ОСНОВЫ ХИМИЧЕСКОЙ МЕТРОЛОГИИ
Лекция 4 Дисперсионный анализ

лектор: Образовский Е. Г.

3 марта 2015 г.

Общая погрешность анализа складывается из погрешностей различных этапов. Если возникает необходимость уменьшения общей погрешности анализа, следует проанализировать вклады отдельных этапов анализа. Этой цели и служит дисперсионный анализ.

Наиболее существенные вклады в общую погрешность в разных ситуациях могут давать различные этапы анализа. Однако, как показывает практика, во многих случаях наиболее заметный вклад дает процедура пробоотбора. Располагая достаточной информацией, в отдельных случаях можно оценить погрешность пробоотбора теоретически.

Например, если исходный материал состоит из отдельных частиц (гранул), каждая из которых либо содержит (с вероятностью p), либо не содержит определяемый компонент, то в отобранной из большого объема пробе из n частиц в среднем $\mu = np$ частиц содержат определяемый компонент. Вероятность $f(m)$ найти m частиц с определяемым компонентом дается распределением Пуассона:

$$\mathbf{f(m) = \frac{\mu^{m}}{m!} \cdot e^{-\mu}}.$$

Тогда погрешность пробоотбора $s_{\text{п}}$ определяется среднеквадратичным отклонением

$$s_{\text{п}} = \sqrt{\mu} = \sqrt{np}, \quad s_{\text{п отн}} = \frac{\sqrt{\mu}}{\mu} = \frac{1}{\sqrt{np}}.$$

Задавая необходимую погрешность пробоотбора и зная приблизительно долю определяемого компонента в анализируемом материале, мы можем рассчитать необходимое количество частиц (гранул) в пробе для анализа.

$$n = \frac{1}{p s_{\text{п}}^2}.$$

Пример Рассмотрим два случая.
а) Для вероятности содержания определяемого компонента $p = 0,1$ и допустимого значения относительного стандартного отклонения пробоотбора $S_{п} = 0,01$ по приведенной формуле получаем, что минимальное необходимое значение числа частиц в пробе для анализа составляет $n = 10^5$.
б) Для 1 дм^3 водного раствора, содержащего $10^{-6}M$ NaCl, получаем $p = 10^{-6}/55,5 = 2 \cdot 10^{-8}$. Тогда для $S_{п} = 0,0001$ минимальное число частиц $n = 5 \cdot 10^{15}$, т. е. необходимый минимальный объем составляет всего $V \approx 2 \cdot 10^{-10}$ дм^3.

Зная стандартные отклонения пробоотбора и анализа, можно оптимизировать общую погрешность. Например, пусть имеется n образцов и у нас есть возможность проанализировать $k(< n)$ проб. Рассмотрим две схемы анализа.

Схема 1. Анализируем k из n образцов и за результат анализа берем среднее значение. Тогда

$$\bar{S}_{a}^{(1)} = \frac{S_a}{\sqrt{k}}, \quad \bar{S}_{п}^{(1)} = \frac{S_п}{\sqrt{k}}, \quad \bar{S}^{(1)} = \frac{\sqrt{S_a^2 + S_п^2}}{\sqrt{k}}.$$

Схема 2. Смешиваем n образцов и выбираем из смеси k проб для анализа. Перемешивание уменьшает дисперсию пробоотбора

$$\bar{S}_{\text{п}} = \frac{S_{\text{п}}}{\sqrt{n}},$$

но не меняет дисперсию анализа, так что при анализе k образцов из смеси среднеквадратичное отклонение для среднего значения равно

$$\bar{S}^{(2)} = \frac{\sqrt{S_a^2 + S_{\text{п}}^2/n}}{\sqrt{k}} < \bar{S}^{(1)}.$$

