Федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»
Факультет математики

Гончаренко Аркадий Александрович

Выбор ковариаций

Выпускная квалификационная работа — бакалаврская работа
по направлению подготовки 01.03.01 — Математика,
образовательная программа «Математика»

Рецензент:
доктор экономических наук,
заведующий кафедрой моделирования и системного анализа факультета информационных технологий и анализа больших данных
Афанасьев Антон Александрович

Научный руководитель:
доктор физико-математических наук,
профессор
Горбунов Василий Геннадьевич

Коруководитель:
PhD,
Вайнберг Аллен Анна Львовна

Москва 2025

Аннотация

В этой работе я разберу метод анализа многомерного нормального распределения, придуманный Демпстером в 1972 году, обсужу его ограничения, и применю его к реальным данным.

1 Введение

Две переменные X и Y называются условно независимыми относительно величины Z, если при любом фиксированном $Z = z_0$ их условные распределения независимы.

При анализе многомерных данных полезно понимать, какие переменные условно независимы относительно других переменных, так как условная независимость означает, что переменные связаны между собой лишь посредственно, через другие переменные.

Демпстер в 1972 году[1] изобрел метод вычисления условных независимостей для многомерного нормального распределения. Этот метод строит граф условных зависимостей переменных, предполагая, что все пары переменных, чьи вершины в графе не соединены ребром, условно независимы относительно всех остальных переменных распределения.

2 Отбор ковариаций

2.1 Введение переменных

Многомерное нормальное распределение $X = (x_1, x_2, \ldots, x_d)$ принадлежит экспоненциальному семейству распределений. Оно однозначно параметризовано его средним и ковариационной матрицей $\Sigma = (\sigma_{ij})$.

Полезное свойство многомерного нормального распределения состоит в том, что две переменные x_i и x_j этого распределения условно независимы тогда и только тогда, когда соответствующий элемент (σ^{ij}) матрицы $\Sigma^{-1} = (\sigma^{ij})$, обратной к ковариационной, равен нулю.

Для краткости обозначений я буду предполагать, что среднее распределения равно нулю.

Тогда плотность d-мерного распределения можно представить в форме

$$f(x, \Phi) = exp(\phi + t(x) + \phi_1 t_1(x) + \phi_2 t_2(x) + \ldots + \phi_r t_r(x)), \tag{1}$$

где $r = \frac{d(d-1)}{2},\ t(x) = 0,\ \phi = \frac{1}{2}(-d\log(2\pi) - \log(|\Sigma|)),\ \Phi = (\phi_1, \ldots, \phi_r) =$
$= (\sigma^{11}, \ldots, \sigma^{1d}, \sigma^{22}, \ldots, \sigma^{2d}, \ldots, \sigma^{dd}),\ t_1 = -\frac{x_1^2}{2},\ t_2 = -x_1x_2,\ \ldots,\ t_d = -x_1x_d,$
$t_{d+1} = -\frac{x_2^2}{2},\ \ldots,\ t_r = -\frac{x_d^2}{2}.$

Введем также

$$\theta_i = \int t_i(x) f(x, \Phi) dx. \tag{2}$$

Тогда $\Theta = (\theta_1, \ldots, \theta_r) = (-\frac{\sigma_{11}}{2}, -\sigma_{12}, \ldots, -\frac{\sigma_{dd}}{2})$

Таким образом, между множествами возможных Σ, Θ и Φ многомерного нормального распределения существует биекции, отвечающие конкретному нормальному распределению (с точностью до среднего). Другими словами, Σ, Θ и Φ однозначно определяют друг друга.

Пусть $\Gamma = (\gamma_{ij})$, $\gamma_{ij} = \int (t_i(x) - \theta_i)(t_j(x) - \theta_j) f(x, \Phi) dx$.

