

Лекции, начиная с MANOVA

Собрано 24 ноября 2017 г. в 00:13

1 Лекция 23.11.2017

1.1 MANOVA. Модель, запись через условные мат. ожидания и мат.ожидания. Разложение ковариационной матрицы. Критерия Лямбда Уилкса.

Этот билет во многом аналогичен текста про ANOVA. Основная разница заключается лишь в том, что здесь η является многомерным вектором.

Общая постановка задачи. Пусть есть k групп и p признаков, мы пытаемся проверить, что группы друг от друга не отличаются. Формально есть k (многомерных) случайных векторов $\eta_1, \dots, \eta_k \in \mathbb{R}^p$ и гипотеза формулируется так:

$$H_0 : \mathcal{P}(\eta_1) = \mathcal{P}(\eta_2) = \dots = \mathcal{P}(\eta_k). \quad (1)$$

Эту задачу можно переформулировать следующим образом. Пусть есть дискретный (м.б. качественный) признак ξ , принимающий ровно k значений: A_1, A_2, \dots, A_k . Рассмотрим случайный вектор $(\eta, \xi)^T \in \mathbb{R}^k \times \{A_1, A_2, \dots, A_k\}$, такой что $\mathcal{P}(\eta | \xi = A_i) = \mathcal{P}(\eta_i)$ для всех $i \in 1 : k$. Тогда гипотеза (??) переписывается в виде:

$$H_0^* : \mathcal{P}(\eta | \xi = A_i) = \mathcal{P}(\eta | \xi = A_j) \text{ для всех } i, j. \quad (2)$$

В MANOVA рассматривается частный случай (модель), когда $\mathcal{P}(\eta_i) = \mathcal{N}(\mu_i, \Sigma)$. В рамках этой модели гипотезам (1) и (2) соответствуют следующие две равносильные гипотезы:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k. \quad (3)$$

$$H_0^* : \mathbb{E}(\eta | \xi = A_1) = \dots = \mathbb{E}(\eta | \xi = A_k).$$

Прежде чем сформулировать гипотезу в рамках Основного Дисперсионного Тождества (??) запишем его на генеральном языке в текущей модели.

$$\text{Cov}\eta = \mathbb{E}[(\eta - \mathbb{E}\eta)(\eta - \mathbb{E}\eta)^T] = \quad (4)$$

$$= \mathbb{E}[(\eta - \mathbb{E}(\eta | \xi))(\eta - \mathbb{E}(\eta | \xi))^T] + \mathbb{E}[(\mathbb{E}(\eta | \xi) - \mathbb{E}\eta)(\mathbb{E}(\eta | \xi) - \mathbb{E}\eta)^T]. \quad (5)$$

Теперь уже можно легко заметить, что в рамках Основного Дисперсионного Тождества гипотезу можно сформулировать следующим образом:

$$H_0 : \mathbb{E}[(\mathbb{E}(\eta | \xi) - \mathbb{E}\eta)(\mathbb{E}(\eta | \xi) - \mathbb{E}\eta)^T] = 0. \quad (6)$$

Эта запись является многомерным обобщением (??) и так же означает, что средние внутри групп не отличаются от общего среднего.

Выборка задается так же как в ANOVA, но теперь индивиды многомерные. Имеется n_i индивидов из i -той группы ($i \in 1 : k$). Обозначим $y_{ij} \in \mathbb{R}^p$ — j -того индивида из i -той группы ($i \in 1 : k$ и $j \in 1 : n_i$). Обозначим \bar{y} — выборочное среднее по всем индивидам, а \bar{y}_i — выборочное среднее индивидов i -той группы.

Запишем (4) на выборочном языке:

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})(y_{ij} - \bar{y})^T = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)(y_{ij} - \bar{y}_i)^T + \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})(\bar{y}_i - \bar{y})^T. \quad (7)$$

Должно быть ясно, что это лишь многомерное обобщение (??).