Пример От большого количества материала отобрано 10 образцов, погрешность пробоотбора $S_{п} = 0,10$, погрешность единичного определения интересующего компонента $S_a = 0,05$. Рассмотрим следующие планы анализа.
План 1. Анализируем 10 проб и за результат анализа принимаем среднее значение. Тогда

$$\bar{S}^{(1)} = \frac{\sqrt{(0,05)^2 + (0,10)^2}}{\sqrt{10}} = 0,035.$$

План 2. Перемешиваем 10 образцов и анализируем 1 пробу. Тогда

$$\bar{S}^{(2)} = \sqrt{(0,05)^2 + (0,10)^2/10} = 0,059.$$

План 3. Перемешиваем 10 образцов и анализируем 10 проб. Тогда

$$\bar{s}^{(3)} = \frac{\sqrt{(0,05)^2 + (0,10)^2/10}}{\sqrt{10}} = 0,019.$$

План 4. Перемешиваем 10 образцов и анализируем 3 пробы. Тогда

$$\bar{s}^{(4)} = \frac{\sqrt{(0,05)^2 + (0,10)^2/10}}{\sqrt{3}} = 0,034.$$

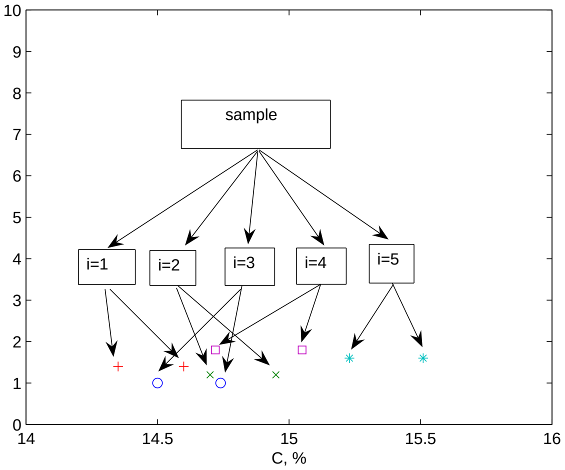
Из приведенных данных видно, что план анализа № 4, будучи значительно экономичнее плана № 1, дает даже немного лучшие результаты.

Сравнение двух схем анализа.

Простой дисперсионный анализ

В подавляющем большинстве случаев разложение ошибок на составляющие проводится экспериментально. Рассмотрим простейший случай разложения ошибки анализа на две составляющие, например на ошибку пробоотбора и ошибку собственно анализа. Для этого используют следующую схему эксперимента. От большой партии исходного материала отбирают m проб, каждая анализируется n_j раз, рис.1.

Простой дисперсионный анализ



Отобранные пробы гомогенизируют, так что среднеквадратичное отклонение при анализе отдельных проб примерно одинаково (что проверяется с помощью статистических критериев, например, критерия Кохрена). Однако из-за неоднородности исходной пробы средние результаты анализов различных проб имеют отклонения, что становится дополнительной причиной ошибок, увеличивая случайную ошибку метода. Дисперсия средних между пробами

$$S_1^2 = \frac{\sum_{i=1}^{m}(\bar{x}_j - \bar{x})^2}{m-1}$$

складывается из дисперсии внутри проб

$$S_2^2 = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_j-1}\sum_{j=1}^{n_j}(x_{ij}-\bar{x}_j)^2$$

и из дисперсии пробоотбора $S_П^2$.

В случае значимости различия между $S_1^2 \cdot n_j$ и S_2^2, проверяемой по F-критерию Фишера, т. е. при условии

$$\mathbf{F = \frac{S_1^2 \cdot n_j}{S_2^2} > F_{\text{табл}}(P = 0,95; \nu_1 = m - 1; \nu_2 = (n_j - 1)m)},$$

погрешность пробоотбора определяется по формуле

$$\mathbf{S_{\text{п}} = \sqrt{S_1^2 - \frac{S_2^2}{n_j}}}.$$

Эффективность дисперсионного анализа зависит также от соотношения погрешности пробоотбора S_s и анализа S_a, Ниже приведены данные относительной доли экспериментов, в которых наблюдаются незначимые отличия дисперсий по критерию Фишера для двух значений $S_s/S_a = 2.0$ и $S_s/S_a = 3.0$ для случая $m = 6,\ n_j = 2.$

Относительная доля экспериментов, в которых наблюдаются незначимые отличия дисперсий по критерию Фишера для двух значений $S_s/S_a = 2.0$ и $S_s/S_a = 3.0$ для случая $m = 6$, $n_j = 2$.

