Отчет идентификация стационара с учетом корреляции параметров

В предыдущих разделах были рассмотрены методы контроля качества продукции, основанные на мониторинге единичного параметра. Однако в сложных многостадийных технологических процессах, характерных для нефтегазовой отрасли (таких как первичная переработка нефти, газофракционирование или производство нефтепродуктов), такой подход демонстрирует принципиальную ограниченность в силу следующих факторов:

1. Многопараметрическая природа процессов

   Качество конечного продукта определяется совокупностью взаимосвязанных технологических параметров (например: давление, температура, состав сырья), требующих комплексного учета.
2. Наличие кросс-корреляций

   Параметры технологического процесса часто статистически зависимы (например, температура и давление в ректификационной колонне связаны уравнением состояния), что приводит к необходимости анализа их совместного распределения, а не изолированных значений.
3. Динамика процессов

   Нестационарные режимы требуют методов, отличающихся от классического статистического контроля, рассчитанного на стабильные условия.

Разработанный алгоритм решает задачу идентификации стационарных/нестационарных участков в многопараметрических процессах нефтегазовой отрасли

Ключевые этапы:

1. Предварительная обработка данных
2. Идентификация стационарного участка

   Программа ищет участок данных, где параметры меняются слабо.
3. Адаптивная модель мониторинга

   На основе стабильного участка строится начальная модель. Для новых

данных рассчитывается статистика, показывающая, насколько они отличаются от нормы. Если отличие слишком велико, система отмечает аномалию. При частых аномалиях алгоритм переходит в режим "переходного процесса". Когда данные снова стабилизируются, модель обновляется.

Обоснование алгоритма

1. Предположение о нормальности распределения данных

1.1. Теоретическое обоснование

Алгоритм базируется на фундаментальном предположении, что на стационарных участках технологического процесса наблюдаемые данные следуют нормальному (Гауссову) распределению. Математически это выражается как:

$$X \sim \mathrm{N}(\mu, \sigma^2)$$

где:

- μ — математическое ожидание (истинное значение параметра)
- σ2 — дисперсия, обусловленная погрешностями измерений

Это предположение справедливо для большинства измерительных систем, где:

- Погрешности датчиков носят случайный характер
- Влияющие факторы многочисленны и независимы (Центральная предельная теорема)
- Отсутствуют систематические смещения измерений

Для уверенного применения нормального распределения и параметрических методов требуется $n \geq 30$ наблюдений. Это обеспечивает:

- Устойчивость оценок среднего и дисперсии
- Достаточную точность расчета медиан и MAD
- Корректность последующих F-тестов и T^2-статистики

2. Выбор стационарного участка для калибровки

Стационарный участок данных является фундаментальной опорой для всей системы адаптивного мониторинга, выполняя три ключевые функции. Во-первых, именно на этом участке рассчитываются все базовые статистические параметры распределения - средние значения, дисперсии и ковариации между параметрами, которые в дальнейшем служат эталоном для сравнения. Во-вторых, анализ этого сегмента позволяет установить естественный разброс каждого параметра в нормальном режиме работы технологического процесса, что особенно важно для многопараметрических систем. В-третьих, на основе этих данных вычисляются пороговые значения, используемые алгоритмом для детектирования аномальных ситуаций.

При этом некорректный выбор опорного стационарного участка может привести к двум принципиально разным последствиям. Если в качестве эталона будет выбран излишне "шумный" сегмент данных с повышенной вариативностью параметров, это неизбежно вызовет многочисленные ложные срабатывания системы, когда нормальные колебания будут ошибочно идентифицированы как аномалии. С другой стороны, если выбранный участок содержит скрытые нестационарности, система мониторинга окажется "слепой" к реальным отклонениям процесса от нормального режима работы, что может привести к пропуску критически важных аномалий. Особую опасность представляет ситуация, когда в эталонный сегмент попадают данные, соответствующие переходным процессам - в этом случае система может принять изменяющийся режим работы за новый "нормальный" стандарт, полностью утратив свою детектирующую способность. Поэтому процесс выбора стационарного участка требует тщательного многоэтапного анализа с применением как статистических методов, так и экспертных знаний о технологическом процессе.

