# Ответы к коллоквиуму по Многомерному Анализу Данных

Собрано 12 января 2016 г. в 13:13

## Содержание

# 1. Первая часть

## 1.1. Билет 1. Многомерное нормальное распределение. Вектор мат.ож. и ковар.матрица при лин. преобразовании (умножении на матрицу).

### Нормальное распределение

***Определение.*** *Говорят, что случайный вектор* $\xi = (\xi_1, \xi_2, \dots, \xi_p)^{\mathrm{T}}$ *имеет* $p$*-мерное нормальное распределение, если для любых* $\{a_i\}_{i=1}^p \subset \mathbb{R}$ *линейная комбинация* $\sum_{i=1}^p a_i\xi_i$ *имеет нормальное распределение.*[1] *Если обозначить* $\mu = \mathbb{E}\xi$, $\Sigma = \mathrm{Cov}\xi$, *то пишут* $\xi \sim \mathcal{N}(\mu, \Sigma)$.

***Теорема.*** *Пусть* $\mu \in \mathbb{R}^p$ *и* $\Sigma \in \mathrm{M}_{p,p}(\mathbb{R})$ *— невырожденная положительно-определенная матрица. Рассмотрим случайный вектор* $\xi \sim \mathcal{N}(\mu, \Sigma)$. *Тогда* $\xi$ *имеет плотность:*

$$p_{\xi}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\sqrt{\det \Sigma}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mu)}, \tag{1}$$

*для любого* $\mathbf{x} \in \mathbb{R}^p$.[2]

Заметим, что в определении 1.1 не требуется невырожденность ковариационной матрицы $\Sigma$. Если же $\Sigma$ вырожденная, то это означает, что распределение сосредоточено на подпространстве в $\mathbb{R}^p$.[3]

### Линейное преобразование

***Теорема.*** *Пусть* $\xi \sim \mathcal{N}(\mu, \Sigma)$, *где* $\mu \in \mathbb{R}^p$, *а* $\Sigma \in \mathrm{M}_{p,p}(\mathbb{R})$. *Рассмотрим матрицу* $\mathbb{A} \in \mathrm{M}_{d,p}(\mathbb{R})$. *Тогда* $\mathbb{A}\xi \sim \mathcal{N}\left(\mathbb{A}\mu, \mathbb{A}\Sigma\mathbb{A}^{\mathrm{T}}\right)$.

*Доказательство.* Утверждение следует из линейности мат. ожидания и того, что $\mathrm{Cov}(\xi) = \mathbb{E}(\xi - \mathbb{E}\xi)(\xi - \mathbb{E}\xi)^{\mathrm{T}}$.[4] □

## 1.2. Билет 2. Оценки вектора средних и ковар.матрицы. Несмещенная оценка ковар. матрицы.

**Генеральный язык.** Пусть дан вектор $\xi = (\xi_1, \dots, \xi_p)^{\mathrm{T}}$. Вектором средних называется $\mathbb{E}\xi = (\mathbb{E}\xi_1, \dots, \mathbb{E}\xi_p)^{\mathrm{T}}$. Ковариационная матрица — $\mathrm{Cov}(\xi) = \mathbb{E}(\xi - \mathbb{E}\xi)(\xi - \mathbb{E}\xi)^{\mathrm{T}}$.

**Выборочный язык.** Генеральную совокупность обозначим $\xi$. Рассмотрим $\mathbb{X} = [X_1 : \dots : X_p] \in M_{n,p}(\mathbb{R})$ — матрица данных. $X_i$ — $i$-тый признак. Тогда $\widehat{\mathbb{E}\xi} = \left(\overline{X}_1, \dots, \overline{X}_p\right)$.

Введем $X_i^{(c)}$–$i$-тый центрированный признак и рассмотрим $\mathbb{X}^{(c)} = [X_1^{(c)}, \dots, X_p^{(c)}]$ — матрицу центрированных данных. Тогда $\widehat{\mathrm{Cov}(\xi)} = \mathbb{X}^{(c)\mathrm{T}}\mathbb{X}^{(c)}/n$. В несмещенной оценке ковариационной матрицы знаменатель дроби равен $n - 1$.

*Здесь также нужно провести доказательство для дисперсии, что несмещенная оценка является несмещенной. Думаю, что все это уже хорошо умеют делать.*

---

[1] *Здесь распределение Дирака тоже считаем нормальным.*

[2] *Убедитесь, что при* $p = 1$ *получается одномерная плотность.*

[3] Чтобы это осознать, представьте себе $\eta \sim \mathcal{N}(0, 1)$. На каком подпространстве (и как) распределен вектор $(\eta, 0)^{\mathrm{T}}$?

[4] В одномерном случае это должно совпасть с обычным определением ковариации.

## 1.3. Билет 3. Распределение вектора средних

***Теорема.*** *Пусть дана выборка (на априорном языке)* $\mathbf{x}_1, \dots, \mathbf{x}_n \dots$ *с ковариационной матрицей* $\Sigma$. *Обозначим* $\overline{\mathbf{x}}_n$ *— выборочное среднее первых* $n$ *индивидов. Тогда для выборочного среднего имеет место следующая (слабая) сходимость:*

$$\sqrt{n}(\overline{\mathbf{x}}_n - \mu) \to \mathcal{N}(\mathbb{0}, \Sigma).$$

*Если же* $\mathbf{x}_1, \dots, \mathbf{x}_n \sim \mathcal{N}(\mu, \Sigma)$, *то выборочное среднее* $\overline{\mathbf{x}}_n$ *имеет распределение* $\mathcal{N}(\mu, \Sigma/n)$.

Заметим, что асимптотическая сходимость есть ни что иное, как многомерное обобщение ЦПТ в форме Леви. Все могут вывести из этого обобщения обычную (одномерную) теорему?

## 1.4. Билет 4. Переход к новым признакам с помощью ортогональной матрицы. Пример про способности по математике и физике (выписать матрицу вращения)

**Переход к новым признакам**. Рассмотрим случайный вектор $\xi \in \mathbb{R}^p$ и детерминированный вектор $a \in \mathbb{R}^p$. Если $\xi$ рассматривать как набор из $p$ признаков, то $\eta = a^{\mathrm{T}}\xi$ — новый признак.

Рассмотрим же теперь матрицу $\mathbb{A} = [A_1 : \dots : A_d] \in \mathrm{M}_{p,d}(\mathbb{R})$. Тогда $\mathbb{A}^{\mathrm{T}}\xi$ — набор из новых $d$ признаков.

На выборочном языке то же самое перепишется так: $Z = \mathbb{X}a$ — для одного признака и $\mathbb{Z} = [Z_1 : \dots : Z_d] = \mathbb{X}\mathbb{A} \in \mathrm{M}_{n,d}(\mathbb{R})$ — для $d$ признаков. Заметим, что в последней записи новыми признакими как раз будут являться столбцы $Z_1, \dots, Z_d$.

**Факторы и факторные нагрузки**. Пусть задана матрица данных $\mathbb{X} = [X_1 : \dots : X_p] \in \mathrm{M}_{n,p}(\mathbb{R})$. Обозначим $d = \mathrm{rk}(\mathbb{X})$. Перейдем с помощью матрицы $\mathbb{A}$ к $d$ ортогональным признакам $\{Z_i\}_{i=1}^d$. Формально, это означает, что $\mathbb{Z} = \mathbb{X}\mathbb{A}$, где $\mathbb{Z} = [Z_1 : \dots : Z_d]$ и $Z_i \perp Z_j$ при $i \neq j$. С точки зрения линейной алгебры $\{Z_i\}_{i=1}^d$ образуют ортогональный базис в пространстве признаков[5].

Превратим этот базис в ортонормированный: для всех $i \in 1 : d$ положим $Q_i = Z_i/\|Z_i\|$. Таким образом, $\{Q_i\}_{i=1}^d$ — ОН-базис в пространстве признаков. Введем матрицу $\mathbb{Q} = [Q_1 : \dots : Q_d] \in \mathbb{M}_{n,d}(\mathbb{R})$.

Разложим исходные признаки по ОН-базису, то есть по всем $j \in 1 : p$

$$X_j = \sum_{k=1}^{d} f_{jk} Q_k,$$

где $f_{jk} = (X_j, Q_k)$ для всех $j \in 1 : p$ и $k \in 1 : d$. Введем матрицу $\mathbb{F} = [F_1 : \dots : F_d]$, где $\left(F_j\right)_i = f_{ij}$. Вектора $Q_{\cdot}$ называют *факторами*, а $f_{\cdot\cdot}$ — *факторными нагрузками*. Тогда ясно, что $F_k = \mathbb{X}^{\mathrm{T}}Q_k \in \mathbb{R}^p$[6]. Но $\mathbb{X} = \sum_{k=1}^d Q_k(\mathbb{X}^T Q_k)^{\mathrm{T}}$[7]. Следовательно, $\mathbb{X} = \mathbb{Q}\mathbb{F}^{\mathrm{T}}$.

---

[5] Формально, $\mathrm{span}(X_1, \dots X_p)$.

[6] Чтобы не запутаться, где тут транспонирование есть один простой трюк. Помните, что вы работаете с признаками! Это же точно столбцы матрицы $\mathbb{X}$. Интерпретируйте алгебраические преобразования именно как преобразования над признаками. Для формальной проверки достаточно обычно проверить, сходятся ли размерности.

[7] Просто разложили элементы пространства по базису, правда?