Для эффективного применения данной схемы разделения ошибок на составляющие необходимо выбрать оптимальное сотношение между числом отбираемых точечных проб m и числом параллельных анализов каждой отобранной пробы n_j.
В этом нам помогают результаты численного моделирования. Рассмотрим в качестве примера три схемы, при которых полное число анализируемых проб $m \cdot n_j$ постоянно:

1. $m = 5, n_j = 8$. Небольшое число параллельных проб приводит к большому разбросу результатов определения S_S

1. $m = 5, n_j = 8$. Небольшое число параллельных проб приводит к большому разбросу результатов определения S_S

2. $m = 10, n_j = 4$. Увеличение числа параллельных проб приводит к уменьшению разброса результатов определения S_s даже при увеличении разброса результатов определения S_a

2. $m = 10, n_j = 4$. Увеличение числа параллельных проб приводит к уменьшению разброса результатов определения S_s даже при увеличении разброса результатов определения S_a

3. $m = 20, n_j = 2$. Дальнейшее увеличение числа параллельных проб приводит к большему уменьшению разброса результатов определения S_s несмотря на заметное увеличении разброса результатов определения S_a

3. $m = 20, n_j = 2$. Дальнейшее увеличение числа параллельных проб приводит к большему уменьшению разброса результатов определения S_s несмотря на заметное увеличении разброса результатов определения S_a

Данные численного моделирования позволяют сделать вывод, что наиболее эффективной является схема 3. $m = 20, n_j = 2$, поскольку именно она дает наиболее узкую функцию распределения для стандартного отклонения погрешности пробоотбора.

Пример От большой партии исходного материала отобрали $m = 6$ проб и каждую проанализировали $n_j = 2$ раза. Результаты приведены в таблице в строках 2 и 3. По этим данным рассчитывают средние

$$\bar{x}_i = \frac{x_{i1} + x_{i2}}{2}, \quad S_i^2 = \frac{(x_{i1} - x_{i2})^2}{2},$$

которые также приведены в таблице в строках 3 и 4.

i	1	2	3	4	5	6
$j = 1$ $j = 2$	14,72 15,05	15,51 15,23	14,60 14,35	15,10 15,23	14,70 14,95	14,74 14,50
x_i	14,885	15,370	14,475	15,165	14,825	14,620
s_i	0,233	0,198	0,177	0,092	0,177	0,170

Проверяем однородность значений относительных стандартных отклонений S_i по критерию Кохрена. Для этого рассчитываем тестовую статистику

$$\mathbf{C} = \frac{\mathbf{S}_{\mathbf{max}}^2}{\sum_{\mathbf{i}=1}^{\mathbf{m}} \mathbf{S}_{\mathbf{i}}^2} = \frac{\mathbf{0,0543}}{\mathbf{0,194}} = \mathbf{0,281}$$

и сравниваем с табличным $C_{5\%}(m = 6, n = 2) = 0,781$. Поскольку рассчитанное значение тестовой статистики не превышает 5 %-го критического значения, совокупность стандартных отклонений можно считать однородной.

Тогда рассчитываем величины

$$\bar{x} = \sum_{i=1}^{m} \bar{x}_i / m = 14,890,$$

$$S_1^2 = \frac{\sum_{i=1}^{m} (\bar{x}_i - \bar{x})^2}{m-1} = 0,111, \quad S_2^2 = \frac{\sum_{i=1}^{m} S_i^2}{m} = 0,0323$$

и проверяем значимость отличия S_1 и S_2 по критерию Фишера:

$$F = \frac{S_1^2 \cdot n_j}{S_2^2} = 6,87 > F_{\text{табл}}(P = 0,95; \nu_1 = 5; \nu_2 = 6) = 4,39.$$

Тогда погрешность пробоотбора равна

$$S_{\text{п}} = \sqrt{S_1^2 - \frac{S_2^2}{2}} = 0,31.$$

Классификация или факторный анализ

Для решения многих практических задач недостаточно определить содержания одной или нескольких компонент, часто необходимо на основе полученных данных классифицировать объекты анализа.

Например

Классификация или факторный анализ

Для решения многих практических задач недостаточно определить содержания одной или нескольких компонент, часто необходимо на основе полученных данных классифицировать объекты анализа.

Например

- классифицировать улики в судебном анализе

Классификация или факторный анализ

Для решения многих практических задач недостаточно определить содержания одной или нескольких компонент, часто необходимо на основе полученных данных классифицировать объекты анализа.

