Калибровка (градуировка)

Алексей Померанцев

Содержание

1 Введение

В этом пособии рассмотрены основные методы, применяемые для решения задач калибровки (называемой также градуировкой). Текст ориентирован, прежде всего, на специалистов в области анализа экспериментальных данных: химиков, физиков, биологов, и т.д. Он может служить пособием для исследователей, начинающих изучение этого вопроса. Продолжить исследования можно с помощью указанной литературы.

В пособии интенсивно используются понятия и методы матричной алгебры – вектор, матрица, и т.п. Читателям, которые плохо знакомы с этим аппаратом, рекомендуется изучить, или, хотя бы просмотреть, пособие *Матрицы и векторы*.

Изложение иллюстрируется примерами, выполненными в рабочей книге Excel Calibration.xlsx, которая сопровождает этот документ. Предполагается, что читатель имеет базовые навыки работы в среде Excel, умеет проводить простейшие матричные вычисления с использованием функций листа, таких как `МУМНОЖ`, `ТЕНДЕНЦИЯ` и т.п. Примеры приведенные в пособии имеют абстрактный, модельный характер, однако, по своей сути, они тесно связаны с задачами, встречающимися на практике.

2 Базовые сведения

2.1 Постановка задачи

Суть задачи *калибровки* состоит в следующем. Пусть имеется некоторая переменная (свойство) y, величину которой необходимо установить. Однако, по ряду обстоятельств (недоступность, дороговизна, длительность), прямое измерение величины y невозможно. В то же время можно легко (быстро, дешево) измерить другие величины: $\mathbf{x} = (x_1, x_2, x_3, \ldots)$, которые связаны с искомой величиной y. Задача калибровки состоит в установлении количественной связи между *переменными* $\mathbf{x}$ и откликом y –

$$y = f(x_1, x_2, x_3, \ldots | a_1, a_2, a_3, \ldots) + \epsilon$$

На практике это означает:

1. подбор вида зависимости f, и ...
2. оценку неизвестных параметров $a_1, a_2, a_3, \ldots$ в этой калибровочной зависимости.

Разумеется, калибровку нельзя построить абсолютно точно. В дальнейшем мы увидим, что это не только невозможно, но и опасно. В калибровочной зависимости всегда присутствуют погрешности (ошибки) ε, источник которых – пробоотбор, измерения, моделирование, и т.д.

Простейший пример калибровки дает общеизвестный прибор, называемый безменом, т.е. пружинные весы. Искомая величина y – это вес образца, а x – это удлинение пружины весов. Закон Гука —

$$y = ax + \varepsilon$$

связывает y и x через коэффициент жесткости пружины a, который априори неизвестен. Процедура калибровки очень проста – взвешиваем стандартный образец весом в 1 кг и отмечаем на шкале удлинение пружины, затем используем образец в 2 кг, и т.д. В результате этой калибровки (точнее, градуировки) получается шкала, по которой можно определить вес нового, нестандартного образца.

Этот элементарный пример демонстрирует основные черты процедуры калибровки, которые подробно будут рассмотрены далее. Во-первых, для калибровки необходимы несколько стандартных образцов,

для которых величины y известны заранее. Во-вторых, диапазон, в котором предполагается измерять показатель y, должен полностью покрываться этими калибровочными образцами. Действительно, нельзя измерять образцы весом более 5 кг, если в калибровке использовались образцы, весом менее чем 5 кг.

Разумеется, на практике все обстоит не так просто, как в этом элементарном примере. Например, в калибровке может участвовать не один показатель y (отклик), а несколько откликов $y_1, y_2, \dots y_K$, которые нас интересуют. Все возможные особенности, различные трудности, сопутствующие процедуре калибровке будут рассмотрены далее. Сейчас же мы подведем первые итоги и сформулируем задачу калибровки в общем виде.

Пусть имеется матрица **Y**, размерностью $(I \times K)$, где I – это число стандартных образцов (сравнения), использованных в калибровке, а K – это число одновременно калибруемых откликов. Матрица **Y** содержит значения откликов y, известные из независимых экспериментов (референтные или стандартные значения). Пусть, с другой стороны, имеется соответствующая матрица переменных **X** размерностью $(I \times J)$, где I – это, естественно, тоже число образцов, а J – это число независимых переменных (каналов), используемых в калибровке. Матрица **X** состоит из альтернативных, как правило, многоканальных ($J >> 1$) измерений. Используя калибровочные данные $(\mathbf{X}, \mathbf{Y})$, требуется построить функциональную связь между **Y** и **X**.

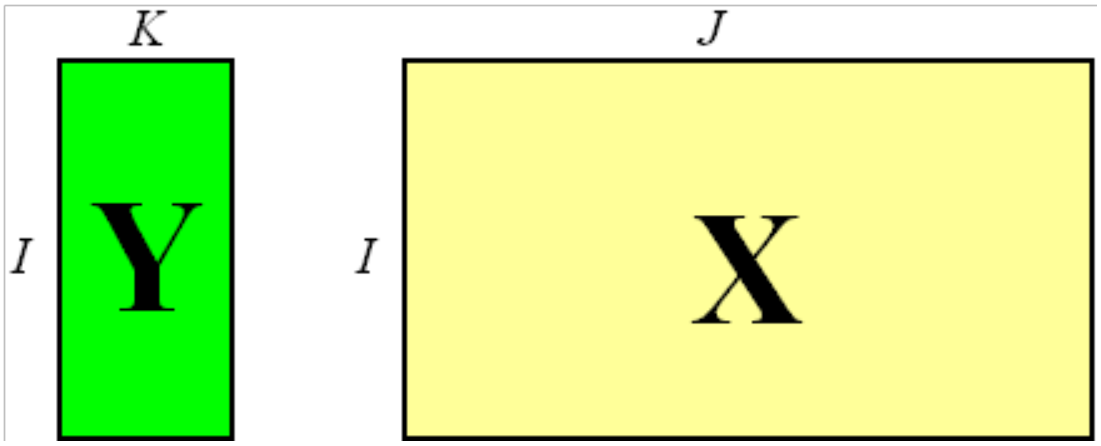


Рис. 2.1. Калибровочный набор данных

Итак, задача калибровки состоит в построении математической модели, связывающей блоки **X** и **Y**, с помощью которой можно в дальнейшем предсказывать значения показателей $y_1, y_2, \dots y_K$, по новой строке значений аналитического сигнала **x**.

2.2 Линейная и нелинейная калибровки

Теоретически функциональная связь между переменными x и y может быть сложной, например,

$$y = b_0 \exp(b_1 x + b_2 x^2) + \varepsilon$$

Однако на практике большинство используемых калибровок являются линейными, т.е. имеют вид –

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_J x_J + \varepsilon$$

Заметим, что термин “линейность” употребляется в контексте этого пособия вместо более правильного термина “билинейность”, означающего линейность как по отношению к переменным x, так и в отношении коэффициентов b. Поэтому, калибровка

$$y = b_0 + b_1 x + b_2 x^2 + \varepsilon$$

является нелинейной, несмотря на то, что она легко “линеаризуется” введением новой переменной $x_2 = x^2$.

Главное преимущество билинейной модели – это единственность, тогда как все прочие калибровки образуют бесконечное множество, выбор из которого затруднителен. В этом пособии рассматриваются только линейные калибровки вида –

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

Читатели, заинтересованные в изучении нелинейных методов анализа данных могут обратиться к описанию программы Fitter.

Предпочтение (би)линейных, формальных моделей, не отягощенных дополнительным физико-химическим смыслом, является магистральным направлением развития хемометрики. Такой подход позволяет исключить влияние субъективного фактора, проявляющегося при выборе калибровочной зависимости. Однако за это приходится расплачиваться – все линейные модели имеют ограниченную область применения.

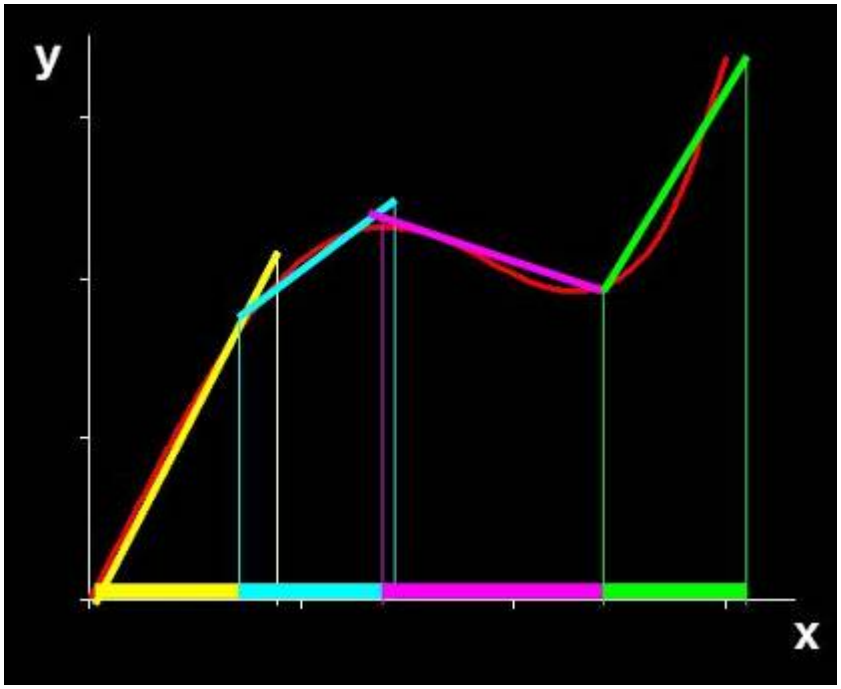


Рис. 2.2. Линеаризация калибровки

Суть дела иллюстрирует Рис. 2.2, где показаны четыре участка, на которых сложная нелинейная зависимость приближается простыми линейными моделями. Каждая из этих моделей работает только на своей области переменной x и выход за ее границы приводит к грубым ошибкам. Принципиальным моментом здесь является то, какую область можно считать допустимой – иначе говоря, насколько широко можно применять калибровочную модель. Ответ на этот вопрос дают методы проверки (валидации).

2.3 Калибровка и проверка

При надлежащем построении калибровочной модели исходный массив данных состоит из двух независимо полученных наборов, каждый из которых является достаточно представительным. Первый набор, называемый *обучающим*, используется для идентификации модели, т.е. для оценки ее параметров. Второй набор, называемый *проверочным*, служит только для проверки модели. Построенная модель применяется к данным из проверочного набора, и полученные результаты сравниваются с проверочными значениями. Таким образом принимается решение о правильности, точности моделирования методом тест-валидации.

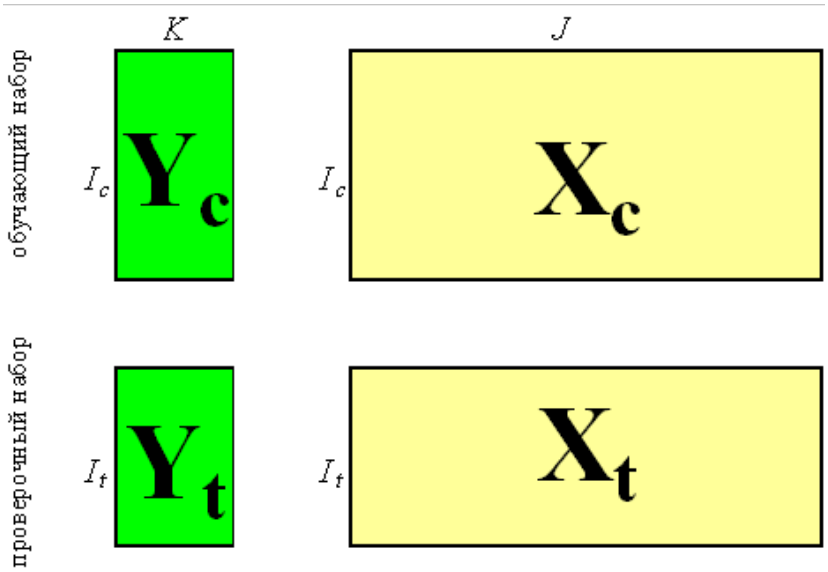


Рис. 2.3. Обучающий и проверочный наборы

В некоторых случаях объем данных слишком мал для такой проверки. Тогда применяют другой метод – *перекрестной проверки* (кросс-валидация, скользящая проверка). В этом методе проверочные значения вычисляют с помощью следующей процедуры.

Некоторую фиксированную долю (например, первые 10% образцов) исключают из исходного набора данных. Затем строят модель, используя только оставшиеся 90% данных, и применяют ее к исключенному набору. На следующем цикле исключенные данные возвращаются, и удаляется уже другая порция данных (следующие 10%), и опять строится модель, которая применяется к исключенным данным. Эта процедура повторяется до тех пор, пока все данные не побывают в числе исключенных (в нашем случае – 10 циклов). Наиболее (но неоправданно) популярен вариант перекрестной проверки, в котором данные исключаются по одному (Leave-One-Out – LOO).

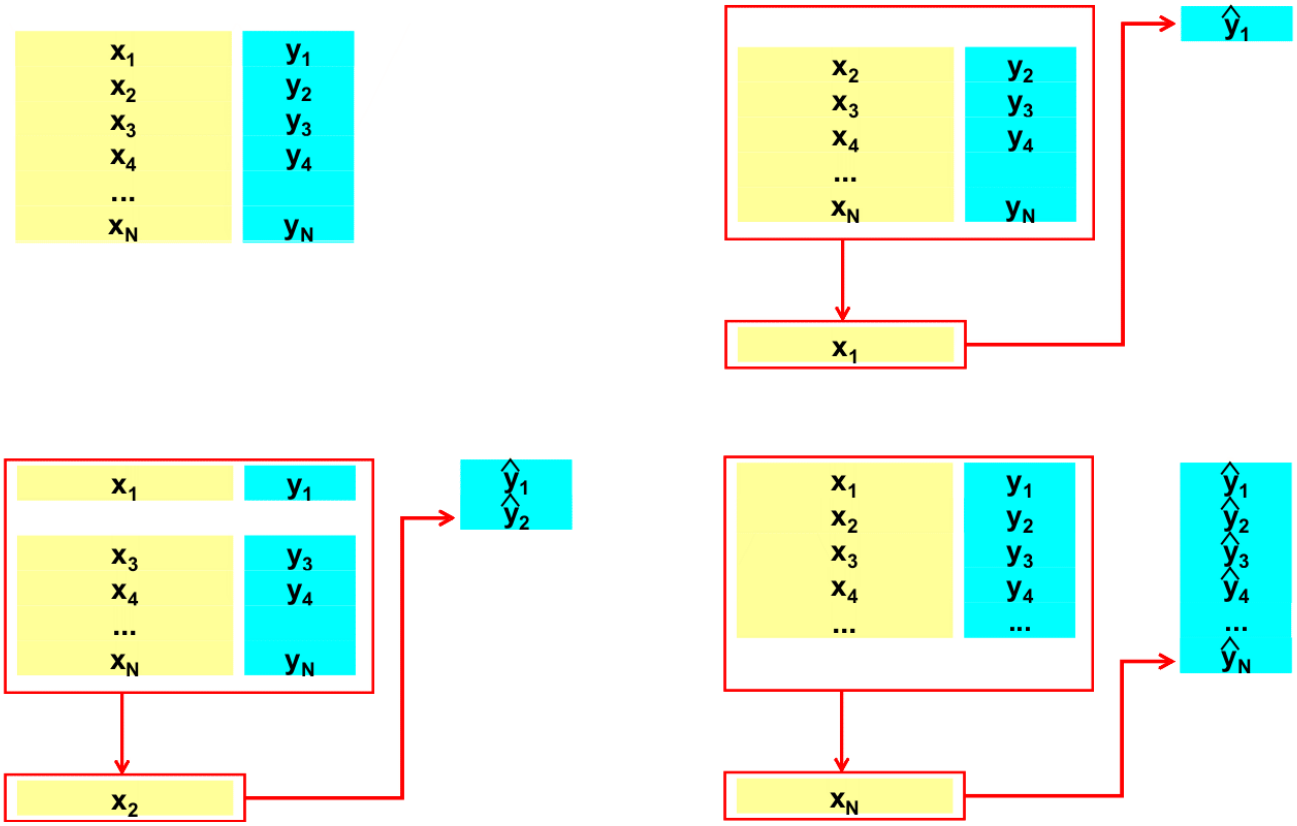


Рис. 2.4. Метод перекрестной проверки

Используется также проверка методом корректировки размахом, суть которой предлагается изучить самостоятельно.

2.4 “Качество” калибровки

Результатом калибровки являются величины $\hat{\mathbf{Y}}_c$ — оценки стандартных откликов $\mathbf{Y}_c$, найденные по модели, построенной на обучающем наборе. Результатом проверки служат величины $\hat{\mathbf{Y}}_t$ – оценки проверочных откликов $\mathbf{Y}_t$, вычисленные по той же модели. Отклонение оценки от стандарта (проверочного значения) вычисляют как матрицу остатков: в обучении $\mathbf{E}_c = \mathbf{Y}_c - \hat{\mathbf{Y}}_c$, и в проверке $\mathbf{E}_t = \mathbf{Y}_t - \hat{\mathbf{Y}}_t$.

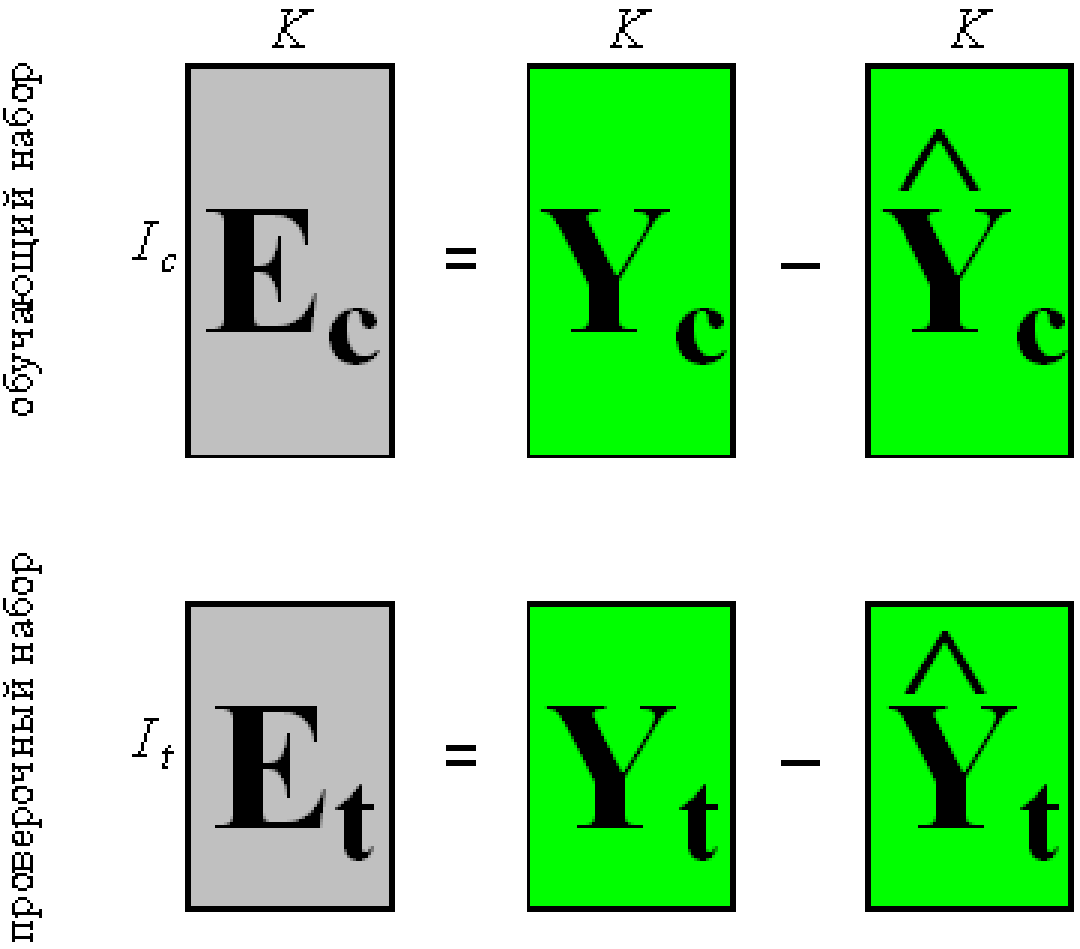


Рис. 2.5. Остатки в обучении и в проверке

Следующие величины характеризуют “качество” калибровки в среднем.

A. *Полная дисперсия* остатков в обучении ($TRVC$) и в проверке ($TRVP$) –

$$TRVC = \frac{1}{KI_{\mathrm{c}}} \sum_{k=1}^{K} \sum_{i=1}^{I_{\mathrm{c}}} e_{ki}^2$$

$$TRVP = \frac{1}{KI_{\mathrm{t}}} \sum_{k=1}^{K} \sum_{i=1}^{I_{\mathrm{t}}} e_{ki}^2$$

Эти величины вычисляются по формуле указанное в пособии *Статистика*, для $m = 0$.

B. *Объясненная дисперсия* остатков при обучении ($ERVC$) и при проверке ($ERVP$) –

$$ERVC = 1 - \sum_{k=1}^{K}\sum_{i=1}^{I_c} e_{ki}^2 \Big/ \sum_{k=1}^{K}\sum_{i=1}^{I_c} y_{ki}^2$$

$$ERVP = 1 - \sum_{k=1}^{K}\sum_{i=1}^{I_t} e_{ki}^2 \Big/ \sum_{k=1}^{K}\sum_{i=1}^{I_t} y_{ki}^2$$

C. *Среднеквадратичные остатки* калибровки (*RMSEC*) и проверки (*RMSEP*) –

$$RMSEC(k) = \sqrt{\frac{1}{I_c}\sum_{i=1}^{I_c} e_{ki}^2}$$

$$RMSEP(k) = \sqrt{\frac{1}{I_t}\sum_{i=1}^{I_t} e_{ki}^2}$$

Величины *RMSE* зависят от k – номера отклика.

D. *Смещение* в калибровке (*BIASC*) и в проверке (*BIASP*) –

$$BIASC(k) = \frac{1}{I_c}\sum_{i=1}^{I_c} e_{ki}$$

$$BIASP(k) = \frac{1}{I_t}\sum_{i=1}^{I_t} e_{ki}$$

Величины *BIAS* зависят от k – номера отклика.

E. *Стандартные ошибки* в калибровке (*SEC*) и в проверке (*SEP*) –

$$SEC(k) = \sqrt{\frac{1}{I_c}\sum_{i=1}^{I_c} \big(e_{ki} - BIASC(k)\big)^2}$$

$$SEP(k) = \sqrt{\frac{1}{I_t}\sum_{i=1}^{I_t} \big(e_{ki} - BIASP(k)\big)^2}$$

Величины *SE* зависят от k – номера отклика.

F. *Коэффициенты корреляции* R^2 между стандартными y_{ki} и оцененными $\hat{y}_{ki}$ откликами. Они также вычисляется отдельно для обучающего $R_c^2(k)$ и проверочного $R_y^2(k)$ наборов.

$$R(k) = \frac{I\sum y_{ki}\hat{y}_{ki} - \sum y_{ki}\sum \hat{y}_{ki}}{\sqrt{I\sum y_{ki}^2 - (\sum y_{ki})^2}\sqrt{I\sum \hat{y}_{ki}^2 - (\sum \hat{y}_{ki})^2}}$$

Величины R^2 зависят от k – номера отклика.

Во всех этих формулах величины e_{ki} – это элементы матриц $\mathbf{E}_c$ или $\mathbf{E}_t$. Для характеристик, наименование которых оканчивается на C (например, *RMSEC*), используется матрица $\mathbf{E}_c$ (обучение), а для тех, которые оканчиваются на P (например, *RMSEP*), берется матрица $\mathbf{E}_t$ (проверка).

2.5 Неопределенность, точность и воспроизводимость

Такое большое число показателей, характеризующих качество калибровки, объясняется не только историческими причинами, но и тем, что они отражают различные свойства неопределенности при обучении, и при проверке (прогнозе). Для объяснения этих показателей необходимо ввести понятия воспроизводимости (прецизионности) и точности.

Воспроизводимость (precision) характеризует то, насколько близко находятся друг от друга независимые повторные измерения.

Точность (accuracy) определяет степень близости оценок к истинному (стандартному) значению y.