**Разложение с помощью ортогональных признаков.** Итак получено разложение $\mathbb{X} = \mathbb{Q}\mathbb{F}^{\mathrm{T}}$, при этом $\mathbb{Q}^{\mathrm{T}}\mathbb{Q}$ — единичная матрица порядка $d$.[8] В этом разложении лишь $Q$ задает ОН-базис. Система, заданная столбцами матрицы $\mathbb{F}$ совершенно не обязана быть нормированной. Рассмотрим $P_j = F_j/\|F_j\|$ и положим $\sigma_j = \|F_j\|$. Введем матрицы $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \dots, \sigma_d)$ и $\mathbb{P} = [P_1 : \ \dots \ : P_d] \in \mathrm{M}_{n,d}(\mathbb{R})$. Тогда $\Sigma\mathbb{P}^{\mathrm{T}} = \mathbb{F}^{\mathrm{T}}$.[9] А значит $\mathbb{X} = \mathbb{Q}\Sigma\mathbb{P}^{\mathrm{T}}$.

С другой стороны, было показано, что $\mathbb{X} = \sum_{i=1}^{d} Q_i F_i^{\mathrm{T}} = \sum_{i=1}^{d} \sigma_i Q_i P_i^{\mathrm{T}}$.

Обратим внимание, что в получившемся разложении $P_i$, обычно[10] не являются ортогональными, а лишь линейно-независимыми. В дальнейшем, утверждается, что единственным биортогональным разложением (то есть таким, когда $P_i$ ортогональны) является SVD.

***Пример.*** *Пусть число признаков $p = 2$, то есть $\mathbb{X} = [X_1 : X_2]$. При этом $X_1$ показывает количество баллов по математике, а $X_2$ — количество баллов по физике. Рассмотрим матрицу поворота:*

$$\mathbb{A} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

*Рассмотрим два новых признака $\mathbb{Z} = [Z_1 : Z_2] = \mathbb{X}\mathbb{A}$, где $Z_1 = (X_1 + X_2)/\sqrt{2}$ — отражает общие способности, $Z_2 = (X_1 - X_2)/\sqrt{2}$ — "разница" между способностями по математике и физике. Очень рекомендуется выписать все буковки, которые встречались раньше в этом примере (в частности, матрицу факторных нагрузок).*

## 1.5. Билет 5. Разложение матрицы данных при переходе к новым признакам в виде суммы и в матричном виде

См. предыдущий билет. По всей видимости интересует представление в виде $\mathbb{X} = \mathbb{Q}\mathbb{F}^{\mathrm{T}}$. Заметим, что справедливо представление $\mathbb{X} = \sum_{k=1}^{d} Q_k F_k^{\mathrm{T}}$. Определим $\mathbb{X}_k = Q_k F_k^{\mathrm{T}}$ для $k \in 1 : d$. Ясно, что $\mathrm{rk}(X_k) = 1$.[11] Тогда $\mathbb{X} = \sum_{k=1}^{d} \mathbb{X}_k$.

## 1.6. Билет 6. Как определяется вклад новых признаков

***Определение.*** *Зафиксируем $t, s \in \mathbb{N}$ (абстрактные). На пространстве $\mathrm{M}_{t,s}(\mathbb{R})$ введем фробениусово скалярное произведение: для любых матриц $\mathbb{X} = \{x_{ij}\}, \mathbb{Y} = \{y_{ij}\} \in \mathrm{M}_{t,s}(\mathbb{R})$ определим $(\mathbb{X}, \mathbb{Y})_F = \sum_{i,j} x_{ij} y_{ij}$.*[12]

*Это скалярное произведение порождает фробениусову норму матрицы:*
$\|\mathbb{X}\| = \sqrt{\sum_{i,j} x_{ij}^2}$.[13]

Вернемся к матрицам $\mathbb{X}_k$ (см. билет 5). Если $\mathbb{X}_i \perp \mathbb{X}_j$ для неравных $i, j$, то $\|X\|^2 = \sum_{k=1}^{d} \|X_k\|^2$[14]. Определим вклад $i$-того признака как отношение $\|\mathbb{X}_i\|^2/\|\mathbb{X}\|^2$.

---

[8] Заметим, что $\mathbb{Q}\mathbb{Q}^{\mathrm{T}}$ не обязано совпадать с единичной матрицей. Разве это удивительно? Вновь, используем язык признаков. Лишь одно из перемножений имеет интерпретируемый смысл.

[9] Смотрим на это на языке столбцов. Тогда все станет ясно.

[10] Я тут сам не понимаю — вроде бы в общем случае линейной независимости не от куда взяться

[11] Почему все строки (или все столбцы) этой матрицы линейно зависимы?

[12] *Несложно проверить, что это, действительно, скалярное произведение.*

[13] *Кто помнит, кажется, 5тую главу вычей второго курса, знает, что это не самая классная норма, потому что она не подчинена никакой векторной, но для наших целей подходит.*

[14] Это стандартное свойство нормы в гильбертовом пространстве.

***Лемма.*** *Пусть* $Q_1, Q_2 \in \mathbb{R}^n$, *а* $F_1, F_2 \in \mathbb{R}^p$ [15].

$$\left(Q_1 F_1^{\mathrm{T}}, Q_2 F_2^{\mathrm{T}}\right)_F = (Q_1, Q_2)(F_1, F_2).$$

Доказательство леммы проводится очень просто, если записать поэлементно, что происходит.

Из этой леммы следует, что для того, чтобы понятие вклада признака имело смысл (то есть чтобы матрицы $\mathbb{X}_k$ были ортогональны) достаточно ортогональности новых признаков $Q_i$.

## 1.7. Билет 7. Сингулярное разложение, как строится

Пусть дана матрица $\mathbb{Y} \in \mathrm{M}_{L,K}(\mathbb{R})$, где $L < K$ [16]. Рассмотрим матрицу $\mathbb{S} = \mathbb{Y}\mathbb{Y}^{\mathrm{T}} \in \mathrm{M}_{L,L}(\mathbb{R})$ [17]. Эта матрица неотрицательно определена и симметрична. Упорядочим ее собственные числа по невозрастанию: $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_L \geqslant 0$. Обозначим $U_i$ — нормированный собственный вектор соответствующий собственному числу $\lambda_i$ матрицы $\mathbb{S}$ (по всем $i \in 1 : L$). Набор $\{U_i\}_{i=1}^L$ образует ОН-базис в $\mathbb{R}^L$.

Следующее утверждение содержит несколько известных фактов из линейной алгебры (доказывать на коллоквиуме их не нужно).

***Теорема.*** *Обозначим* $d = \mathrm{rk}(\mathbb{S})$ [18].

1. $d \leqslant L, K$.
2. $d = \mathrm{rk}(\mathbb{Y}\mathbb{Y}^{\mathrm{T}}) = \mathrm{rk}(\mathbb{Y}^{\mathrm{T}}\mathbb{Y}) = \mathrm{colrank}(\mathbb{Y}) = \mathrm{rowrank}(\mathbb{Y})$.
3. $\lambda_d > 0$, $\lambda_{d+1} = 0$.
4. $\{U_i\}_{i=1}^d$ — *образует ОН-базис в* $\mathrm{colspan}(\mathbb{Y})$.

Следующая теорема играет ключевую роль в первой части.

***Теорема*** (The SVD). *Введем вектора* $V_i = \mathbb{Y}^{\mathrm{T}} U_i / \sqrt{\lambda_i}$ *для* $i \leqslant d$.

1. $\{V_i\}_{i=1}^d$ — *образуют ОН-базис в* $\mathrm{rowspan}(\mathbb{Y})$. *При этом* $\mathbb{Y}^{\mathrm{T}} U_i = \mathbb{0}$ *при* $i > d$. [19].
2. $V_i$ — *собственный вектор матрицы* $\mathbb{Y}^{\mathrm{T}}\mathbb{Y}$, *соответствующий собственному числу* $\lambda_i$ *(для* $i \in 1 : d$*). Все остальные собственные вектора соответствуют нулевым собственным числам.*
3. $U_i = \mathbb{Y} V_i / \sqrt{\lambda_i}$ *для* $i \leqslant d$.
4. $\mathbb{Y} = \sum_{k=1}^d \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$ — *The SVD (Singular Value Decomposition, сингулярное разложение). Терминология:* $\sqrt{\lambda_i}$ — *сингулярные числа матрицы* $Y$, $U_i$ — *левые сингулярные вектора матрицы* $\mathbb{Y}$, $V_i$ — *правые сингулярные вектора матрицы* $\mathbb{Y}$.

*Доказательство.* 1. Пусть $1 \leqslant i, j \leqslant d$. Тогда $(V_i, V_j) = \left(\mathbb{Y}^{\mathrm{T}} U_i, \mathbb{Y}^{\mathrm{T}} U_j\right) / \sqrt{\lambda_i \lambda_j} =$
$= \left(U_i, \mathbb{Y}\mathbb{Y}^{\mathrm{T}} U_j\right) = \lambda_j (U_i, U_j) \sqrt{\lambda_i \lambda_j} = \delta_{i,j}$.

---

[15] *Обозначения выбраны так, чтобы было ясно видно, причем тут матрицы* $\mathbb{X}_k$.

[16] Обозначения здесь вводятся с расчетом на Гусеницу в следующем семестре

[17] Матрица потом будет обозначать ковариационную, поэтому обозначение правильно ее напоминает.

[18] *Именно новое количество признаков всегда записывалось* $d$.

[19] *Кто-нибудь умеет аккуратно это доказывать?*

2. То, что $V_i$ — собственный вектор, соответствующий $\lambda_i$ при $i \leqslant d$ проверяется непосредственно. Докажем, что $V_i$ при $i > d$ соответствуют нулевым собственным числам. Действительно, пусть некоторый вектор $V$ такой, что $V \perp V_i$ для всех $i \leqslant d$. Это означает, что $0 = (\mathbb{Y}^{\mathrm{T}} U_i, V) = (U_i, \mathbb{Y}U)$ для всех $i \in 1 : d$, то есть $\mathbb{Y}V$ соответствует нулевому собственному числу матрицы $\mathbb{Y}\mathbb{Y}^{\mathrm{T}}$, а значит по первому пункту[20] $\mathbb{Y}^{\mathrm{T}}(\mathbb{Y}V) = 0$.