2.1. Использование медианы вместо среднего

В отличие от среднего арифметического, которое рассчитывается как сумма всех значений, деленная на их количество, медиана представляет собой центральное значение в упорядоченном ряду данных. Это свойство делает медиану значительно более устойчивой к выбросам - экстремальным значениям, которые могут существенно исказить оценку центра распределения. Например, если в выборке из 100 значений температуры, составляющих в основном 50±2°C, появляется ошибочное значение 200°C (из-за сбоя датчика), среднее сместится на 1.5°C, в то время как медиана останется практически неизменной. Такая робастность особенно важна в промышленных системах мониторинга, где ложные срабатывания датчиков - нередкое явление.

$$Med(\mathrm{X}) = \begin{cases} \mathrm{X}(\mathrm{k}+1) & \text{если } n=2k+1, \\ \dfrac{\mathrm{X}(\mathrm{k}) + \mathrm{X}(\mathrm{k}+1)}{2} & \text{если } n=2k, \end{cases}$$

где X(1),…,X(n)X(1),…,X(n) — упорядоченная выборка.

Применение медианы особенно оправдано в системах с несколькими разнородными параметрами, где:

1. Разные датчики могут иметь различную природу шумов (например, термопары vs. тензодатчики)
2. Отдельные параметры могут демонстрировать тяжелые хвосты распределения
3. Возможны кратковременные сбои в работе измерительного оборудования

Медианная оценка обеспечивает устойчивость системы мониторинга к подобным аномалиям, сохраняя при этом способность корректно отражать центральную тенденцию параметров. Это особенно критично на этапе

инициализации алгоритма, когда по стабильному участку определяются базовые характеристики технологического процесса.

2.2. Относительная медианная

При мониторинге промышленных процессов ключевой проблемой является сопоставимость различных технологических параметров, измеряемых в принципиально разных единицах и масштабах. Например, в нефтепереработке одновременно контролируются температура (обычно в диапазоне 50-500°C), давление (0.1-10 МПа) и концентрации компонентов (0.01-100%). Если использовать абсолютные значения MAD, параметры с большими численными значениями будут доминировать в анализе, маскируя важные отклонения в низкоамплитудных, но критически значимых показателях. Именно поэтому вводится понятие относительной MAD (rMAD) - безразмерной величины, вычисляемой как отношение медианного абсолютного отклонения к модулю медианы распределения. Такой подход обеспечивает корректное сравнение стабильности параметров независимо от их физической природы и единиц измерения.

$$\mathrm{MAD}=\mathrm{Med}(|X_i-\mathrm{Med}(X)|),$$

$$rMAD = \frac{\mathrm{MAD}}{|\ Med(\mathrm{X})\ |}$$

Это позволяет сравнивать стабильность разных параметров, независимо от их единиц измерения.

Использование rMAD приобретает особую важность при работе с датчиками различного типа и точности. Например, современная система мониторинга газоперерабатывающего завода может одновременно получать данные от высокоточных хроматографов (погрешность 0.1%) и более грубых датчиков давления (погрешность 1%). Применение rMAD позволяет:

1) автоматически учитывать различную точность измерительных приборов через нормировку отклонений;

2) устанавливать единые критерии стабильности для всей системы (например, 5% отклонение);

3) корректно выявлять аномалии в параметрах с разными порядками значений.

1.3. Критерии стабильности участка

Для надежной идентификации стационарного участка необходимо, чтобы подавляющее большинство контролируемых параметров демонстрировали устойчивое поведение. Это условие обеспечивает, что процесс действительно находится в стабильном состоянии, а не просто часть датчиков временно показывает плавные изменения.

Математически это условие формулируется как:

$$\frac{N_{стаб}}{N_{общ}} \geq 0.9$$

где:

- Nстаб — количество параметров с rMAD < порога
- Nобщ — общее количество контролируемых параметров

Второе ключевое условие стабильности требует, чтобы медианное значение относительных медианных абсолютных отклонений (rMAD) всех параметров не превышало полуторократного значения установленного порога. Это дополнительное ограничение вводится для защиты от ситуаций, когда формально большинство параметров могут оставаться в пределах допустимых отклонений (удовлетворяя первому условию), но при этом отдельные критические параметры демонстрируют катастрофические отклонения, что может свидетельствовать о серьезных нарушениях в технологическом процессе.