Это равно ковариации $\mathrm{cov}(x_{i_1} x_{i_2}, x_{j_1} x_{j_2}) = \sigma_{i_1,j_1}\sigma_{i_2,j_2} + \sigma_{i_1,j_2}\sigma_{i_2,j_1}$, умноженной на $1/2$, если только $i_1 = i_2$, или только $j_1 = j_2$, и на $1/4$, если оба условия выполняются.

Лемма. $\gamma_{ij} = \partial\theta_i / \partial\phi_j$.

◀ Так как f – плотность вероятности, $\int f(x, \Phi) dx = 1$. Выразим $f(x, \Phi)$ через (1) и возьмем дифференциал обоих частей равенства по Φ. Получим

$$d\phi + \left(\int t_1(x) f(x, \Phi) dx \right) d\phi_1 + \ldots + \left(\int t_r(x) f(x, \Phi) dx \right) d\phi_r = 0 \tag{3}$$

Выразим $f(x, \Phi)$ в (2) через (1) и возьмем дифференциал обоих частей равенства по Φ. Получим

$$\left(\int t_i(x) f(x, \Phi) dx \right) d\phi + \left(\int t_i(x) t_1(x) f(x, \Phi) dx \right) d\phi_1 + \ldots + \\ + \left(\int t_i(x) t_r(x) f(x, \Phi) dx \right) d\phi_r = d\theta_i \tag{4}$$

Умножим обе части (3) на $\int t_i(x) f(x, \Phi) dx$ и вычтем их из (4). Благодаря (2) получим

$$\gamma_{i1} d\phi_1 + \ldots + \gamma_{ir} d\phi_r = d\theta_i \tag{5}$$

Из (5) следует $\gamma_{ij} = \partial\theta_i / \partial\phi_j$. ▶

2.2 Отбор ковариаций

Алгоритм начинает с пустого графа, в котором все пары переменных считаются условно независимыми относительно всех остальных.

Это дает строгую структуру данных – все переменные независимы – но логарифм правдоподобия у матриц, подчиняющихся такой структуре данных, невелик по сравнению с другими матрицами ковариации.

Затем алгоритм добавляет по ребру, находит матрицу ковариации с максимальным логарифмом правдоподобия среди матриц, подчиняющихся структуре графа, и проверяет эффективность добавления этого ребра. Если логарифм правдоподобия получившейся матрицы значимо увеличился по сравнению с предыдущим, то мы его оставляем в графе. Иначе мы его выбрасываем и выбираем другое ребро, пока ребра не закончатся.

В итоге получается структура, дающая нам полезную информацию за счет ограничений на переменные, снятие которых не даст нам ощутимого улучшения в логарифме правдоподобия.

2.3 Поиск ковариационной матрицы

Демпстер в [1] доказал следующее утверждение:

Теорема. Пусть $\hat{\Sigma}$ – симметричная матрица, совпадающая с выборочной матрицей ковариаций на диагонали и парах коэффициентов, отвечающих ребрах графа, обратная $\hat{\Sigma}^{-1}$ к которой равна 0 на всех парах разных коэффициентов, не отвечающим ребрам графа. Пусть она положительно определена. Тогда

1. $\hat{\Sigma}$ существует и единственна, если существует положительно определенная симметричная матрица, совпадающая с выборочной матрицей ковариаций на диагонали и парах коэффициентов, отвечающих ребрах графа.

2. Среди матриц, совпадающих с выборочной матрицей ковариаций на диагонали и парах коэффициентов, отвечающих ребрах графа, $\hat{\Sigma}$ имеет максимальную энтропию.

3. Среди ковариационных матриц, обратные к которым равны 0 на всех парах разных коэффициентов, не отвечающим ребрам графа, $\hat{\Sigma}$ является максимально правдоподобной оценкой.

Таким образом, $\hat{\Sigma}$ является оценкой максимального правдоподобия среди матриц с соответствующими графу условными независимостями.

Искать $\hat{\Sigma}$ мы будем с помощью метода Ньютона для нелинейных уравнений, аналогичному градиентному спуску в машинном обучении.