3. Подстаортонормированныйвьте и все будет хорошо.

4. Внешний факт: $\mathbb{E}_L = \sum_{i=1}^{L} U_i U_i^{\mathrm{T}}$[21]. Дальше все сводится к простой подстановке, так как $\mathbb{Y} = \mathbb{E}_L \mathbb{Y}$.

□

Заметим, что доказанный факт является очень мощным — на матрицу $\mathbb{Y}$ не наложено никаких ограничений! SVD является биортогональным разложением матрицы (на самом деле, SVD — единственное биортогональное разложение).

Перепишем SVD в матричном виде: $\mathbb{Y} = \mathbb{U}\Lambda^{1/2}\mathbb{V}$, где $\mathbb{U} = [U_1 : \ \dots U_L]$, $\mathbb{V} = [V_1 : \ \dots V_L]$ и

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 & \dots & 0 & 0 & \dots & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 & 0 & \dots & \dots & 0 \\ 0 & 0 & \lambda_3 & \dots & 0 & 0 & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \lambda_d & 0 & \dots & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 0 & \dots & \dots & 0 \end{pmatrix} \in \mathrm{M}_{L,K}(\mathbb{R}).$$

Из этого разложения легко получить $\Lambda^{1/2} = \mathbb{U}^{\mathrm{T}}\mathbb{Y}\mathbb{V}$ — квазидиагональное разложение матрицы $\mathbb{Y}$ [22]

Есть еще eigenvalue decomposition (спектральное разложение): $\mathbb{Y}\mathbb{Y}^{\mathrm{T}} = \mathbb{U}\Lambda\mathbb{U}^{\mathrm{T}}$.
**НЕ НАДО ПУТАТЬ!**

## 1.8. Билет 8. Сингулярное разложение. В каком смысле оно единственно.

***Обозначение*** (1). $\sqrt{\lambda_i}$ — *сингулярные числа*

***Обозначение*** (2). $U_i$ — *левые сингулярные вектора*

***Обозначение*** (3). $V_i$ — *правые сингулярные вектора*

Пусть $\mathbb{Y} = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$.

Если сделать замену сингулярной тройки $(\sqrt{\lambda_i}, U_i, V_i)$ на $(\sqrt{\lambda_i}, -U_i, -V_i)$, то разложение $\sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$ не поменяется.

Пусть $\lambda_1 = \lambda_2 > \lambda_3$. Тогда $\forall\, U \in span(U_1, U_2)$ — тоже сингулярный вектор, соответствующий $\lambda_1 = \lambda_2$. Таким образом, $U_1, U_2$ можно заменить на $\forall$ ортонормированные вектора из $span(U_1, U_2)$.

---

[20] Здесь пока непонятно.

[21] Такое разложение справедливо для любой ОН-системы.

[22] Это очень интересное утверждение. Для того, чтобы это осознать, нужно вспомнить, что матрицу можно рассматривать, как отображение (линейное). Тогда утверждение состоит в том, что можно подобрать такие базисы в dom и codom нашей матрицы, что она сама примет почти диагональный вид.

***Предложение.*** *Пусть* $\mathbb{Y} = \sum_{i=1}^{L} c_i P_i Q_i^{\mathrm{T}}, c_1 \geqslant c_2 \geqslant \ldots \geqslant 0$, *при этом* $\{P_i\}$ *и* $\{Q_i\}$ *— ортонормированные, тогда разложение* $\mathbb{Y}$ *— SVD.*

Без доказательства.

***Замечание.*** *Как только имеет место биортогональность — тогда SVD.*

## 1.9. Билет 9. Разложение Шмидта

Пусть $(D_1, \mathfrak{A}_1, \mu_1), (D_2, \mathfrak{A}_2, \mu_2)$ — измеримые пространства с мерой. Введем гильбертово пространство вещественных функций $f \in L^2 \leftrightarrow \int\limits_D |f|^2 d\mu < +\infty$

$$L_i^2 = L^2(D_i,\mu_i), i = 1,2$$

$$L_{1,2}^2 = L^2(D_1 \times D_2,\mu_1 \otimes \mu_2), < \cdot,\cdot >_{1,2}, || \cdot ||$$

$g$ — ядро интегрального оператора. $G : L_2^2 \longrightarrow L_1^2$

Оператор Гильберта–Шмидта: $Gh = \int\limits_{D_2} g(\cdot,s)h(s)\mu_2(ds)$

Сопряженный оператор: $G^* : L_1^2 \longrightarrow L_2^2$;
$G^* = \int\limits_{D_2} g(x,\cdot)f(x)\mu_1(dx)$
$< f, Gh >=< G^* f, h >$

Самосопряженный оператор: $GG^* : L_1^2 \longrightarrow L_1^2$ и $G^*G : L_2^2 \longrightarrow L_2^2$ с ядрами:
$g_{22}(u,v) = \int\limits_{D_1} g(x,u)g(x,v)\mu_1(dx)$
$g_{11}(x,y) = \int\limits_{D_2} g(x,s)g(y,s)\mu_2(ds)$

***Теорема.*** *1.* $GG^*$ *имеет* $\geqslant 1$ *ненулевых собственных чисел*

2. $GG^*$ *имеет н.б.ч.с. число ненулевых вещественных положительных собственных чисел конечной кратности.*

Без доказательства.

***Свойство*** (1). *Пусть* $\{\lambda_n\}, n \geqslant 1$ *— положительные собственные числа* $GG^*$, $\{\phi_n\}$ *— соответствующие собственные функции.*
*Тогда* $\{\phi_n\}$ *— ортонормированная система в* $L_1^2$. *И если* $\phi \perp \phi_n \forall n$, *тогда* $\phi$ *соответствует с. ч.* $\lambda = 0$ *на* $GG^*$

***Свойство*** (2). *Пусть* $\psi_n = \frac{G^*\phi_n}{\sqrt{\lambda_n}}$, *тогда* $\{\psi_n\}$ *— ортонормированная система в* $L_2^2$. *Если* $\psi \perp \psi_n \forall n$, *то* $\psi$ *соответствует нулевому собственному числу* $G^*G$

***Свойство*** (3). $\phi_n = \frac{G\psi_n}{\sqrt{\lambda_n}}$

Разложение Шмидта функции $g$:
$g(\cdot,x) = \sum\limits_n \sqrt{\lambda_n}\phi_n(\cdot) \otimes \psi_n(\cdot) \forall g \in L_{1,2}^2 \Rightarrow ||g||_{1,2}^2 = \sum\limits_n \lambda_n < +\infty$

## 1.10. Билет 10. Выборочный анализ главных компонент и сингулярное разложение, общее и различия.

Разложение Шмидта функции $g$:

$g(\cdot,x) = \sum\limits_n \sqrt{\lambda_n}\phi_n(\cdot) \otimes \psi_n(\cdot) \forall g \in L^2_{1,2} \Rightarrow ||g||^2_{1,2} = \sum\limits_n \lambda_n < +\infty$

### 1.10.1. Сингулярное разложение

$g(\cdot,\cdot) \leftrightarrow Y_{ij}$

$g \leftrightarrow \mathbb{Y}$

$D_1 = \{1,\dots,L\}$, $\mu_1$ — считающая мера

$D_2 = \{1,\dots,K\}$, $\mu_2$ — считающая мера

$g \leftrightarrow \mathbb{Y}\mathbb{Y}^{\mathrm{T}}$

### 1.10.2. Выборочный анализ главных компонент

$\mathbb{X}, \mathbb{Y} = \mathbb{X}^{\mathrm{T}}$

$D_1 = \{1,\dots,L\}$, $\mu_1$ — считающая мера

$D_2 = \{1,\dots,K\}$, $\mu_2$ — вероятностная мера такая, что $\mu_2(\{i\}) = \frac{1}{K}$

$||\mathbb{Y}||^2_{1,2} = \frac{||\mathbb{Y}||^2_F}{K}$

Предположим, что $\int\limits_{D_2} g(x,s)\mu_2(ds) = 0$, т.е. признаки $X_i$ — центрированы.

$g_{11} \leftrightarrow \frac{\mathbb{Y}\mathbb{Y}^{\mathrm{T}}}{K}$ — выборочная ковариационная матрица.

1. $\lambda_i = \frac{\hat{\lambda}_i}{K}$
2. $U_i = \hat{U}_i$
3. $V_i = \hat{V}_i\sqrt{K}$

SVD:

$$\mathbb{Y} = \sum_i \sqrt{\hat{\lambda}_i}\hat{U}_i\hat{V}_i^{\mathrm{T}}$$

PCA:

$$\mathbb{Y} = \sum_i \sqrt{\lambda_i}U_iV_i^{\mathrm{T}}$$

## 1.11. Билет 11. Анализ главных компонент на генеральном языке как частный случай разложения Шмидта.

Разложение случайного вектора $\xi = (\xi_1,\dots,\xi_L)^{\mathrm{T}}$.