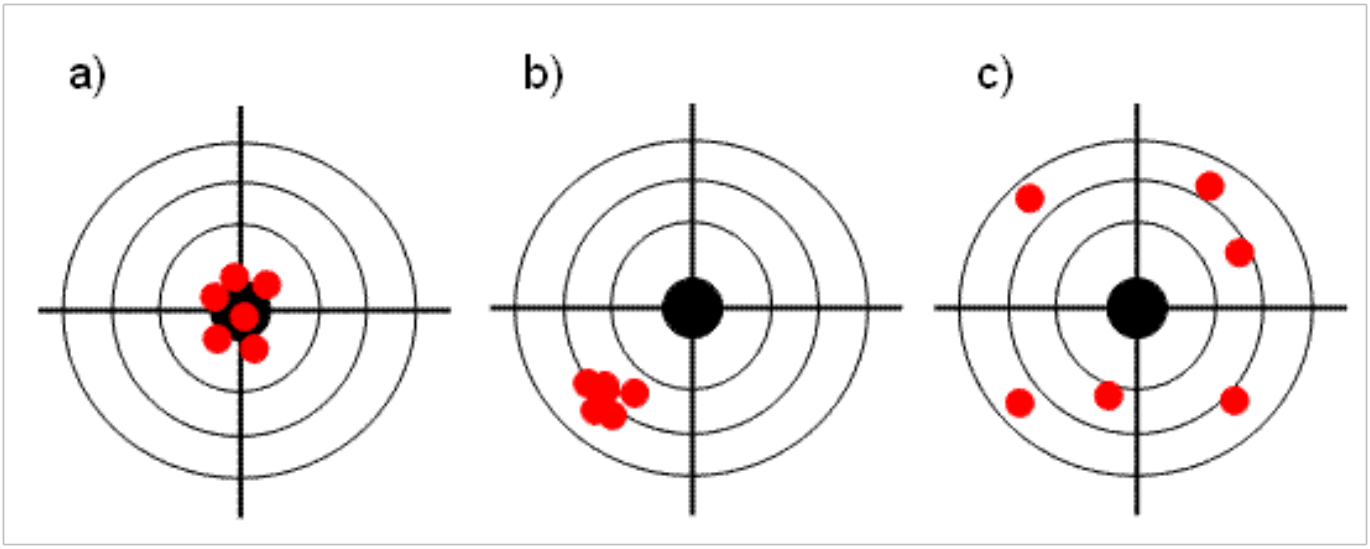


Рис. 2.6. Точность и воспроизводимость

Рис. 2.6 объясняет суть дела. Графики a) и b) представляют оценки с хорошей воспроизводимостью. Вариант a) , кроме того, обладает и высокой точностью, что, разумеется, нельзя сказать о графике c). В последнем случае и воспроизводимость, и точность оставляют желать лучшего.

Показатели SEC и SEP характеризуют воспроизводимость калибровки, тогда как $RMSEC$ и $RMSEP$ показывают ее точность. Величины $BIASC$ и $BIASP$ определяют смещение калибровки относительно истинного значения (Рис. 2.6b). Можно показать, что –

$$RMSE^2 = SE^2 + BIAS^2$$

Таким образом, при построении калибровки предпочтение следует отдавать показателям $RMSE$, а не SE.

Показатели TRV и ERV характеризуют ситуацию “в целом”, без различения калибруемых откликов, т.е.

$$TRV = \frac{1}{K}\sum_{k=1}^{K} RMSE^2(k)$$

2.6 Недооценка и переоценка

При построении калибровки исследователь часто имеет возможность последовательного усложнения модели. Приведем простейший пример, иллюстрирующий этот процесс.

Из курса школьной физики известно, что расстояние L, на которое летит снаряд, выпущенный со скоростью v из орудия, наклоненного под углом α к горизонту, равно –

$$L = v^2 \sin(2\alpha)/g$$

Предположим, что у нас имеются экспериментальные данные $L(\alpha)$.

		α	L
calibration	1	0	0.000
	2	5	1.852
	3	10	3.896
	4	15	4.017
	5	20	7.681
	6	25	6.013
	7	30	8.810
	8	35	10.068
	9	40	7.628
	10	45	11.266
	11	50	8.121
	12	55	7.476
	13	60	9.218
	14	65	7.522
test	15	70	6.010
	16	75	4.723
	17	80	4.467
	18	85	0.764
	19	90	0.725

Рис. 2.7. Данные для артиллерийского примера

Забудем о том, что функциональная связь между L (т.е., y) и α (т.е., x) нам известна, и попробуем построить формальную полиномиальную калибровку –

$$L = b_0 + b_1\alpha + b_2\alpha^2 + \dots b_n\alpha^n + \varepsilon$$

по этим данным. На Рис. 2.8а показано, как описываются обучающие и проверочные данные моделями при $n = 0, 1, \dots, 5$. Видно, что при недостаточной сложности модели ($n = 0, 1$) обучающие данные лежат далеко от модели. Это – случай *недооценки*. Когда сложность модели увеличивается, то согласие между обучающими данными и моделью улучшается. Однако, при излишней сложности ($n = 4, 5$), модель все хуже работает при прогнозе на проверочный набор. Это – случай *переоценки*.

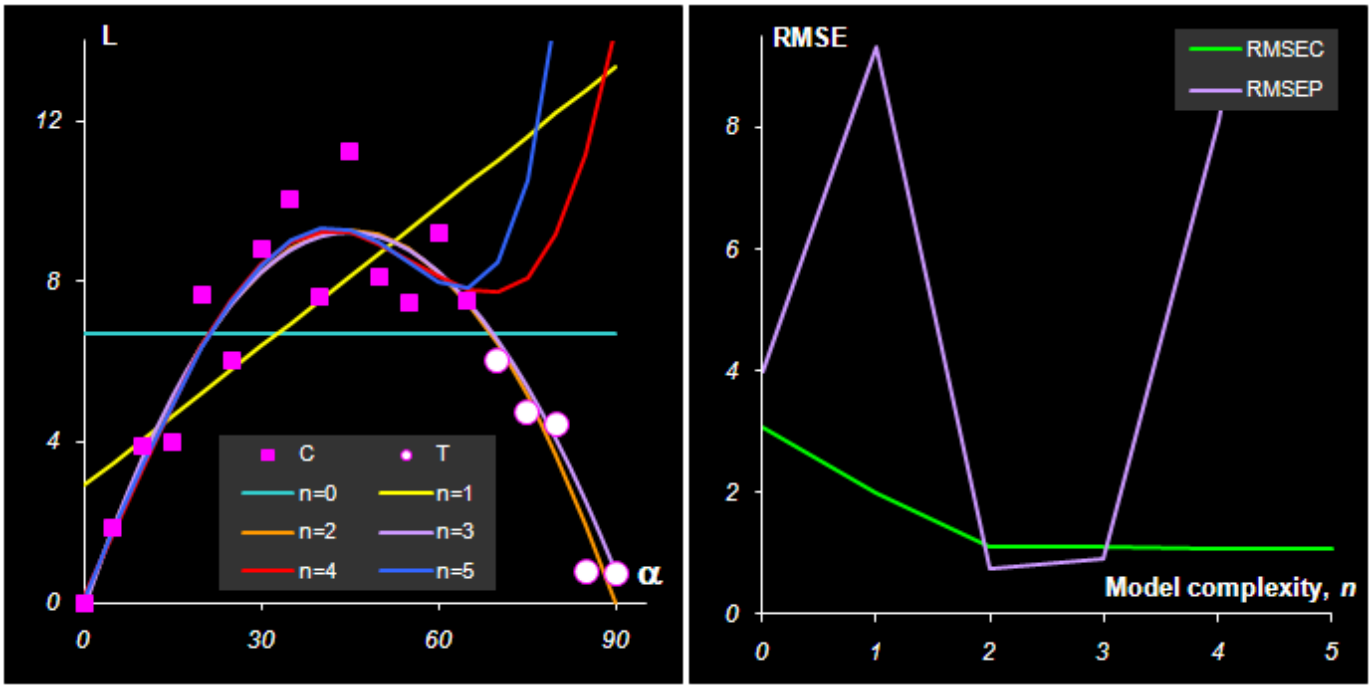


Рис. 2.8. Артиллерийский пример. a) Калибровка, b) Среднеквадратичные остатки обучения (RMSEC) и проверки (RMSEP)

На Рис. 2.8b показано, как изменяются показатели *RMSEC* и *RMSEP* при увеличении сложности модели. Это – типичный график, в котором *RMSEC* монотонно падает, а *RMSEP* проходит через минимум. Именно точка минимума *RMSEP* позволяет определить оптимальную сложность модели. В нашем случае – это 2. Теперь мы можем вспомнить о содержательной модели, которая хорошо аппроксимируется полиномом второй степени по α –

$$L = v^2 \sin(2\alpha)/2g \approx Const\alpha(\alpha - \pi/4)$$

В задачах многомерной калибровки недооценка и переоценка проявляются через выбор числа скрытых, главных компонент. Когда их число мало, модель плохо приближает обучающий набор и при увеличении сложности модели *RMSEC* монотонно уменьшается. Однако качество прогноза на проверочный набор может при этом ухудшаться (U-образная форма *RMSEP*). Точка минимума *RMSEP*, или начало плато, соответствует оптимальному числу главных компонент. Проблема сбалансированности описания данных рассматривается во многих работах А. Хоскюлдссона, который в 1988 году ввел новую концепцию моделирования – так называемый H-принцип. Согласно этому принципу точность моделирования (*RMSEC*) и точность прогнозирования (*RMSEP*) связаны между собой. Улучшение *RMSEC* неминуемо влечет ухудшение *RMSEP*, поэтому их нужно рассматривать совместно. Именно по этой причине множественная линейная регрессия, в которой всегда участвует явно избыточное число параметров, неизбежно приводит к неустойчивым моделям, непригодным для практического применения.

2.7 Мультиколлинеарность

Рассмотрим задачу множественной линейной калибровки

$$\mathbf{Y} = \mathbf{XB}$$

Разумеется, здесь мы имеем дело с обучающим набором данных. Мы не пишем индекс у матриц ($\mathbf{Y}_c$, $\mathbf{X}_c$) только для простоты обозначений. Классическое решение этой задачи находится с помощью метода наименьших квадратов. Минимизируя сумму квадратов отклонений $(\mathbf{Y}-\mathbf{XB})^t(\mathbf{Y}-\mathbf{XB})$, получаем оценки неизвестных коэффициентов $\mathbf{B}$ –

$$\hat{\mathbf{B}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{Y}$$

Далее находится –

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}$$

– оценка искомых откликов. Главная проблема при таком классическом подходе – это обращение матрицы $\mathbf{X}^t\mathbf{X}$. Очевидно, что если число стандартных образцов меньше, чем число переменных ($I < J$) то обратной матрицы не существует. Более того, даже при достаточно большом числе образцов ($I > J$), обратной матрицы может и не быть. Рассмотрим простейший пример.

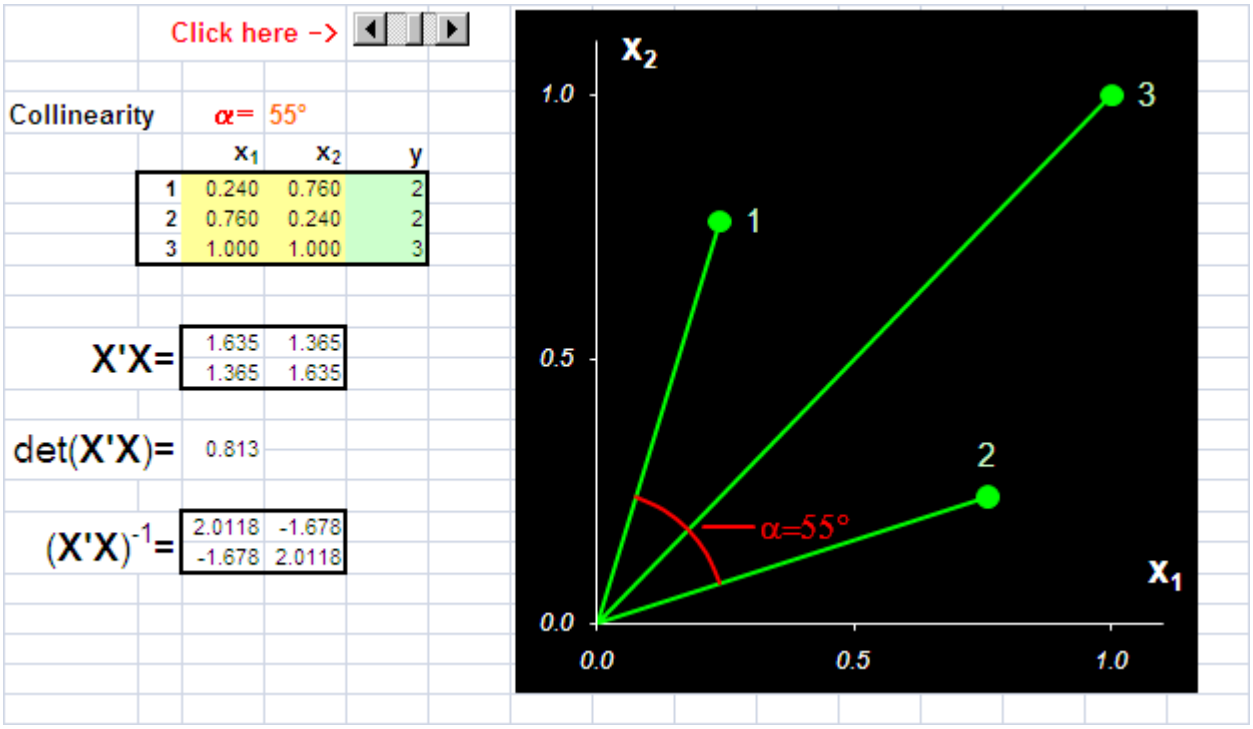


Рис. 2.9. Пример коллинеарных данных

На Рис. 2.9 представлены данные, в которых всего три образа ($I = 3$) и две переменные ($J = 2$). Используя активный элемент на листе, угол между векторами 1 и 2 можно изменять от 90 до 0. Чем меньше угол, тем меньше детерминант матрицы $\mathbf{X}^t\mathbf{X}$ и тем хуже определяется матрица $(\mathbf{X}^t\mathbf{X})^{-1}$. В предельном случае, когда угол между векторами равен 0, матрица $\mathbf{X}^t\mathbf{X}$ не может быть обращена. При этом векторы 1 и 2 колинеарны, т.е. они лежат на одной прямой, совпадающей с вектором 3. Рис. 2.10 иллюстрирует суть проблемы – невозможность построения калибровки классическим способом, когда данные коллинеарны.

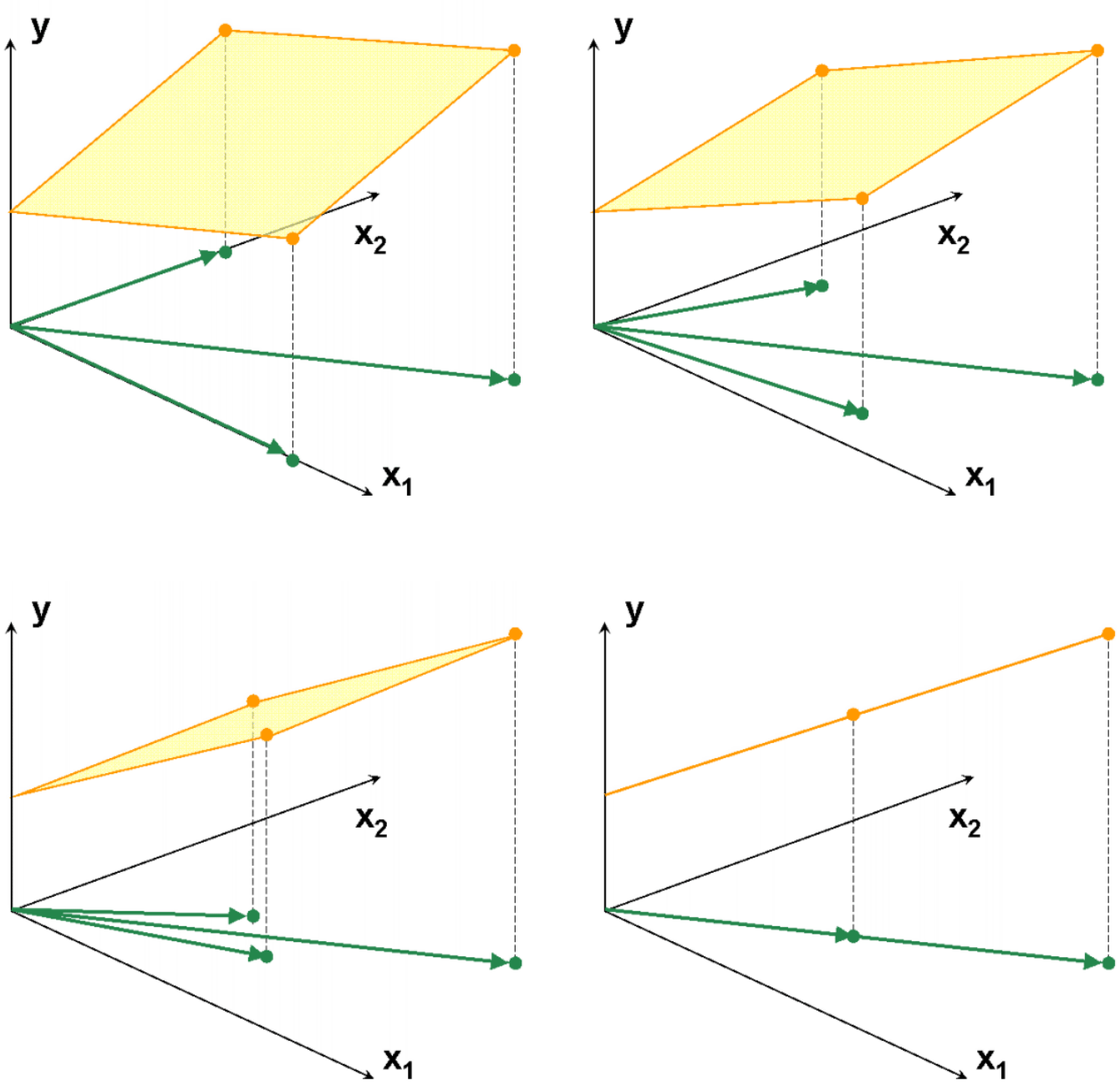


Рис. 2.10. Пример данных с разной степенью коллинеарности

В многомерной калибровке (т.е. при $J >> 1$), мы все время сталкиваемся с мультиколлинеарностью – множественными связями между векторами X-переменных, приводящими к той же проблеме – невозможности обращения матрицы $\mathbf{X}^t\mathbf{X}$. Далее мы увидим, что можно сделать в случае мультиколлинеарности данных.

2.8 Подготовка данных

Важным условием правильного моделирования и, соответственно, успешного анализа, является предварительная подготовка данных, которая включает различные преобразования исходных, “сырых” экспериментальных значений. Простейшими преобразованиями является центрирование и нормирование.

Центрирование – это вычитание из исходной матрицы $\mathbf{X}$ матрицы $\mathbf{M}$, т.е.

$$\tilde{\mathbf{X}} = \mathbf{X} - \mathbf{M}$$

Обычно усреднение проводится по столбцам: для каждого вектора $\mathbf{x}_j$ вычисляется среднее значение –

$$m_i = \left(x_{1j} + \cdots + x_{Ij}\right) \Big/ I$$

Тогда $\mathbf{M} = (m_1\mathbf{1}, \ldots, m_J\mathbf{1})$, где $\mathbf{1}$ – это вектор из единиц размерности I.

Центрирование – это почти обязательная процедура перед применением проекционных методов. Второе простейшее преобразование данных, *нормирование*, не является обязательным. Нормирование, в отличие от центрирования, не меняет структуру данных, а просто изменяет вес различных частей данных при обработке. Чаще всего применяется нормирование по столбцам – это умножение исходной матрицы $\mathbf{X}$ справа на матрицу $\mathbf{W}$, т.е.

$$\tilde{\mathbf{X}} = \mathbf{X}\mathbf{W}$$

Матрица $\mathbf{W}$ – это диагональная матрица размерности $J \times J$. Обычно диагональные элементы w_{jj} равны обратным значениям стандартного отклонения –

$$d_i = \sqrt{\sum_{i=1}^{I} \left(x_{ij} - m_j\right)^2 \Big/ I}$$

вычисленным для каждого столбца $\mathbf{x}_j$. Нормирование по строкам (называемое также нормализацией) – это умножение матрицы $\mathbf{X}$ слева на диагональную матрицу $\mathbf{W}$, т.е.

$$\tilde{\mathbf{X}} = \mathbf{X}\mathbf{W}$$

При этом размерность $\mathbf{W}$ равна $I \times I$, а ее элементы w_{ii} – это обратные значения стандартных отклонений строк $\mathbf{x}_i^{\mathrm{t}}$.

Комбинация центрирования и нормирования по столбцам

$$\tilde{x}_{ij} = \left(x_{ij} - m_j\right)\Big/ d_j$$

называется *автошкалированием*. Нормирование данных часто применяют для того, чтобы уравнять вклад в модель от различных переменных (например, в гибридном методе ЖХ-МС), учесть неравномерные погрешности, или для того, чтобы обрабатывать совместно разные блоки данных. Нормирование также можно рассматривать как метод, позволяющий стабилизировать вычислительные алгоритмы. В тоже время, к этому преобразованию нужно относиться с большой осторожностью, т.к. оно может сильно исказить результаты качественного анализа. Любое преобразование данных – центрирование, шкалирование, и т.п. – всегда делается сначала на обучающем наборе. По этому набору находятся значения m_j и d_j, которые затем применяются и к обучающему, и к проверочному набору

3 Модельные данные

3.1 Принцип линейности

Для иллюстрации и сравнения различных методов калибровки будем использовать модельный пример, в котором “экспериментальные” данные мы создадим сами. Прообразом для этого примера служит популярная задача анализа спектральных данных. Известно, что в таких экспериментах хорошо выполняется принцип линейности. Пусть имеются два вещества A и B, смешанные в концентрациях C_A и C_B. Тогда спектр смеси есть –

$$X = C_A \cdot S_A + C_B \cdot S_B$$

где S_A и S_B – спектры чистых веществ. Заметим, что тот же принцип линейности выполняется и в хроматографии, где в роли “спектров” выступают хроматографические профили чистых компонент смеси.

3.2 “Чистые” спектры

Для моделирования “чистых” спектров $S(\lambda)$ мы использовали гауссовы пики –

$$S(\lambda) = S_0 \exp\left[-\left(\frac{\lambda - m}{\nu}\right)^2 \right]$$

где S_0, m, ν – это параметры спектров, а $\lambda = 0, 1, 2, \ldots, 100$ – это условный номер канала (длина волны, время удерживания, и т.п.). На Рис. 3.1 показаны спектры трех веществ: A, B и C, которые участвуют в примере. Вещества A и B – это искомые величины (отклики), а компонент C добавлен в систему как шум, т.е. нежелательная примесь.

				λ–>							
	S_0	m	v	0	1	2	3	4	5	6	7
A	1	45	40	0.282	0.298	0.315	0.332	0.350	0.368	0.386	0.406
B	0.909	55	50	0.271	0.283	0.296	0.308	0.321	0.334	0.348	0.362
C	1.053	40	15	0.001	0.001	0.002	0.002	0.003	0.005	0.006	0.008

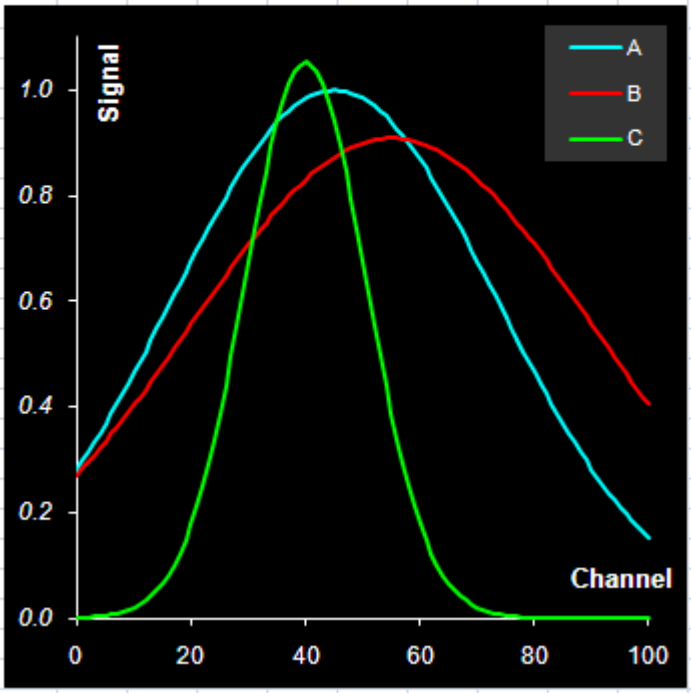


Рис. 3.1. Чистые спектры

Видно, что спектры сильно перекрываются, особенно *A* и *B*, у которых нет участков, где бы присутствовал только один сигнал. Это обстоятельство сильно мешает построению калибровок классическими способами.

Чистые спектры можно изменить, задавая новые величины в ячейках `B4:D6` на Экселевском листе, которые соответствуют параметрам формы гауссовых пиков. Например “растащить” спектры *A* и *B* так чтобы они мало перекрывались и посмотреть, что из этого получится.

3.3 “Стандартные” образцы

Для создания модельных данных мы должны задать концентрации всех компонентов в системе для различных образцов. Будем считать, что у нас имеется 14 образцов, из которых девять – это *обучающий* набор, а пять – это *проверочный* набор. Значения концентраций представлены на Рис. 3.2.