Следуя обобщенным обозначениям, предложенным в предварительных материалах про ANOVA и регрессию, можно ввести следующие:

$$\mathbf{E} \stackrel{\text{def}}{=} \mathbb{E}[(\boldsymbol{\eta} - \mathbb{E}(\boldsymbol{\eta} | \xi))(\widehat{\boldsymbol{\eta}} - \mathbb{E}(\boldsymbol{\eta} | \xi))^T] = \sum_{i=1}^k \sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)^T;$$

$$\mathbf{H} \stackrel{\text{def}}{=} \mathbb{E}[(\mathbb{E}(\boldsymbol{\eta} | \xi) - \widehat{\mathbb{E}\boldsymbol{\eta}})(\mathbb{E}(\boldsymbol{\eta} | \xi) - \mathbb{E}\boldsymbol{\eta})^T] = \sum_{i=1}^k n_i (\bar{\mathbf{y}}_i - \bar{\mathbf{y}})(\bar{\mathbf{y}}_i - \bar{\mathbf{y}})^T.$$

Стоит сразу отметить, что, когда в дальнейшем мы будем строить критерии для проверки значимости, мы всегда будем их строить, как некие преобразования над парой матриц (\mathbf{H}, \mathbf{E}) .

Теорема. Пусть матрица $\mathbf{H} \sim \mathcal{W}_p(\Sigma, \nu_{\mathbf{H}})$ и $\mathbf{E} \sim \mathcal{W}_p(\Sigma, \nu_{\mathbf{E}})$, причем \mathbf{H} и \mathbf{E} независимы. Потребуем так же, чтоб $\nu_{\mathbf{H}} < \nu_{\mathbf{E}}$ и $\nu_{\mathbf{E}} > p$. Обозначим λ_i — i -тое по упорядоченности по невозрастанию собственное число матрицы $\mathbf{E}^{-1}\mathbf{H}$. Тогда

$$t = \frac{|\mathbf{E}|}{|\mathbf{H} + \mathbf{E}|} = \frac{1}{|1 + \mathbf{E}^{-1}\mathbf{H}|} = \prod_{i=1}^s \frac{1}{1 + \lambda_i} \sim \Lambda_p(\nu_{\mathbf{H}}, \nu_{\mathbf{E}}), \quad (8)$$

где $s = \min(\nu_{\mathbf{H}}, p)$.

Таким образом, (8) дает нам статистику критерия, где в случае дискриминантного анализа $\nu_{\mathbf{H}} = k - 1$, $\nu_{\mathbf{E}} = n - k$. В случае многомерной множественной регрессии $\nu_{\mathbf{H}} = k$, $\nu_{\mathbf{E}} = n - k - 1$.

Значение s — это граница сверху для числа ненулевых чисел, которое определяется рангом матрицы \mathbf{H} . Во-первых, ранг не превосходит размерности p матрицы \mathbf{H} . Во-вторых, для дисперсионного анализа матрица \mathbf{H} соответствует ковариационной матрице, построенной на основе k индивидов с весами, пропорциональными размерам групп. Поэтому ее ранг не может быть больше k . А так как данные исходно центрированы, то ранг не превосходит $k - 1$.

Заметим, что построенный критерий для проверки соответствующей гипотезы является далеко не единственным. О построении других критериев речь пойдет дальше.

1.2 Про дискриминантный анализ целом

Постановка задачи, отличие от кластерного анализа.

Задачи - построить классифицирующее правило, объяснить его (интерпретация), feature selection, feature extraction.

Первый вопрос - а вообще группы отличаются?

Рассмотрим модель MANOVA, т.е. распределения в группах нормальные и отличаются только векторами средних. Тогда MANOVA используется для проверки значимости отличия. (гипотеза о том, что группы не отличаются.) В этом случае получаем модель линейного дискриминантного анализ (LDA).

Посмотрим на MANOVA с точки зрения отличия групп в LDA и того, как увидеть это отличие.

1.3 Канонические дискриминантные функции (коэффициенты) и переменные (feature extraction)

Напомним, что имеется p признаков и k групп: $\boldsymbol{\eta} \in \mathbb{R}^p$, причем $\mathcal{P}(\boldsymbol{\eta} | \xi = A_i) = \mathcal{N}(\boldsymbol{\mu}_i, \Sigma)$. Попытаемся на основе имеющихся p признаках построить новые, по которым

группы «наиболее бы отличались», причем было бы удобно, чтобы эти новые признаки были ортогональны. Эта неформальная задача на самом деле очень просто формализуется, как мы увидим несколько далее.