$D_1 = \{1,\dots,L\}$, $\mu_1$ — считающая мера

$(D_2, \mathfrak{A}_2, \mu_2)$ — вероятностное пространство

$g(x,s) \leftrightarrow \xi_i(\omega), g \leftrightarrow \xi, x \leftrightarrow i, s \leftrightarrow \omega$

$g \in L^2_{1,2} \Leftrightarrow \sum\limits_{i=1}^{L} \mathbb{E}{\xi_i}^2 < \infty$

Пусть $\int\limits_{D_2} g(x,s)\mu_2(ds) = 0$, т.е. $\mathbb{E}\xi_i = 0 \forall i$

$g_{11} = \int\limits_{D_2} g(x,s)g(y,s)d\mu_2$

Значит $g_{11}(i,j) = \mathbb{E}\xi_i\xi_j = cov(\xi_i, \xi_j)$, т.е. $g_{11}$ — ковариационная матрица $\{\mathbb{E}\xi_i\xi_j\}_{i,j}$

$\lambda_i, U_i (U_i \leftrightarrow \phi_i)$ — с. ч. и с. в. матрицы вектора $\xi$

$\psi_i \leftrightarrow \epsilon_i$ (белый шум)

$\xi(\omega) = \sum\limits_n \sqrt{\lambda_n} U_n \epsilon_n(\omega)$

## 1.12. Билет 12. Почему главные компоненты так называются, в каком смысле они главные.

В силу третьего свойства оптимальности SVD:

Пусть $Y_1, \dots, Y_K \in \mathbb{R}^r$. $P \in \mathbb{R}^L$ задает направление ($|P| = 1$, $P$ — главное направление, которое задается 1–м с. в.).

$$\sum_{i=1}^{K} < Y_i, P >^2 \longrightarrow \max_P$$

***Предложение.*** *1.* $\max\limits_P \sum\limits_{i=1}^{K} < Y_i, P >^2 = \lambda_1$, *и достигается на* $P = U_1$

*2.* $\max\limits_{P: P \perp U_j} \sum\limits_{i=1}^{K} < Y_i, P >^2 = \lambda_i$, *и достигается на* $U_i \forall j = 1, \dots, k-1$, *где* $U_i$ — *i–й главный вектор, задающий i–е главное направление*

$< Y_j, U_i >$ — $i$–я главная компонента $j$–го индивида.

***Обозначение.*** $Z_i$ — *новые признаки*

$Z_i = (< Y_1, U_i >, \dots, < Y_k, U_i >)^{\mathrm{T}}$ — вектор $i$–х главных компонент.

$Z_i = \mathbb{X}U_i = \mathbb{Y}^{\mathrm{T}}U_i = \sqrt{\lambda_i}V_i$

***Обозначение.*** $V_i$ — *факторный вектор или вектор факторных значений.*

***Замечание.*** *Если исходные признаки были центрированы (*$\mathbb{E}X_i = 0$*), то все остальное тоже будет центрированным (*$\mathbb{E}U_i = 0, \mathbb{E}V_i = 0$*).*

***Обозначение.*** $\{V_i\}_{i=1}^d$ — *базис пространства признаков.*

$X_i \perp \mathbb{I}, < X_i, \mathbb{I} >= 0 \implies \mathbb{E}X_i = 0$

Если $\forall i X_i \perp \mathbb{I} \implies$ линейная комбинация $X_i \perp \mathbb{I}$.

## 1.13. Билет 13. Оптимальность сингулярного разложения в смысле аппроксимации матрицей ранга r.

Пусть

$$\mathbb{Y} = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^T$$

—- некоторое разложение SVD.

Определим множество матриц $M_r \subset \mathbb{R}^{L\times K}$ ранга $\leq r$.

***Предложение***. *1. Аппроксимация матрицей меньшего ранга:*

$$\min_{\widetilde{\mathbb{Y}}\in M_r} ||\mathbb{Y} - \widetilde{\mathbb{Y}}||_F^2 = \sum_{i=r+1}^{d} \lambda_i;$$

*2. Минимум достигается на первых r элементах сингулярного разложения:*

$$\widetilde{\mathbb{Y}}_0 = \sum_{i=1}^{r} \sqrt{\lambda_i} U_i V_i^T.$$

Без доказательства.

## 1.14. Билет 14. Оптимальность сингулярного разложения в смысле аппроксимации подпространством размерности r

Пусть $\alpha_r \in \mathbb{R}^L$ —- подпространство размерности $r$.

***Предложение***. *1.*

$$\min_{\alpha_r} \sum_{i=1}^{K} dist^2(\mathbb{Y}_i, \alpha_r) = \sum_{i=r+1}^{d} \lambda_i;$$

*2. Минимум достигается на подпространстве натянутом на первые r с. в. $(span(U_1, \dots, U_r))$.*

Набирал Вася 15-21. Кажется обозначения должны совпадать с 22-35, хотя мог что-нибудь пропустить.

## 1.15. Билет 15. Оптимальность в анализе главных компонент в статистической терминологии (через дисперсии)

***Определение***. *Будем говорить, что $w \in \mathbb{R}^p$ задает первое главное направление, если*

1. $||w||_2 = 1$
2. $w = \arg\max_{w\in\mathbb{R}^p} \sum_{i=1}^{n} < y_i, w >^2$

***Определение***. *Пусть $w_1, \dots, w_{s-1}$ — главные направления. Будем говорить, что $w \in \mathbb{R}^p$ задает s-ое главное направление, если*

1. $||w||_2 = 1$
2. $w_s = \arg\max_{w} \sum_{i=1}^{n} < y_i, w >^2 \; s.t. w \perp w_i \forall i = 1 \dots (s-1)$

***Теорема***. *В обозначениях, введенных выше:*

1. $\max_{w} \sum_{i=1}^{n} < y_i, w >^2 = \lambda_1$

2. $\arg\max_{w} \sum_{i=1}^{n} < y_i, w >^2 = u_1$

3. *Для $s$-ого главного направления аналогично*

$$\arg\max_{\substack{w_s \in \mathbb{R}^p \\ w_s \perp w_i \\ i=1\ldots(s-1)}} \sum_{i=1}^{n} < y_i, w_s >^2 = u_s$$

$$\max_{\substack{w_s \in \mathbb{R}^p \\ w \perp w_i \\ i=1\ldots(s-1)}} \sum_{i=1}^{n} < y_i, w_s >^2 = \lambda_s$$

Данное утверждение — переформулировка экстремальной задачи, определяющей SVD.

Заметим, что за счет центрирования $\sum_{j=1}^{n} < y_j, u_i >^2$ с точностью до константы $(\frac{1}{n})$ совпадает с выборочной дисперсией признака $z_j =< y_j, u_i >$ ($z_i$ — линейная комбинация изначальных признаков). Таким образом, первое главное направление в статистической терминологии — такая прямая, при проекции на которую получившийся признак будет иметь максимальную выборочную дисперсию. $s$-ые направления аналогично, при условии, что ищем прямую, перпендикулярную всем предыдущим.

## 1.16. Билет 16. Оптимизация в АГК в терминах ковариационных матриц.

***Теорема.*** *Следующие две задачи эквивалентны:*

1. $$\arg\min_{\widetilde{Y}, rk\widetilde{Y} \le d} ||Y - \widetilde{Y}||_F$$

2. $$\arg\min_{\widetilde{Y}, rk\widetilde{Y} \le d} ||YY^\intercal - \widetilde{Y}\widetilde{Y}^\intercal||_F = \arg\min_{\widetilde{S}, \widetilde{S} sym., p.d., rk(\widetilde{S}) \le d} ||S - \widetilde{S}||_F$$

Почему это так? Мы знаем решение как первой, так и второй задачи. Для первой — через SVD, для второй — через жорданову форму (или как частный случай SVD) – записываем матрицу в виде нужного разложения, обнуляем наименьшие собственные числа и получаем матрицу необходимого ранга, которая решает либо первую, либо вторую задачу:

Для первой задачи:

$$Y = U\Lambda^{\frac{1}{2}}V^\intercal$$

Для второй:

$$YY^\intercal = U\Lambda U^\intercal$$

## 1.17. Билет 17. В двух статистических пакетах получились разные главные компоненты. Отчего так могло получиться?

Смотрим на разложение матрицы.

$$\mathbb{Y} = \sum_{i=1}^{p} \sqrt{\lambda_i} u_i v_i^\intercal$$

Как мы знаем, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$

Данное разложение определено не совсем единственно: Для собственных чисел кратности 1 $u_i$ и $v_i$ можно одновременно умножить на $-1$ и ничего не поменяется (а для комплексных чисел поворотов еще больше...), а для собственных чисел, у которых кратность больше 1, можно выбрать любой базис линейного подпространства, соответствующего данному собственному числу и будут получаться различные разложения.

В стат. пакетах выч. методы, с помощью которых ищутся главные компоненты различаются, поэтому и найденные ими компоненты могут отличаться (а некоторые выч. методы могут и от запуска к запуску выдавать разные результаты, так что на самом деле и в одном стат. пакете теоретически можно получать разные результаты.)

## 1.18. Билет 18. Смысл первой ГК, если все ковариации (корреляции) исходных признаков положительны.

***Теорема*** (Перрон-Фробениус). *Пусть у матрицы $M$ все элементы строго-положительны. Тогда у матрицы $M$:*

- *Максимальное собственное число больше нуля*
- *Максимальное собственное число имеет кратность 1 (т.е. ему соответвует одномерное подпространство, а все остальные с.ч. меньше его)*
- *У данного собственного числа существует собственный вектор с строго-положительными компонентами*

Данная теорема, примененная для ковариационной матрицы с положительными корреляциями, дает нам, что первое собственное направление будет взвешенной суммой исходных признаков. Таким образом, смысл первой главной компоненты — «среднее всех признаков».

## 1.19. Билет 19. Разница между АГК по корреляционной и по ковариационной матрице на примере двух признаков. Когда что использовать.