Математически это условие записывается как:

$$\text{Med}(\text{rMAD}) < 1.5 \times \text{порог}$$

где Med(rMAD) - медиана относительных отклонений по всем контролируемым параметрам, а порог - заранее установленное предельное значение относительного отклонения (например, 5%). Такая форма критерия обеспечивает дополнительный уровень надежности при идентификации действительно стабильных участков.

Обоснование выбора порога стабильности

Порог rel_threshold = 0.05 (5%) определяет максимально допустимое относительное отклонение параметра от его медианного значения на стационарном участке и выбирается как компромисс между чувствительностью к реальным аномалиям и устойчивостью к ложным срабатываниям из-за шума. При определении порога учитывается паспортная погрешность датчиков: если датчик имеет погрешность 2%, минимальный теоретический порог составляет 2%, а рекомендуемое значение устанавливается на уровне 3% (1.5× погрешность) для обеспечения запаса на кратковременные всплески. Для процессов с разным уровнем шума порог может адаптивно корректироваться в диапазоне 2-10% на основе анализа исторических данных и эмпирической проверки количества ложных и пропущенных срабатываний.

Предложенный алгоритм опирается на несколько статистических и вероятностных принципов, которые делают его устойчивым к шуму и способным корректно идентифицировать стационарные участки. При отсутствии подходящего сегмента заданной длины алгоритм генерирует исключение, предотвращая работу с некачественными данными.

Адаптивная модель мониторинга:

На первом этапе адаптивного мониторинга критически важно получить надежные начальные параметры распределения данных.

Вместо традиционного выборочного среднего, которое крайне восприимчиво к аномальным значениям, алгоритм использует **медиану** как меру центра распределения. Математически это выражается следующим образом:

$$Med(\mathrm{X}) = \begin{cases} \mathrm{X}(\mathrm{k}+1) & \text{если } n=2k+1, \\ \dfrac{\mathrm{X}(\mathrm{k}) + \mathrm{X}(\mathrm{k}+1)}{2} & \text{если } n=2k, \end{cases}$$

где:

- Xstable — матрица данных стационарного участка размером n×p (n наблюдений по p параметров)
- X(k) — k-я порядковая статистика (значение, стоящее на k-м месте в упорядоченном ряду)

2. Робастная оценка ковариационной матрицы

Для определения ковариации между параметрами будет использована ковариационная матрица.

Ковариационная матрица (Σ) — это квадратная матрица, которая описывает:

- Дисперсии каждого параметра (на диагонали).
- Ковариации между всеми парами параметров (вне диагонали).

Формула для ковариации между двумя параметрами X и Y:

$$\mathrm{Cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y)$$

где:

- Yμ_X,μ_Y — средние значения XX и YY, X_2,X_3))
- nn — количество наблюдений.

Пример для 3 параметров (расход, давление, температура):

$$\Sigma = \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & Cov(X_1, X_3) \\ Cov(X_2, X_1) & Var(X_2) & Cov(X_2, X_3) \\ Cov(X_3, X_1) & Cov(X_3, X_2) & Var(X_3) \end{bmatrix}$$

Ковариационная матрица помогает находить стационарные участки, анализируя изменения средних значений и корреляций между параметрами. Стационарный участок характеризуется постоянством средних значений параметров (отсутствием тренда) и стабильностью корреляций между ними (неизменностью ковариаций).

Ковариационная матрица учитывает взаимосвязи между параметрами, что особенно важно в технологических процессах, где параметры часто связаны физическими законами. Если эти связи нарушаются (например, ковариация меняется), это может сигнализировать о переходном процессе или аномалии, позволяя детектировать отклонения от стационарности.

Однако для оценки ковариационной структуры данных применяется усовершенствованный метод Minimum Covariance Determinant. Обычная ковариационная матрица чувствительна к выбросам т.к. один аномальный пункт данных может сильно исказить оценки μ и Σ.