		A	B	C
calibration	1	0.90	0.85	0.15
	2	0.80	0.35	0.25
	3	0.70	0.25	0.10
	4	0.60	0.95	0.15
	5	0.50	0.45	0.20
	6	0.40	0.15	0.15
	7	0.20	0.55	0.10
	8	0.30	0.65	0.25
	9	0.10	0.75	0.20
test	10	0.55	0.70	0.25
	11	0.75	0.50	0.15
	12	0.25	0.30	0.20
	13	0.35	0.80	0.15
	14	1.45	0.20	0.20
mean		0.50	0.55	0.17

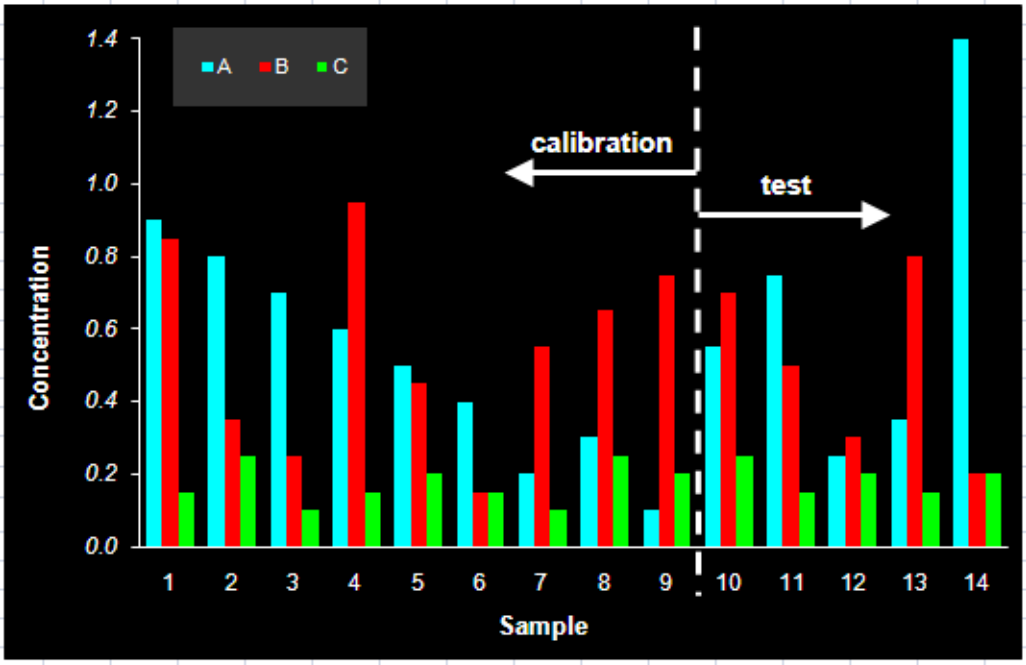


Рис. 3.2. Концентрации в обучающем и проверочном наборах

Заметим, что среди этих образцов нет ни одного “чистого”, т.е. такого в котором все концентрации, кроме одной (например, A) , равны нулю. Это тоже “мешает” использованию традиционных методов калибровки.

3.4 Создание X данных

Для получения “экспериментальных спектров” надо матрицу концентраций **C** умножить на матрицу чистых спектров **S** и добавить к результату случайные ошибки (погрешности) –

$$\mathbf{X} = \mathbf{CS} + \mathbf{E}$$

На Рис. 3.3 представлены полученные спектры в обучающем и проверочном наборах. Погрешности моделировались со стандартным отклонением 0.015.

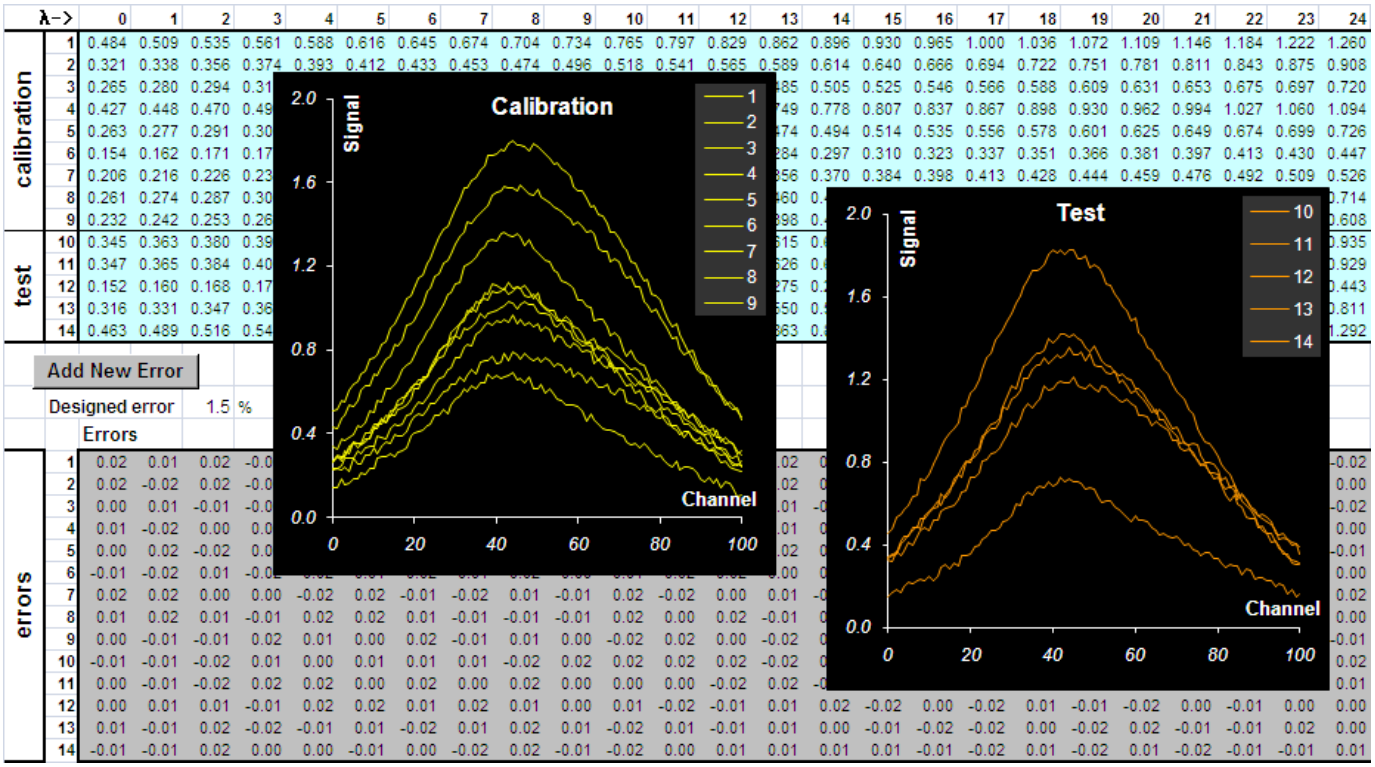


Рис. 3.3. Создание модельных данных

Генерация случайных ошибок в книге *Calibration.xlsx* производится с помощью простого VBA макроса, который запускается кнопкой `Add New Error`. Величина желаемой погрешности (СКО) предварительно задается в ячейке `E18` с именем `RMSE`, озаглавленной `Designed error`. Погрешность, которая получилась в результате генерации случайных чисел выводится в ячейку `J18`, озаглавленную `Obtained error`.

Генерируя новые данные можно узнать много интересного о методах калибровки.

3.5 Центрирование данных

В соответствие с концепцией, изложенной в предыдущей главе, данные нужно правильно подготовить. В рассматриваемом модельном примере нет необходимости в выравнивании переменных – все спектральные каналы имеют схожие величины сигналов. А вот центрирование, как спектров **X**, так и откликов **Y** , бывает необходимо для построения некоторых моделей.

Для центрирования концентраций **Y** вычисляются средние по обучающему набору. Эти средние затем вычитаются из всех значений **Y**. Аналогично проводится и центрирование в данных **X** – вычисляются средние значения для каждого канала по обучающему набору и затем эти значения вычитаются из всех величин **X** – по столбцам.

3.6 Обзор данных

Итак, мы построили модельные данные (**Y**, **X**): концентрации – матрицу **Y** размером (14 × 2) и спектры – матрицу **X** размером (14 × 101). Исследуя эти данные, мы "забудем" (т.е. не будем использовать в расчетах) то, что в системе присутствует еще одно "скрытое" вещество C. Интересно, сможем ли мы обнаружить его присутствие? Кроме того, не будут использоваться и спектры чистых веществ A и B, примененные для построения данных. Мы постараемся их восстановить и сравнить с исходными спектрами.

Все данные мы разделили на два блока: обучающий (или калибровочный) – 9 образцов, и проверочный (или тестовый) – 5 образцов. Мы будем строить калибровки с помощью разных методов, используя только первый, обучающий набор. Второй, проверочный набор послужит для оценки качества получаемых моделей.

Данные используются для иллюстрации и сравнения различных методов калибровки. Они размещены в рабочей книге Excel с именем Calibration.xls. Эта книга включает в себя следующие листы:

- `Intro`: краткое введение
- `Layout`: схемы, объясняющая имена массивов, используемых в примере
- `Gun`: иллюстрация пере - и недооценки в калибровке
- `Multicollinearity`: иллюстрация проблемы мультиколлинеарности
- `Pure Spectra`: истинные чистые спектры S
- `Concentrations`: истинные концентрационные профили C
- `Data`: модельные данные, используемые в примере.
- `UVR`: одноканальная калибровка
- `Vierordt`: калибровка методом Фирордта
- `Indirect`: непрямая калибровка
- `MLR`: множественная линейная регрессия
- `SWR`: калибровка пошаговой реегрессией
- `PCR`: регрессия на главные компоненты
- `PLS1`: метод проекция на латентные переменные 1
- `PLS2`: метод проекция на латентные переменные 2
- `Compare` сравнение различных методов

4 Классическая калибровка

4.1 Введение

Классическая калибровка опирается на использовании того же принципа линейности, по которому строились модельные данные:

$$\mathbf{X} = \mathbf{CS}$$

Здесь $\mathbf{X} = \mathbf{X}_c$ и $\mathbf{C} = \mathbf{Y}_c$ – это обучающая часть исходных данных (не центрированных), а $\mathbf{S}$ – это матрица "чистых спектров" (она же матрица чувствительности).

Если матрица $\mathbf{S}$ известна априори, то концентрации определяются так

$$\mathbf{C} = \mathbf{XS}^{+}$$

где $\mathbf{S}^{+} = \mathbf{S}^{\mathrm{t}}(\mathbf{SS}^{\mathrm{t}})^{-1}$ – это псевдообратная матрица. Этот случай называется *прямой* калибровкой. Однако, на практике матрица чистых спектров, как правило, неизвестна и ее приходится восстанавливать из обучающих данных.

4.2 Калибровка по одному каналу

Это самый простой, наивный вид калибровки. Если для каждого аналита из всех данных $\mathbf{X}$ выделить один канал (длину волны), то получится несколько векторов $\mathbf{x}$. Тогда, используя обучающие данные, можно построить для каждого вещества простейшую одномерную регрессию –

$$x = sc + b$$

одну для вещества A ($C = C_A$), а другую для B ($C = C_B$). Формулы для оценивания коэффициентов s (наклон)и b (отсечение) можно найти в любом пособии по линейной регрессии, например, здесь. В Excel это проще всего сделать с помощью функции `ЛИНЕЙН` (`LINEST`). После того, как эти коэффициенты найдены, значения концентраций можно вычислять по уравнению

$$c = (x - b)/s$$

На Рис. 4.1 приведен пример построения двух этих регрессий для каналов 31 (A) и 98 (B). График показывает, как проходят линии регрессий для веществ A и B.

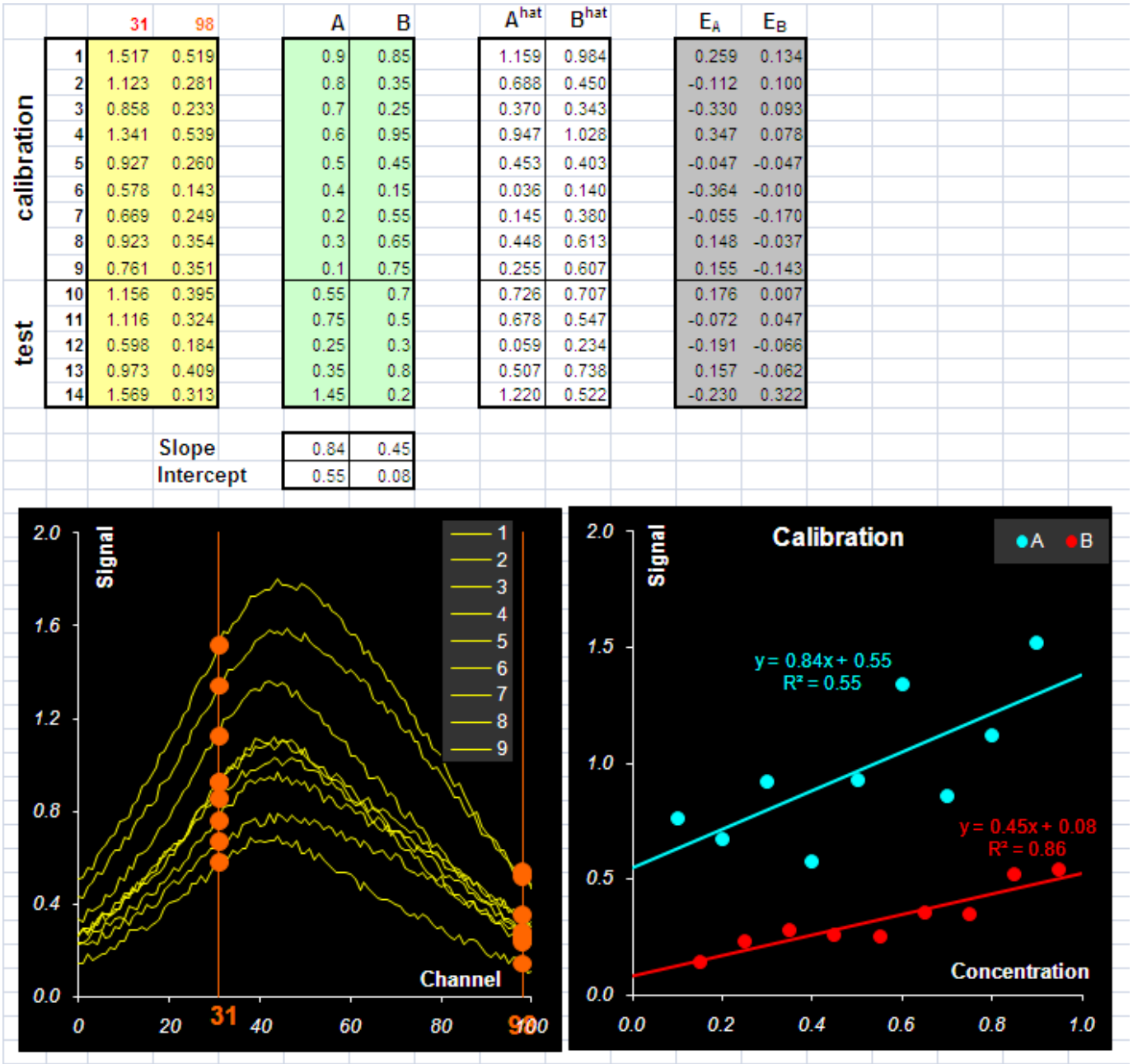


Рис. 4.1. Калибровка по одному каналу

Обратите внимание, что в каждом калибровочном уравнении присутствует свободный член b. Это

противоречит исходному уравнению, но зато позволяет учесть влияние фона, или, как в нашем случае, присутствие посторонней примеси (вещества C). Величины коэффициентов s (наклон) соответствуют значениям “чистых” спектров. На Рис. 4.2 показаны “чистые” спектры веществ A и B и соответствующие значения коэффициентов m.

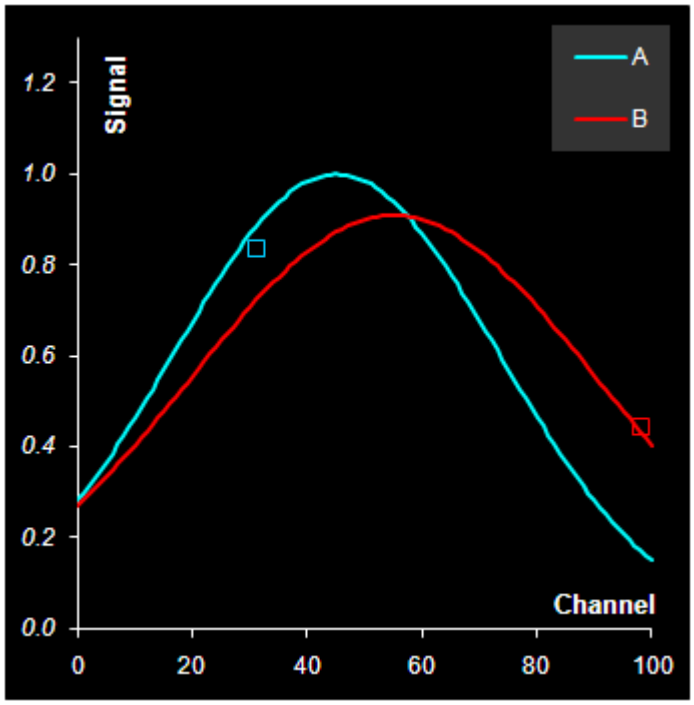


Рис. 4.2. “Чистые” спектры и их оценки для выбранных каналов

Для визуальной оценки качества калибровки традиционно используются графики “измерено-предсказано”, в которых по оси абсцисс откладываются известные (стандартные) концентрации y, а по оси ординат соответствующие им величины оценок $\hat{y}$, найденные (предсказанные) с помощью построенной калибровки. Такие графики показаны на Рис. 4.3. Помимо вышеупомянутых данных (точки), на таких графиках обычно показывают линии регрессии $ky + b$. Коэффициенты k (наклон) и b (отсечение) характеризуют “качество” калибровки. В идеальном случае $k = 1$ и $b = 0$. Чем дальше от идеала, тем хуже калибровка. Кроме того, на таких графиках обычно приводится значение коэффициента корреляции R^2 между y и $\hat{y}$.

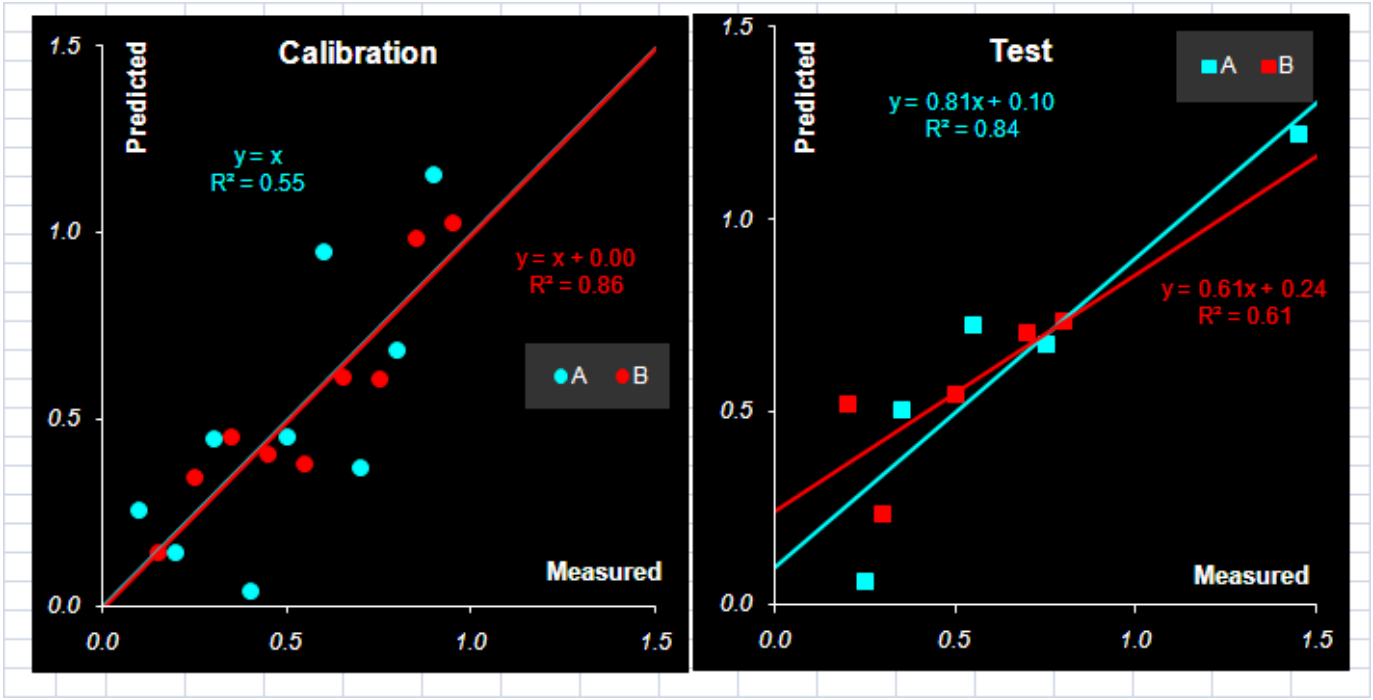


Рис. 4.3. Графики “измерено-предсказано” для однофакторной калибровки. Обучающий и проверочный наборы”

В таблице на Рис. 4.4 приведены характеристики качества однофакторной калибровки веществ A и B, вычисленные в соответствие с формулами раздела 1.4.

A		B	
R_c^2=	0.549	R_c^2=	0.862
RMSEC=	0.234	RMSEC=	0.103
BIASC=	0.000	BIASC=	0.000
SEC=	0.234	SEC=	0.103
R_t^2=	0.842	R_t^2=	0.609
RMSEP=	0.173	RMSEP=	0.151
BIASP=	-0.032	BIASP=	0.050
SEP=	0.170	SEP=	0.151
TRVC= 0.033			
ERVC= 0.905			
TRVP= 0.026			
ERVP= 0.994			

Рис. 4.4. Характеристики качества однофакторной калибровки.

Разумеется одномерную калибровку можно проводить по любому каналу. На листе `UVR` для этого достаточно изменить значения в ячейках `C2` и `D2`, которые соответствуют двум каналам – для веществ A и B. и проверить как меняется качество калибровки. Так наилучшее значение $RMSEP = 0.173$ для вещества A достигается при калибровке по 31 каналу, а наихудшее значение $RMSEP = 0.702$ получается при $\lambda = 97$. Соответственно для

вещества B имеем: наилучшее значение $RMSEP = 0.151$ для $\lambda = 98$ и наихудшее $RMSEP = 0.611$ для $\lambda = 20$.

Существует мнение, что калибровку лучше проводить в точке, где соответствующий “чистый спектр” имеет максимум. Из рассмотренного примера видно, что это не так. В точке $\lambda = 45$ (максимум A) $RMSEP = 0.205$, а в точке $\lambda = 55$ (максимум B) $RMSEP = 0.484$.

4.3 Метод Фирордта

Впервые задача анализа спектрофотометрических данных была рассмотрена в работе Карла фон Фирордта (Karl von Vierordt) в 1873 г. Исследуя изменения гемоглобина в крови, он предложил метод калибровки, который и теперь, спустя 140 лет, еще очень популярен у некоторых аналитиков.

Метод Фирордта похож на метод одноканальной калибровки. В нем также выбирается по одному каналу для каждого вещества, но регрессионные уравнения рассматриваются не независимо, а совместно

$$\tilde{\mathbf{X}}_c = \mathbf{Y}_c\mathbf{S}$$

Здесь матрица $\tilde{\mathbf{X}}_c$ имеет столько столбцов, сколько определяется веществ (у нас 2), матрица $\mathbf{Y}_c$ – это известные значения концентраций в обучающем наборе (у нас 2), а матрица $\mathbf{S}$ – это неизвестная квадратная (у нас 2×2) матрица чувствительности, которую нужно оценить. Заметим, что в традиционном методе Фирордта свободный член отсутствует в полном согласии с уравнением $\mathbf{X} = \mathbf{CS} + \mathbf{E}$,

Для оценки матрицы чувствительности $\mathbf{S}$ умножим это уравнение слева на транспонированную матрицу концентраций $\mathbf{Y}_c$:

$$\mathbf{Y}_c^t\tilde{\mathbf{X}}_c = \mathbf{Y}_c^t\mathbf{Y}_c\mathbf{S}$$

Тогда матрица –

$$\hat{\mathbf{S}} = \left(\mathbf{Y}_c^t\mathbf{Y}_c\right)^{-1}\mathbf{Y}_c^t\tilde{\mathbf{X}}_c$$

будет оценкой матрицы чувствительности.

На Рис. 4.5 показано применение метода Фирордта для модельного примера, где для калибровки выбраны два канала 29 и 100.