Пусть сначала у нас признаки стандартизированы. Тогда корреляционная матрица выглядит следующим образом:

$$\mathbb{S} = \mathbb{X}^\intercal \mathbb{X} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

Ее собственные числа и вектора легко посчитать:

$$\lambda_1 = 1 + \rho, u_1 = (2^{-1/2}, 2^{-1/2})^\intercal$$

$$\lambda_2 = 1 - \rho, u_2 = (2^{-1/2}, -2^{-1/2})^\intercal$$

Главные компоненты соответственно[23]:

$$z_1 = \frac{x_1 + x_2}{\sqrt{2}}$$

$$z_2 = \frac{x_1 - x_2}{\sqrt{2}}$$

[23] на свякий случай: $x_i$ — вектор-столбец для 1-ого признака

А теперь пусть у нас дисперсия первой компоненты будет «большой»:

$$\mathbb{S} = \mathbb{X}^\intercal \mathbb{X} = \begin{pmatrix} a^2 & a\rho \\ a\rho & 1 \end{pmatrix}$$

Матрица $2 \times 2$, для нее можно выписать формулы для собственных чисел и векторов. Теперь если устремить $a \to \infty$, то будет выполнено следующее:

$$\frac{u_{11}}{u_{12}} \to \frac{a}{\rho}$$

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} \to 1$$

Вспомним, что $u_1$ — координаты первой главной компоненты в старом базисе и получим, что первый признак будет «перетягивать» на себя главное направление. И чем меньше корреляция, тем сильнее будет вклад первого признака в первую компоненту.

Как выбрать:

1. Признаки измеряны в «разных» единицах или вообще ничего про них не знаем — стандартизуем
2. Считаем, что признаки в «примерно одинаковых» единицах измерения — можем не стандартизовать.

## 1.20. Билет 20. Способы выбора числа главных компонент

Все методы во-многом эвристики. Кроме того, у АГК отсутствует модель, из-за чего никаких гипотез особо не проверить.

1. Scree plot — рисуем график: собственное число по оси $y$, номер с.ч. по оси $x$. Смотрим, где график начинает напоминать прямую, параллельную оси $x$ и берем компоненты до этого места в качестве результаты.
2. Воспользуемся тем, что $\sum \lambda_i$ равняется общей дисперсии (total variance). Поэтому $\frac{\sum_{i=1}^{d} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$ можно трактовать, как процент «объясненной» дисперсии. Фиксируем, хотя бы сколько процентов хотим объяснить и на основе этого выбираем $d$.
3. Правило Кайзера. Берем $\lambda_i > \frac{tr(S)}{p}$, где $S$ — ковариационная матрица. Т.к. след матрицы равняется сумме ее собственных чисел, мы на самом деле берем те собственные числа, которые «выше среднего». В случае, когда анализ ведется по корреляционной матрицы, диагональ $S$ равна 1, поэтому критерий упрощается до $\lambda_i > 1$.
4. Правило сломанной трости. Пусть $l_i = \frac{\lambda_i}{\sum \lambda_i}$. $\sum l_i = 1$. Кинем $n - 1$ случайную точку в отрезке [0,1] и получим $n$ отрезков. Отсортируем отрезки по длине и получим $L_1 \geq L2, \ldots, \geq L_n$. Можно посчитать $EL_i = \frac{1}{n} \sum_{j=i}^{n} \frac{1}{j}$. Найдем максимальное $d$ такое, что выполнено $l_i > EL_i \forall i = 1 \ldots d$ и возьмем в качестве главных компонент $u_1 \ldots u_d$.
5. Выбираем столько, сколько можем объяснить.

## 1.21. Билет 21. Почему доля собственного числа по отношению к сумме собственных чисел называется объясненной долей общей дисперсии?

Запишем разложение по главным компонентам:

$$\mathbb{X}^\intercal = \mathbb{Y} = \sum_{i=1}^{p} \sqrt{\lambda_i} u_i v_i^\intercal$$

Как мы знаем, в этом разложении $u_i$ — собственные вектора, а $\lambda_i$ — собственные числа матрицы ковариаций (или корреляций), равной $\frac{\mathbb{X}^\intercal X}{n}$.

Запишем суммарную дисперсию (данные центрированы):

$$\sum_{i=1}^{p} var_i = \sum_{i=1}^{p} \frac{1}{n} \sum_{j=1}^{n} \mathbb{X}_{ji}^2 = \frac{trX^\intercal X}{n} = total\ variance$$

В АГК мы ищем собственные числа и собственные вектора как раз матрицы $\frac{1}{n} X^\intercal X$, след матрицы при смене базиса не меняется, поэтому $\sum \lambda_i = total\ variance$. Поэтому $\frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i}$ можно называть объясненной долей общей дисперсии.

Начиная отсюда и до конца 28 билета набирала Лиза.

***Напоминание.*** *$\mathbb{Y} \in \mathbb{R}^{K \times L}$. SVD $\mathbb{Y}$ называется*

$$\mathbb{Y} = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^T,$$

*где*

1. *$\lambda_i$ — сингулярные числа $\mathbb{Y}$ (собственные числа $\mathbb{Y}\mathbb{Y}^T$),*
2. *$U_i$ — левые сингулярные вектора (собственные вектора $\mathbb{Y}\mathbb{Y}^T$),*
3. *$V_i$ — правые сингулярные вектора ($V_i = \mathbb{Y}^T U_i / \sqrt{\lambda_i}$),*

***Напоминание.*** *$\mathbb{X} \in \mathbb{R}^{K \times L}$ — матрица признаков. $\mathbb{Y} = \mathbb{X}^T$.*

*Рассмотрим измеримые пространства $(D_1, \mathfrak{A}_1, \mu_1)$ и $(D_2, \mathfrak{A}_2, \mu_2)$, где*

1. *$D_1 = \{1, \dots, L\}$, $D_2 = \{1, \dots, K\}$,*
2. *$\mathfrak{A}_{1,2}$ — множества всех подмножеств $D_{1,2}$,*
3. *$\mu_1$ — считающая мера, $\mu_2$ — вероятностная, $\mu_2(\{i\}) = 1/K$.*

*Если*

$$\mathbb{Y} = \sum_{i=1}^{d} \sqrt{\widetilde{\lambda}_i} \widetilde{U}_i \widetilde{V}_i^T$$

*— SVD $\mathbb{Y}$, то PCA $\mathbb{Y}$ называется*

$$\mathbb{Y} = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^T,$$

*где $\lambda_i = \widetilde{\lambda}_i / K$, $U_i = \widetilde{U}_i$, $V_i = \sqrt{K} \widetilde{V}_i$.*

***Замечание.*** *Об обозначениях. Считается, что $\mathbb{A}$ — это матрица, $\{A_i\}$ — столбцы матрицы $\mathbb{A}$, $\{a_{ij}\}$ — ее элементы.*

*Также $p$ — это количество признаков, $n$ — количество индивидов. В ссылках на SVD $L$ соответствует $p$, $K$ соответствует $n$.*

## 1.22. Билет 22. На основе каких элементов сингулярного разложения интерпретируются главные компоненты как линейные комбинации исходных признаков? Привести формулу и пример

Пусть $\mathbb{X} \in \mathbb{R}^{K \times L}$ — матрица исходных признаков, $\mathbb{Y} = \mathbb{X}^T$.

Главные компоненты — это $Z_i = \sqrt{\lambda_i} V_i$, $i = 1, \dots, d$, где $\sqrt{\lambda_i}, V_i$ из PCA $\mathbb{Y}$.

Можно показать, что $Z_i = \mathbb{X} U_i$, где $U_i$ — собственные вектора ковариационной матрицы $\mathbb{X}^T\mathbb{X}/K$. Таким образом, главные компоненты — это линейные комбинации исходных признаков, где в качестве коэффициентов выступают собственные вектора ковариационной матрицы исходных признаков.

***Пример.*** *Двумерный случай со стандартизованной матрицей, $U_i, Z_i$ легко считаются.*

$$\frac{1}{K}\mathbb{X}^T\mathbb{X} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

*Пусть $\rho > 0$, тогда $U_1 = (1,1)^T/\sqrt{2}$ — первое главное направление. $U_1 \perp U_2$, значит, $U_2 = (1,-1)^T/\sqrt{2}$.*

*Получаем, что $Z_1 = (x_{11} + x_{12}, x_{21} + x_{22})^T$, то есть сумма исходных признаков, а $Z_1 = (x_{11} - x_{12}, x_{21} - x_{22})^T$ – их разность.*

*Вклад первой компоненты $(1+\rho)/2$, вклад второй — $(1-\rho)/2$.*

## 1.23. Билет 23. АГК с точки зрения построения базиса в пространстве индивидов и в пространстве признаков. Координаты в новых базисах

$\mathbb{X} = [X_1 : \dots : X_L] \in \mathbb{R}^{K \times L}$ — матрица исходных признаков, $\mathbb{Y} = [Y_1 : \dots : Y_K] = \mathbb{X}^T$. $L = p$, $K = n$, также считаем, что $d = p$, то есть матрица $\mathbb{Y}$ — полного ранга.[24] Вектора $U_i, V_i$, $i = 1, \dots, d$ из PCA $\mathbb{Y}$.

С точки зрения построения базиса в пространстве индивидов:

$\{U_j\}_{j=1}^d$ — это ортонормированный базис в пространстве индивидов, $z_{ij} = Y_i^T U_j$ — это координаты $i$-го индивида в этом базисе.[25]

С точки зрения построения базиса в пространстве признаков:

$\{V_j\}_{j=1}^d$ — это ортонормированный базис в пространстве признаков,

$$f_{ij} = \langle X_i, V_j \rangle = \begin{cases} cov(X_i, V_j), \text{ если считать по ковариационной матрице,} \\ \rho(X_i, V_j), \text{ если считать по корреляционной матрице,} \end{cases}$$

— это координаты $i$-го признака в этом базисе.

## 1.24. Билет 24. Как выявить индивидов, которые плохо описываются плоскостью первых двух главных компонент?

Рассмотрим $Y_i$ — $i$-индивид и первые два главных направления $U_{1,2}$.

