# Лекции, начиная с MANOVA

Собрано 1 декабря 2017 г. в 22:52

# 1 Лекция 23.11.2017

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

Этот билет во многом аналогичен текста про ANOVA. Основная разница заключается лишь в том, что здесь $\eta$ 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 $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_k \in \mathbb{R}^p$ и гипотеза формулируется так:

$$\mathrm{H}_0 : \mathcal{P}(\boldsymbol{\eta}_1) = \mathcal{P}(\boldsymbol{\eta}_2) = \ldots = P(\boldsymbol{\eta}_k). \tag{1}$$

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

$$\mathrm{H}_0^* : \mathcal{P}(\boldsymbol{\eta} \,|\, \xi = A_i) = \mathcal{P}(\boldsymbol{\eta} \,|\, \xi = A_j) \text{ для всех } i, j. \tag{2}$$

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

$$\begin{gathered} \mathrm{H}_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_k. \\ \mathrm{H}_0^* : \mathbb{E}(\boldsymbol{\eta} \,|\, \xi = A_1) = \ldots = \mathbb{E}(\boldsymbol{\eta} \,|\, \xi = A_k). \end{gathered} \tag{3}$$

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

$$\mathrm{Cov}\boldsymbol{\eta} = \mathbb{E}\big[(\boldsymbol{\eta} - \mathbb{E}\boldsymbol{\eta})(\boldsymbol{\eta} - \mathbb{E}\boldsymbol{\eta})^{\mathrm{T}}\big] = \tag{4}$$

$$= \mathbb{E}\big[(\boldsymbol{\eta} - \mathbb{E}(\boldsymbol{\eta} \,|\, \xi))(\boldsymbol{\eta} - \mathbb{E}(\boldsymbol{\eta} \,|\, \xi))^{\mathrm{T}}\big] + \mathbb{E}\big[(\mathbb{E}(\boldsymbol{\eta} \,|\, \xi) - \mathbb{E}\boldsymbol{\eta})(\mathbb{E}(\boldsymbol{\eta} \,|\, \xi) - \mathbb{E}\boldsymbol{\eta})^{\mathrm{T}}\big]. \tag{5}$$

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

$$\mathrm{H}_0 : \mathbb{E}\big[(\mathbb{E}(\boldsymbol{\eta} \,|\, \xi) - \mathbb{E}\boldsymbol{\eta})(\mathbb{E}(\boldsymbol{\eta} \,|\, \xi) - \mathbb{E}\boldsymbol{\eta})^{\mathrm{T}}\big] = 0. \tag{6}$$

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 $\mathbf{y}_{ij} \in \mathbb{R}^p$— $j$-того индивида из $i$-той группы ($i \in 1 : k$ и $j \in 1 : n_i$). Обозначим $\overline{\mathbf{y}}$ — выборочное среднее по всем индивидам, а $\overline{\mathbf{y}_i}$ — выборочное среднее индивидов $i$-той группы.

Запишем (4) на выборочном языке:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{y}_{ij} - \overline{\mathbf{y}})(\mathbf{y}_{ij} - \overline{\mathbf{y}})^{\mathrm{T}} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{y}_{ij} - \overline{\mathbf{y}_i})(\mathbf{y}_{ij} - \overline{\mathbf{y}_i})^{\mathrm{T}} + \sum_{i=1}^{k} n_i(\overline{\mathbf{y}_i} - \overline{\mathbf{y}})(\overline{\mathbf{y}_i} - \overline{\mathbf{y}})^{\mathrm{T}}. \tag{7}$$

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

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

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

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

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* $\lambda_i$ *— i-тое по упорядоченности по невозрастанию собственное число матрицы* $\mathbf{E}^{-1}\mathbf{H}$*. Тогда*

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

*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 увидем несколько далее.

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

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

$$F = F(A) = C\frac{\mathrm{SSH}_\zeta}{\mathrm{SSE}_\zeta} = C\frac{A^{\mathrm{T}}\mathbf{H}A}{A^{\mathrm{T}}\mathbf{E}A} \sim \mathrm{F}_{\nu_{\mathbf{H}},\nu_{\mathbf{E}}}.$$

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.

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

Обозначим собственные числа матрицы $\mathbf{E}^{-1}\mathbf{H}$ в порядке невозрастания: $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_s$, где $s = \min\{p, \nu_H\}$, и собственные вектора той же матрицы — $A_i$ для $i \in 1:s$. Причем $A_i^{\mathrm{T}}\mathbf{E}A_j = 0$ 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: $Z_i^{\mathrm{T}}Z_j = 0$ для $j \neq i$.

