Количественное определение индивидуальных компонентов в смеси с использованием многомерных градуировочных моделей

Оглавление

1. Теоретическая часть

1.1. Градуировочные модели и проблема коллинеарности

Градуировочные модели используют для проведения косвенного измерения. Типичным примером является линейная модель, где предсказываемая (зависимая переменная) прямо пропорциональна одной или нескольким известным (независимым) переменным. Теоретически, нелинейную зависимость также возможно учесть путём построения более сложной модели, но для них может быть гораздо сложнее доказать единственность и существование оптимального набора параметров, а также найти его; кроме того, на практике нелинейность обычно выражается в отклонениях от приближённых физических принципов, на основе которых строят линейную модель, которые сложнее зафиксировать и воспроизвести в последующем эксперименте, что сужает применимость такого рода моделей.

Многомерные градуировочные модели привлекательны своей способностью компенсировать примеси в наблюдаемых независимых переменных: если каждый вектор $\boldsymbol{x}$ является спектром смеси двух перекрывающихся веществ, концентрацию одного из которых нужно предсказать, оптимальный вектор параметров $\boldsymbol{b}$, для которого $y \approx \boldsymbol{x}^{\mathrm{T}} \boldsymbol{b}$, сложит фрагменты спектра, принадлежащие искомому компоненту, с положительными коэффициентами, после чего вычтет фрагменты, где вещества перекрываются. Решать такую задачу значительно проще, чем раскладывать матрицу независимых переменных на неизвестные профили в неизвестных концентрациях $\boldsymbol{X} \approx \boldsymbol{C}\,\boldsymbol{S}^{\mathrm{T}}$, не только потому, что что количество неизвестных примесей в такой задаче также неизвестно, но и потому, что в двух измерениях её решение является принципиально неединственным[1].

1.2. Регуляризация

Формально, для задачи поиска коэффициентов линейной модели $\min\limits_{\boldsymbol{b}} \|\boldsymbol{y} - \boldsymbol{X}\,\boldsymbol{b}\|^2$ решением является $\hat{\boldsymbol{b}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$. Однако, на практике обратная матрица $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$ зачастую либо не существует вовсе, либо даёт некачественные $\hat{\boldsymbol{b}}$, которые плохо обобщаются на другие образцы с таким же составом. Чтобы получить более надёжное решение, можно применять методы регуляризации.

Например, добавление штрафа на норму решения $\min\limits_{\boldsymbol{b}} \|\boldsymbol{y} - \boldsymbol{X}\,\boldsymbol{b}\|^2 + \lambda \|\boldsymbol{b}\|^2$ [2] позволяет получить надёжное решение $\hat{\boldsymbol{b}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda\,\boldsymbol{I})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$ в одно действие, не испытывая проблем с обратимостью $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$.

Вместо добавления штрафа на Евклидову норму решения, можно штрафовать 1-норму, т. е. сумму модулей коэффициентов: $\min\limits_{\boldsymbol{b}} \|\boldsymbol{y} - \boldsymbol{X}\,\boldsymbol{b}\|^2 + \lambda \|\boldsymbol{b}\|_1$ [3]. Решение такой задачи требует использования итеративных методов, зато в $\hat{\boldsymbol{b}}$ оказывается «встроен» автоматический выбор

---

1 Для трёх и более измерений, например, спектров возбуждения-испускания флуоресценции, см. метод параллельного факторного анализа.

2 Hastie, Tibshirani, Friedman. The Elements of Statistical Learning. (2009) Секция 3.4.1, Ridge Regression.

3 Там же, секция 3.4.2, The Lasso.

переменных, потому что для «лишних» независимых переменных коэффициент регрессии b_j оказываются в точности равны 0.

Данные методы требуют выбора оптимального значения λ: слишком маленькие значения эквивалентны полному отсутствию регуляризации, тогда так слишком большие значения дадут смещённую оценку $\boldsymbol{b}$, занижая предсказываемые концентрации.