$\cos^2(\angle(Y_i, span(U_1, U_2))) = \left(\frac{Y_i^T U_1}{\|Y_i\|\|U_1\|}\right)^2 + \left(\frac{Y_i^T U_2}{\|Y_i\|\|U_2\|}\right)^2 = (z_{i1}^2 + z_{i2}^2)/\|Y_i\|^2$.

Те индивиды, для которых $\cos^2$ мал, плохо описываются плоскостью первых двух главных компонент.

---

[24] **UPD**. Добавила разъяснение.

[25] **UPD**. Здесь исправила определение $z_{ij}$.

## 1.25. Билет 25. Как вычислить значения главных компонент для индивида, которого не было в исходной выборке. А как вычислить значения факторных значений?

Пусть[26] $A \in \mathbb{R}^p$ — новый индивид. Тогда $B = A^T\mathbb{U}$ — значения[27] главных компонент для этого индивида.

Напомню, что факторные значения $\{V_i\}$ — это нормированные главные компоненты, поэтому $\widetilde{B} = A^T\widetilde{\mathbb{U}}$, где $\widetilde{\mathbb{U}} = \mathbb{U}/\sqrt{\lambda_i}$.

## 1.26. Билет 26. В каком случае координаты в ортонормированном базисе можно назвать корреляциями?

Нужно, чтобы признак, который расскладывают, был стандартизован, а вектора в базисе — центрированы (они уже отнормированы).

## 1.27. Билет 27. Чему равны суммы по строкам и по столбцам в матрице, составленной из собственных векторов в АГК?

Матрица, составленная из собственных векторов, это $\mathbb{U} = [U_1 : \ldots : U_p] = \{u_{ij}\} \in \mathbb{R}^{p\times p}$.

Сумма квадратов по строкам — $\sum_{i=1}^p u_{ij}^2 = \|U_j\|^2 = 1$. Сумма квадратов по столбцам — $\sum_{j=1}^p u_{ij}^2 = 1$, так как $\mathbb{U}^T = \mathbb{U}^{-1}$.

## 1.28. Билет 28. Чему равны суммы по строкам и по столбцам в матрице факторных нагрузок в АГК?

Матрица факторных нагрузок — это $\mathbb{F} = [F_1 : \ldots : F_d] = \{f_{ij}\} \in \mathbb{R}^{p\times d}$. Вспомним, что $V_j \in \mathbb{R}^n$, как и столбцы матрицы $\mathbb{X} = [X_1,\ldots,X_p]$.. Далее из $SVD$-разложения мы знаем, что $U_i = \frac{YV_i}{\sqrt{\lambda_i}}$. Поэтому $\sqrt{\lambda_i}U_i = YV_i$ и $(F_i)_j = \langle X_j, U_i\rangle$. $F_i$ — i-ый столбец, поэтому $f_{ij} =< X_i, V_j$.

Сумма квадратов по строкам — $\sum_{i=1}^p f_{ij}^2 = \|F_j\|^2 = \lambda_j$. Сумма квадратов по столбцам

$$\sum_{j=1}^d f_{ij}^2 = \sum_{j=1}^d \langle X_i, V_j\rangle^2 = \begin{cases} \|X_i\|^2, \text{ если считать по ковариационной матрице,} \\ 1, \text{ если считать по корреляционной матрице.} \end{cases}$$

## 1.29. Билет 29. Как интерпретировать скалярное произведение строк в матрице факторных нагрузок в АГК?

Смотрим на билет 28. Из него мы знаем, как интерпретировать скалярное произведение строки с самой собой — это либо норма $X_i$, если анализ ведется по ковариационной матрице, либо 1, если анализ ведется по корреляционной матрице. Естественно предположить, что примерно так же интерпретируется и скалярное произведение строк в матрице: в самом деле, пусть как и до этого $\mathbb{F} = [F_1 : \ldots : F_d] = \{f_{ij}\} \in \mathbb{R}^{p\times d}$, $f_{ij} = \langle X_i, V_j\rangle$ — матрица факторных нагрузок. Тогда $i$-ая строчка $\mathbf{F}$ интерпретируется как координаты $i$-ого признака в ортонормированном базисе $V_1, \ldots, V_d$. Отсюда, за счет ОНБ, без всяких формул получаем, что скалярное произведение строчек $i$ и $j$ это тоже самое, что скалярное прозведение векторов

[26] Здесь $L = p$ и обычно $d = p$

[27] **UPD**. Транспонирование $\mathbb{U}$ было лишним. Исправила.

$X_i$ и $X_j$. Формально:

$$\sum_{j=1}^{d} f_{ij} f_{kj} = \sum_{j=1}^{d} \langle X_i, V_j \rangle \langle X_k, V_j \rangle = \left( \sum_{j=1}^{d} \langle X_i, V_j \rangle \right) \left( \sum_{j=1}^{d} \langle X_k, V_j \rangle \right) = \begin{cases} \langle X_i, X_k \rangle, \text{ если считать по ковариац} \\ \rho(X_i, X_k), \text{ если считать по корреля} \end{cases}$$

Признаки у нас центрированы, поэтому $< X_i, X_j >= cov(X_i, X_j)$[28].

## 1.30. Билет 30.Как нарисовать исходные орты в плоскости двух первых главных компонент?

Матрица $\mathbb{U} = [U_1, \ldots, U_p]$ — ортогональная матрица, составленная из собственных векторов ковариационной/корреляционной матрицы. Вспоминаем линейную алгебру или один из множества предыдущих вопросов и понимаем, что столбец — координаты вектора $U_i$ в исходном базисе. Следовательно (матрица то ортогональная), строчка $U_i$ — координаты старого базиса в новом (новый = составленный из $U_1, \ldots, U_p$). Поэтому $U_{i1}, U_{i2}$ — координаты $i$-ого орта в плоскости первых двух главных компонент.

## 1.31. Билет 31.Зачем и когда первые две координаты факторных нагрузок рисуются в единичном круге?

Если АГК строился по корреляционной матрице, то

$$\sum_{j=1}^{d} f_{ij}^2 = 1, \tag{2}$$

где соответственно $d = \operatorname{rk} \mathbb{Y}$, $f_{ij}$ — факторные нагрузки.

Таким образом, $f_{i1}^2 + f_{i2}^2$ (эта сумма, естественно, меньше 1) показывает, насколько хорошо первые две компоненты отражают $i$-ый признак. Поэтому в таком случае на единичной окружности отображается вектор, выходящий из нуля, с концом в $(f_{i1}; f_{i2})$ и длина этого вектора показывает, насколько хорошо $i$-ый признак описывается в плоскости первых двух главных компонент.

## 1.32. Билет 32.Чему равна норма i-го вектора из главных компонент?

Главные компоненты — координаты индивидов в базисе из главных направлений. Вспомним, что $Y = \sum_{i=1}^{d} \sqrt{\lambda_i} u_i v_i^{\mathrm{T}} = \sum u_i z_i^{\mathrm{T}}$, где $z_i = \sqrt{\lambda_i} v_i^{\mathrm{T}}$. Таким образом, $||z_i||^2 = \sum_{j=1}^{n} (v_i)_j = \lambda_i \sum_{i=1}^{n} |v_i|^2$. $v$ — ортонормированный вектор, поэтому сумма превращается в 1 и норма равна корню из $i$-го собственного числа матрицы $\mathbb{Y}$.

## 1.33. Билет 33. Как формализовать веса для признаков и для индивидов в АГК?

Иногда мы хотим, чтобы некоторые индивиды давали вклад больше, чем другие. Для этого нужно каждому индивиду придать определенный вес (чем больше вес, тес больше вклад индивида). Если хотим придать каждому индивиду вес, то в разложении Шмидта вводим меру $\mu_2$: $\mu_2(i) = \omega_i$. В итоге получаем по прежнему биортогональное разложение, но с

[28] Рекомендуется убедиться, что вы понимаете, что подразумевается под скалярным произведение

весами: $\langle V_i, V_j \rangle = \frac{1}{\sum \omega_k} \sum\limits_{k=1}^{n} \omega_k (V_i)_k (V_j)_k$. Веса на признаках — это масштаб (когда мы осознанно придаем больший вес индивиду, у которого больше разброс).

## 1.34. Билет 34. Какова модель в факторном анализе?

Модель факторного анализа

$$\xi = \mathbb{F}\eta + \varepsilon,$$

где $\xi$ — случайный вектор размерности $p$, $\mathbb{F}$ — матрица размерности $p \times r$, $\eta$ —случайный вектор размерности $r$, $\varepsilon$ — случайный вектор размерности $p$.

При этом $\operatorname{cov} \xi = \Sigma, \operatorname{cov} \eta = \mathbb{I}, \operatorname{cov} \varepsilon = \operatorname{diag}(\sigma_1^2, \dots, \sigma_p^2) = \Psi$.

Можно переписать все в виде:

$$\Sigma = \mathbb{F}\mathbb{F}^{\mathrm{T}} + \Psi.$$

***Замечание.*** *$F_i$ не может иметь вид $(0, \dots, a, \dots, 0)^{\mathrm{T}} \neq 0$, то есть не может быть факторов, уникальных для признаков!*

Стандартно факторный анализ делается по стандартизованным признакам, то есть $\operatorname{cov} \xi = \operatorname{cor} \xi$. $\mathbb{F}$ называется факторными нагрузками, $\eta$ – факторными значениями.

Общностью будем называть $\sum\limits_{j=1}^{r} f_{ij}^2 = 1 - \sigma_i^2$.