		29	100	A	B	A^{hat}	B^{hat}	E_A	E_B	X^{hat}	
calibration	1	1.424	0.465	0.9	0.85	0.795	0.867	-0.105	0.017	1.506	0.473
	2	1.073	0.238	0.8	0.35	0.950	0.254	0.150	-0.096	1.013	0.256
	3	0.827	0.223	0.7	0.25	0.608	0.337	-0.092	0.087	0.841	0.201
	4	1.260	0.480	0.6	0.95	0.490	1.010	-0.110	0.060	1.312	0.471
	5	0.855	0.272	0.5	0.45	0.501	0.494	0.001	0.044	0.819	0.254
	6	0.534	0.104	0.4	0.15	0.517	0.076	0.117	-0.074	0.486	0.118
	7	0.637	0.250	0.2	0.55	0.224	0.537	0.024	-0.013	0.625	0.252
	8	0.885	0.298	0.3	0.65	0.467	0.569	0.167	-0.081	0.797	0.307
	9	0.715	0.319	0.1	0.75	0.133	0.738	0.033	-0.012	0.694	0.319
test	10	1.116	0.354	0.55	0.7	0.656	0.642	0.106	-0.058	1.066	0.363
	11	1.053	0.315	0.75	0.5	0.678	0.539	-0.072	0.039	1.088	0.310
	12	0.520	0.166	0.25	0.3	0.303	0.302	0.053	0.002	0.470	0.157
	13	0.935	0.392	0.35	0.8	0.253	0.875	-0.097	0.075	0.963	0.375
	14	1.486	0.303	1.45	0.2	1.399	0.257	-0.051	0.057	1.487	0.287

$(Y_c'Y_c)^{-1}$= 0.965 -0.713 / -0.713 0.828

S^{hat}= $(Y_c'Y_c)^{-1}Y'X_c$= 0.915 0.142 / 0.804 0.407

$V=(S^{hat})^{-1}$= 1.576 -0.550 / -3.114 3.546

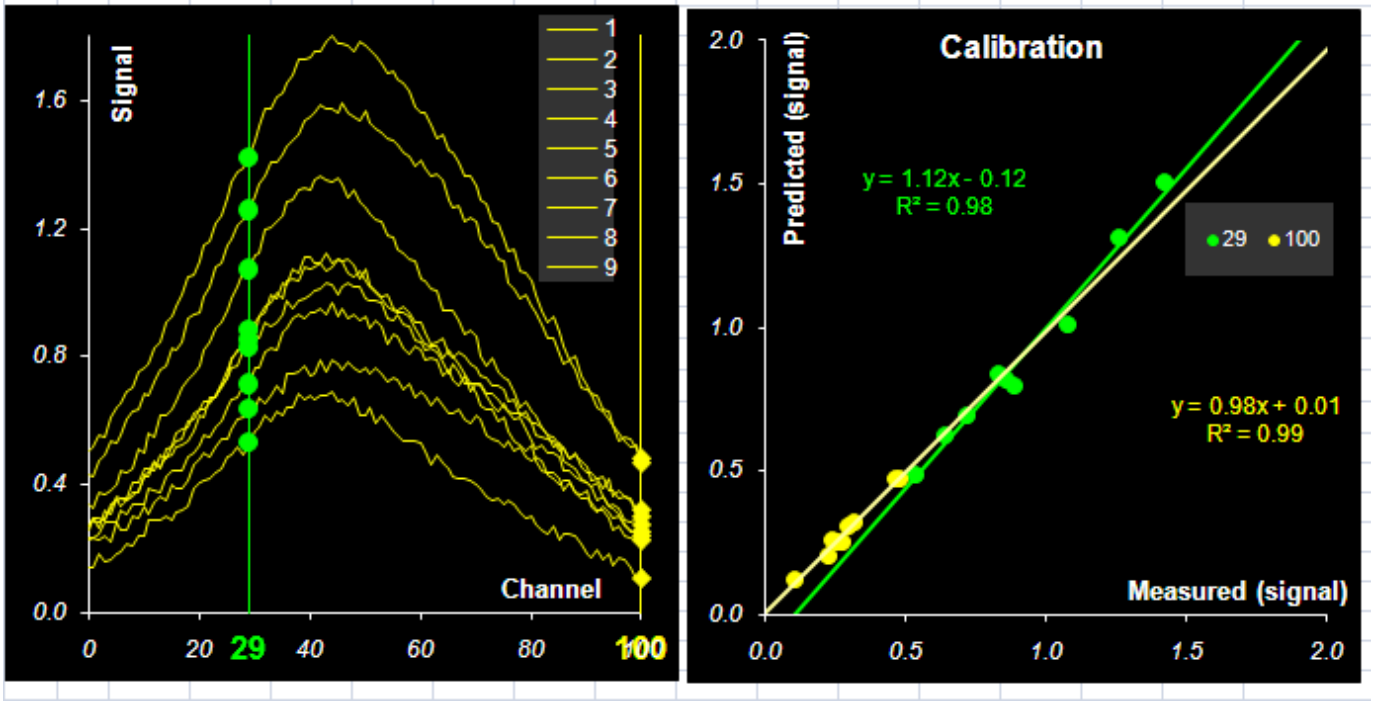


Рис. 4.5. Применение метода Фирордта

График “предсказано-измерено” представляет результаты калибровки сигнала (спектров **x**) для выбранных каналов. Элементы матрицы чувствительности **S** соответствуют значениям матрицы “чистых” спектров. На Рис. 4.6 показано это соответствие.

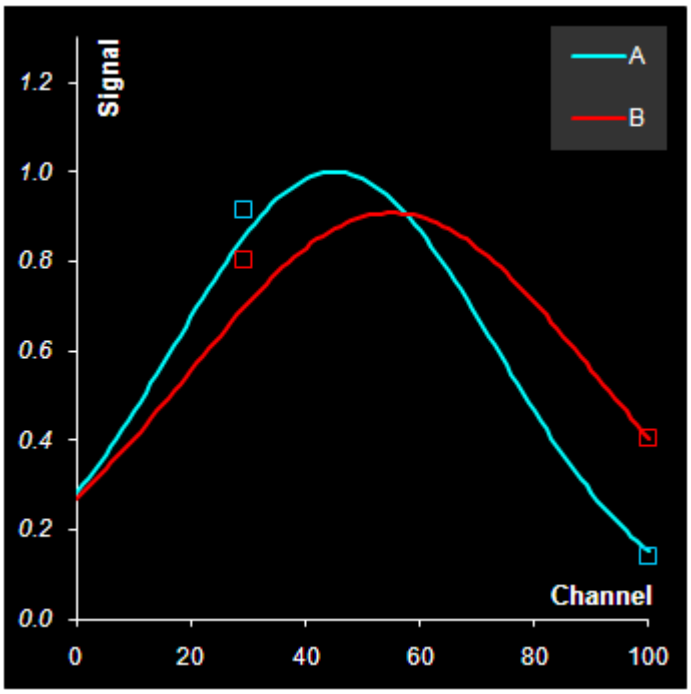


Рис. 4.6. Элементы матрицы чувствительности **S** (точки) в методе Фирордта и “чистые” спектры

Для оценки и предсказания значений концентрации веществ *A* и *B*. нужно вычислить матрицу обратную к матрице чувствительности **S**. Сделать это просто, т.к. эта матрица квадратная. Тогда матрица **V**:

$$\mathbf{V} = \hat{\mathbf{S}}^{-1}$$

будет линейным “оценивателем” матрицы концентраций **Y**. Для того, чтобы найти оценки величин концентраций в обучающем наборе, надо матрицу **V** умножить на матрицу соответствующих спектров $\tilde{\mathbf{X}}_c$ –

$$\hat{\mathbf{Y}}_c = \tilde{\mathbf{X}}_c \mathbf{V}$$

Аналогично, для оценки концентраций в проверочном наборе, матрица **V** умножается на $\mathbf{X}_t$ –

$$\hat{\mathbf{Y}}_t = \tilde{\mathbf{X}}_t \mathbf{V}$$

Графики “предсказано-измерено” показаны на Рис. 4.7.

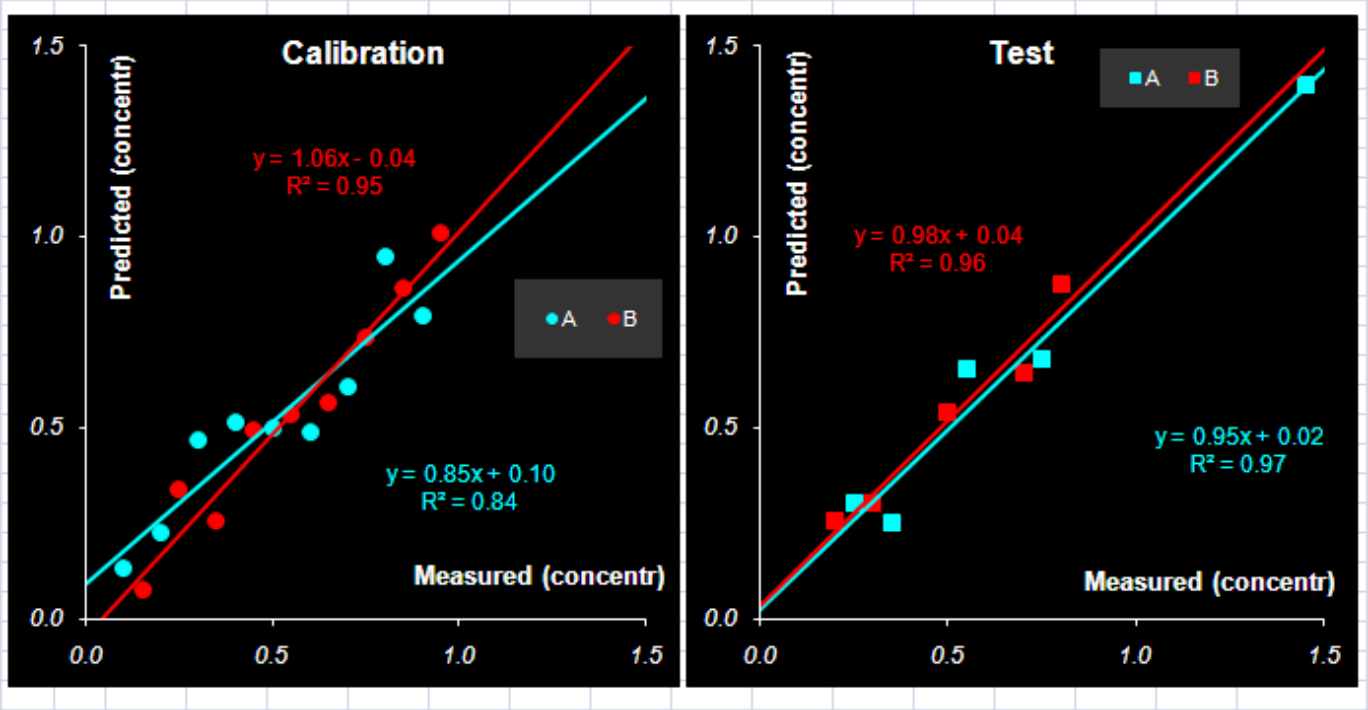


Рис. 4.7. Графики “измерено-предсказано” для метода Фирордта. Обучающий и проверочный наборы.

В таблице на Рис. 4.8 приведены характеристики качества калибровки по методу Фирордта веществ A и B, вычисленные в соответствие с формулами главы 2.

A	B
R_c^2= 0.844	R_c^2= 0.954
RMSEC= 0.104	RMSEC= 0.062
BIASC= 0.020	BIASC= -0.007
SEC= 0.102	SEC= 0.062
R_t^2= 0.967	R_t^2= 0.957
RMSEP= 0.079	RMSEP= 0.053
BIASP= -0.012	BIASP= 0.023
SEP= 0.078	SEP= 0.047
TRVC= 0.007	
ERVC= 0.979	
TRVP= 0.005	
ERVP= 0.999	

Рис. 4.8. Характеристики качества калибровки по методу Фирордта

Качество калибровки сильно зависит от того, какие каналы выбраны в качестве аналитических. Часто рекомендуется действовать таким образом. Строится график $F(\lambda) = S_A/S_B$, где S_A и S_B – это “чистые спектры”. На нем определяют такие точки λ_1 и λ_2, в которых разница $|F(\lambda_1)-F(\lambda_2)|$ максимальна. Способ правильный, но построить график $F(\lambda)$, не зная чистых спектров, невозможно.

5 Обратная калибровка

5.1 Введение

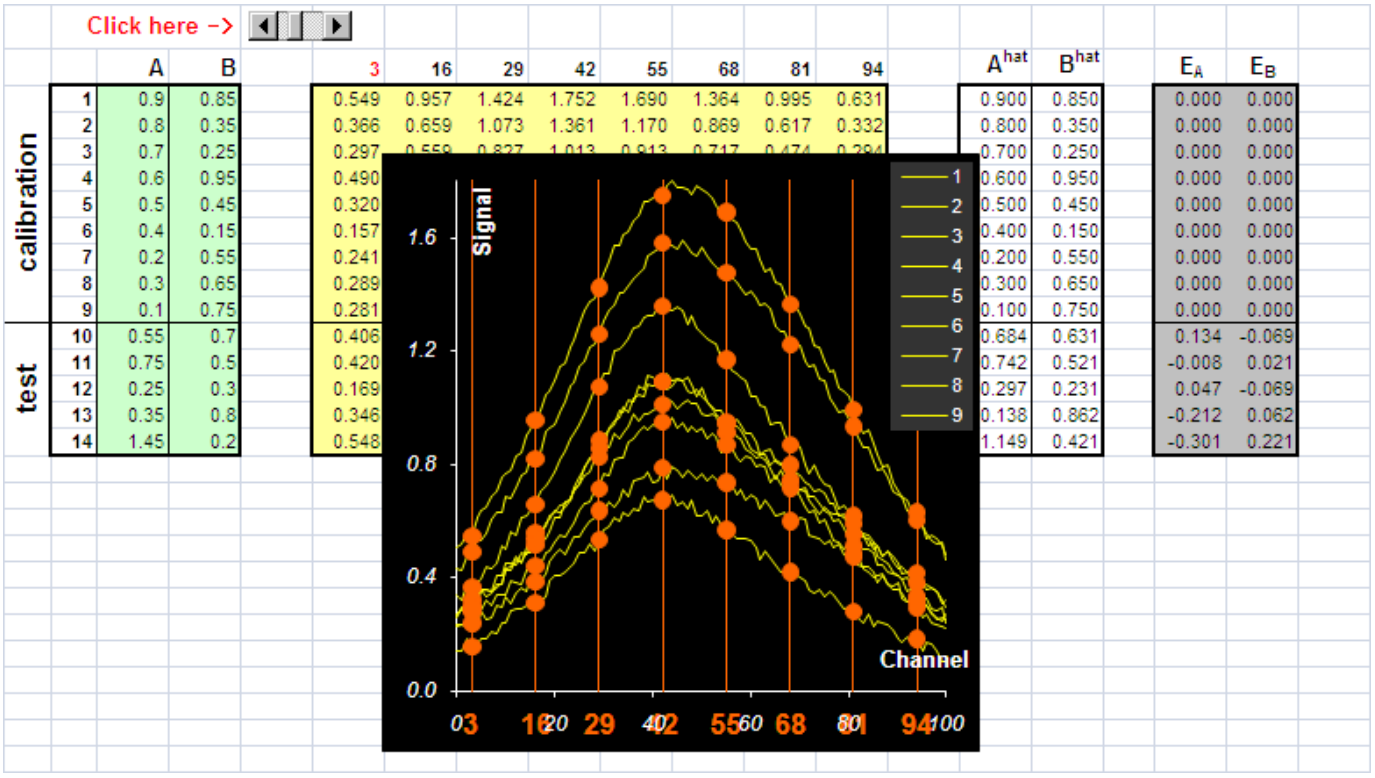
В обратной калибровке основное уравнение имеет вид:

$$\mathbf{Y} = \mathbf{XB}$$

в котором. искомая величина **Y** (концентрация) прямо выражается через известную матрицу спектров **X**. Хотя такое представление калибровочного уравнения и противоречит основному соотношению $\mathbf{X} = \mathbf{CS}$, такой подход обеспечивает лучшее качество моделирования.

5.2 Множественная калибровка

Простейшим вариантом обратной калибровки является *множественная линейная регрессия* (MLR). В главе 3 мы уже обсуждали свойства MLR в связи с проблемой мультиколлинеарности. В частности отмечалось, что во множественной регрессии число переменных должно быть меньше числа образцов. В нашем модельном примере число калибровочных образцов равно 9, поэтому для использования MLR необходимо отобрать из 101 канала только 8 и по ним строить калибровку. Больше переменных взять нельзя, но можно меньше. На листе `MLR` есть активный элемент, с помощью которого можно быстро сменить первый канал; остальные изменятся автоматически.



	A	B	3	16	29	42	55	68	81	94	A^{hat}	B^{hat}	E_A	E_B
1	0.9	0.85	0.549	0.957	1.424	1.752	1.690	1.364	0.995	0.631	0.900	0.850	0.000	0.000
2	0.8	0.35	0.366	0.659	1.073	1.361	1.170	0.869	0.617	0.332	0.800	0.350	0.000	0.000
3	0.7	0.25	0.297	0.559	0.827	1.013	0.813	0.717	0.474	0.284	0.700	0.250	0.000	0.000
4	0.6	0.95	0.490								0.600	0.950	0.000	0.000
5	0.5	0.45	0.320								0.500	0.450	0.000	0.000
6	0.4	0.15	0.157								0.400	0.150	0.000	0.000
7	0.2	0.55	0.241								0.200	0.550	0.000	0.000
8	0.3	0.65	0.289								0.300	0.650	0.000	0.000
9	0.1	0.75	0.281								0.100	0.750	0.000	0.000
10	0.55	0.7	0.406								0.684	0.631	0.134	-0.069
11	0.75	0.5	0.420								0.742	0.521	-0.008	0.021
12	0.25	0.3	0.169								0.297	0.231	0.047	-0.069
13	0.35	0.8	0.346								0.138	0.862	-0.212	0.062
14	1.45	0.2	0.548								1.149	0.421	-0.301	0.221

Рис. 5.1. Множественная линейная калибровка

На Рис. 5.1 показано как отбираются эти каналы – равномерно, с шагом 13. Первый канал можно выбрать произвольно: от 0 до 9, тогда все последующие определяются однозначно. В результате этого отбора получается матрица независимых переменных **X** размерностью (14 × 8), состоящая из двух частей: обучающего набора (9 × 8) и проверочного (5 × 8). Используя обучающий набор переменных можно построить множественную регрессию: между **X** и **Y**. Для этого можно применить формулы из раздела про мульиколлинеарность, но проще воспользоваться функцией Excel `ТЕНДЕНЦИЯ`.

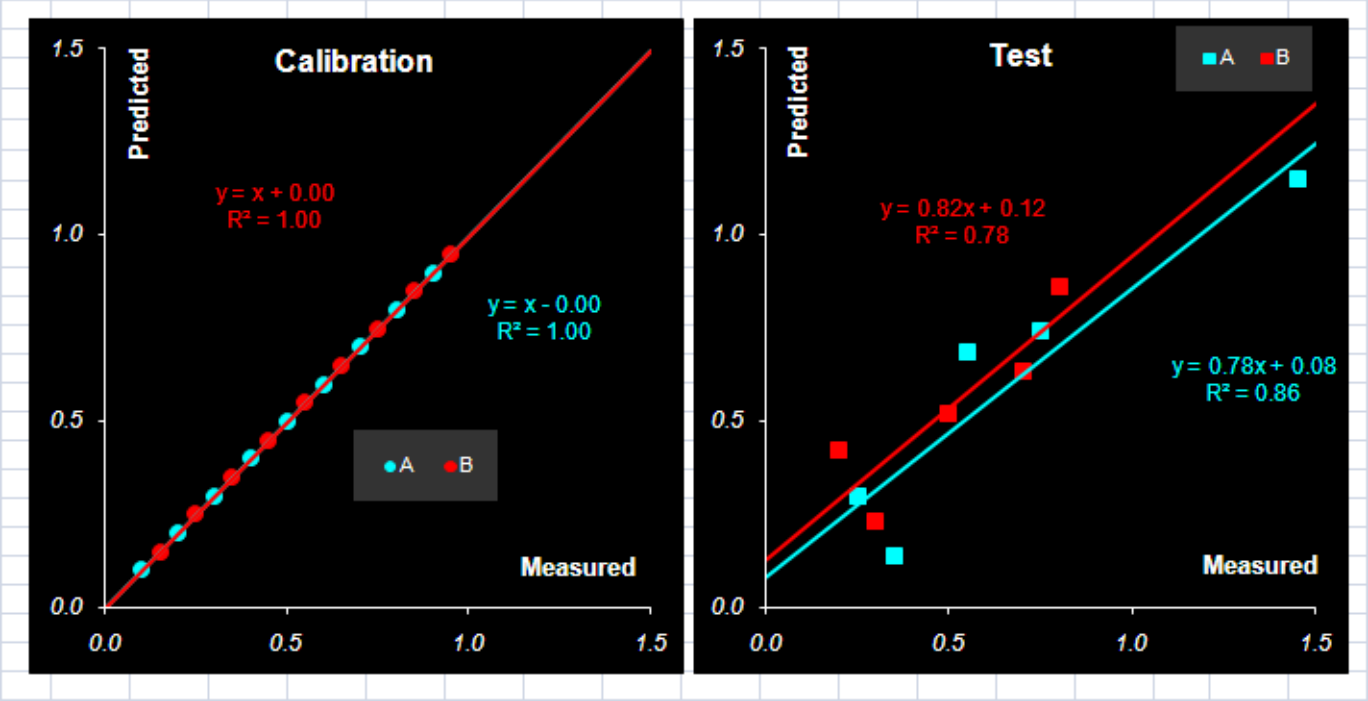


Рис. 5.2. Графики “измерено-предсказано” для множественной калибровки. Обучающий (a) и проверочный (b) наборы

На Рис. 5.2 показаны графики “измерено-предсказано” для множественной калибровки. Видно, что обучающий набор “слишком хорошо” описывается моделью. А вот проверка неудовлетворительна. Здесь заметно, и смещение, и малая корреляция.

В таблице на Рис. 5.3 приведены характеристики качества множественной калибровки веществ *A* и *B*.

A	B
R_c^2= 1.000	R_c^2= 1.000
RMSEC= 0.000	RMSEC= 0.000
BIASC= 0.000	BIASC= 0.000
SEC= 0.000	SEC= 0.000
R_t^2= 0.863	R_t^2= 0.782
RMSEP= 0.176	RMSEP= 0.112
BIASP= -0.068	BIASP= 0.033
SEP= 0.163	SEP= 0.107
TRVC= 0.000	
ERVC= 1.000	
TRVP= 0.022	
ERVP= 0.995	

Рис. 5.3. Характеристики качества множественной линейной калибровки

Видно, что в этом случае мы получили типичную переоценку модели (см. раздел 2.6) – число отобранных

переменных слишком велико. Попытки сменить набор переменных ситуацию не улучшают. Таким образом множественная калибровка является неприемлемым методом. Она приводит к переоценке модели и дает неудовлетворительные результаты при использовании на новом (проверочном) наборе образцов.

5.3 Пошаговая калибровка

Как мы только что видели множественная линейная калибровка неудовлетворительна – она представляет явный пример переоценки. В этом разделе мы рассмотрим пошаговую калибровку (stepwise regression, SWR), в которой отбор переменных является способом справится с переоценкой. Идея метода состоит в следующем.

Пусть имеется калибровочная модель, построенная по M отобранным каналам. Добавим к ним еще один $M + 1$-ый канал. Выбор этого дополнительного канала основан на простом принципе – добавляется тот, который дает минимум величины *RMSEC*. Добавление новых каналов продолжается до тех пор, пока не наступает риск переоценки, т.е. до начала роста величины *RMSEP*.

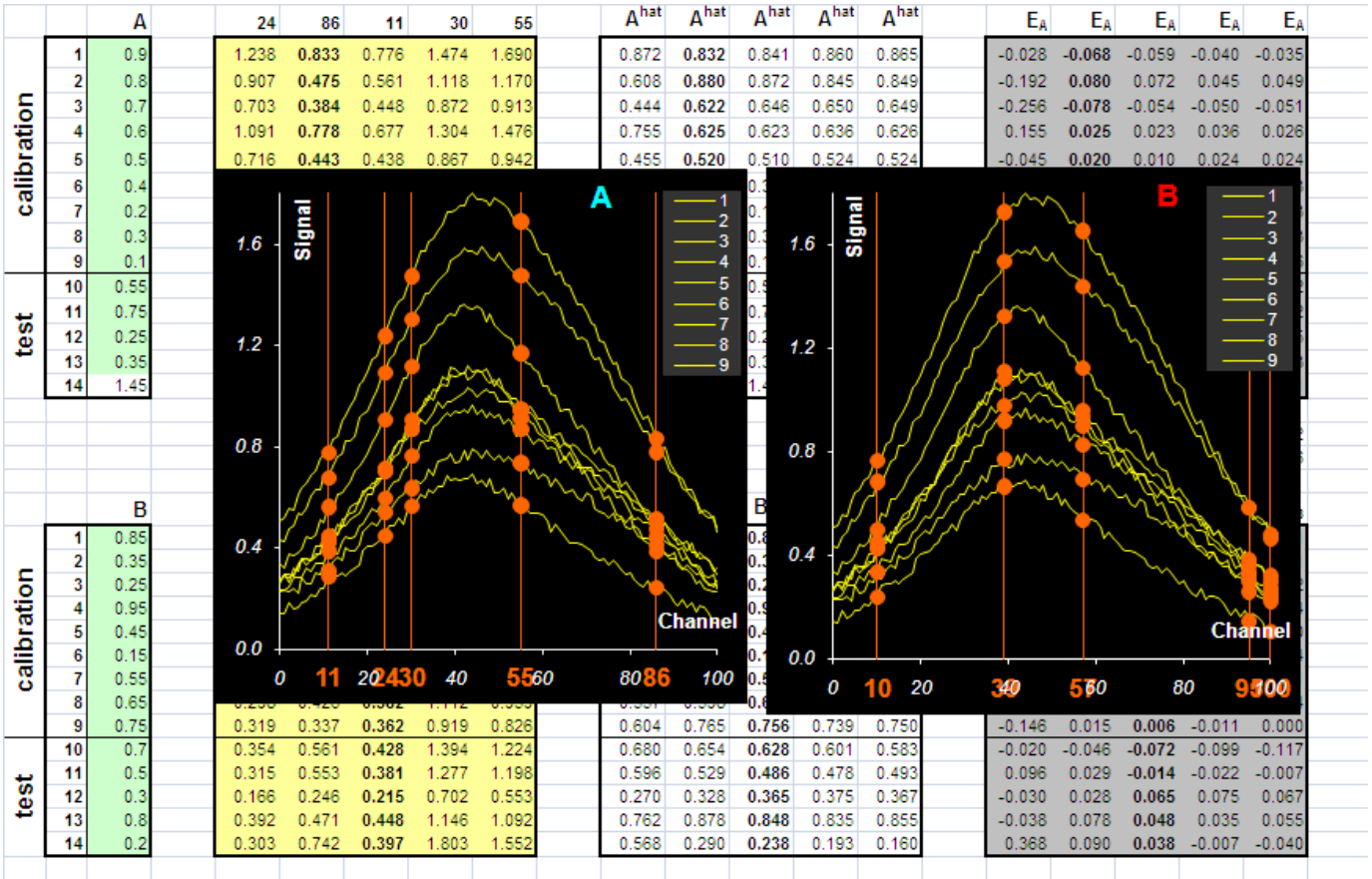


Рис. 5.4. Пошаговая калибровка

Очевидно, что наилучший результат для веществ A и B достигается для разных каналов. Поэтому

“оптимальные” наборы для A и B отличаются. Для A – это каналы $24, 86, 11, 30, \ldots$, а для B – это каналы $100, 10, 95, 39, 57,$ Именно в таком порядке каналы добавляются в соответствующие наборы. Отбор этих каналов – простая, но трудоемкая процедура, которую можно упростить, написав небольшой макрос в Excel.