Метод MinCovDet находит подмножество точек размера h (где h≈0.75n), для которого определитель ковариационной матрицы минимален. Это эквивалентно поиску наиболее "компактного" облака точек.

1. Ищется подмножество данных, дающее ковариацию с наименьшим определителем (минимизирует "разброс").
2. Выбросы автоматически исключаются из расчёта.
3. Оценки μ и Σ становятся робастными.

$$(\mu, \Sigma\ ) = {}_{\mu,\,\Sigma}\mathrm{argmin}\det(\Sigma)$$

$$\text{при условии}\frac{1}{h}\sum (x_i - \mu)^T \Sigma^{-1}(x_i - \mu) \le \chi^2_{p,0.975}$$

где H — подмножество из h точек, χp,0.9752 — критическое значение распределения хи-квадрат.

3. Подтверждение устойчивости MinCovDet

Метод основан на работе Rousseeuw & Van Driessen (1999), где доказано, что MinCovDet имеет **высокую точку пробоя** (breakdown point) — до 50% выбросов не разрушают оценку.

В то время как обычная ковариация имеет точку пробоя **0%** (один выброс может полностью её исказить).

Метод	Устойчивость	Скорость	Требования к данным
Стандартная ковариация	Низкая	Быстрая	Нет выбросов
MinCovDet	Высокая	Средняя	До 25% выбросов

3.2. Расчет статистики Хотеллинга

1. Вычисление вектора отклонений

Для каждого нового поступающего наблюдения x_t (вектор размерности p×1, где p – количество контролируемых параметров) в первую очередь рассчитывается его отклонение от текущего оцененного центра распределения:

Формула отклонения:

$$Dt = xt - \mu dt = x_t - \mu$$

Где:

- μ – вектор медианных значений параметров (p×1), полученный на стадии инициализации
- x_t – вектор текущих измеренных значений параметров (p×1)

Данный шаг позволяет перейти от абсолютных значений измерений к их отклонениям от "нормального" состояния процесса, что является фундаментом для последующего многомерного анализа.

2. Регуляризация ковариационной матрицы

Для устойчивого обращения ковариационной матрицы применяется процедура регуляризации:

Формула регуляризованной матрицы:

$$\sum_{reg}^{-1} = (\Sigma + \epsilon \mathrm{I})^{-1}$$

Где:

- **Σ** – исходная ковариационная матрица (p×p)
- **I** – единичная матрица (p×p)
- **ε** – малый параметр регуляризации (10^{-6})
- Регуляризация предотвращает вычислительные проблемы при обращении плохо обусловленных матриц
- Добавление εI гарантирует положительную определенность матрицы
- Величина ε подобрана эмпирически как компромисс между устойчивостью и точностью

3. Расчет T^2-статистики Хотеллинга

Статистика T2 Хотеллинга — это многомерный аналог Z- или t-статистики, расширенный для работы с коррелированными параметрами. Она измеряет, насколько текущее наблюдение x отклоняется от эталонного распределения (стационарного режима), учитывая ковариации между переменными.

Формула T2 фактически вычисляет квадрат расстояния Махаланобиса между точкой x и центром μ в метрике, заданной ковариационной матрицей Σ:

$$T^2 = (\mathrm{x}i - \mu)^T \Sigma^{-1} (\mathrm{x}i - \mu)$$

В отличие от обычного евклидова расстояния, эта метрика учитывает две ключевые характеристики данных:

1. Масштаб параметров – через диагональные элементы ковариационной матрицы (Σ), которые отражают дисперсии каждого параметра. Чем больше естественный разброс параметра, тем менее значительным считается одинаковое по величине отклонение.
2. Корреляции между параметрами – через недиагональные элементы Σ. Если два параметра сильно связаны, их совместное отклонение оценивается иначе, чем независимое.

В стационарной системе поведение T^2-статистики предсказуемо:

- Значения колеблются вокруг теоретически ожидаемого среднего, соответствующего эталонному распределению.
- При нормальности данных статистика следует распределению Хотеллинга, которое связано с F-распределением. Это позволяет устанавливать точные вероятностные границы для принятия решений.