## 1.35. Билет 35. Что делает АГК в модели факторного анализа при равных общностях?

Перепишем модель факторного анализа на выборочном языке.

$$\mathbb{X} = \mathbb{V}\mathbb{F}^{\mathrm{T}} + \varepsilon.$$

$\mathbb{S}$ — выборочная ковариационная матрица. Хотим

$$||\mathbb{S} - (\mathbb{F}\mathbb{F}^T + \Psi)||_F^2 \to \min_{\mathbb{F},\Psi}$$

Обозначим $\widetilde{\mathbb{S}} = \mathbb{F}\mathbb{F}^T + \Psi$. Оказывается, что эта задача эквивалентна следующей:

$$\begin{cases} \sum\limits_{i \neq j} (s_{ij} - \sum\limits_{k=1}^{r} f_{ik} f_{jk})^2 \to \min \\ (\mathbb{F}\mathbb{F}^{\mathrm{T}})_{ii} \leq 1) \end{cases}$$

То есть, минимизации по всем элементам, кроме диагональных. Данный метод поиска факторов называется MINRES. Известно, что АГК эквивалентно задаче

$$||\mathbb{S} - \widetilde{\mathbb{S}}|| \to \min \tag{3}$$

Поэтому, если общность одинаковая, то АГК и MINRES решают одну и ту же задачу.

Набирала Белла 36-42. Есть вопросы, выделены жирным, кое-что не нашла вовсе.

***Напоминание.*** *Модель в факторном анализе имеет вид*

$$\xi = \mathbb{F}\eta + \varepsilon, \tag{4}$$

*где*

1. $\xi$ — *исходный вектор признаков.*
2. $\mathbb{F} = [F_1 : \ldots : F_r] = \{f_{ij}\} \in \mathbb{R}^{p \times r}$ — *матрица факторных нагрузок.*
3. $\eta \in \mathbb{R}^r$ — *факторное значение.*
4. $\varepsilon \in \mathbb{R}^p$ — *вектор индивидуальных (характерных/специфических) факторов* ***(можно ли так говорить?)****.*
5. $p$ — *исходное число признаков,* $r$ — *число общих факторов* ***(можно ли так говорить?),*** $r << p$.

*Предполагаемые условия:*

1. $\mathbb{E}\eta = 0,\ \mathbb{D}\eta = 1,\ \eta_i$ — *некорр.*
2. $\varepsilon$ *и* $\eta$ *некорр.,* $\varepsilon_i$ *некорр. между собой.*
3. $\mathbb{E}\varepsilon = 0,\ cov(\varepsilon) = diag(\sigma_1^2,..,\sigma_p^2),\ cov(\xi) = \Sigma,\ \tilde{\xi} = \mathbb{F}\eta,\ cov(\varepsilon) = \Psi,\ cov(\tilde{\xi}) = \mathbb{F}\mathbb{F}^{\mathrm{T}}$

*Модель* (4) *переписывается как* $\Sigma = \mathbb{F}\mathbb{F}^{\mathrm{T}} + \Psi$,
$\xi_i = f_{i1}\eta_1 + \cdots + f_{ir}\eta_r + \varepsilon_i$.

***Замечание****. Почти всегда предполагается, что признаки стандартизованы, то есть* $cov(\xi) = corr(\xi)$.

## 1.36. 36. Какая разница между АГК и факторным анализом?

1. В факторном анализе есть модель.
2. В факторном анализе не может быть факторов, которые уникальны по одному признаку, то есть $F_i$ не может иметь вид $\begin{pmatrix} 0 \\ 0 \\ \vdots \\ a \\ 0 \end{pmatrix}$. Такие факторы относятся к уникальным, то есть входят в $\varepsilon$, а мы интересуемся общей частью.

## 1.37. 37. Связь между числом факторов и числом признаков для корректности задачи.

Из $\Sigma = \mathbb{F}\mathbb{F}^{\mathrm{T}} + \Psi$ получаем $\frac{p(p+1)}{2}$ равенств (помним, что $\Sigma$ — симметр.) и число параметров $\leq pr + p$.

Условие корректности задачи $pr + p \leq \frac{p(p+1)}{2} \Rightarrow r \leq \frac{p-1}{2}$.

Получили условие, когда число уравнение не меньше числа параметров. Тут была заминка на лекции (2015.X.08), примерно 40-50 минута . Сошлись на том, что если модель верна, то лишние уравнения не сделают ее неверной. Если бы знак стоял в обратную сторону, то вышло бы, что равенства модель не характеризуют.

## 1.38. 38. Что минимизируется в методе MINRES? В чем разница с тем, что минимизируется в АГК?

Задача на выборочном языке имеет вид $\mathbb{X} = \mathbb{V}\mathbb{F}^{\mathrm{T}} + \varepsilon$, где $\mathbb{X} \in \mathbb{R}^{n\times p}$, $\mathbb{V} \in \mathbb{R}^{n\times r}$, $\mathbb{F} \in \mathbb{R}^{p\times r}$.

Пусть $\mathbb{S}$ — выборочная ковариационная матрица (известна). $\| \mathbb{S} - (\mathbb{F}\mathbb{F}^{\mathrm{T}} + \Psi) \|_F^2 \to \min\limits_{\mathbb{F},\Psi}$ (Метод Наименьших Квадратов).

Пусть $\widetilde{\mathbb{S}} := \mathbb{F}\mathbb{F}^{\mathrm{T}} + \Psi$. Тогда $\sum\limits_{i,j}(S_{ij} - \widetilde{S}_{ij})^2 \to \min$.

$$\begin{cases} \sum\limits_{i\neq j}(S_{ij} - \sum\limits_{k=1}^{r} f_{ik}f_{jk})^2 \to min, \\ (\mathbb{F}\mathbb{F}^{\mathrm{T}})_{ii} \leq 1 \Rightarrow \sigma_i^2 = 1 - (\mathbb{F}\mathbb{F}^{\mathrm{T}})_{ii}. \end{cases}$$

Minres — minimization residual correlations (минимизация разницы известных корреляций и той их частью, что объясняется факторами).

АГК: $\| \mathbb{Y} - \widetilde{\mathbb{Y}} \| \to \min\limits_{rank\widetilde{\mathbb{Y}}\leq r} \Leftrightarrow \| \mathbb{Y}\mathbb{Y}^{\mathrm{T}} - \widetilde{\mathbb{Y}}\widetilde{\mathbb{Y}}^{\mathrm{T}} \| \to \min\limits_{rank\widetilde{\mathbb{Y}}\leq r}$

**Тут надо все пояснить про АГК, но меня не было на лекции той, простите-помогите.**

## 1.39. 39. Какой вид имеет функция правдоподобия в ФА?

Пусть $\xi \sim N(0, \Sigma)$.

$\mathcal{L}(\mathbb{X}; \mathbb{F}, \Psi) = \prod\limits_{i=1}^{n} \frac{1}{(2\pi)^{n/2}\det\Sigma^{1/2}} e^{-\frac{1}{2}X_i^{\mathrm{T}}\Sigma^{-1}X_i} = \prod\limits_{i=1}^{n} \frac{1}{(2\pi)^{n/2}\det(\mathbb{F}\mathbb{F}^{\mathrm{T}}+\Psi)^{1/2}} e^{-\frac{1}{2}(X_i^{\mathrm{T}}(\mathbb{F}\mathbb{F}^{\mathrm{T}})+\Psi^{-1})X_i}$.

Вместо $\mathbb{X}$ рассм. выборочная ковар. матрица $\mathbb{S}$. $\mathcal{L}(\mathbb{S}; \mathbb{F}, \Psi) \sim W_p(\Sigma)$.

Так как решений бесконечно много с точностью до вращений, и чтобы как-то зафиксировать$\mathcal{L}(\mathbb{S}; \mathbb{F}$ $\max\limits_{\mathbb{F},\Psi}$ Здесь кажется что-то потерялось???

## 1.40. 40. Проверка значимости модели ФА.

**Не уверена, что это тот самый вопрос.**

$H_0 : \Sigma = \mathbb{F}\mathbb{F}^{\mathrm{T}} + \Psi$

Статистика критерия $t = \left(n - 1 - \frac{2p+4r-5}{6}\right) ln\left(\frac{|\hat{\mathbb{F}}\hat{\mathbb{F}}^{\mathrm{T}}++\hat{\Psi}|}{|\hat{\mathbb{S}}|}\right) \sim \chi^2\left(\frac{(p+r)^2-(p+r)}{2}\right)$.