1.3. Проекционные методы

Если значения большого количества независимых переменных обусловлены небольшим количеством химических компонент, имеет смысл снизить размерность задачи, найдя направления в пространстве независимых (но скоррелированных между собой) переменных, вдоль которых они выстроены, и перейдя к новым координатам, выраженным в терминах этих направлений. В этом и есть суть метода главных компонент[4]. У задачи $\boldsymbol{X} \approx \boldsymbol{T}\,\boldsymbol{P}^{\mathrm{T}}$ есть единственное решение, если наложить на $\boldsymbol{T}$ условие ортогональности ($\boldsymbol{T}^{\mathrm{T}}\boldsymbol{T} = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$), а на $\boldsymbol{P}$ — ортонормированности ($\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P} = \boldsymbol{I}$); обычно его получают через сингулярное разложение. Обычно выбирают ограниченное количество столбцов в матрицах $\boldsymbol{T}$ и $\boldsymbol{P}$, имеющих физический смысл, а остальные отбрасывают, считая их шумом.

Спроецировав новый образец на полученные главные компоненты, можно получить две величины, на основании которых его можно счесть выбросом. Одна из них характеризует расстояние данного образца от системы, описываемой данными главными компонентами: $Q_i = \left\| (\boldsymbol{x}_i \boldsymbol{P})\,\boldsymbol{P}^{\mathrm{T}} - \boldsymbol{x}_i \right\|^2$. Иными словами, это та часть данных, которая остаётся после проекции образца на главные компоненты. Другая из них характеризует расстояние данного образца от центра системы главных компонент: $T_i^2 = \boldsymbol{t}_i \boldsymbol{\Sigma}^{-1} \boldsymbol{t}_i^{\mathrm{T}}$. Иными словами, это та часть данных, которая описывается моделью, но лежит вдалеке от типичного диапазона значений.

На счетах ($\boldsymbol{T}$) главных компонент легко строить регрессионные модели, потому что у произведения ортогональной матрицы на транспонированную всегда будет обратная, а размерность пространства счетов значительно меньше исходной.

Поскольку метод главных компонент не работает с зависимыми переменными, он не может гарантировать, что самыми главными окажутся компоненты, наилучшим образом объясняющие дисперсию искомой зависимой переменной. (Экспериментаторы должны и так к этому стремиться при планировании эксперимента, но достижение этой цели не всегда возможно.) По аналогии с методом главных компонент можно предложить метод проекции на латентные структуры[5], где (в наиболее общей формулировке) проекции подвергаются обе матрицы: и зависимых переменных ($\boldsymbol{X} = \boldsymbol{T}\,\boldsymbol{P}^{\mathrm{T}}$), и независимых переменных ($\boldsymbol{Y} = \boldsymbol{U}\,\boldsymbol{Q}^{\mathrm{T}}$), таким образом, чтобы максимизировать ковариацию $\boldsymbol{T}$ и $\boldsymbol{U}$ (в одномерном случае — столбца зависимых переменных $\boldsymbol{y}$), сохраняя ортогональность $\boldsymbol{T}$. При этом получается возможно найти коэффициенты регрессии для нахождения $\boldsymbol{U}$ из $\boldsymbol{T}$, после чего пересчитать их в терминах исходных переменных.

---

4 Karl Pearson F.R.S., On lines and planes of closest fit to systems of points in space, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 1901.

5 Svante Wold, Paul Geladi, Kim Esbensen, Jerker Öhman. Multi-way principal components-and PLS-analysis. J. Chemometrics, 1(1987): 41-56.

Оба метода требуют подбора правильного количества столбцов в усечённых матрицах проекции, соответствующих количеству главных компонент / латентных структур в модели. Выбор слишком маленького количества компонентов не опишет достаточно компонент, чтобы правильно предсказать концентрации. Выбор слишком большого количества компонент приведёт к тому, что модель начнёт описывать шум и вернёт некачественные коэффициенты регрессии, плохо обобщаемые на другие образцы с таким же составом.