**Теорема.**

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

*Доказательство.* $\langle Z_i, Z_j \rangle = \langle \mathbf{Y}A_i, \mathbf{Y}A_j \rangle = A_i^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}A_j = A_i^{\mathrm{T}}(\mathbf{H}+\mathbf{E})A_j = A_i^{\mathrm{T}}\mathbf{E}(\mathbf{I}+\mathbf{E}^{-1}\mathbf{H})A_j$. Так как $\forall \mathbf{C} \geq 0$ с с. ч. $\lambda_1, \ldots, \lambda_s$ с. ч. матрицы $\mathbf{I}+\mathbf{C}$ — это $1+\lambda_1, \ldots, 1+\lambda_s$, 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, $\langle Z_i, Z_j \rangle = (1+\lambda_j)A_i^{\mathrm{T}}\mathbf{E}A_j = 0$. □

Исходя из описания, смысл $\lambda_i$ — 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 $\lambda_i$.

## 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.

[1] Чтобы 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 средних[2]

$$\mathrm{H}_0 : \mathbb{E}\boldsymbol{\eta}_1 = \mathbb{E}\boldsymbol{\eta}_2 = \ldots = \mathbb{E}\boldsymbol{\eta}_k.$$

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 корреляции [3]. Потому они на самом деле имеют интерпретируемый смысл.

3. Pillai's:
$$V^{(s)} = \mathrm{tr}\left(\mathbf{H}(\mathbf{E}+\mathbf{H})^{-1}\right) = \sum_{i=1}^{s} r_i^2.$$

4. Hotelling:
$$U^{(s)} = \mathrm{tr}\left(\mathbf{E}^{-1}\mathbf{H}\right) = \sum_{i=1}^{s}\lambda_i.$$

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 критериев более мощные?».

Посмотрим на это на примере первых двух критериев. Первый критерий включает в статистику все (!) направления, по которым разброс максимален (иными словами, все $\lambda_i$). Второй же критерий включает в себя лишь первое направление, с максимальной степенью разброса, измеряемой $\lambda_1$. 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 $\lambda_i$ 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.

[2]На 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 $\mathbf{H} = \sum_{i=1}^{k} n_i \left(\overline{\mathbf{y}}_i - \overline{\mathbf{y}}\right)\left(\overline{\mathbf{y}}_i - \overline{\mathbf{y}}\right)^{\mathrm{T}}$, матрица внутригрупповых отклонений $\mathbf{E} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\mathbf{y}_{ij} - \overline{\mathbf{y}}_i\right)\left(\mathbf{y}_{ij} - \overline{\mathbf{y}}_i\right)^{\mathrm{T}}$.

В случае, если $\mathbf{E} = \sigma^2\mathbf{I}$ (все группы сферические и одинакового размера), дискриминантные функции данных, т.е. $A_1, \ldots, A_s$ – 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.

## 2 30 ноября 2017

### 2.1 Значимое число дискриминантных функций (размерность пространства, где группы различаются).

Модель дискриминантного анализа (на генеральном языке): $\xi$ – дискретная с.в., принимающая значения $\{A_i\}_{i=1}^{k}$, распределение $\boldsymbol{\eta}$ при условии $\xi = A_i$ совпадает с распределением $\boldsymbol{\eta}_i$. Определены матрицы

$$\mathbf{H}_{th} = \mathbb{E}\left(\left(\mathbb{E}\left(\boldsymbol{\eta}|\xi\right) - \mathbb{E}\boldsymbol{\eta}\right)\left(\mathbb{E}\left(\boldsymbol{\eta}|\xi\right) - \ E\boldsymbol{\eta}\right)^{\mathrm{T}}\right) \in \mathbb{R}^{p\times p},$$

$$\mathbf{E}_{th} = \mathbb{E}\left(\left(\boldsymbol{\eta} - \mathbb{E}\left(\boldsymbol{\eta}|\xi\right)\right)\left(\boldsymbol{\eta} - \mathbb{E}\left(\boldsymbol{\eta}|\xi\right)\right)^{\mathrm{T}}\right) \in \mathbb{R}^{p\times p}.$$

Дискриминантные функции $A_1, \ldots, A_s$ — это с.в. матрицы $\mathbf{E}^{-1}\mathbf{H}$, $s = \min\{p, \nu_H\}$.