В пошаговой регрессии существует много способов отбора “оптимальных” переменных. Тот, который использован здесь, самый простой – выбирать тот канал, на котором достигается минимум среднеквадратичной ошибки в обучении, *RMSEC*.

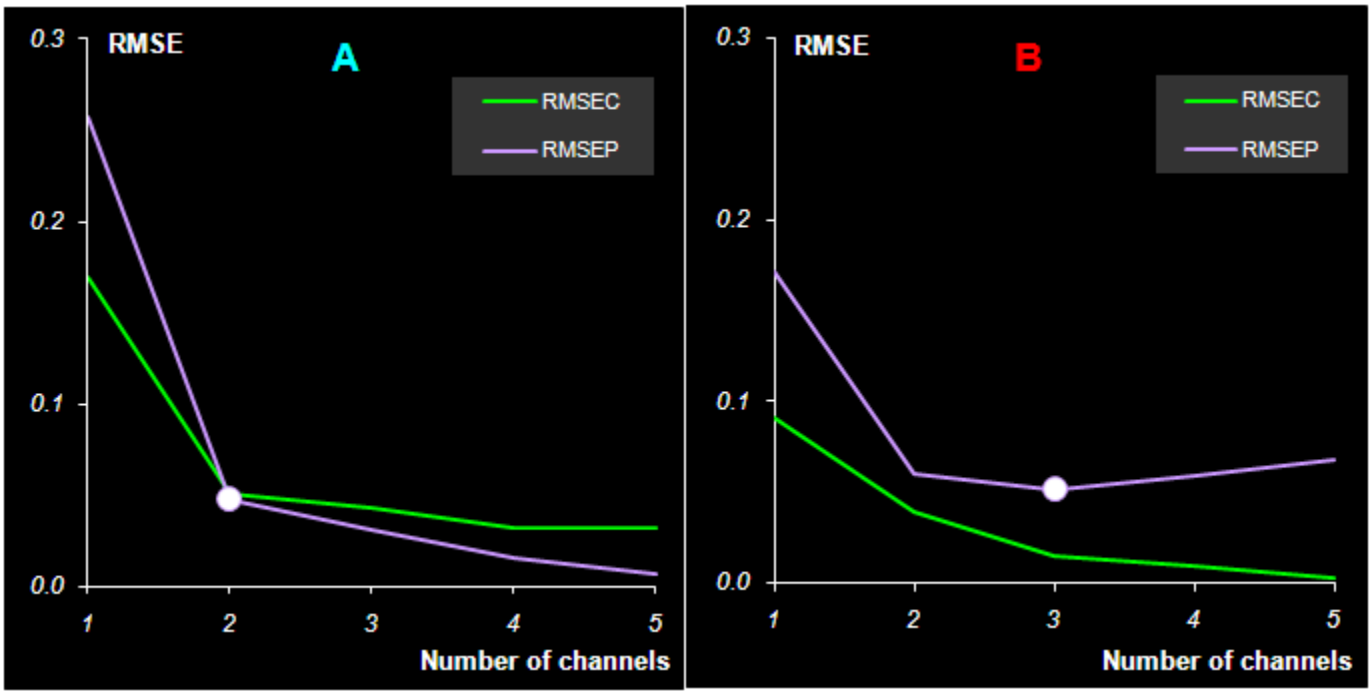


Рис. 5.5. Среднеквадратичные остатки обучения (RMSEC) и проверки (RMSEP) в пошаговой калибровке

На Рис. 5.5 показано, как изменяются среднеквадратичные остатки в обучении (*RMSEC*) и в проверке (*RMSEP*) при увеличении числа каналов в SWR. В соответствие с принципом минимума *RMSEP*, оптимальное число каналов для вещества B – три. Это четко видно на графике. А вот выбор числа каналов для вещества A затруднителен. На соответствующем графике кривая *RMSEP* не имеет минимума. Так часто случается при анализе сложных данных. В рассматриваемом примере оптимальные каналы для вещества A располагаются по краям “спектральной” области – там, где влияние скрытой примеси C не существенно. Сравните Рис. 3.1 и Рис. 5.4. Поэтому SWR калибровка для вещества A никак не может “заметить” наличие вещества C. В таком сомнительном случае следует выбирать точку излома на графике *RMSEP*. Именно поэтому мы выбираем только два канала для вещества A.

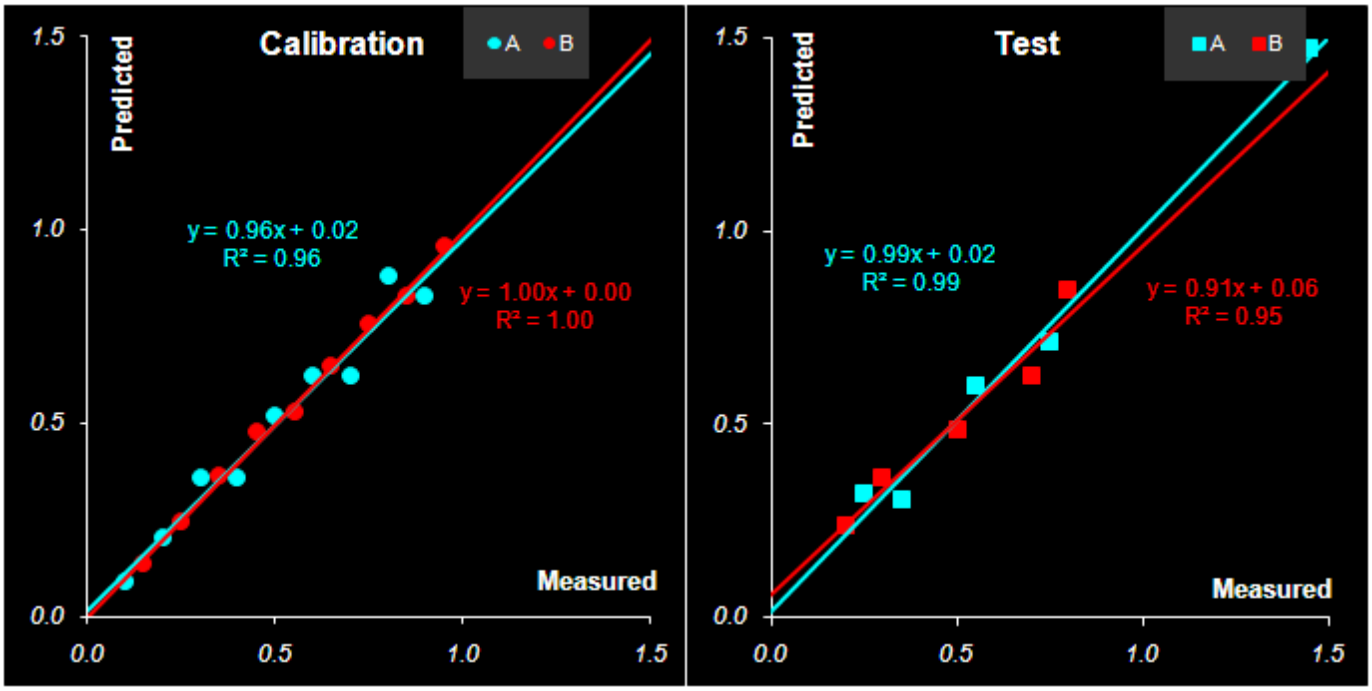


Рис. 5.6. Графики “измерено-предсказано” для пошаговой калибровки. Обучающий и проверочный наборы.

На Рис. 5.6 показаны графики “измерено-предсказано” для пошаговой калибровки. Здесь заметно, что описание сбалансировано уже гораздо лучше – отличие точности обучения от проверки не так существенно, как во множественной калибровке. В таблице на Рис. 5.7 приведены характеристики пошаговой калибровки веществ *A* и *B*.

A	B
R_c^2= 0.961	R_c^2= 0.996
RMSEC= 0.051	RMSEC= 0.015
BIASC= 0.000	BIASC= 0.000
SEC= 0.051	SEC= 0.015
R_t^2= 0.989	R_t^2= 0.954
RMSEP= 0.031	RMSEP= 0.052
BIASP= 0.013	BIASP= 0.013
SEP= 0.046	SEP= 0.050
TRVC= 0.001	
ERVC= 0.996	
TRVP= 0.002	
ERVP= 0.995	

Рис. 5.7. Характеристики качества пошаговой калибровки

Подводя итог можно заметить, что пошаговая регрессия дала наилучший результат среди всех исследованных нами методов калибровки. Но есть и более точные методы.

6 Калибровка на латентных переменных

6.1 Проекционные методы

В методе пошаговой калибровки использовались только два или три (наиболее информативных) канала из 101, и это позволило достичь приемлемого результата. Прочие каналы были отброшены как “излишние”, не нужные. Такой подход представляется несколько расточительным – ведь отброшенные данные содержали полезную информацию, которую хорошо было бы использовать. В частности, применяя SWR, мы не заметили присутствие третьего вещества – примеси C, которая сильно проявляется как раз в этих исключенных каналах. Проблема, которую мы сейчас будет исследовать, состоит в следующем. Как использовать все имеющиеся данные (каналы), и, в то же время, избежать переоценки и мультколлинеарности, неизбежных при большом числе регрессионных переменных.

Решение этой дилеммы было предложено К. Пирсоном (Karl Pearson) в 1901 году – надо использовать новые, латентные переменные $\mathbf{t}_a$, $(a = 1, \dots, A)$, являющиеся линейной комбинацией исходных переменных $\mathbf{x}_j$ $(j = 1, \dots, J)$ –

$$\mathbf{t}_a = \mathbf{p}_{a1}\mathbf{x}_1 + \cdots + \mathbf{p}_{aJ}\mathbf{x}_J$$

или в матричном виде

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{t}} + \mathbf{E} = \sum_{a=1}^{A} \mathbf{t}_a \mathbf{p}_a^{\mathrm{t}} + \mathbf{E}$$

В этом уравнении $\mathbf{T}$ называется матрицей счетов (scores). Ее размерность – $(I \times A)$. Матрица $\mathbf{P}$ называется матрицей нагрузок (loadings). Ее размерность $(A \times J)$. $\mathbf{E}$ – это матрицей остатков, размерностью $(I \times J)$.

Новые латентные переменные $\mathbf{t}_a$ называются главными компонентами (Principal Components), поэтому и сам метод называется *методом главных компонент* (PCA). Число столбцов – $\mathbf{t}_a$ в матрице $\mathbf{T}$, и $\mathbf{p}_a$ в матрице $\mathbf{P}$ – равно эффективному (химическому) рангу матрицы $\mathbf{X}$. Эта величина обозначается A называется числом главных компонент (PC). Она заведомо меньше числа переменных J и числа образцов I.

Для иллюстрации метода PCA, мы опять вернемся к тому, как строились модельные данные. Матрица спектров смесей $\mathbf{X}$ равна произведению матрицы концентраций $\mathbf{C}$ и матрицы спектров чистых

компонентов **S**. Число строк в матрице **X** равно числу образцов (I), и каждая ее строка соответствует спектру одного образца, снятому для J длин волн. Число строк в матрице **C** также равно I, а вот число столбцов соответствует числу компонентов в смеси ($A = 3$). У матрицы чистых спектров **S** число строк равно числу каналов (длин волн) J, а число столбцов равно A. При анализе экспериментальных данных **X**, отягощенных погрешностями, представленными матрицей **E**, эффективный ранг A может не совпадать с реальным числом компонентов в смеси.

Метод главных компонент часто применяется при исследовательском анализе химических данных. В общем случае матрицы счетов **T** и нагрузок **P** уже нельзя интерпретировать как спектры и концентрации, а *число главных компонент* A – как число химических компонентов, присутствующих в исследуемой системе. Тем не менее, даже формальный анализ счетов и нагрузок оказывается очень полезным для понимания устройства данных. Дадим простейшую двумерную иллюстрацию метода PCA.

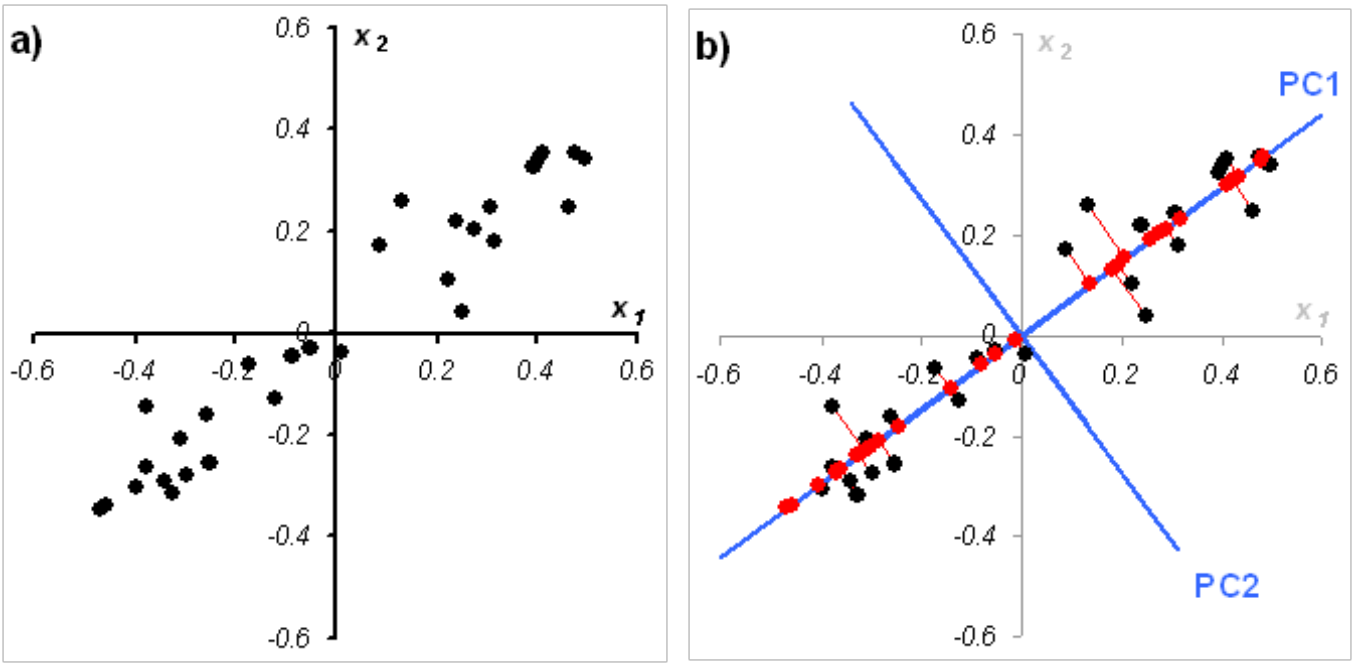


Рис. 6.1. Графическая иллюстрация метода главных компонент. a) данные в исходных координатах, b) данные в координатах главных компонент

На Рис. 6.1 показаны данные, состоящие только из двух переменных x_1 и x_2, которые связаны сильной корреляцией. На соседнем рисунке те же данные представлены в новых координатах. Вектор нагрузок $\mathbf{p}_1$ первой главной компоненты (PC1) определяет направление новой оси, вдоль которой происходит наибольшее изменение данных. Проекции всех исходных точек на эту ось составляют вектор $\mathbf{t}_1$. Вторая главная компонента $\mathbf{p}_2$ ортогональна первой, и ее направление (PC2) соответствует наибольшему изменению в остатках, показанных отрезками, перпендикулярными оси $\mathbf{p}_1$. Этот тривиальный пример показывает, что метод главных компонент осуществляется последовательно, шаг за шагом. На каждом

шаге исследуются остатки $\mathbf{E}_a$, среди них выбирается направление наибольшего изменения, данные проецируются на эту ось, вычисляются новые остатки, и т. д. Этот алгоритм называется NIPALS.

Метод главных компонент можно трактовать как проецирование данных на подпространство меньшей размерности A. Возникающие при этом остатки $\mathbf{E}$ рассматриваются как шум, не содержащий значимой химической информации. На Рис. 6.2 представлена графическая иллюстрация этого тезиса.

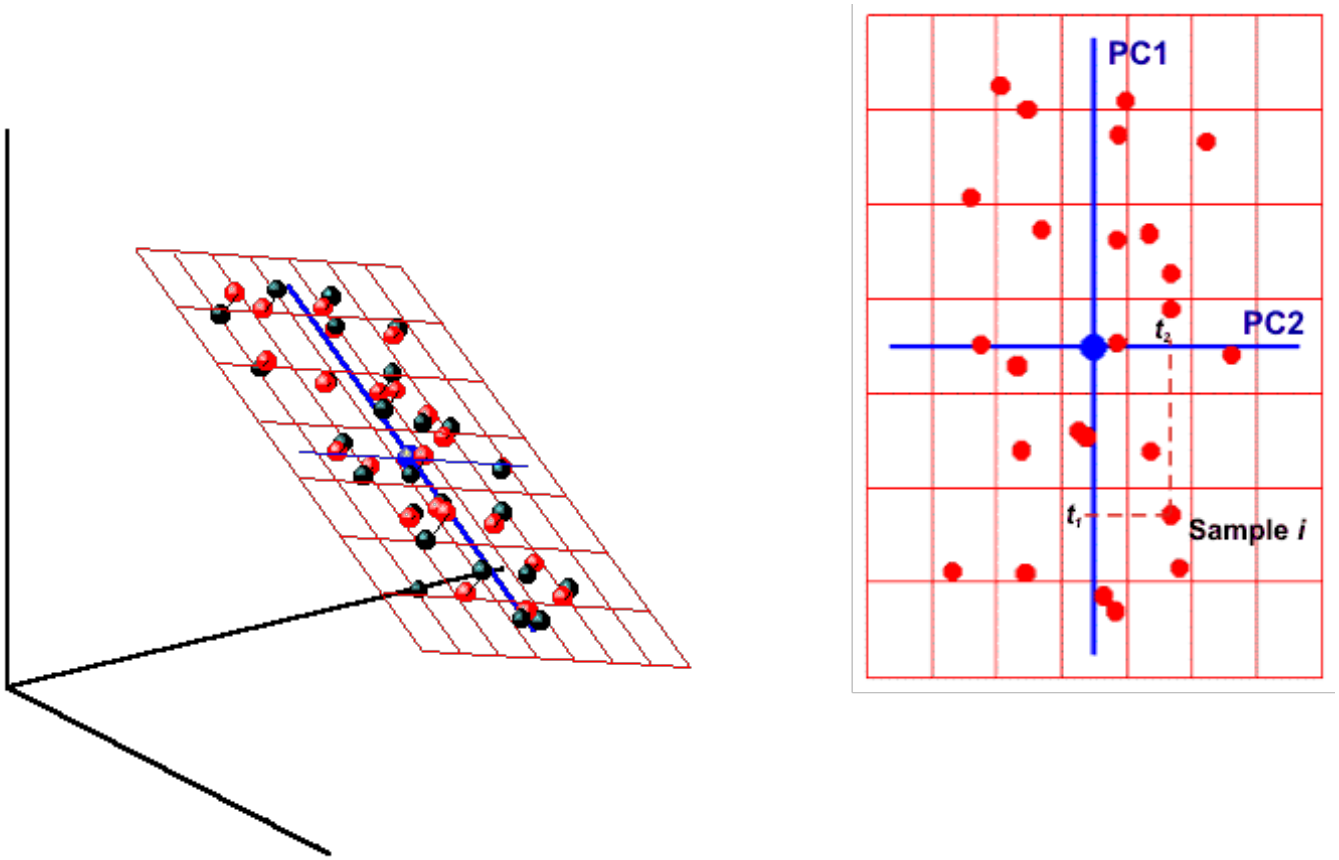


Рис. 6.2. Метод главных компонент – как проекция на подпространство

В случае, когда метод главных компонент применяется для решения калибровочной (регрессионной) задачи он носит название регрессии на главные компоненты (Principal Component Regression, PCR).

В методе PCA проекция строится только по данным $\mathbf{X}$. Значения откликов $\mathbf{Y}$ никак не используются. Обобщением метода PCA является метод проекций на латентные структуры (Projection on Latent Structures, PLS), который сейчас является самым популярным методом многомерной калибровки. Он во многом похож на метод PCR, с тем существенным отличием, что в PLS проводится одновременная декомпозиция матриц $\mathbf{X}$ и $\mathbf{Y}$

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{t}} + \mathbf{E},$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathrm{t}} + \mathbf{F},$$

$$\mathbf{T} = \mathbf{X}\mathbf{W} + \mathbf{G}$$

Проекция строится согласованно – так, чтобы максимизировать корреляцию между соответствующими векторами X-счетов $\mathbf{t}_a$ и Y-счетов $\mathbf{u}_a$. Поэтому PLS декомпозиция гораздо лучше описывает сложные связи, используя при этом меньшее число главных компонент.

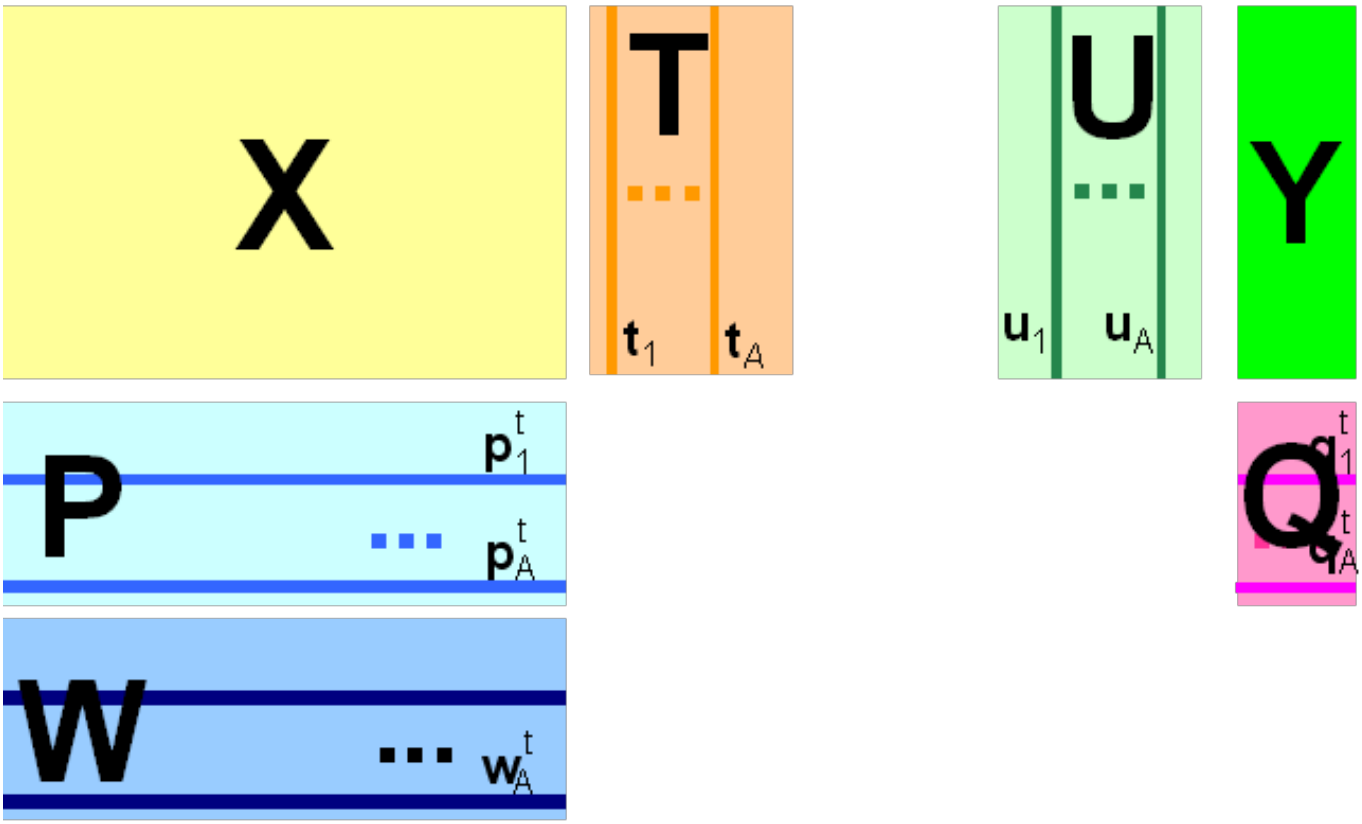


Рис. 6.3. Графическое представление метода проекций на латентные структуры

В том случае, когда имеется несколько откликов $\mathbf{Y}$ (т.е. $K > 1$), можно построить две проекции исходных данных – PLS1 и PLS2. В первом случае для каждого из откликов $\mathbf{y}_k$ строится свое проекционное подпространство. При этом и счета $\mathbf{T}$ ($\mathbf{U}$) и нагрузки $\mathbf{P}$ ($\mathbf{W}$, $\mathbf{Q}$), зависят от того, какой отклик используется. Этот подход называется PLS1. Для метода PLS2 строится одно общее проекционное пространство, которое является общим для всех откликов.