1.4. Предварительная обработка данных

Даже в тех случаях, когда условия проведения эксперимента нарушают прямую пропорциональность между зависимыми и независимыми переменными, простые преобразования данных могут восстановить такие зависимости. Например, в ИК-спектроскопии коррекцию сигнала рассеяния, который смещает или умножает спектр на неизвестную константу, выполняют при помощи стандартизации всего спектра (приведения к среднему 0 и стандартному отклонению 1).[6]

Традиционным способом борьбы со смещением спектра на наклонную прямую («дрифт» базовой линии) является взятие от него численных производных. Поскольку численные производные чувствительны к шуму в данных, эту операцию обычно сочетают со сглаживанием спектра по Савицкому-Голею. Эта процедура может изменить форму спектров, но обычно сохраняет пропорциональные зависимости между площадью под пиком и концентрацией аналита.

Базовую линию, которая не описывается простой функцией, но является гладкой, можно попытаться оценить и вычесть при помощи сглаживания по Уиттэкеру с асимметричным штрафом[7], но это уже менее точный метод, и он требует подбора нескольких неизвестных параметров.

1.5. Подбор параметров и оценка модели

Все методы линейной регрессии уже подбирают вектор коэффициентов $\hat{\boldsymbol{b}}$ таким образом, чтобы минимизировать ошибку предсказания концентраций известных образцов, т.е. максимизировать объяснительную способность модели. Если варьировать штраф регуляризационного метода или количество компонентов — проекционного, глядя только на описательную способность модели, результат получится во многом бесполезным: максимальной описательной способности можно добиться, не накладывая штраф или используя все доступные компоненты. К сожалению, предсказательная способность модели при этом очень сильно страдает.

Предсказательную способность, в которой мы заинтересованы, демонстрирует поведение модели на образцах, не использованных для её обучения. Чтобы её оценить, из обучающего набора данных тем или иным способом извлекают часть образцов, а после обучения модели их предсказывают. Для надёжности этот процесс можно повторять много раз[8]. В результате

---

6 Harald Martens, Edward Stark. Extended multiplicative signal correction and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy. Journal of Pharmaceutical and Biomedical Analysis, 9(8), 1991.

7 Eilers, P. H. C. (2004). Parametric Time Warping. Analytical Chemistry, 76(2), 404–411. См. приложение, где задан сам этот метод, а не статью, где говорится о другом.

8 Efron, Bradley (1979). Bootstrap methods: Another look at the jackknife. The Annals of Statistics. 7: 1–26.

получается вполне надёжная оценка предсказательной способности модели, которую можно использовать для выбора оптимального значения гиперпараметров (например, количества компонент).

Важно отметить, что точно так же, как после обучения модели получается вектор коэффициентов регрессии, минимизирующий ошибку предсказания обучающего набора (и поэтому эту минимизированную величину нельзя использовать для оценки модели), так и после перекрёстной проверки получаются значения гиперпараметров, минимизирующие ошибку предсказания данных перекрёстной проверки. Это значит, что ошибку на перекрёстной проверке нельзя использовать для оценки модели, если она уже была использована для выбора параметров. Вместо этого нужно либо выделять независимый проверочный набор данных, образцы из которого не войдут ни в обучающий, ни в оптимизационный наборы, и использовать его в самом конце процедуры, либо выполнять эту процедуру много раз[9]. В последнем случае придётся как-то решать проблему, в результате которой на разных итерациях процесса могут получиться разные оптимальные значения гиперпараметров.

2. Практическая часть

Настоящая задача посвящена использованию регрессии на латентных структурах для предсказания количественного содержания нескольких компонентов в для модельного набора данных.