Для того, чтобы узнать, сколько функций являются значимыми (сколько новых признаков интерпретировать), мы последовательно проверяем гипотезы $H_0$: «дискриминантные функции с номерами $m, \ldots, s$ не описывают различия в данных» для возрастающих $m \in 1:s$. Формально, гипотеза эквивалентна тому, $\mathrm{rk}\mathbf{H}_{th} \leq m-1$, т.е. все собственные числа матрицы, начиная с $m$-го равны нулю. Мы можем позволить себе проверять её последовательно, так как знаем, что дискриминантные функции, будучи отсортированы по убыванию соответствующих собственных чисел, оказываются в порядке убывания качества объяснения различий.

Если $H_0$ верна, то может быть показано, что

$$\Lambda'_m = \prod_{i=m}^{s} \frac{1}{1+\lambda_i} \sim \Lambda_p\left(\nu_H + (m-1), \nu_E - (m-1)\right).$$

Здесь уже матрицы, полученные по выборке.

С помощью статистики $\Lambda'_m$ (lambda prime) мы для каждого $m$ можем проверить, незначимы ли эта и все последующие дискриминантные функции. Если они значимы, то все перед ними тоже значимы и $m$ нужно увеличить.

## 2.2 Интерпретация разделения: стандартизованные дискриминантные функции и факторная структура

Пусть $\mathbf{Y}$ — матрица данных (строки $\mathbf{y}_i, i \in 1:n$ – 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 $\mathbf{H} = \sum_{i=1}^{k} n_i \left(\overline{\mathbf{y}}_i - \overline{\mathbf{y}}\right)\left(\overline{\mathbf{y}}_i - \overline{\mathbf{y}}\right)^{\mathrm{T}}$, матрица внутригрупповых отклонений $\mathbf{E} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\mathbf{y}_{ij} - \overline{\mathbf{y}}_i\right)\left(\mathbf{y}_{ij} - \overline{\mathbf{y}}_i^{\mathrm{T}}\right)$.

$A_1, \ldots, A_s$ — собственные вектора $\mathbf{E}^{-1}\mathbf{H}$, называемые дискриминантными функциями, $\lambda_i$ — упорядоченные по убыванию собственные числа. Канонические переменные — вектора данных в новых координатах $Z_i = \mathbf{Y}A_i$.

Задача: проинтепретировать канонические переменные (новые признаки).

### 2.2.1 Стандартизованные дискриминантные функции

Первый способ проинтерпретировать разложение по дискриминантным переменным — посмотреть на коэффициенты, с которыми исходные переменные входят в дискриминантные. Если исходные переменные измерены в различных шкалах, то коэффициенты в векторе $A_i$ одновременно ещё и приводят показатели к нужной шкале. Чтобы избежать этого эффекта, можно посмотреть на стандартизованные дискриминантные функции.

Пусть $\mathbf{S}$ – матрица взвешенных ковариаций (pooled covariance matrix), $s_1^2, \ldots, s_p^2$ – элементы на её диагонали (взвешенные дисперсии признаков, pooled variance), $A_i = \begin{pmatrix} a_{i1} \\ \vdots \\ a_{ip} \end{pmatrix}$.

Тогда

$$Z_i = \mathbf{Y}A_i = \sum_{j=1}^{p} Y_j a_{ij} = \sum_{j=1}^{p} \frac{Y_j}{s_j} \cdot s_j a_{ij},$$

и $\widetilde{A}_i = (\mathrm{diag}\mathbf{S})^{\frac{1}{2}} A_i$ — $i$-я стандартизованная дискриминантная функция (стд. д. ф.). Коэффициенты стд. д. ф. показывают вклады исходных признаков в дискриминантные переменные.

### 2.2.2 Факторная структура

Факторная структура — матрица корреляций между исходными и каноническими переменными. Так как дискриминантные функции, вообще говоря, не ортогональны, то получатся другие числа. Что правильнее - сложный вопрос.

## 2.3 Уменьшение числа признаков (feature selection)

«Плохие» признаки, это признаки, которые:

1. Являются линейной комбинацией других признаков, т.е. имеют большой коэффициент множественной корреляции $R^2 = R^2\left(\eta^{(i)}; \left\{\eta^{(i)} \middle| j \in 1:p \ \{i\}\right\}\right)$. Соответствующая характеристика – tolerance $= 1 - R^2$.