Следует обратить внимание на то, что и PCA, и PLS методы не учитывают свободного члена при разложении матрицы $\mathbf{X}$. Это видно из формул выше. Исходно предполагается, что все столбцы матриц $\mathbf{X}$ и $\mathbf{Y}$ имеют нулевое среднее, т.е.,

$\sum_{i=1}^{I} x_{ij} = 0$ и $\sum_{i=1}^{I} y_{ik} = 0$ для всех $j = 1, \ldots, J$ и $k = 1, \ldots, K$.

Этого легко можно достичь, проведя центрирование данных. Но, построив калибровочную модель на центрированных данных, надо не забыть пересчитать полученные оценки $\hat{\mathbf{y}}_{centre}$, добавив к ним вектор средних значений.

$\hat{\mathbf{y}} = \hat{\mathbf{y}}_{centre} + m_k\mathbf{1}, m_k = \frac{1}{I}\sum_{i=1}^{I} y_{ik}$

Приведем некоторые полезные формулы, доказательство которых можно найти во многих учебниках по анализу многомерных данных.

Во всех методах: $\mathbf{T}^t\mathbf{T} = \mathrm{diag}(\lambda_1, \ldots, \lambda_A)$

В методе PCR: $\mathbf{P}^t\mathbf{P} = \mathbf{I} = \mathrm{diag}(1, \ldots, 1)$

В методе PLS: $\mathbf{W}^t\mathbf{W} = \mathbf{I} = \mathrm{diag}(1, \ldots, 1)$

6.2 Регрессия на латентных переменных

После того, как данные $\mathbf{X}$ спроецированы на подпространство размерности A, исходная калибровочная задача превращается в серию регрессий на латентных переменных $\mathbf{T}$ –

$$\mathbf{T}_k\mathbf{b}_k = y_k, k = 1, \ldots, K$$

Здесь индекс k нумерует отклики в матрице $\mathbf{Y}$. Заметим, что в методах PCR и PLS2 имеется только одна матрица латентных переменных $\mathbf{T} = \mathbf{T}_1 = \cdots = \mathbf{T}_K$. Векторы $\mathbf{b}_k$ – это неизвестные коэффициенты, а $\mathbf{T}_k$ – матрицы счетов. Их общая размерность A существенно меньше числа переменных J, поэтому матрица $\mathbf{T}_k^t\mathbf{T}_k$ обращается устойчиво и проблема мультиколлинеарности не существенна. Но вот другая трудность – возможность переоценки остается, и с ней нужно уметь справляться.

Для проекционных методов сложность модели целиком определяется числом главных компонент A и выбор этой величины является основной трудностью в проекционных методах. Существует много методов определения величины A – эффективной размерности многомерных данных. Выше мы уже отмечали, что часто она связана с химическим рангом системы, т.е. с числом веществ, присутствующих в системе. Однако самым универсальным способом оценки размерности A является исследование графиков величин *RMSEC* и *RMSEP*, так как это было уже сделано в предыдущих разделах.

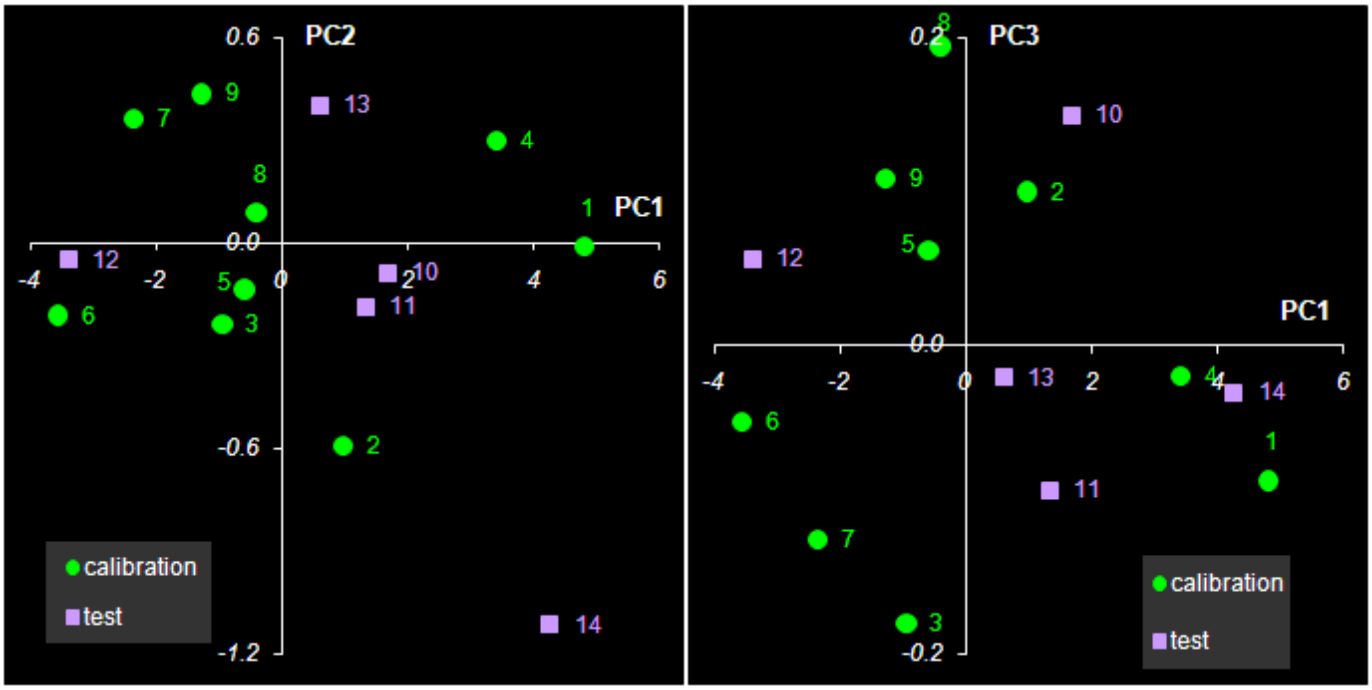


Рис. 6.4. Графики T счетов в методе PCR

При построении многомерных калибровок, большое внимание уделяется графикам счетов и нагрузок. Они несут в себе информацию, полезную для понимания того, как устроены данные. На графике счетов (Рис. 30) каждый образец изображается в координатах ($\mathbf{t}_i$, $\mathbf{t}_j$), чаще всего – ($\mathbf{t}_1$, $\mathbf{t}_2$). Близость двух точек означает их схожесть. Образцы, расположенные близко к началу координат (например, образцы 5, 11), являются самыми типичными – образцовыми. Напротив, образцы, которые находятся далеко (например, образцы 1 и 14), являются подозрительными маргиналами, и, может быть, даже выбросами.

Рис. 6.5. Графики P нагрузок в методе PCR

Если график счетов используется для анализа взаимоотношений образцов, то график нагрузок применяется для исследования роли переменных. На Рис. 31 показано как меняются нагрузки в зависимости от номера канала *j*. Можно обратить внимание на следующую закономерность. Чем больше номер главной компоненты, тем более зашумленным выглядит соответствующий нагрузки. График первой компоненты почти гладкий. По форме он похож на спектры чистых веществ *A* и *B* (см. Рис. 3.1). Вторая и третья главные компоненты отличаются более сложной формой, но они все еще имеют некоторые систематические тренды. А вот четвертая и пятые компоненты представляют просто случайный шум. Так, исследуя графики нагрузок, можно установить, что в нашем примере нужно использовать только 2 или 3 главные компоненты, но ни никак не больше трех.

6.3 Практическое применение

Калибровка на латентных переменных обычно строится на центрированном обучающем наборе ($\mathbf{X}_c$, $\mathbf{Y}_c$). Для того, чтобы получить прогноза на проверочный или новый набор образцов ($\mathbf{X}_t$, $\mathbf{Y}_t$), их надо также сцентрировать и спроецировать на уже имеющееся подпространство. Иными словами, нужно найти счета $\mathbf{T}_t$ новых образцов. В методе PCR это делается очень просто –

$$\mathbf{T}_t = \mathbf{X}_t\mathbf{P}$$

где $\mathbf{P}$ – это матрица нагрузок, построенная по обучающему набору.

Для методов PLS дело обстоит гораздо сложнее. Вычислительные аспекты многомерной калибровки выходят за пределы настоящего пособия.

Для работы с проекционными методами существуют большие коммерческие программы, например, The Unscrambler, SIMCA, которые представляют специальную среду для проведения различных манипуляций с многомерными данными. В частности, в них реализованы и все обсуждаемые методы калибровки. Среди свободно распространяемых программ можно отметить Multivariate Analysis Add-in – надстройка для Excel, разработанная в Хемометрическом центре Бристольского университета под руководством проф. Р. Бреретона.

В этом пособии все проекционные вычисления (PCA/PLS) проводятся в рабочей книге *Calibration.xls* с помощью специальной надстройки (Add-In) к программе Excel, которое называется Chemometrics.xla. Оно дополняет список стандартных функций Excel и позволяет проводить разложение на листах рабочей книги. От том как ее установить написано в пособии Инсталляция Chemometrics Add-In, а о том как ее использовать, рассказано в пособии Проекционные методы в системе Excel.

6.4 Регрессия на главные компоненты (PCR)

Рассмотрим, как применяется PCR в модельном примере. Все вычисления приведены на листе PCR.

Сначала, по обучающим данным $\mathbf{X}_c$, находится матрица PCA счетов $\mathbf{T}_c$ размерностью (9×5). Она состоит из пяти столбцов – главных компонент $\mathbf{t}_a$, $a = 1, 2, ..., 5$. С регрессионной точки зрения $\mathbf{T}_c$ – это матрица предикторов, т.е. независимых переменных. Двумя откликами будут центрированные значения концентраций веществ *A* и *B*. Затем, с помощью функции ТЕНДЕНЦИЯ (TREND) для каждого отклика строятся по пять регрессий. В этих регрессиях участвуют, соответственно, одна, две, … , пять главных PCA компонент. Так строится PCR калибровка на обучающем наборе.

В этих простых вычислениях есть одна тонкость. Регрессия строится на центрированных данных $\mathbf{A}_c$ и $\mathbf{B}_c$ только потому, что в версии Excel 2003 имеется критическая ошибка, исправленная в последующих версия, начиная с Excel 2007.

Проверочные данные тоже проецируются на пространство главных компонент и для них тоже определяются PCA счета $\mathbf{t}_1, ..., \mathbf{t}_5$. К этим счетам применяется построенная калибровка, и находятся оценки центрированных концентраций A_c^{hat}. Окончательные оценки откликов A^{hat} для каждого числа главных компонент a получаются после учета центрирования по концентрациям.

Соответствующие значения среднеквадратичных остатков обучения (*RMSEC*) и проверки (*RMSEP*) показаны на Рис. 6.6.

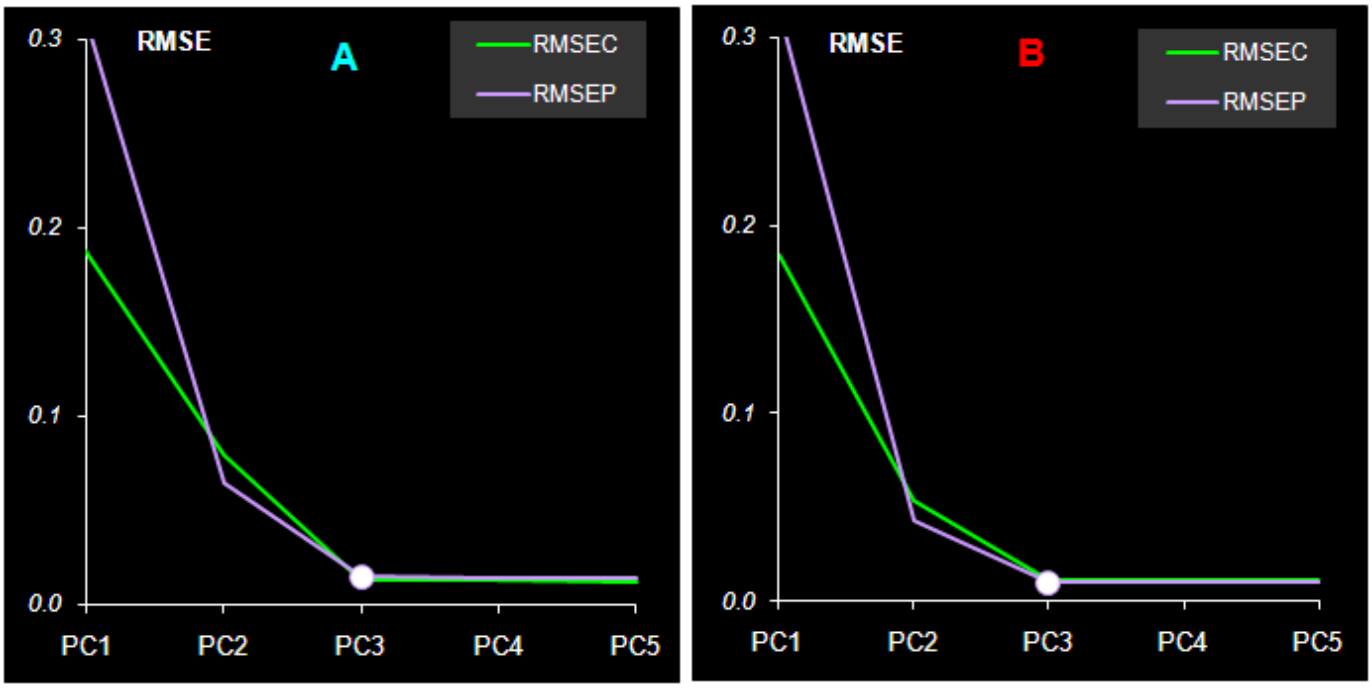


Рис. 6.6. Среднеквадратичные остатки обучения (RMSEC) и проверки (RMSEP) в регрессии на главные компоненты

Из Рис. 6.6 видно, что для обоих веществ минимум *RMSEP* достигается для трех PC ($A = 3$). Таким образом, применив PCR, мы легко смогли установить, что в исследуемой системе присутствуют не два, а три вещества.

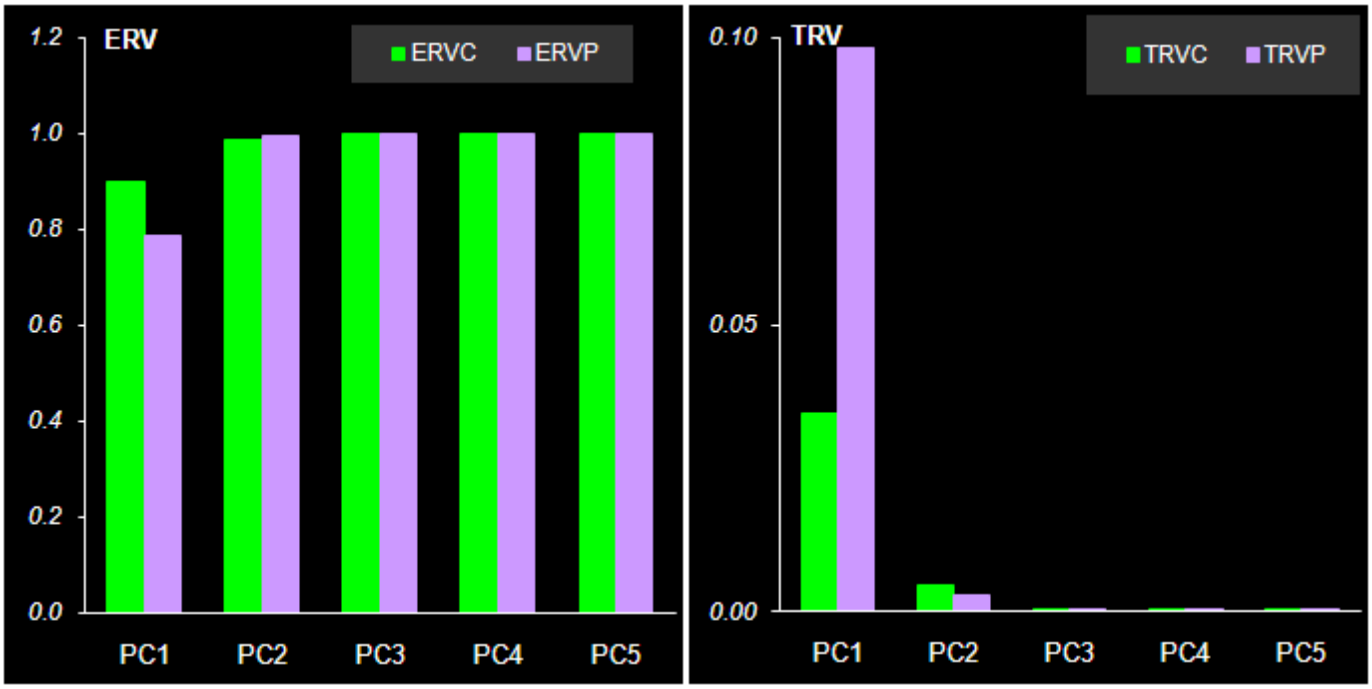


Рис. 6.7. Полная (TRV) и объясненная (ERV) дисперсии остатков в регрессии на главные компоненты

На Рис. 6.7 показаны полная (*TRV*) и объясненная (*ERV*) дисперсии остатков. Эти графики также свидетельствуют о том, что эффективная размерность системы – три. Для трех PC объясненная дисперсия практически равна 1, а полная дисперсия близка к нулю. В тоже время, видно, что графики среднеквадратичных остатков (Рис. 6.6) по сравнению с графиками дисперсий (Рис. 6.7) более наглядны.

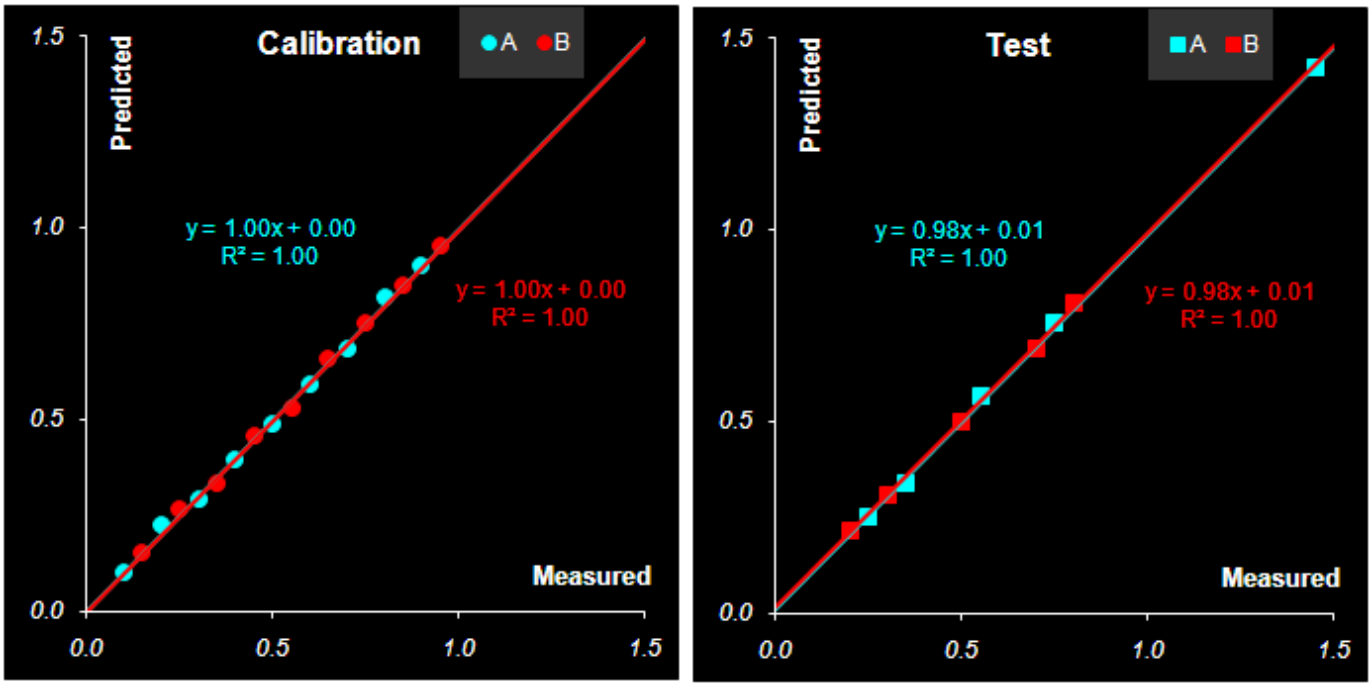


Рис. 6.8. Графики “измерено-предсказано” в регрессии на главные компоненты. Обучающий и проверочный наборы.

На Рис. 6.8 показаны графики “измерено-предсказано” для PCR при трех главных компонентах. Результат выглядит просто идеально – высокие коэффициенты корреляции, малые сдвиги, как в обучающем, так и проверочном наборах. Сравнивая эти графики с их аналогами в других метода – Рис. 4.3, Рис. 4.7, Рис. 5.2, Рис. 5.6 – можно увидеть очевидные преимущества регрессии на главные компоненты.

В таблице на Рис. 6.9 приведены характеристики качества калибровки веществ *A* и *B*, построенной методом главных компонент.

A	B
R_c^2= 0.997	R_c^2= 0.998
RMSEC= 0.013	RMSEC= 0.012
BIASC= 0.000	BIASC= 0.000
SEC= 0.013	SEC= 0.012
R_t^2= 0.999	R_t^2= 0.999
RMSEP= 0.015	RMSEP= 0.010
BIASP= -0.003	BIASP= 0.004
SEP= 0.015	SEP= 0.009
TRVC= 0.002	
ERVC= 1.000	
TRVP= 0.000	
ERVP= 1.000	

Рис. 6.9. Характеристики качества регрессии на главные компоненты

Таким образом, мы получили, что регрессия на главные компоненты (PCR) имеет явные преимущества

в сравнении с методами классической калибровки. Этот способ моделирования точнее, имеет меньшее смещение. Это объясняется тем, что в многомерной калибровке используются все имеющиеся экспериментальные данные. При этом они модифицированы так, чтобы избежать как мультиколлинеарности, так и переоценки.

6.5 Регрессия на латентные структуры (PLS1)

Регрессия на латентные структуры (PLS1) очень похожа на метод PCR.

Отличие состоит в том, что при построении проекционного пространства учитываются не только значения предикторов **X**, но и откликов **Y**. В результате получается не одна, а две матрицы **T** счетов – для каждого отклика (*A* и *B*) в отдельности.