Разделим Φ на (Φ_A, Φ_B), где Φ_B – все компоненты, которые соответствуют парам разных вершин без ребер между ними, соответственно Φ_B приравняем нулю. Аналогично разделим Θ на (Θ_A, Θ_B), соответственно будем использовать алгоритм, который приближает Θ_A к значениям, соответствующим выборочной матрице. Аналогично разделим матрицу $\Gamma = (\gamma_{ij})$ на 4 подматрицы Γ_{AA}, Γ_{AB}, Γ_{BA} и Γ_{BB}:

$$\Gamma = \begin{bmatrix} \Gamma_{AA} & \Gamma_{AB} \\ \Gamma_{BA} & \Gamma_{BB} \end{bmatrix}$$

Пусть Σ_i – матрица на i-м шаге, Φ_A^i и Θ_A^i – соответствующие ей векторы, Γ^i – соответствующая ей матрица Γ. Пусть Θ_A^0 – значение Θ_A, соответствующее выборочной матрице. Пользуясь Леммой, разложим Θ_A в ряд Тейлора как функцию от Φ_A:

$$\Theta_A^0 = \Theta_A^i + \Gamma_{AA}^i(\Phi_A^0 - \Phi_A^i) + o(\Phi_A^0 - \Phi_A^i) \tag{6}$$

Применим метод Ньютона к отображению $\Theta_A(\Phi_A)$, чтобы найти точку, в которой $\Theta_A = \Theta_A^0$: $\Phi_A^i + s$ возьмем в качестве Φ_A^{i+1}, где s определяется уравнением $\Theta_A^0 - \Theta_A^i = \Gamma_{AA}^i s$. Пользуясь тем, что $\Phi^{i+1} = \Sigma^{-1}$ для соответствующего распределения, вычислим Θ_A^{i+1} и Γ_{AA}^{i+1}, соответствующие Φ^{i+1}.

Будем повторять процесс, пока изменение s не станет достаточно маленьким, что произойдет, когда Θ_A^i будет находиться в окрестности Θ_A^0. Получим приближение $\hat{\Sigma}$.

2.4 Проверка значимости

Точных тестов значимости не существует, но в [1] предлагается χ^2- тест с одной степенью свободы для умноженного на удвоенный размер выборки изменения в логарифме правдоподобия.

Предложение. Логарифм правдоподобия равен $-\frac{1}{2}(d\log(2\pi)+\log(|\Sigma|)+tr(\hat{\Sigma}\Sigma^{-1}))$

◀ $\log(f(x,\Phi))$ по (1) равно

$$\phi + t(x) + \phi_1 t_1(x) + \phi_2 t_2(x) + \ldots + \phi_r t_r(x) = \frac{1}{2}(-d\log(2\pi) - \log(|\Sigma|)) + \Phi t(x) =$$
$$= \frac{1}{2}(-d\log(2\pi) - \log(|\Sigma|)) - \frac{1}{2}\sum_{i,j=1}^{d}(\Sigma^{-1}_{i,j}\hat{\Sigma}_{i,j}) = -\frac{1}{2}(d\log(2\pi) + \log(|\Sigma|) + tr(\hat{\Sigma}\Sigma^{-1})) \blacktriangleright \tag{7}$$

3 Ограничения метода

Метод выбора ковариации работает только на многомерном нормальном распределении. Он использует то, что условная независимость выражается нулем в матрице, обратной матрице ковариации.

Теоретически вышеописанная теорема имеет аналог для других непрерывных многомерных распределений экспоненциального семейства, так что возможно выработать свой метод поиска аналога матрице ковариации в многомерном нормальном распределении, но другие распределения сильно менее распространены, и проблема поиска этого аналога(и его существования) является совсем другой проблемой. Я доказал возможность применения этого метода к распределению Дирихле, например, но понял, что там вопрос относительной независимости переменных не востребован, насколько я понимаю.