2. При удалении из модели не влияют на качество разделения. Соответствующая гипотеза $H_0$: «удаление признака $i$ не влияет на качество разделения», т.е. от удаления $i$-го признака теоретическая $\Lambda$ не меняется. Статистика:

$$(\text{Partial } \Lambda)_i = \frac{\Lambda\left(Y_1, \ldots, Y_p\right)}{\Lambda\left(Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_p\right)} \sim \Lambda_1\left(\nu_H, \nu_E - p + 1\right) \tag{9}$$

### 2.3.1 Пошаговый дискриминантный анализ

Пошаговый дискриминантный анализ подбирает тот набор признаков, который лучше всего будет разделять переменные (аналогично пошаговой множественной регрессии) жадным образом. На каждом шаге добавляется одна переменная, которая максимально увеличивает качество разделения групп (качество меряется статистикой partial lambda из (9), точнее, из-за наличия точного преобразования из распределения $\Lambda_1$ в $F$-распределение, эквивалентной статистикой с распределением Фишера). После этого набранная модель пересматривается на предмет наличия избыточных переменных. Процедура останавливается, когда максимальное значение $F$ для вновь добавляемых переменных не превосходит наперёд заданного порога.

Первой переменной в модели становится та, для которой $F$-статистика из ANOVA оказалась наибольшей.

Выше описан вариант forward с добавлением признаков. Аналогично можно рассмотреть backward версию.

## 2.4 Общий подход к классификации через апостериорные вероятности

Общая подход к классификации: строятся классифицирующие функции $f_i$, такие что классификация проводится так: индивид с признаками $x$ относится к группе с максимальным значением на нем классифицирующей функции: $\arg\max \mathrm{f}_i(x)$.

Откуда берутся эти классифицирующие функции? Естественная идея взять в качестве $f_i$ вероятность (ее оценку) принадлежности к $i$-му классу. Пусть $\xi$ – дискретная с.в., принимающая значения $\{A_i\}_{i=1}^k$, $\eta \sim \mathcal{P}_i$ и имеет плотность $p_i(x)$, если $\xi = A_i$. Тогда было бы логично взять $f_i = p_i$. Для практического применения надо было бы оценить плотности, либо непараметрически (например, по числу точек, попавших в дельта-окрестность — типа метода ближайших соседей), либо параметрически (если известно, что распределение нормальное, тогда просто оцениваем векторы средних и ковариационные матрицы).

Более сложный подход — через апостериорные вероятности. Если у нас есть априорное знание вероятности того, что индивид из того или иного класса, то мы можем его учесть. Введем понятие класса $C_i = \{\xi = A_i\}$. Чтобы классифицировать наблюдение $x$, необходимо найти

$$\arg\max \mathrm{P}\left(\xi \in A_i | \eta = x\right) = \arg\max \mathrm{P}\left(C_i | x\right).$$

Пусть известны априорные вероятности принадлежности нового наблюдения к $i$-му классу $\pi_i = \mathrm{P}\left(C_i\right)$. Тогда апостериорные вероятности по формуле Байеса будут иметь вид

$$\mathrm{P}\left(C_i | x\right) = \frac{\mathrm{P}\left(x | C_i\right)\pi_i}{\sum_{j=1}^k \mathrm{P}\left(x | C_j\right)\pi_j}.$$

Поэтому в качестве классифицирующих функций берут

$$f_i\left(x\right) = \frac{p_i(x)\pi_i}{\sum_{j=1}^k p_j(x)\pi_j}.$$

Так как знаменатель у всех $f_i$ одинаковый, его можно отбросить, и итоговые классифицирующие функции будут выглядеть как $f_i\left(x\right) = \mathrm{P}\left(x | C_i\right)\pi_i$.

Как выбрать априорные вероятности?

1. Равномерно, $\forall i \in 1 : k\ \pi_i = 1 \,/\, k$.
2. По соотношениям в обучающей выборке: $\pi_i = n_i \,/\, \sum_{j=1}^k n_j$.

3. На основе другой дополнительной информации о данных (результаты предыдущих исследований, etc.)

**Свойство.** *Построенный метод классификации* $\text{predict}(x) = \arg\max_i \pi_i p_i(x)$ *минимизирует среднюю апостериорную ошибку:*

$$\sum_{i=1}^{k} \pi_i P(\text{predict}(x)! = i \mid C_i).$$

Видно, что можно с помощью априорных вероятностей формально задавать важность ошибочных классификаций для разных классов.