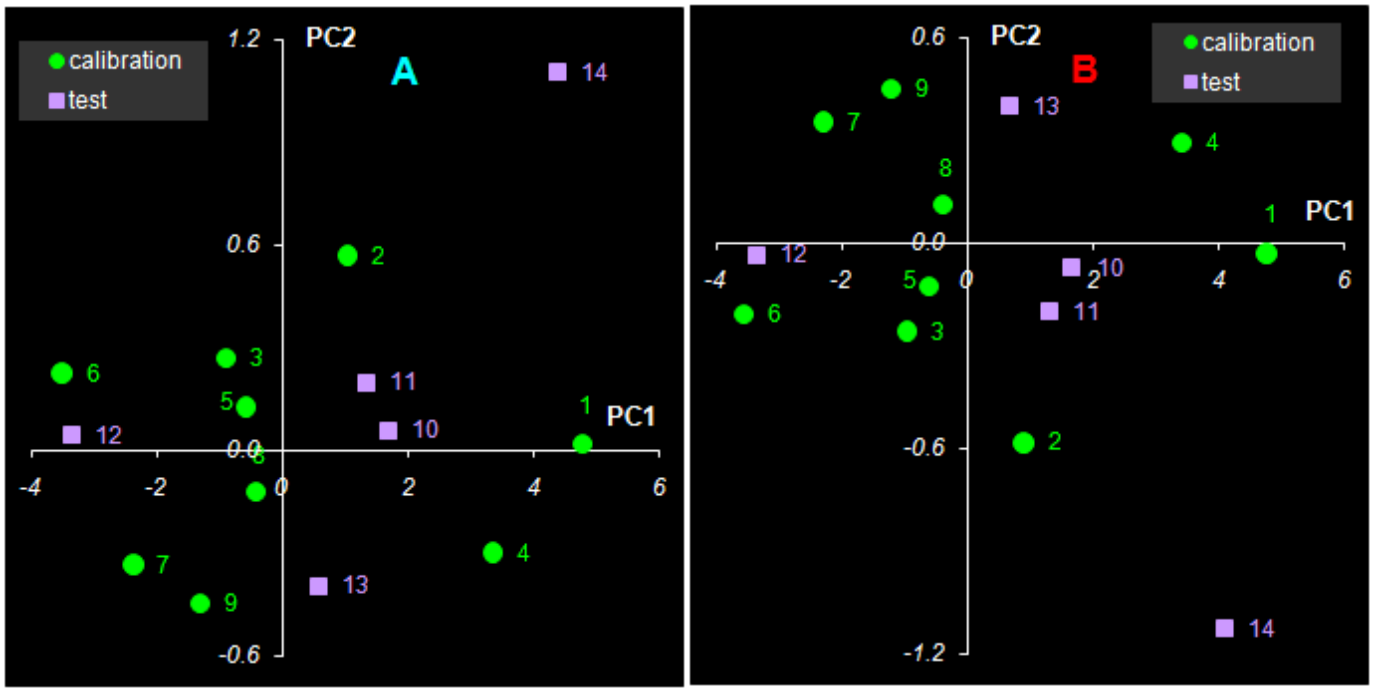


Рис. 6.10. Графики T счетов в методе PLS1

На Рис. 6.10 показаны графики первых двух **T** счетов в регрессии PLS1 для откликов *A* и *B*. При кажущемся различии, они очень схожи между собой и с аналогичным графиком PC1–PC2 на Рис. 6.4. Чтобы это понять, достаточно изменить знаки у первой главной PLS компоненты для отклика *A*. Такое преобразование главных компонент вполне законно. Дело в том, что, и PCA, и PLS декомпозиции определяются неоднозначно. В эти разложения всегда можно вставить произвольную, дополнительную матрицу **O**, такую, что $\mathbf{OO}^t = \mathbf{I}$. Действительно –

$$\mathbf{X} = \mathbf{TP}^t = \mathbf{TIP}^t = \mathbf{T}(\mathbf{OO}^t)\mathbf{P}^t = (\mathbf{TO})(\mathbf{PO})^t = (\mathbf{T}_{new})(\mathbf{P}_{new})^t$$

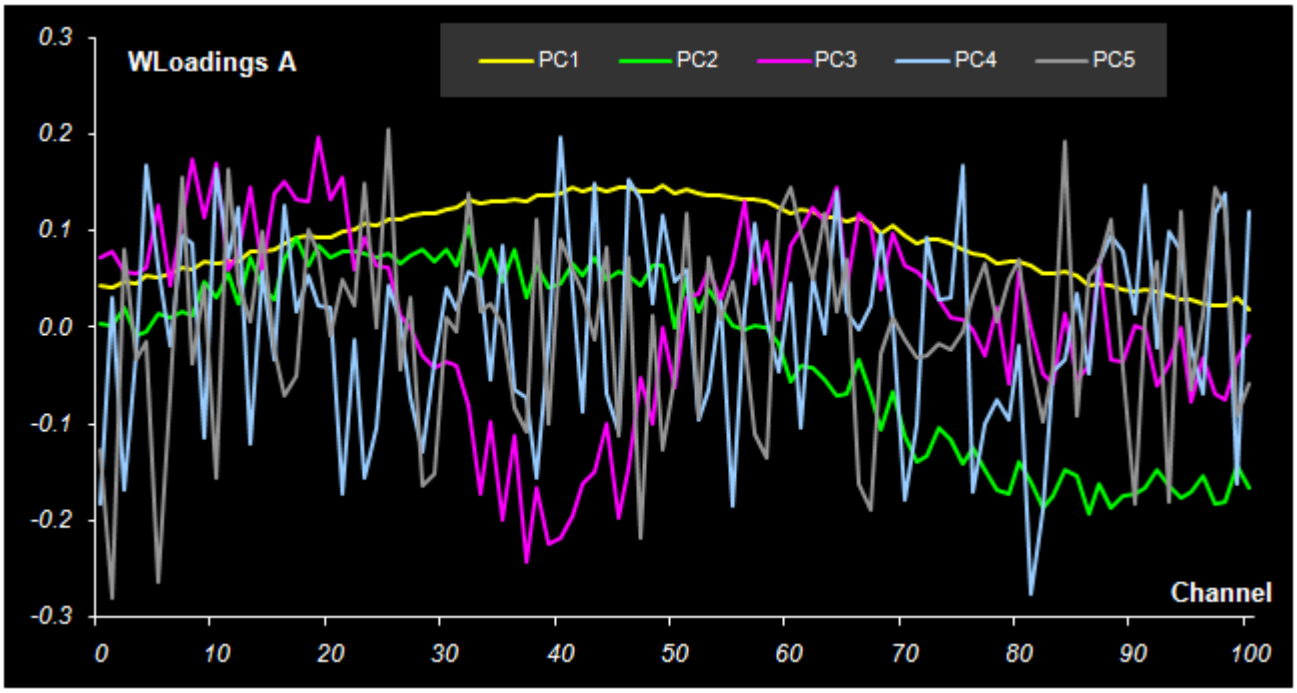


Рис. 6.11. Графики весовых нагрузок **W** для отклика *A* в методе PLS1

Аналогичная картина наблюдается и для нагрузок. На Рис. 6.11 показаны весовые нагрузки **W** для отклика *A*. Мы используем весовые нагрузки **W**, а не просто нагрузки **P**, потому, что они, по своей сути, ближе к **P** нагрузкам в методе PCA.

Калибровка по методу PLS1 строится вполне аналогично методу PCR, только в этом случае для каждого отклика используется своя матрица счетов.

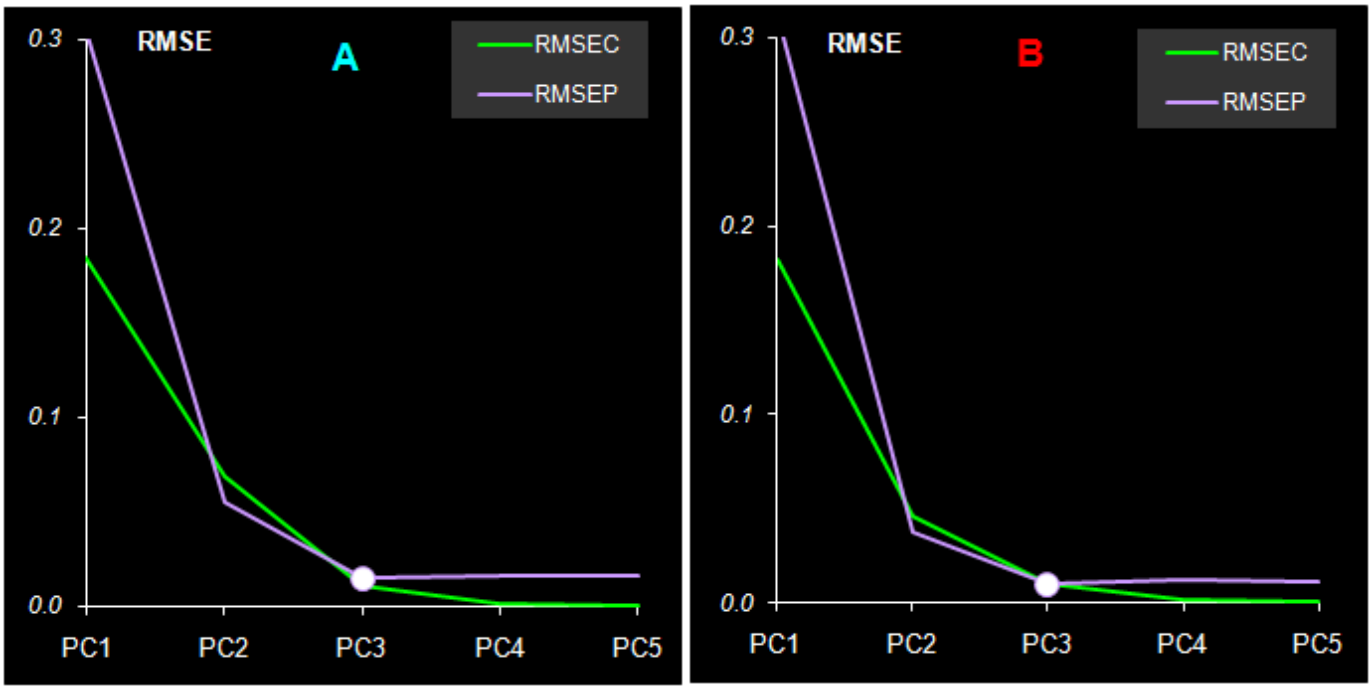


Рис. 6.12. Среднеквадратичные остатки обучения (RMSEC) и проверки (RMSEP) в регрессии PLS1

На Рис. 6.12 показаны среднеквадратичные остатки обучения (*RMSEC*) и проверки (*RMSEP*) в регрессии PLS1. Этот график похож на свой аналог в методе PCR (Рис. 6.6). Здесь тоже очевидно, что минимум *RMSEP* достигается для трех PC ($A = 3$). Некоторое отличие можно заметить только в поведении *RMSEC* — график не выходит на предел при трех PC, а продолжает падать с ростом числа компонент. В этом проявляется особенность метода проекций на латентные структуры – нацеленность на поддержание максимальной корреляции между **T** и **U** счетами.

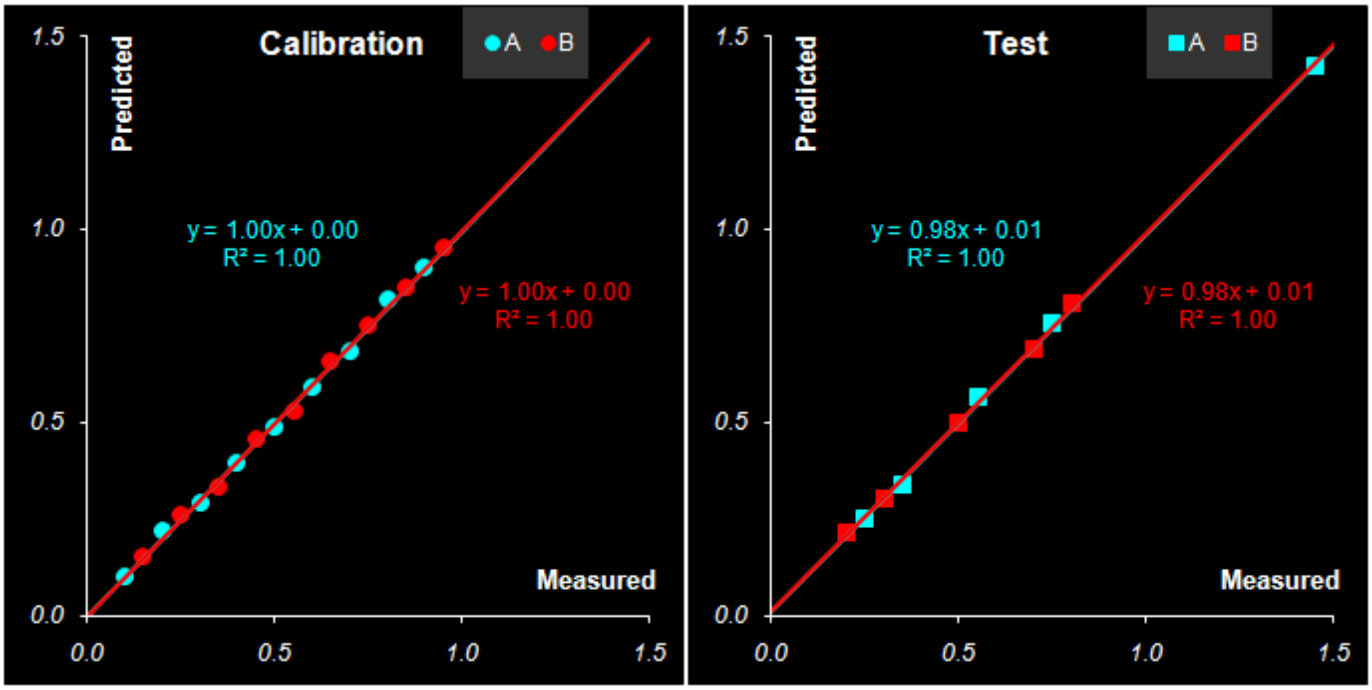


Рис. 6.13. Графики “измерено-предсказано” в PLS1 регрессии. Обучающий и проверочный наборы

Естественно, что и графики “измерено-предсказано” для калибровки методом PLS1 (Рис. 6.13) похожи на аналогичные графики для метода PCR (Рис. 6.8). То же можно сказать и о характеристиках качества калибровок, построенных методом PLS1, которые приведенные в таблице на Рис. 6.14.

A	B
R_c^2= 0.998	R_c^2= 0.998
RMSEC= 0.011	RMSEC= 0.010
BIASC= 0.000	BIASC= 0.000
SEC= 0.011	SEC= 0.010
R_t^2= 0.999	R_t^2= 0.998
RMSEP= 0.015	RMSEP= 0.010
BIASP= -0.003	BIASP= 0.003
SEP= 0.015	SEP= 0.010
TRVC= 0.001	
ERVC= 1.000	
TRVP= 0.000	
ERVP= 1.000	

Рис. 6.14. Характеристики качества PLS1 регрессии

Таким образом, калибровка методом PLS1 для рассматриваемого модельного примера очень близка по своим свойствам к калибровке методом PCR.

6.6 Регрессия на латентные структуры (PLS2)

Если метод PLS1 был похож на метод PCR, то отличие между PLS1 и PLS2 в нашем примере еще менее заметно.

При построении проекционного пространства PLS2 также учитываются значения и **X**, и **Y**. Однако все отклики **Y** рассматриваются совместно, поэтому получаются не несколько (как в PLS1), а одно общее проекционное подпространство (как в PCR) В результате мы имеем пару матриц счетов **T** и **U**, и три матрицы нагрузок **P**, **W** и **Q**.

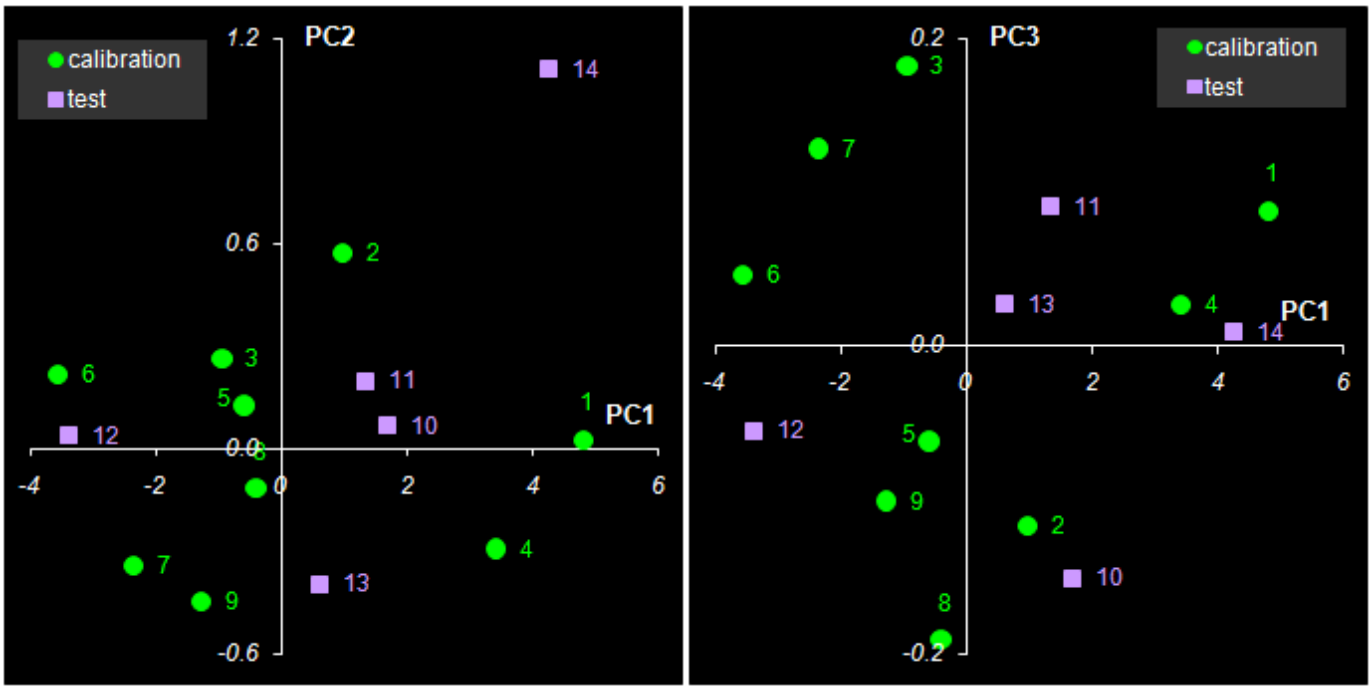


Рис. 6.15. Графики **T** счетов в методе PLS2

На Рис. 6.15 показаны графики первых трех **T** счетов в регрессии PLS2. Сравнивая их с аналогичными графиками на Рис. 6.4 и Рис. 6.10, легко заметить сходство. То же самое видно и для **W** нагрузок, представленных на Рис. 6.16.

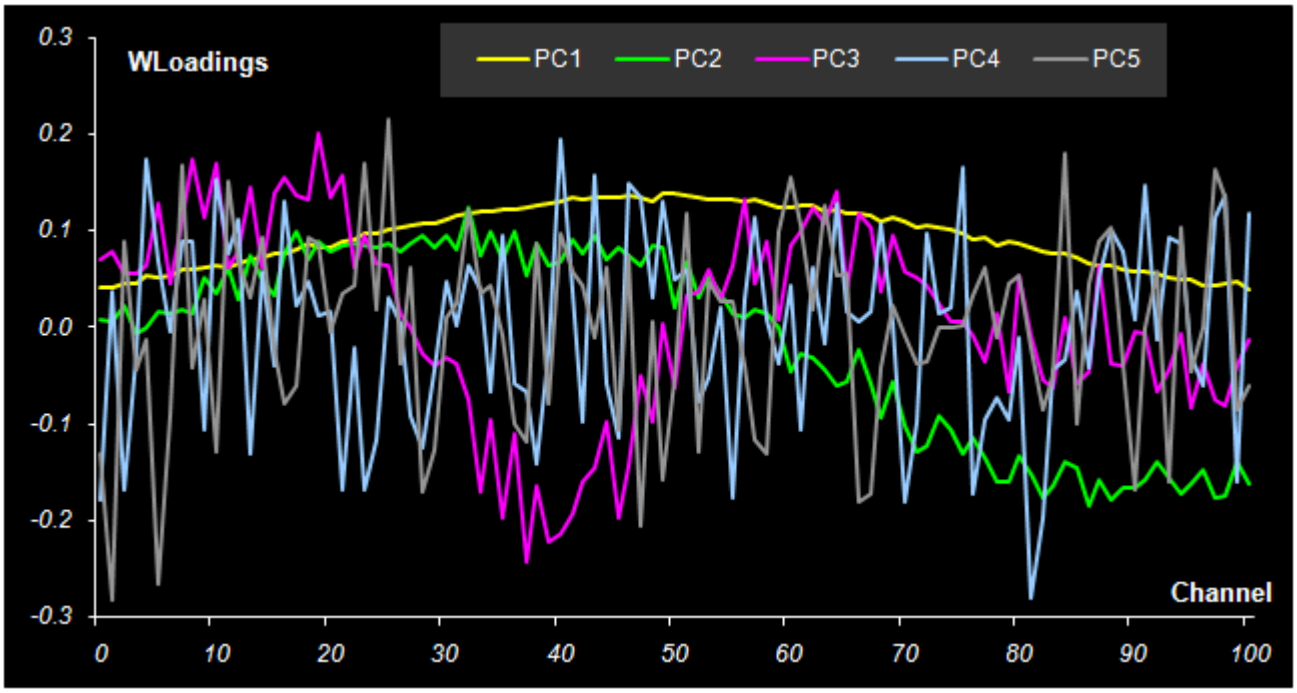


Рис. 6.16. Графики весовых нагрузок **W** в методе PLS2

Сравнивая **P** нагрузки в методе PCR на Рис. 6.5 и **W** нагрузки в методе PLS2, можно заметить, что между первыми тремя компонентами (нагрузками) есть хорошая линейная корреляция – для PC1 положительная ($R = 0.9999$), а для PC2 и PC3 отрицательная ($R = -0.990$, $R = -0.987$). Но вот четвертые компоненты уже никак не связаны друг с другом (R = 0.221). Смотри Рис. 6.17а). Это еще одно, дополнительное подтверждение того, что эффективная размерность исследуемой системы – три.

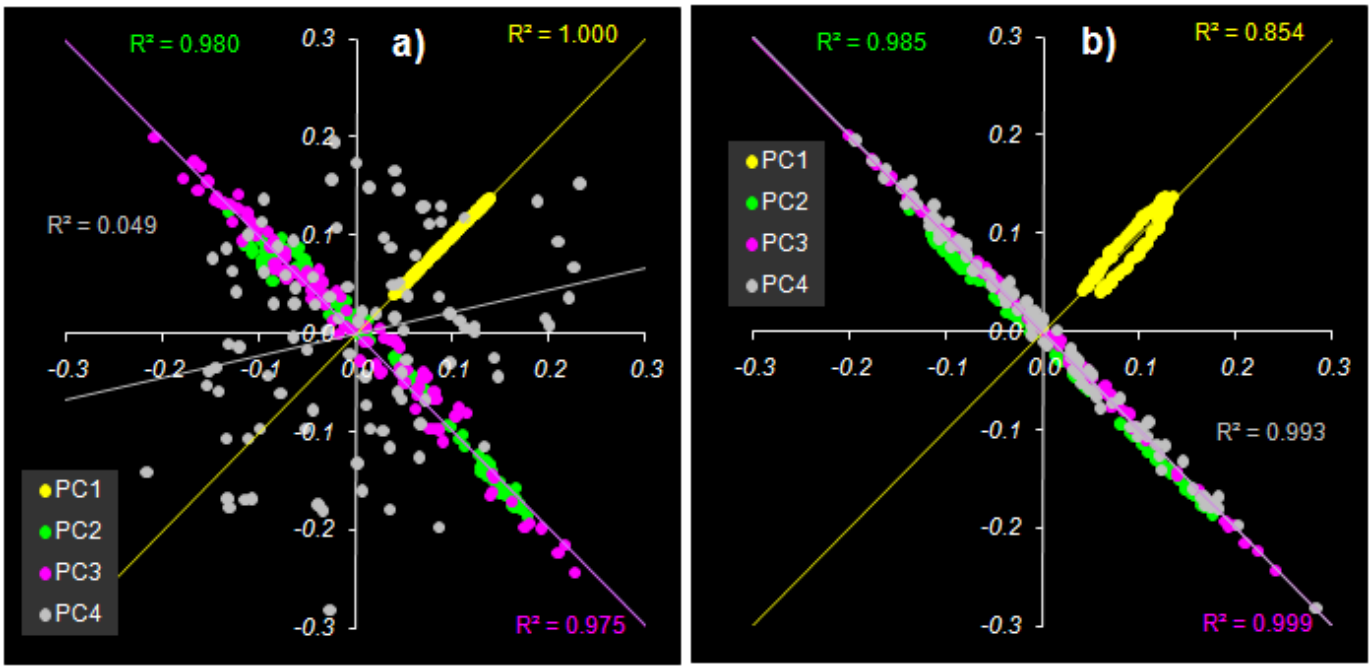


Рис. 6.17. a) Корреляция между P нагрузками в методе PCR и W нагрузками в методе PLS2 b) Корреляция между W нагрузками вещества B в методе PLS1 и W нагрузками в методе PLS2

Однако между методами PLS1 и PLS2 такой интересной связи не наблюдается. На Рис. 6.17b) показано, как связаны между собой **W** нагрузки, найденные методами PLS1 и PLS2. В первом случае использовались PLS1 **W** нагрузки для вещества *B*, но и для вещества *A* наблюдается аналогичная картина.

После того, как построена проекция на PLS2 подпространство, калибровка строится также как в методе PCR.