Например

- классифицировать улики в судебном анализе
- идентифицировать источники загрязнения окружающей среды по результатам анализа проб воздуха, почвы, воды

Классификация или факторный анализ

Для решения многих практических задач недостаточно определить содержания одной или нескольких компонент, часто необходимо на основе полученных данных классифицировать объекты анализа.
Например

- классифицировать улики в судебном анализе
- идентифицировать источники загрязнения окружающей среды по результатам анализа проб воздуха, почвы, воды
- классифицировать археологические находки на основе анализа микроэлементов

Классификация или факторный анализ

Имеется множество объектов, в каждом из которых определена совокупность некоторых свойств (например, концентрации нескольких элементов). Задача заключается в нахождении или предсказании свойств объекта, которое непосредственному измерению не подвергалось, но так что оно считалось косвенно связанным с измерением через неизвестные или неопределенные соотношения.

Например, имеются археологические изделия и требуется установить из какого из многих возможных источников они получены. Для этой цели определяется элементный состав возможных источников и проводится классификация этих источников (разбиение на некоторое число классов). Затем по элементному составу неизвестного образца относят его к тому или иному классу (источнику).
В данном примере совокупностью объектов являются образцы из возможных источников, а свойствами – концентрации элементов в этих образцах.
Исходные данные для классификации удобно представить в виде матрицы

$$\hat{\mathbf{X}} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2k} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{i1} & x_{i2} & \dots & x_{ik} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk} & \dots & x_{np} \end{pmatrix},$$

где i – номер объекта, k – элемент. Содержания различных элементов могут существенно различаться и для того чтобы они все могли быть использованы для классификации необходимо провести масштабирование данных.

Наиболее естественным является так называемое автомасштабирование, когда данные преобразуются согласно

$$x'_{ik} = \frac{x_{ik} - \bar{x}_k}{s_k},$$

где

$$\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik}, \quad s_k = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}.$$

Существует множество методов проведения классификации. В качестве примера мы рассмотрим *проекционный метод* и *кластерный анализ*.

Проекционный метод

Проекционный метод основан на выполнении вращения матрицы данных так, чтобы первая новая ось отвечала направлению наибольшей дисперсии данных, а каждая последующая ось — максимуму остаточной дисперсии.

В качестве примера рассмотрим возможность отнесение объектов к трем классам по следующим результатам определения их элементного состава (концентрации элементов в исходных данных приведена в ppm).

Проекционный метод

$$\hat{\mathbf{X}} = \begin{array}{c} \begin{matrix} Cu & Mn & Cl & Br & I \end{matrix} \\ \begin{pmatrix} 9.2 & 0.30 & 1730 & 12.0 & 3.6 \\ 12.4 & 0.39 & 930 & 50.0 & 2.3 \\ 7.2 & 0.32 & 2750 & 65.3 & 3.4 \\ 10.2 & 0.36 & 1500 & 3.4 & 5.3 \\ 10.1 & 0.50 & 1040 & 39.2 & 1.9 \\ 6.5 & 0.20 & 2490 & 90.0 & 4.6 \\ 5.6 & 0.29 & 2940 & 88.0 & 5.6 \\ 11.8 & 0.42 & 867 & 43.1 & 1.5 \\ 8.5 & 0.25 & 1620 & 5.2 & 6.2 \end{pmatrix} \end{array}.$$

После автомасштабирования матрица данных принимает следующий вид

$$\hat{\mathbf{X}} = \begin{pmatrix} 0.06 & -0.40 & -0.04 & -0.97 & -0.13 \\ 1.44 & 0.58 & -1.05 & 0.18 & -0.89 \\ -0.80 & -0.18 & 1.25 & 0.64 & -0.25 \\ 0.49 & 0.25 & -0.33 & -1.23 & 0.87 \\ 0.45 & 1.78 & -0.91 & -0.15 & -1.13 \\ -1.10 & -1.49 & 0.92 & 1.39 & 0.46 \\ -1.48 & -0.51 & 1.49 & 1.33 & 1.04 \\ 1.18 & 0.91 & -1.13 & -0.03 & -1.36 \\ -0.24 & -0.95 & -0.18 & -1.17 & 1.40 \end{pmatrix}.$$