При выходе из стационарности наблюдаются характерные изменения:

- Нарушается либо центр распределения (μ), либо структура ковариаций (Σ), либо их совместное поведение.
- T^2-статистика демонстрирует резкий рост, так как новые наблюдения перестают соответствовать исходной модели. Сильные отклонения от μ увеличивают числитель в формуле T^2. Изменение Σ делает эталонную метрику неадекватной для новых данных.

Преобразование в F-статистику

Преобразование T^2-статистики в F-статистику имеет важный статистический смысл: оно позволяет использовать стандартные критические значения F-распределения, что значительно упрощает интерпретацию результатов и делает метод универсальным. Это преобразование учитывает как размерность задачи (количество параметров p), так и объем доступных данных (эффективный размер выборки n), обеспечивая согласованность статистических выводов при различной размерности данных и разном количестве наблюдений. Благодаря такому подходу алгоритм сохраняет свою надежность и сравнимость результатов даже при изменении числа

контролируемых параметров или при работе с выборками разного объема, что особенно важно для промышленных систем, где эти характеристики могут варьироваться.

Формула преобразования:

$$\mathrm{F} = \frac{(n-p)}{p(n-1)} T^2$$

Где:

- n – эффективный объем выборки (адаптивно рассчитывается как max(2/α, p+1))
- p – количество контролируемых параметров

Определение порогового значения

Критическое значение для детектирования аномалий:

$$\mathrm{F}_{threshold} = F_{1-\alpha}(p, n-p) \times 2$$

Где:

- α – уровень значимости (0.05)
- $F_{1-\alpha}$ – квантиль F-распределения
- Множитель 2 введен для повышения чувствительности

Практические аспекты:

- Уровень α=0.05 соответствует 95% доверительному интервалу
- Множитель 2 является эмпирической поправкой для промышленных данных
- При необходимости может быть адаптирован под конкретный процесс

Интерпретация результатов

- Если F > F_threshold – фиксируется аномалия
- Последовательность аномалий указывает на переходный процесс
- Величина превышения характеризует степень отклонения

Интерпретация на код

Метод find_stable_segment

Алгоритм проверяет данные через скользящее окно фиксированной длины (по умолчанию 30 точек), последовательно анализируя каждый отрезок. Для каждого такого отрезка сначала вычисляются медианные значения всех параметров и их абсолютные медианные отклонения (MAD). Чтобы сравнивать параметры разных масштабов, используется относительный MAD (rMAD), получаемый делением MAD на абсолютное значение медианы.

Критерии стабильности включают два основных условия. Во-первых, не менее 90% параметров в анализируемом отрезке должны иметь относительное отклонение rMAD меньше заданного порога (по умолчанию 5%). Во-вторых, медианное значение rMAD по всем параметрам не должно превышать 1.5-кратного значения порога. Такой подход позволяет отсеять участки с локальными аномалиями или нестабильностью параметров.

Возвращает первый подходящий отрезок или ошибку, если стабильный участок не найден.

Метод adaptive_monitoring

Для каждого нового наблюдения x_t:

1. Расчет статистики:
    - Отклонение: diff = x_t - mean.
    - T^2-статистика Хотеллинга:
    - F-преобразование
2. Детектирование аномалий:
    - Аномалия: F > F_threshold.
    - При transition_window последовательных аномалиях активируется переходный режим (transition_flag = True).
3. Обновление модели:
    - Экспоненциальное сглаживание (для плавной адаптации):

1. Среднее: mean = l1 * x_t + (1 - l1) * mean (l1 = 0.05).
2. Ковариация: cov = l2 * np.outer(residual, residual) + (1 - l2) * cov (l2 = 0.01).

4. Пересчет ковариационной матрицы при смене режима

При обнаружении нового стационарного участка:

- Условия:
  - Система в переходном режиме.
  - Все параметры стабильны (MAD < var_threshold) в течение transition_window точек.

5. Сохранение данных:
   - Записываются: F-статистика, порог, метки аномалий и переходов.
   - Возвращается DataFrame с временными метками.