Запишем на генеральном языке, что значит «новый признак». Есть $A \in \mathbb{R}^p$ и новый признак $\zeta = A^T \boldsymbol{\eta}$: $\mathcal{P}(\zeta | \xi = A_i) = \mathcal{N}(A^T \boldsymbol{\mu}_i, A^T \Sigma A)$. На выборочном языке — $Z = \mathbf{Y}A$ и выборочная ковариационная матрица (с точностью до коэффициента имеет вид): $A^T \mathbf{Y}^T \mathbf{Y} A = A^T (\mathbf{H} + \mathbf{E}) A$. Из этого следует, что «аналогом» \mathbf{H} для нового признака является $A^T \mathbf{H} A$, а аналогом \mathbf{E} является $A^T \mathbf{E} A$.¹

Посмотрим теперь на F -статистику, следуя обобщенному определению статистики, данному в ANOVA:

$$F = F(A) = C \frac{SSH_\zeta}{SSE_\zeta} = C \frac{A^T \mathbf{H} A}{A^T \mathbf{E} A} \sim F_{\nu_{\mathbf{H}}, \nu_{\mathbf{E}}}.$$

где $C = \nu_{\mathbf{E}} / \nu_{\mathbf{H}}$ — коэффициент, не зависящий от A .

Исходно ставилась задача — найти такие признаки, по которым группы «наиболее бы отличались», причем желательно, чтобы признаки были ортогональны. В терминах максимизации статистики F — это означает, что решается обобщенная задача на собственные числа и собственные вектора.

$$\frac{A^T \mathbf{H} A}{A^T \mathbf{E} A} \rightarrow \max_A.$$

Обозначим собственные числа матрицы $\mathbf{E}^{-1} \mathbf{H}$ в порядке невозрастания: $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_s$, где $s = \min\{p, \nu_{\mathbf{H}}\}$, и собственные вектора той же матрицы — A_i для $i \in 1 : s$. Причем $A_i^T \mathbf{E} A_j = 0$ для $j \neq i$. Вектора A_i — канонические коэффициенты (дискриминантные функции). Новые признаки $Z_i = \mathbf{Y} A_i$ — канонические переменные. Можно показать, новые признаки являются ортогональными: $Z_i^T Z_j = 0$ для $j \neq i$.

Теорема.

$$\forall i \neq j \in 1 : s \quad Z_i \perp Z_j$$

Доказательство. $\langle Z_i, Z_j \rangle = \langle \mathbf{Y} A_i, \mathbf{Y} A_j \rangle = A_i^T \mathbf{Y}^T \mathbf{Y} A_j = A_i^T (\mathbf{H} + \mathbf{E}) A_j = A_i^T \mathbf{E} (\mathbf{I} + \mathbf{E}^{-1} \mathbf{H}) A_j$. Так как $\forall \mathbf{C} \geq 0$ с с. ч. $\lambda_1, \dots, \lambda_s$ с. ч. матрицы $\mathbf{I} + \mathbf{C}$ — это $1 + \lambda_1, \dots, 1 + \lambda_s$, а собственные вектора те же самые, $(\mathbf{I} + \mathbf{E}^{-1} \mathbf{H}) A_j = (1 + \lambda_j) A_j$ по определению собственного вектора. Так как собственные вектора разных с.ч. \mathbf{E} -ортогональны, $\langle Z_i, Z_j \rangle = (1 + \lambda_j) A_i^T \mathbf{E} A_j = 0$. \square

Исходя из описания, смысл λ_i — степень различия групп по i -тому направлению, задаваемом i -той дискриминантной функцией. A_i — коэффициенты с которыми нужно взять исходные признаки, чтобы получить новый признак, соответствующий наибольшей разнице групп и ортогональный предыдущим. А сами канонические переменные тем самым — ортогональные новые признаки, по которым соответствующий наибольшее различие равен λ_i .

1.4 MANOVA и значимость LDA. Разные критерии, чем отличаются.

Использование MANOVA для проверки значимости дискриминантного анализа совершенно естественно. Один из критериев 1.1 уже был построен — это критерий Лямбда Уилкса.

¹Чтобы в этом убедиться, нужно написать оценкой чего является \mathbf{H} и \mathbf{E} . Дальше домножить это на A нужным образом и ввести обозначение для нового признака. Дальше все станет ясно.

На самом деле это далеко не единственный критерий, который может быть построен, в частности, для проверки значимости дискриминантного анализа. Как уже отмечалось, главными объектами, с которыми приходится работать, являются матрицы \mathbf{H} и \mathbf{E} . На основе разных способов «совмещения» этих объектов могут получаться разные критерии. Рассмотрим несколько способов. Напомним, что $s = \min\{\nu_{\mathbf{H}}, p\}$.

