения основан на алгоритме Грама – Шмидта, имеющем две разновидности: классическую и модифицированную.

Особенно хорошую вычислительную устойчивость (примерно, ту же, что и модифицированный алгоритм Грама - Шмидта) имеет ортогональное преобразование Хаусхолдера, а также метод сингулярного разложения.

Сравнивая по вычислительной устойчивости три группы методов, на первое место надо поставить методы ортогонального разложения, затем - методы треугольного разложения, и, наконец, - метод исключения Гаусса (его предпочтительный вариант с выбором ведущего элемента).

3. Алгоритмы оценивания параметров и структурно-параметрической идентификации

3.1. Алгоритмы оценивания параметров

К сожалению, все разнообразные вычислительные схемы МНК (в условиях соблюдения остальных предположений) в той или иной степени чувствительны к плохой обусловленности информационной матрицы.

1. Если данные предварительно не центрированы, ни один из трех рассмотренных методов не дает удовлетворительных результатов. Наихудшие оценки соответствуют методу Гаусса, несколько лучший результат получается при выборе ведущего элемента, а метод ортогонального разложения по Хаусхолдеру дает весьма малое улучшение.

3. Алгоритмы оценивания параметров и структурно-параметрической идентификации

3.1. Алгоритмы оценивания параметров

2. Предварительное центрирование резко улучшает точность оценок. Резко уменьшается квадратичная ошибка оценки, а сами оценки становятся гораздо ближе к истинным. При этом снижается как дисперсия оценок регрессионных коэффициентов, так и коэффициенты парной корреляции между x_j.

Наиболее предпочтительными схемами МНК для задач вида (2) являются «скоростные» и «экономичные» методы, что вызвано необходимостью обработки огромного наблюдательного материала. Перспективными для исследования можно считать метод Грама - Шмидта и Хаусхолдера. Итерационные методы, хоть и являются достаточно «экономичными», имеют малую скорость сходимости к решению.

3. Алгоритмы оценивания параметров и структурно-параметрической идентификации

3.2. Алгоритмы структурно-параметрической идентификации

До сих пор структура модели (2) считалась жестко заданной. Однако при поиске оптимальной модели для прогноза в режиме адаптации (с использованием одного или нескольких критериев качества) необходимо идентифицировать соответствующую оптимальную структуру.

Алгоритмы структурной идентификации

Ниже рассматриваются несколько перспективных для использования (или уже внедренных) алгоритмов поиска оптимальных структур на базе исходного разложения. Конечная структура будет считаться оптимальной либо по одному внешнему глобальному критерию, либо по двум (по минимуму степени взаимозависимости регрессоров и отсутствию незначимых гармоник).

3. Алгоритмы оценивания параметров и структурно-параметрической идентификации

3.2. Алгоритмы структурно-параметрической идентификации

Однокритериальные алгоритмы структурной идентификации (СИ)

Полный перебор. Наиболее точным, естественно, является метод полного перебора (ПП), при котором расчеты ведутся для всех моделей, возможных на основе исходного математического описания вида (2,3). Последние сравнивают и выбирают наилучшую по заданной мере (например, внешней стандартной ошибке). При невозможности из-за чрезмерных затрат машинного времени использовать этот метод прибегают к одному из ниже описанных алгоритмов.

Случайный поиск. Из всех методов дискретного (целочисленного) программирования (ДП) наиболее пригодным для РА оказывается метод случайного поиска с адаптацией (СПА). Его разновидностями являются метод случайного поиска с возвратом (СПВ) для задач условной и безусловной оптимизации.

3. Алгоритмы оценивания параметров и структурно-параметрической идентификации

3.2. Алгоритмы структурно-параметрической идентификации

Двухкритериальные алгоритмы СИ

Пошаговая регрессия. Основной вклад в ошибки прогноза вносят шумовые слагаемые и взаимозависимость (мультиколлинеарность) регрессоров (и параметров). Поэтому наряду с однокритериальным поиском используются и предлагаются методы, адаптивные к нарушениям <2.1> и <3.1>. Из них наиболее известен метод пошаговой регрессии (ПР) с разновидностями: метод включения, метод исключения, метод включения с исключением. Применение метода ПР в принципе не ограничено количеством регрессоров в модели, так как число перебираемых вариантов порядка *p*. Однако точность идентификации метода при увеличении *p* падает, а взаимозависимость полностью не устраняется.

3. Алгоритмы оценивания параметров и структурно-параметрической идентификации

3.2. Алгоритмы структурно-параметрической идентификации

Метод ступенчатой (последовательной) ортогонализации базиса (метод ступенчатого оценивания – МСО).