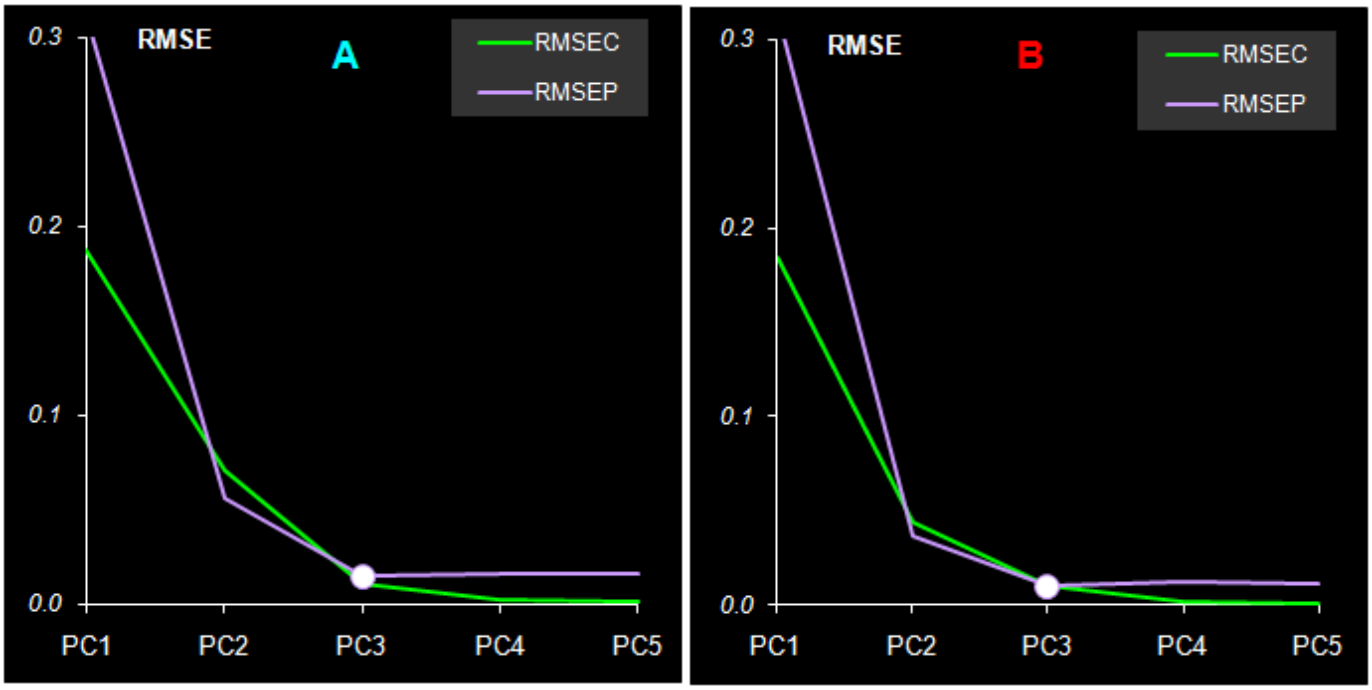


Рис. 6.18. Среднеквадратичные остатки обучения (*RMSEC*) и проверки (*RMSEP*) в методе PLS2

По Рис. 6.18, на котором показаны среднеквадратичные остатки обучения (*RMSEC*) и проверки (*RMSEP*) в регрессии PLS2, легко определить нужное число главных компонент $A = 3$. Отличия от аналогичного графика на Рис. 6.12 в методе PLS1 глазом не различимы.

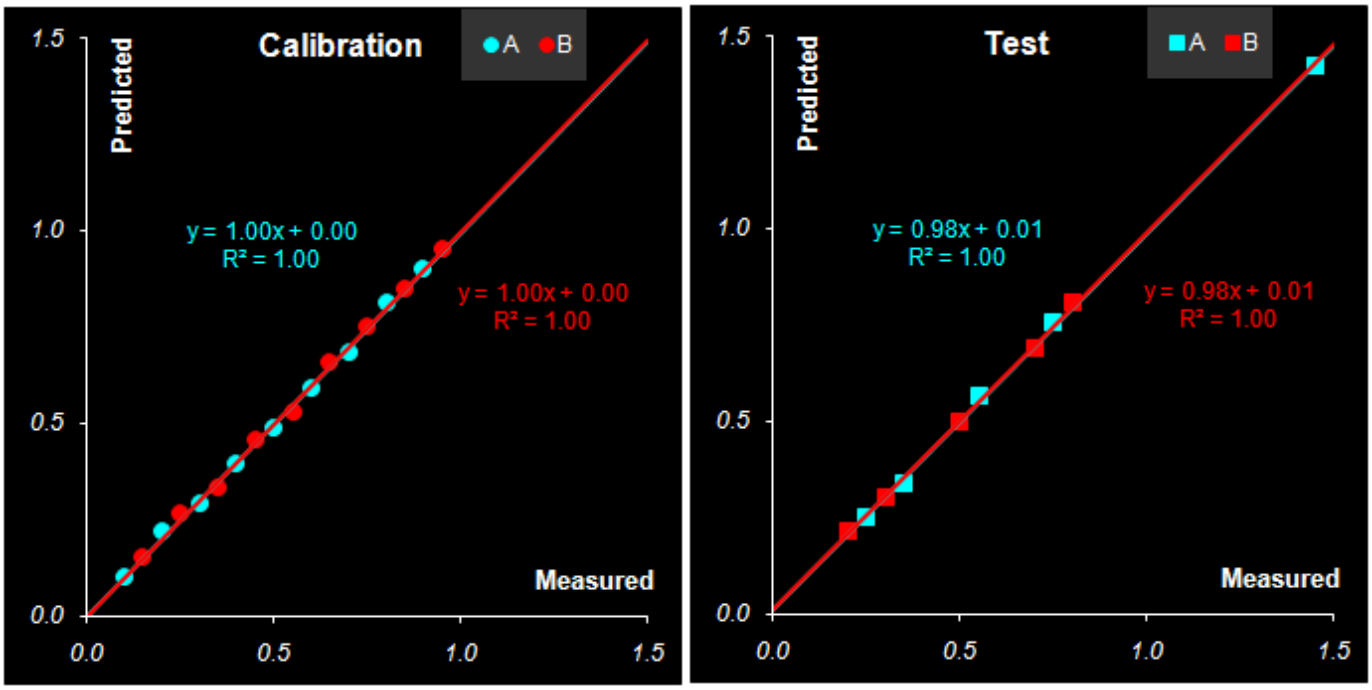


Рис. 6.19. Графики “измерено-предсказано” в PLS2 регрессии. Обучающий и проверочный наборы

Также неотличимы и графики “измерено-предсказано” для PLS1 (Рис. 6.13) и PLS2 (Рис. 6.19). Характеристики качества PLS2 калибровки, которые приведены в таблице на Рис. 6.20 совпадают с аналогичными для PLS1 (Табл. 2.7).

A	B
R_c^2= 0.998	R_c^2= 0.998
RMSEC= 0.011	RMSEC= 0.010
BIASC= 0.000	BIASC= 0.000
SEC= 0.011	SEC= 0.010
R_t^2= 0.999	R_t^2= 0.998
RMSEP= 0.015	RMSEP= 0.010
BIASP= -0.003	BIASP= 0.004
SEP= 0.015	SEP= 0.010
TRVC= 0.001	
ERVC= 1.000	
TRVP= 0.000	
ERVP= 1.000	

Рис. 6.20. Характеристики качества PLS2 регрессии

Может сложиться впечатление, что PLS2 регрессия не дает ничего нового по сравнению с PLS1 регрессией. Действительно, в большинстве случаев так и происходит. Более того, если между откликами в матрице Y нет корреляции, то PLS1 обычно дает лучшие результаты, нежели PLS2. Для прояснения этой ситуации бывает полезно построить PCA декомпозицию для матрицы **Y**. Если это разложение обнаруживает связи между

откликами, то метод PLS2 может оказаться очень полезным. Дальнейшее обсуждение этой интересной проблемы выходит за рамки настоящего пособия.

7 Заключение

7.1 Сравнение разных методов

Сопоставление различных методов начнем с характеристик *RMSEC* и *RMSEP*, которые наиболее полно отражают качество моделирования.

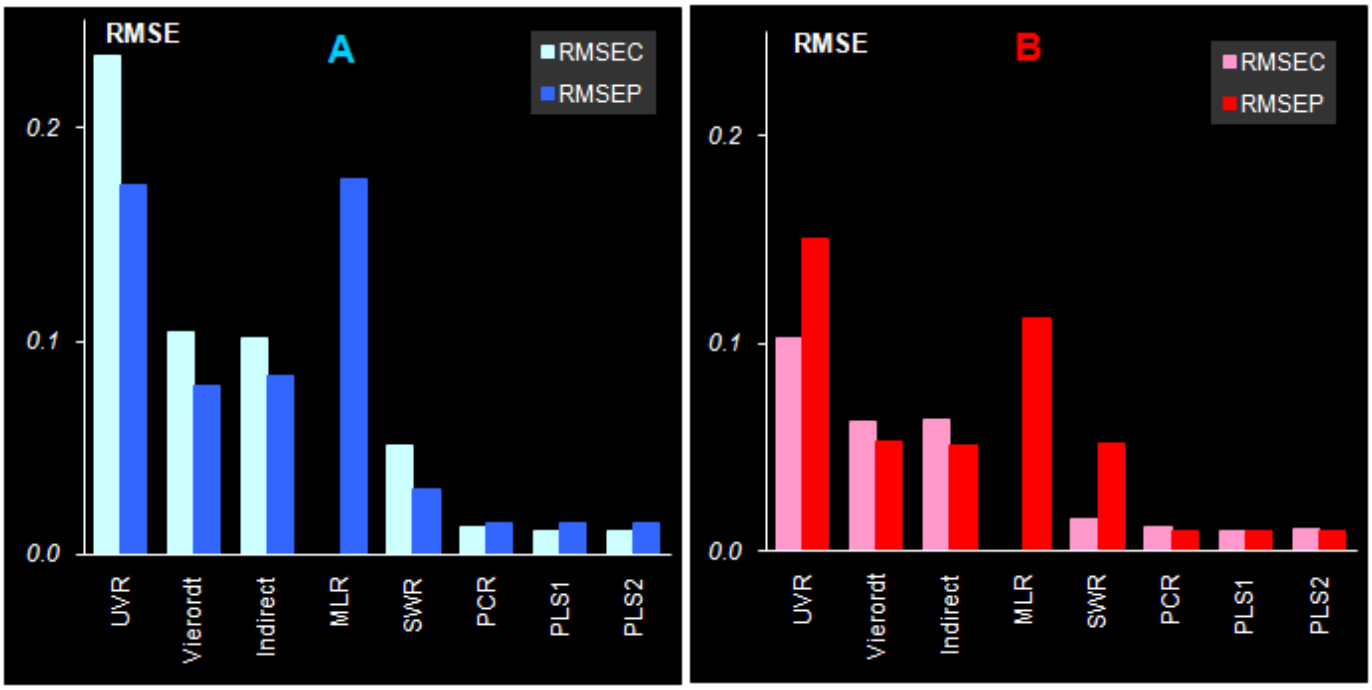


Рис. 7.1. Среднеквадратичные остатки обучения (*RMSEC*) и проверки (*RMSEP*) в различных методах калибровки веществ *A* и *B*

На Рис. 7.1 показаны среднеквадратичные остатки обучения и проверки, вычисленные для различных методов калибровки. Первое, что можно заметить на этих графиках – явное превосходство калибровки на латентных переменных: PCR, PLS и PLS2, перед традиционными подходами. Вторая интересная особенность состоит в том, что, часто, *RMSEP* больше, чем *RMSEC*. Это типично для правильной калибровки, если только различие между *RMSEP* и *RMSEC* не так разительно, как в случае множественной калибровки (MLR) или пошаговой регрессии (SWR) вещества *B*.

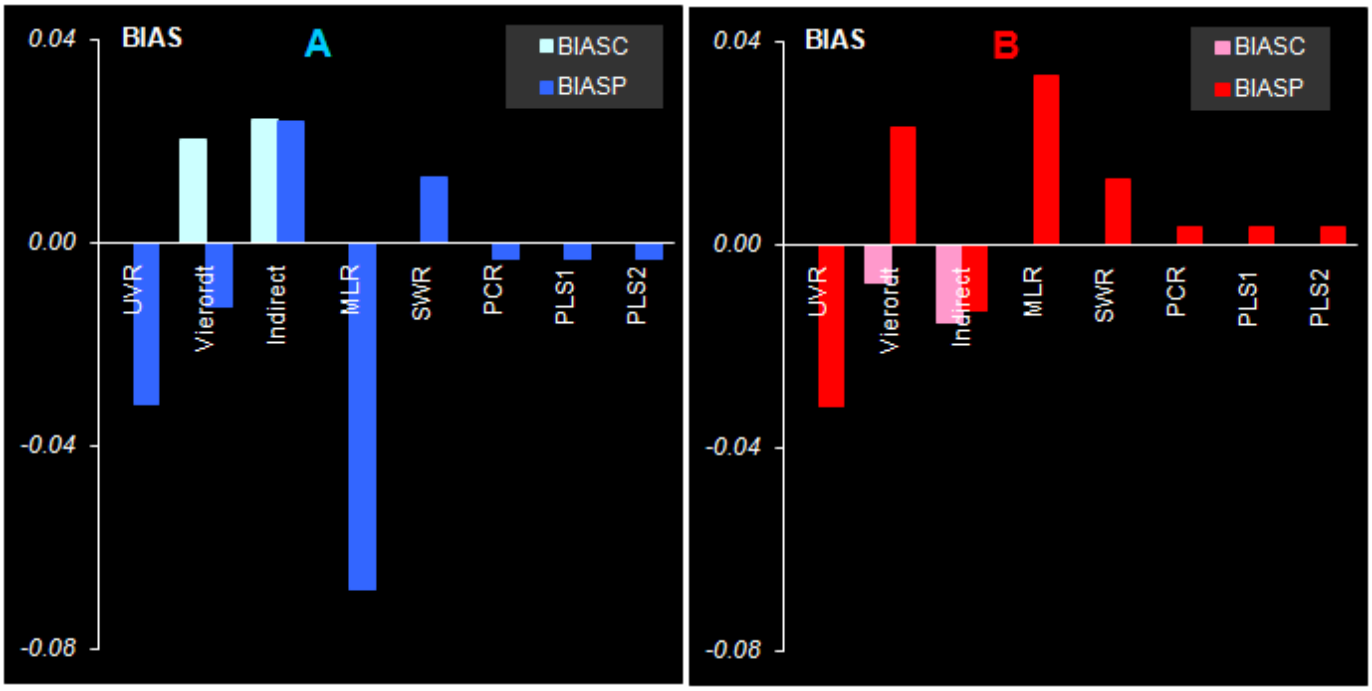


Рис. 7.2. Смещение в обучении (*BIASC*) и в проверке (*BIASP*) для различных методов калибровки веществ *A* и *B*

На Рис. 7.2 представлены величины смещений, полученные на обучающей выборке (*BIASC*) и на проверочной (*BIASP*) выборке. Здесь опять мы видим превосходство проекционных методов, которые дают значительно меньшие величины систематических отклонений. Аналогично, почти всегда *BIASC* меньше, чем *BIASP*, а для UVR и MLR отличие обучения от проверки особенно заметно.

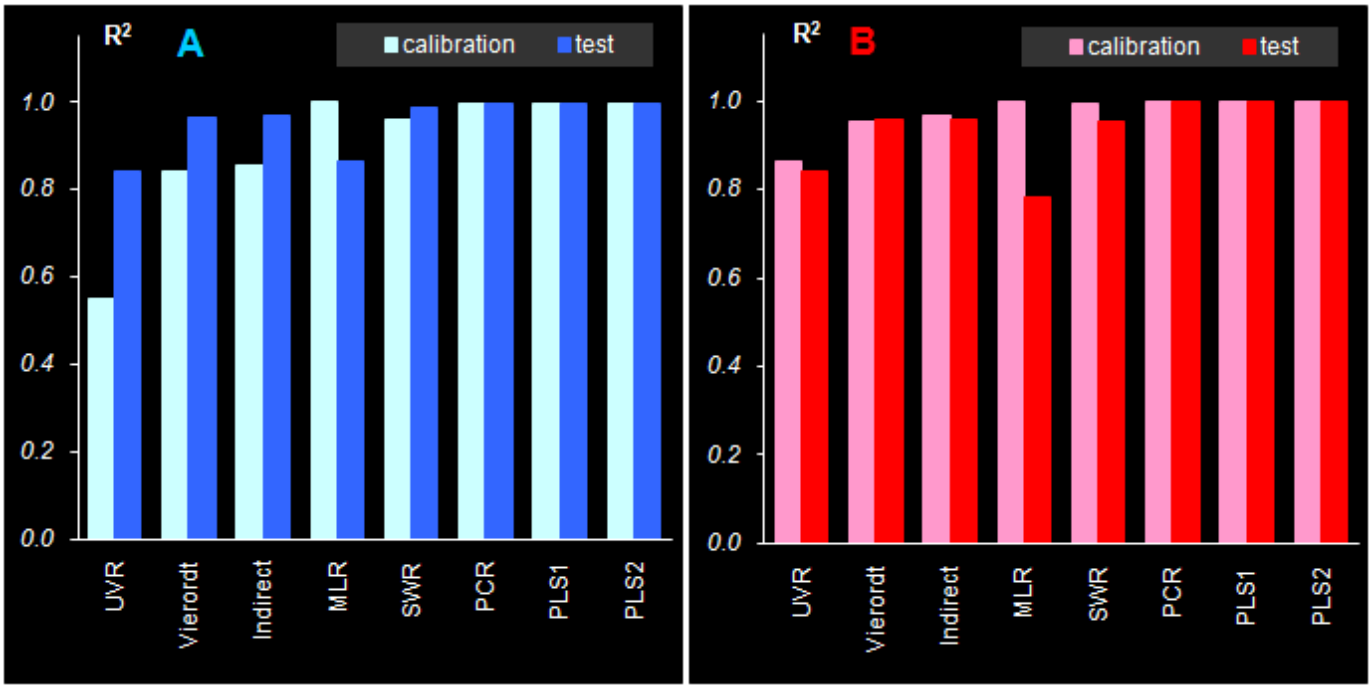


Рис. 7.3. Коэффициенты корреляции в обучении и в проверке для различных методов калибровки веществ *A* и *B*

На Рис. 7.3 показаны величины коэффициентов корреляции между стандартными и оцененными откликами. Видно, что проекционные методы не только дают лучшие значения R^2 (ближе к единице), но и, кроме того, представляют правильный баланс между обучением и проверкой.

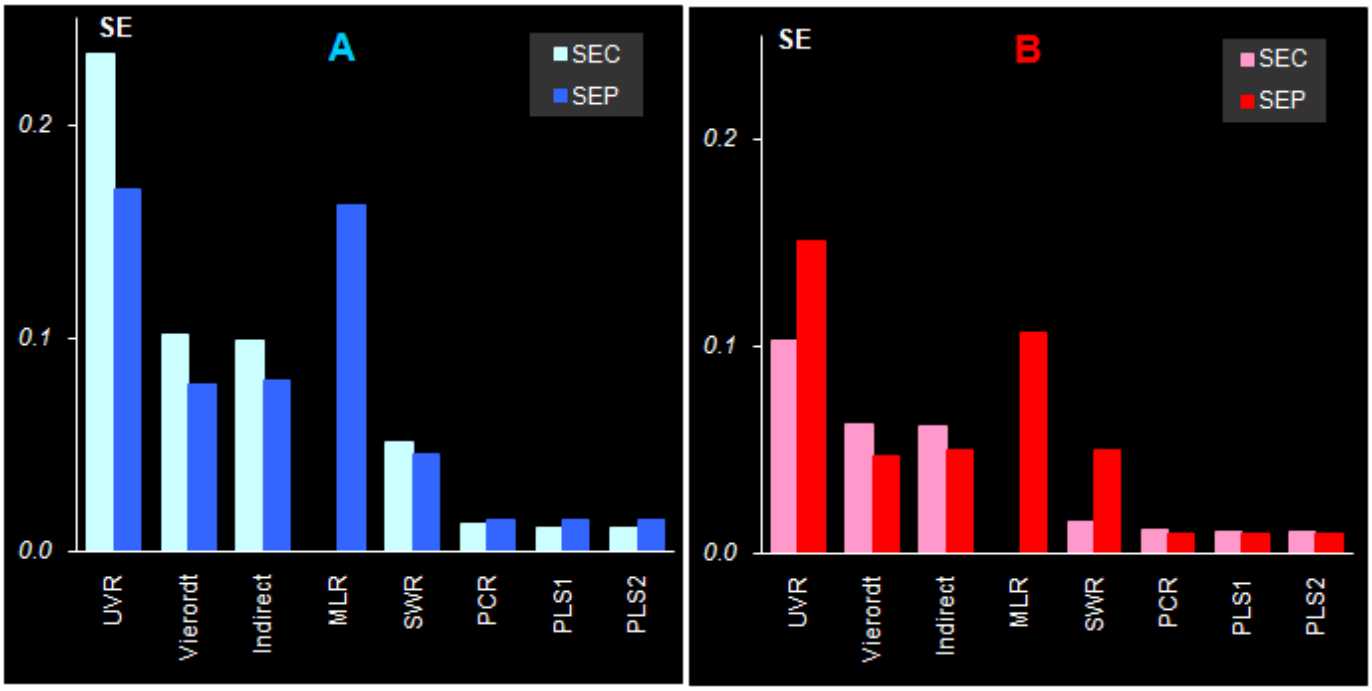


Рис. 7.4. Стандартные ошибки в обучении (*SEC*) и в проверке (*SEP*) для различных методов калибровки веществ A и B

На Рис. 7.4 показаны стандартные ошибки в обучении (*SEC*) и в проверке (*SEP*). Из-за того, что в нашем примере все смещения малы по сравнению с *RMSE*, эти графики не добавляют нам новых открытий по сравнению с уже исследованными зависимостями показанными на Рис. 7.1.

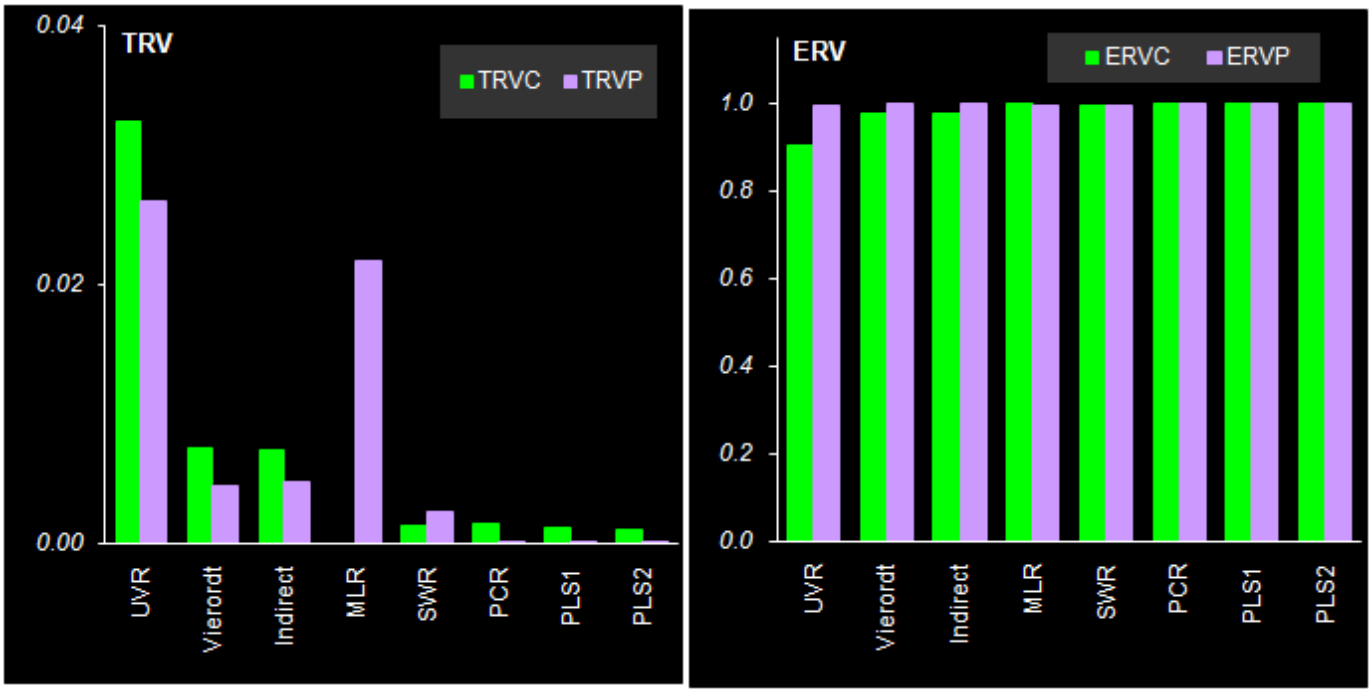


Рис. 7.5. Полная дисперсия остатков в обучении ($TRVC$) и в проверке ($TRVP$). Объясненная дисперсия остатков в обучении ($ERVC$) и в проверке ($ERVP$) для различных методов калибровки веществ A и B

На Рис. 7.5 слева приведена полная дисперсия остатков в обучении ($TRVC$) и в проверке ($TRVP$), а справа – объясненная дисперсия остатков в обучении ($ERVC$) и в проверке ($ERVP$). На этих графиках мы снова видим те же закономерности, что и ранее – методы калибровки на латентных переменных лучше традиционных. Можно также отметить, что графики ERV неудобны для анализа качества моделирования – на них плохо заметны недостатки или преимущества различных методов.

7.2 Выводы

Подведем итоги исследования различных методов калибровки. Итак, мы видели, что:

- калибровка по одному каналу приводит к недооценке – величины $RMSEC$ и $RMSEP$ слишком велики;
- множественная калибровка, напротив, ведет к переоценке – величина $RMSEC$ значительно меньше, чем $RMSEP$;
- наилучшие результаты дает калибровка на латентных переменных (PCR и PLS) – достигается правильный баланс между величинами $RMSEC$ и $RMSEP$.