Проекционный метод

Матрица поворота $\hat{R}$ составлена из собственных векторов матрицы ковариации исходных преобразованных данных

$$\hat{\mathbf{C}} = \frac{1}{\mathbf{p}-1}\hat{\mathbf{X}}^{\mathbf{T}} \cdot \hat{\mathbf{X}},$$

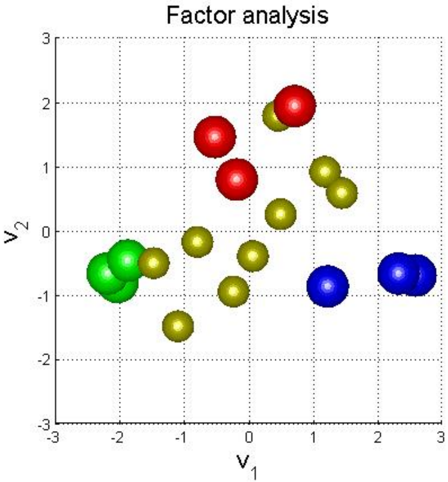
и имеет вид

$$\hat{\mathbf{R}} = \begin{pmatrix} -0.52 & 0.11 & 0.41 & 0.21 & 0.71 \\ -0.46 & -0.27 & -0.79 & 0.28 & 0.07 \\ 0.52 & -0.17 & -0.31 & -0.36 & 0.69 \\ 0.28 & -0.76 & 0.28 & 0.52 & 0.00 \\ 0.41 & 0.56 & -0.17 & 0.69 & 0.12 \end{pmatrix},$$

а собственные значения

$$\lambda = \begin{pmatrix} 26.8 & 9.5 & 2.3 & 1.1 & 0.4 \end{pmatrix}.$$

Преобразованные с помощью проекционного метода данные, в отличие от исходных, однозначно разбиваются на три класса в пространстве двух параметров.



Кластерный анализ

Кластерный анализ использует для классификации понятие расстояния между объектами, определяемого как

$$\mathbf{d_{ij}} = \sqrt{\sum_{\mathbf{k}=1}^{\mathbf{p}} (\mathbf{x_{ik}} - \mathbf{x_{jk}})^2}.$$

Совокупность всех расстояний можно представить в виде матрицы

$$\hat{\mathbf{D}} = \begin{pmatrix} 0 & 0 & \dots & \dots & 0 \\ d_{21} & 0 & \dots & \dots & 0 \\ d_{31} & d_{32} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & 0 \\ d_{n1} & d_{n2} & \dots & d_{nn-1} & 0 \end{pmatrix}.$$

Среди всех расстояний находится минимальное, например d_{ij}, и тогда объекты i и j объединяются в один кластер. Все расстояния пересчитываются, например, новое расстояние от полученного кластера ij до объекта k вычисляют как

$$\mathbf{d}_{\mathbf{k(ij)}} = \frac{\mathbf{d}_{\mathbf{ki}} + \mathbf{d}_{\mathbf{kj}}}{\mathbf{2}}.$$

Процедуру повторяют снова до тех пор пока не останется необходимое число кластеров. Это удобно представить в виде дендрограммы, рис.2.

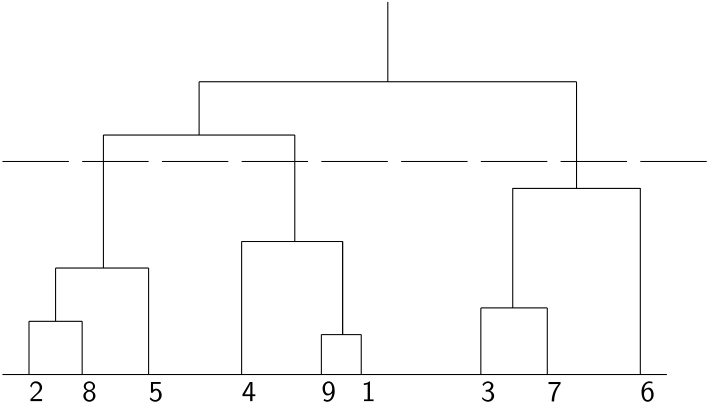


Рис.: Дендрограмма