Вам были даны файлы с независимыми переменными (`train_X.txt`) и соответствующими им значениями зависимых переменных (`train_y.txt`), а также отдельный файл (`test_X.txt`) со значениями независимых переменных для неизвестных образцов, зависимые переменные для которых предстоит предсказать.

Запустите студенческий модуль и убедитесь в том, что необходимые для его работы компоненты успешно загружаются. Задайте имена выданных Вам файлов и выполните ячейки, которые считывают эти файлы. Рассмотрите график, содержащий спектры всех образцов, содержавшихся в обучающем наборе данных.

2.1. Метод главных компонент

Глядя на график значений сингулярных чисел и объяснённой дисперсии, а также на графики нагрузок, выберите количество компонент, максимально адекватно описывающее обучающий набор данных. «Лишние» компоненты будут выглядеть как шум, иметь низкие сингулярные числа и описывать пренебрежимо малую долю дисперсии. Полезные, содержащие информацию компоненты будут, наоборот, иметь форму спектральных пиков и объяснять значимую долю дисперсии.

2.2. Проекция на латентные структуры

Здесь будет проведена перекрёстная проверка моделей ПЛС-1 методом деления обучающего набора данных на 10 частей и расчёта средней ошибки предсказания по десяти поднаборам. Рассмотрите графики средней ошибки предсказания для обучающего набора и набора

9 https://stats.stackexchange.com/a/65156

перекрёстной проверки. Выберите минимальное количество компонент, для которого ошибка предсказания на перекрёстной проверке мала по сравнению с другими представленными значениями ошибки. В качестве альтернативного критерия можно предложить сравнивать значения ошибки предсказания для двух поднаборов: у недо- и как раз-обученной модели эти значения приблизительно совпадают, а у переобученной — ошибка на обучающем наборе значительно меньше, чем на наборе перекрёстной проверки.

2.3. Порядок представления результатов

1. График сингулярных чисел и доли дисперсии, объясняемой различными количествами главных компонентов.
2. Графики нагрузок для выбранного количества главных компонентов.
3. Графики ошибки предсказания на обучающем наборе и наборе перекрёстной проверки.
4. Графики коэффициентов регрессии и весов ПЛС для моделей для всех пяти компонент.

Используйте файл студенческого модуля, чтобы получить все необходимые графики.

2.4. Контрольные вопросы

1. Какие задачи решает многомерная градуировка?
2. Какие проблемы возникают при попытке решения задачи многомерной градуировки напрямую?
3. Какие методы решают задачу многомерной градуировки? Как именно они работают?
4. Каким образом можно определить наличие выбросов в новых порциях независимых переменных?
5. Чем отличаются регрессия на главных компонентах и регрессия на латентных структурах? Чем один из этих методов может быть принципиально лучше другого?
6. В каких ситуациях может нарушаться линейная зависимость между значениями спектра и предсказываемыми концентрациями?
7. Какие методы предварительной обработки данных можно применить, чтобы восстановить линейность зависимости между значениями спектра и предсказываемыми концентрациями?
8. Что такое гиперпараметры метода многомерной линейной регрессии? Приведите примеры.
9. Как правильно задавать значения гиперпараметров?

3. Рекомендуемая литература

[1] Родионова О. Е. Хемометрика в аналитической химии [Электронный ресурс] / О. Е. Родионова, А. Л. Померанцев // Хемометрика в России : [сайт]. – 61 с. : ил. –

Библиогр.: 245 назв. – URL: http://pca.narod.ru/chemometrics_review.pdf https://web.archive.org/web/20110815064403if_/http://www.chemometrics.ru:80/materials/articles/chemometrics_review.pdf (27.03.2023).

[2] А. Л. Померанцев. Калибровка (Градуировка) [Электронный ресурс] / А. Л. Померанцев // Российское хемометрическое общество : [сайт]. – URL: https://chemometrics.ru/ru/books/kalibrovka/ (27.03.2023).