Поэтому для обоснованного применения этого метода нужно проверить многомерную нормальность данных. Тут я руководствовался статьей[2], и написал BHEP-тест, придуманный для одномерного случая (Epps, Pulley)[3] в 1983 и обобщенный для многомерного (Baringhaus, Henze)[4] в 1988. Этот тест сравнивает эмпирическую характеристическую функцию преобразованной выборки и сравнивает её с характеристической функцией многомерного нормального распределения(преобразование выборки приводит её к стандартному виду).

4 Эксперимент на реальных данных

В качестве реальных данных я использовал один из файлов метеорологических данных Нидерландов[5].

Отбор переменных

Для того, чтобы BHEP-тест работал, мне нужно было сделать все переменные примерно равными численно и отбросить линейно зависимые переменные (я это определял по маленькому детерминанту матриц ковариации пар или троек переменных). Затем я отбрасывал переменные, плохо проходящие BHEP-тест нормального распределения, пока BHEP-тест не потвердил нормальность выборки.

В итоге у меня остались следующие переменные:

- wd^{av}, $wd^{\min}$, $wd^{\max}$, wd^{σ} – Среднее, минимум, максимум и стандартное отклонение направления ветра за последние 10 минут (если происходила резкая смена направления или скорости ветра, станции отбрасывали все, кроме последних 2 минут и считали от этого);

- $ws^{\max}$, ws^{σ} – максимум и стандартное отклонение скорости ветра за последние 10 минут(с той же оговоркой);
- p_1^{sea} – среднее атмосферное давление на уровне моря за последнюю минуту;
- p_{10} – среднее атмосферное давление на уровне мет. станции за последние 10 минут;
- $rd_{10}^{\min}$, $rd_{10}^{\max}$, rd_h^{sum}, rd_d^{sum} – минимум и максимум солнечного излучения за последние 10 минут, а также суммарное излучение за последний час и за последние 24 часа;
- rh_1 – средняя относительная влажность за последнюю минуту
- $ws_{6h}^{\max}$ – максимальная скорость ветра за последние 6 часов
- sd – продолжительность солнечного сияния;
- $wg_{6h}^{\max}$ – Максимум порывов ветра за последние 6 часов;
- gt_{10}^{av}, $gt_{10}^{\min}$, $t_{12h}^{\min}$ – средняя и минимальная температура воздуха за последние 10 минут и минимальная температура воздуха за последние 12 часов.