Метод предложен как альтернатива пошаговой регрессии, исключающей в ряде случаев чрезмерно большое количество регрессоров.

При необходимости адаптироваться к большему количеству нарушенных условий РА-МНК возникает проблема оптимальности сценария адаптации – выбора оптимального маршрута обработки данных. После получения двухкритериальным способом оптимальной структуры ее можно улучшить, адаптируясь к остальным нарушениям в той или иной последовательности.

3. Алгоритмы оценивания параметров и структурно-параметрической идентификации

3.3. Сценарии адаптации

Стационарные модели (модели в пространстве факторов)

В самом простом случае можно найти оптимальную структуру по глобальному критерию, затем последовательно её улучшать, проверяя соблюдение всех условий РА-МНК и осуществляя адаптацию на каждом шаге (по убыванию степени искажения).

3. Алгоритмы оценивания параметров и структурно-параметрической идентификации

3.3. Сценарии адаптации

Модели динамики

При обработке временных рядов на основе ДРМ-подхода основным инструментом адаптации является структура функций $f(t)$, $\varphi(t)$, $\psi(t)$, которую можно менять: а) непосредственно, извлекая из базы функций или формируя алгоритмически конкурентное описание; б) изменением последовательности формирования функций; в) применением для гармонических составляющих того или иного метода структурной идентификации; г) в пределах возможностей принятого для обработки метода идентификации.

4. Базовое программное обеспечение.

На сегодняшний день подготовленными к распространению является ряд программных систем математической обработки наблюдений:

1. **СПОР 2.0** – система поиска оптимальных регрессий, предназначенная для получения прецизионных моделей прогноза (редукционных моделей в фотографической астрометрии, моделей трансформации координат и т.д.); *математическое наполнение*: формирование алгебраического полинома произвольного порядка по заданному количеству переменных, множественная регрессия, гребневая регрессия, полный и ограниченный перебор структур, перебор нормальных систем, пошаговая регрессия, корреляционный алгоритм, случайный поиск;

4. Базовое программное обеспечение.

2. **СПО 2.0** – система параметрического оценивания, используемая прежде всего для получения оптимальных оценок параметров моделей в различных областях астрономии (например, констант теорий в уравнениях дифференциальных поправок, получаемых разложением в ряд Тейлора разностей O-C); *математическое наполнение*: метод ступенчатого оценивания (МСО1 – выбор на каждой стадии только ортогональных параметров, МСО2 – только значимых по t-статистике параметров, МСО3 – выбор ортогональных и одновременно значимых по t-статистике параметров), метод характеристического корня, методы робастного оценивания (Андрюса, Хубера), пошаговая регрессия, гребневая регрессия;

4. Базовое программное обеспечение.

3. **АСНИ «СФЕРА» 3.0** – автоматизированная система научных исследований, используемая для описания (по всей сфере или её фрагментам) потенциального поля исследуемой характеристики (высота объекта, аномалия силы тяжести, напряженность магнитного поля и т.д.) ортогональными разложениями по сферическим функциям; *математическое наполнение*: формирование разложения по сферическим функциям произвольной степени и порядка, множественная регрессия, пошаговая регрессия, случайный поиск, алгоритм расширения для фрагментов сферической поверхности, методы картирования;

4. Базовое программное обеспечение.

4. **АС ДРМ 2.0** – автоматизированная система динамического регрессионного моделирования, предназначенная для прецизионной обработки временных рядов; *математическое наполнение*: тесты на нормальность выборки, на стационарность, построение автокорреляционной функции, спектральный анализ ряда по временной и частотной областям, совместный спектральный анализ, методы вейвлет-анализа, построение трендов, гармонической модели с автоматическим выбором базовых гармоник методом пошаговой регрессии или случайного поиска, АРСС-моделей, ARCH-, HARCH-моделей различных версий, фильтра Калмана, мартингалов.

4. Базовое программное обеспечение.

В описанных пакетах реализована схема АРМ-подхода: - оценка качества модели по трем классам критериев; - диагностика нарушений условий РА-МНК, при которых гарантируются свойства наилучших линейных оценок; - адаптация схемы обработки методами оценивания и структурно-параметрической идентификации в случае нарушения условий.

Список литературы, иллюстративный пример (три статьи из журнала «Известия вузов. Геодезия и аэрофотосъемка»), демоверсия пакета СПОР – располагаются на сайте http://pmi.ulstu.ru