[3] Родионова О. Е. Хемометрический подход к исследованию больших массивов химических данных // Рос. хим. журн. – 2006. – Т. 1, № 2. – С. 128-144 : ил. – Библиогр.: 195 назв. ; То же [Электронный ресурс]. – URL: http://www.chem.msu.ru/rus/jvho/2006-2/128.pdf (13.11.2011).

[4] Rasmus Bro, Age K. Smilde. Principal component analysis. // Analytical Methods. – 2014. – V. 6. – P. 2812-2831. – DOI: 10.1039/C3AY41907J.

[5] José Manuel Amigo. Data Mining. Machine Learning. Deep Learning. Chemometrics. Definitions, Common Points and Trends (Spoiler Alert: VALIDATE your models!). // Brazilian Journal of Analytical Chemistry. – 2021. – V. 8, № 32. – P. 45-61. – DOI: 10.30744/brjac.2179-3425.AR-38-2021.

[6] Oxana Rodionova, Sergey Kucheryavskiy, Alexey Pomerantsev. Efficient tools for principal component analysis of complex data — a tutorial. // Chemometrics and Intelligent Laboratory Systems. – 2021. – V. 213. – P. 104304. – DOI: 10.1016/j.chemolab.2021.104304

4. Приложение: Установка Python, scikit-learn и Jupyter Notebook / JupyterLab

Поскольку программирование на Python не является целью практикума, выполнять ПЛС-регрессию должно быть возможно любым удобным способом. Ниже можно найти краткие инструкции по установке и запуску среды программирования, которую можно использовать для выполнения ПЛС-регрессии на Python:

Установка Python

Для Windows и macOS: загрузите Python с официального сайта https://www.python.org/downloads/ (используйте «installer», а не «source tarball»). Для GNU/Linux и других POSIX-совместимых ОС: скорее всего, Python уже входит в поставку. Если он не установлен, используйте системный менеджер пакетов, чтобы его установить.

Подробнее см. Using Python on Windows, Unix platforms, Mac.

Виртуальные окружения

Если Python планируется использовать на одном и том же ПК для разных проектов, работа над которыми требует установки не совместимых между собой пакетов, решить эту проблему позволят виртуальные окружения. Пакеты, установленные внутри одного окружения, не затрагивают другие окружения, предотвращая возникновение конфликтов.

Подробнее см. Creation of virtual environments, PEP 405.

Установка пакетов для Python

Чтобы установить нужные пакеты, используйте командную строку:

```
python -m pip install -U scikit-learn jupyterlab matplotlib
```

Для запуска JupyterLab используйте команду `jupyter-lab`. Чтобы использовать Jupyter Notebook вместо JupyterLab, замените имя пакета `jupyterlab` на `notebook`, а команду запуска — на `jupyter notebook`. На GNU/Linux и других POSIX-совместимых ОС предварительно проверьте наличие scikit-learn и Jupyter в системном менеджере пакетов, поскольку установка оттуда, вероятно, пройдёт быстрее.

Подробнее см. Installing scikit-learn, Installing Jupyter.

Использование Anaconda

Сборку «Anaconda» от Anaconda Inc для некоторых версий Windows, Linux и macOS и некоторых процессоров можно загрузить по адресу https://www.anaconda.com/products/distribution. Она включает в себя Python, scikit-learn, JupyterLab и другие пакеты.

5. Приложение: Другие ресурсы, которые можно использовать для выполнения регрессии на латентных структурах

Графические интерфейсы:

- OriginPro ⩾ 2016, Windows, доступна бесплатная версия для студентов с почтовым адресом от Химического факультета МГУ [проверено 2023-02-09]
- Regression Toolbox for MATLAB, бесплатно (CC BY-NC-ND 4.0), требует MATLAB

Языки программирования:

- R, пакеты pls и mdatools, бесплатно (GPL-2 и MIT)
- MATLAB и Octave (бесплатно, GPL-3+), функция plsregress

Разное:

- Chemometrics Add-In для Excel