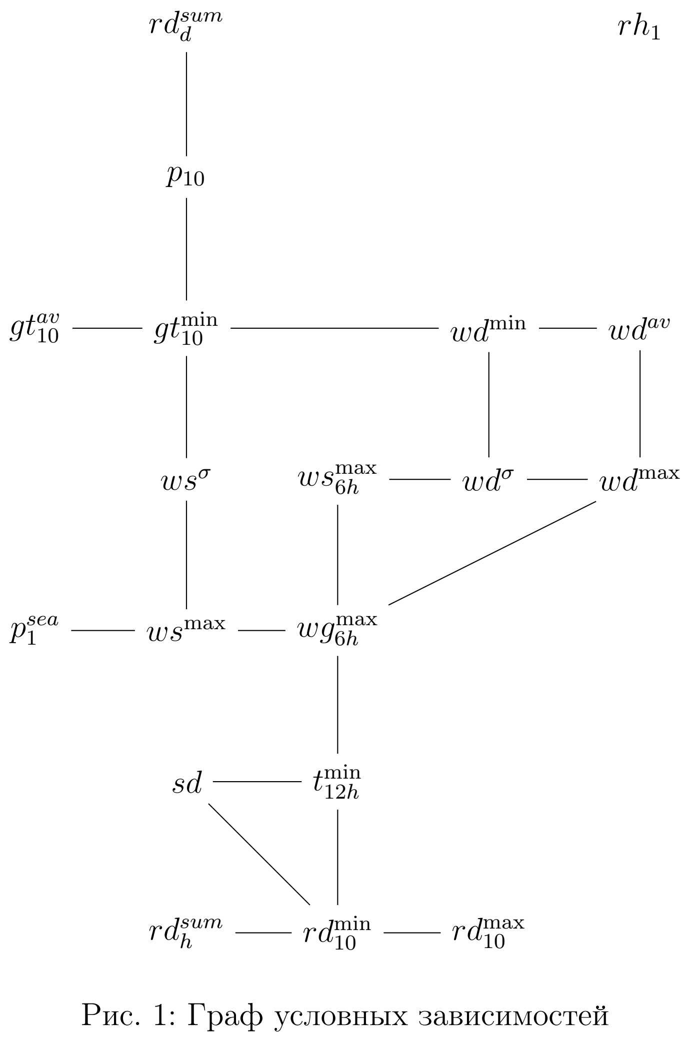

Рис. 1: Граф условных зависимостей

Всего получается 19 переменных и 22 станции, которые их зарегистрировали. Запуская алгоритм, получаем граф условных зависимостей (Рис. 1).

4.1 Интерпретация результатов

Я не сильно знаком с метрологией, но могу заметить следующее:

- В четверке переменных, отвечающих за направление ветра, стандартное отклонение и среднее, а также минимум и максимум условно независимы относительно другой пары, что логично. Например, отклонение и среднее задают распределение, соответственно максиму от минимума при заданном распределении не зависят.

- То, что максимум направления ветра связано со скоростью порывов ветра, означает, что с того направления дуют более резкие ветра – иначе для порывов ветра было бы важнее среднее направление, а не крайние значения.

- То, что максимальная скорость ветра за последние 6 часов связана с максимальной скоростью порывов ветра за последние 6 часов, и лишь через неё связана с максимальной скоростью ветра за последние 10 минут, означает, что для максимальной скорости ветра за последние 6 часов лучшим кратковременным (10-минутным) показателем служит не скорость ветра, а разброс направления ветра, с которым она непосредственно связана. Конечно, это может быть лишь причудой этой конкретной выборки, например, связанной со временем дня.

- Суммарное солнечное излучение за час связано непосредственно только с минимумом солнечного излучения за 10 минут. Возможно, минимальное солнечное излучение является хорошим показателем качества погоды в кратком диапазоне, что отражается на суммарном солнечном излучении за час. Та же обособленность есть и у максимума излучения за 10 минут, возможно, потому что самое ясное небо за 10 минут не зависит от каких-то факторов помимо хмурости неба, которая выражается минимум солнечной радиации.

Это всё, конечно, гипотезы, требующие отдельных исследований, и выборка у меня небольшая, но этим примером хорошо демонстрируется возможность анализа связей между переменными по графу условных зависимостей.

5 Код

Код я выложил в [6]. Там же лежит курсовая прошлого года, на которой я строил эту работу.

6 Список литературы

[1] Dempster, A. P. (1972). Covariance Selection. Biometrics, 28(1), 157–175.
[2] Norbert Henze (2002). Invariant tests for multivariate normality: a critical review. , 43(4), 467–506. doi:10.1007/s00362-002-0119-6

[3] T. W. EPPS, LAWRENCE B. PULLEY, A test for normality based on the empirical characteristic function, Biometrika, Volume 70, Issue 3, December 1983, Pages 723–726, https://doi.org/10.1093/biomet/70.3.723
[4] Baringhaus, L., Henze, N. A consistent test for multivariate normality based on the empirical characteristic function. Metrika 35, 339–348 (1988). https://doi.org/10.1007/BF02613322
[5] https://dataplatform.knmi.nl/dataset/preview/actuele10mindataknmistations-2
[6] https://github.com/grigonal/Covariance_Selection
[7] А. Л. Вайнберг Аллен. Графы для анализа структурных соотношений между переменными и их приложение к изучению российских регионов (Части 1 и 2), Прикладная эконометрика, №2 (10) и №4 (12),2008.