Проверяется гипотеза о равенстве средних²

$$H_0 : \mathbb{E}\eta_1 = \mathbb{E}\eta_2 = \dots = \mathbb{E}\eta_k.$$

1. Lambda Wilks's

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}|} = \frac{1}{|1 + \mathbf{E}^{-1}\mathbf{H}|} = \prod_{i=1}^s \frac{1}{1 + \lambda_i} \sim \Lambda_p(\nu_{\mathbf{H}}, \nu_{\mathbf{E}}).$$

2. Roy's largest root

$$Q = \frac{\lambda_1}{1 + \lambda_1} = r_1^2.$$

Этот тест использует только первое собственное число матрицы $\mathbf{E}^{-1}\mathbf{H}$. Напомним, что r_i^2 — называют i -тым каноническим корнем. На самом деле r_i — так называемые канонические корреляции³. Потому они на самом деле имеют интерпретируемый смысл.

3. Pillai's:

$$V^{(s)} = \text{tr}(\mathbf{H}(\mathbf{E} + \mathbf{H})^{-1}) = \sum_{i=1}^s r_i^2.$$

4. Hotelling:

$$U^{(s)} = \text{tr}(\mathbf{E}^{-1}\mathbf{H}) = \sum_{i=1}^s \lambda_i.$$

Как всегда, когда мы видим много критериев для проверки одной гипотезы, нужно научиться их сравнивать. Сравнивают критерии по мощности. Но, чтобы говорить о мощности нужно фиксировать альтернативу. Таким образом, вопрос можно поставить следующим образом: «Для каких альтернатив предложенные критерии более мощные?».

Посмотрим на это на примере первых двух критериев. Первый критерий включает в статистику все (!) направления, по которым разброс максимален (иными словами, все λ_i). Второй же критерий включает в себя лишь первое направление, с максимальной степенью разброса, измеряемой λ_1 . Из этого становится ясно, что если у нас на самом деле лишь одно направление определяет разброс (например, k шариков лежат на одной прямой), то остальные λ_i при $i > 2$ уже не отображают различие между группами и потому в такой ситуации следует ожидать, что второй критерий окажется мощнее первого. В обратном случае, когда все направления описывают различие между группами, следует ожидать, что первый критерий окажется мощнее второго.

Про третий и четвертый критерий в литературе утверждается, что они «где-то по середине» между первым и вторым.

В заключение заметим, что у всех критериев кроме первого критическая область находится «слева» и около 0 (все статистики неотрицательны), у первого критерия все наоборот: носитель статистики от 0 до 1 и критическая область расположена около 1.

²На самом деле о равенстве распределений в рамках нормальной модели при условии гомоскедастичности.

³ Позднее, если успею, расскажу про то, между чем это корреляции.

1.5 Частный случай — ошибки сферические

Пусть $\mathbf{Y} \in \mathbb{R}^{n \times p}$ — матрица данных (строки \mathbf{y}_k , $k \in 1:n$ (или \mathbf{y}_{ij} в другой нумерации по группам) — наблюдения, столбцы Y_j , $j \in 1:p$ — признаки), наблюдения принадлежат к одной из k групп, в каждой группе n_i наблюдений, $n = \sum_{i=1}^k n_i$. Матрица межгрупповых отклонений $\mathbf{H} = \sum_{i=1}^k n_i (\bar{\mathbf{y}}_i - \bar{\mathbf{y}})(\bar{\mathbf{y}}_i - \bar{\mathbf{y}})^T$, матрица внутригрупповых отклонений $\mathbf{E} = \sum_{i=1}^k \sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)^T$.

В случае, если $\mathbf{E} = \sigma^2 \mathbf{I}$ (все группы сферические и одинакового размера), дискриминантные функции данных, т.е. A_1, \dots, A_s — собственные вектора $\mathbf{E}^{-1} \mathbf{H}$, становятся собственными векторами матрицы $\sigma^2 \mathbf{H}$. То есть дискриминантные функции являются собственными векторами матрицы межгрупповых ковариаций. Так они совпадают с главными направлениями в терминах анализа главных компонент, если каждую группу рассматривать как одного индивида (в АГК главные направления — это собственные вектора ковариационной матрицы), а дискриминантные переменные $\mathbf{Y} A_i$ при такой постановке являются главными компонентами.