## 1.41. 41. Критерий сферичности Бартлетта, для чего нужен

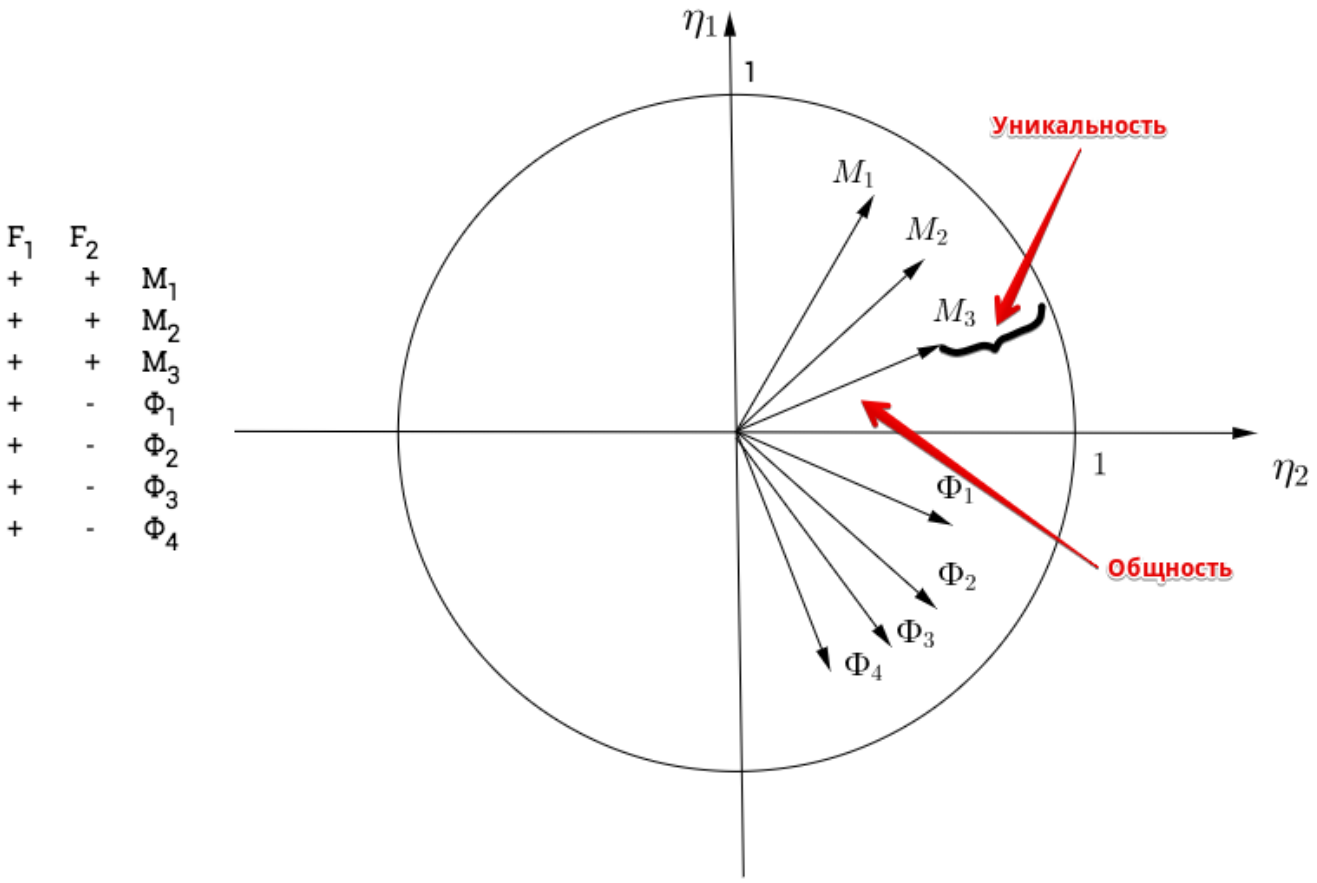
Частный случай критерия (Вопрос 40 ) при $r = 0$, имеет смысл проверять перед поиском факторов, вдруг общих факторов совсем нет.

$H_0 : \Sigma = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & 1 & \vdots \\ 0 & \cdots & 1 \end{pmatrix}$ — сферичность (т.к. данные выглядят как сфера).

Статистика критерия $t = \left(n - 1 - \frac{2p-5}{6}\right) ln\left(\frac{1}{\hat{\mathbb{S}}}\right) \sim \chi^2\left(\frac{p^2-p}{2}\right)$.

## 1.42. 42. Что такое общность и уникальность признака? Какие факторы не находит факторный анализ?

$\sum\limits_{j=1}^{r} f_{ij}^2 = 1 - D(\varepsilon_i)$ — communality(общность), $D(\varepsilon_i)$ — уникальность.




***Пример.*** *При повороте осей примерно на* $45^\circ$ *получим скрытые факторы — это способности по математике и по физике.* ***Не нашла материла про то, какие факторы не находит факторый анализ.***

Вспомним, что в $F$ не бывает признаков вида $(0,\dots,0,a,0,\dots,0)^{\mathrm{T}}$. Поэтому «уникальные» признаки и шум нам разделить не удастся. В частности, может оказаться так, что то, что мы трактуем как шум — на самом деле не шум, а естественный признак объекта.

## 1.43. Билет 43. Общность как множественный коэффициент корреляции.

Тут вроде даже сама НЭ не может ответить что надо говорить. А я уж тем более...

## 1.44. Билет 44. Как интерпретируются признаки в ФА?

$$\mathbb{F} = \begin{pmatrix} f_{11} & \cdots & f_{1r} \\ \vdots & \vdots & \vdots \\ f_{p1} & \cdots & f_{pr} \end{pmatrix}, f_{ij} = \rho(\xi_i, \eta_j)$$

Специфика факторного анализа в том, что факторные значения нам вообще не нужны, в базовой постановке (для интерпретации). Для их интерпретации хватает и $\mathbb{F}$. То есть можно сказать, что элементы матрицы факторных нагрузок — это корреляция между исходными признаками и факторными значениями. С чем коррелирует фактор, то и объясняет.

## 1.45. Билет 45. Зачем нужны вращения в ФА? Как устроены ортогональные вращения?

Предположим, что мы нашли $\tilde{\xi} = \mathbb{F}\eta$. Пусть $\mathbb{W}$ — ортогональна матрица вращения в $R^r$, $\eta' = \mathbb{W}\eta$. При этом после ортогонального вращения факторных значений они так и останутся ортогональными (если были исходно ортогональными). Тогда

$$\tilde{\tilde{\xi}} = \mathbb{F}\mathbb{W}\mathbb{W}^{\mathbb{T}}\eta^{29} = \mathbb{F}'\eta. \tag{5}$$

Модель при этом остается верной. То есть факторы определяются не единственным образом. При этом $\eta'$ удовлетворяет тем же условиям. Выбирая матрицу вращения, мы можем упрощатиь интерпретацию факторов. Если мы возьмем $r = p$ при вращении, то мы получим тоже самое (не корректный результат).

## 1.46. Билет 46. Вращение по методу varimax.

**Интерпретация факторов в ФА.**

$$\mathbb{F} = \begin{pmatrix} x & 0 & x & \dots \\ x & 0 & x & \dots \\ x & x & 0 & \dots \\ x & x & 0 & \dots \\ x & x & 0 & \dots \\ x & x & 0 & \dots \\ x & x & 0 & \dots \end{pmatrix}$$

Первый и второй столбец — такие факторы плохие. Второй и третий идеально. Желательно, чтобы факторы не пересекались. И то, что у нас есть мы можем вращать. $\widetilde{\mathbb{F}} = \mathbb{F}\mathbb{W}$. $\mathbb{X} = \mathbb{V}\mathbb{F}^{\mathbb{T}} = \widetilde{\mathbb{V}}\widetilde{\mathbb{F}}$, $\widetilde{\mathbb{V}} = \mathbb{V}\mathbb{W}$. Если мы будем изменять $\mathbb{W}$ то мы будем изменять матрицу $\mathbb{F}$.

Осталось только понять как и что нам надо улучшать? Хорошая характеристика для столбца (1) и (2) это стандартное отклонение. Для первого будет равна нулю. Чем больше «контрастность», тем характеристика будет больше. Метод varimax:

$$\sum_{j=1}^{r}\left[\frac{1}{p}\sum_{i=1}^{p}(f_{ij}^2)^2 - \left(\frac{1}{p}\sum_{i=1}^{p}\tilde{f}_{ij}^2\right)^2\right] \to \max_{w} \tag{6}$$

*Varimax*: ищем простые факторы (то есть максимальная контрастность).*Quartimax*: простая интерпретация признаков (берем sd по строчкам). *Equimax*: и то и то. Важно, что после вращения $\widetilde{\mathbb{V}}$ будут тоже ортогональны (столбцы).

Для лучшей интерпретации факторов иногда разрешают неортогональные вращения: *oblique, oblimin* — косоугольные вращения. Но при таком типе вращения можно получить фактор там, где их нет.

## 1.47. Билет 48. Факторная структура (корреляции исходных признаков с факторами) и факторный паттерн (коэффициенты лин. комбинации, с которыми исходные признаки выражаются через факторы) в случае ортогональных и неортогональных факторов.

***Определение.*** *Factor structure — корреляция между исходными признаками и факторными значениями.*

[29] Ясно, что это вектор

***Определение***. *Factor pattern — коэффициенты линейной комбинации, как исходные признаки выражаются через факторы.*

Если факторы ортогональны, то это одно и тоже. То есть это будет просто матрица $\mathbb{F}$. Но может быть неортогональны (например вращение неортогонально), то эти вещи разные. Набирала Стася

## 1.48. Билет 47. Методы нахождения факторных значений: LS и WLS (метод Бартлетта)

Матрица наблюдений $\mathbb{X} \in \boldsymbol{M}_{n,p}(\mathbb{R})$ в факторном анализе представляется как

$$\mathbb{X} = \mathbb{V}\mathbb{F}^{\mathrm{T}} + \varepsilon, \mathbb{V} \in \boldsymbol{M}_{n,r}(\mathbb{R}), \mathbb{F} \in \boldsymbol{M}_{p,r}(\mathbb{R}), \varepsilon \in \boldsymbol{M}_{n,p}(\mathbb{R}).$$

Перепишем для $\mathbb{Y} = \mathbb{X}^{\mathrm{T}}, \mathbb{W} = \mathbb{V}^{\mathrm{T}}$ (матрица $\mathbb{Y} = \begin{bmatrix} Y_1 : \cdots : Y_n \end{bmatrix}$ содержит наблюдения в столбцах):

$$\mathbb{Y} = \left(\mathbb{V}\mathbb{F}^{\mathrm{T}} + \varepsilon\right)^{\mathrm{T}} = \mathbb{F}\mathbb{V}^{\mathrm{T}} + \varepsilon^{\mathrm{T}} \implies \forall i \in 1 : n \; Y_i = \mathbb{F}W_i + \varepsilon_i.$$

$$\begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix} = \mathbb{F} \begin{pmatrix} w_{i1} \\ \vdots \\ w_{ir} \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{ip} \end{pmatrix} \tag{7}$$

### 1.48.1. Ordinary Least Squares

При наличии уже оцененной матрицы факторных нагрузок $\mathbb{F}$ выражение 7 может быть рассмотрено как задача линейной регрессии $W_i = (V^{\mathrm{T}})_i$ на $\mathbb{F}$. Тогда для каждого наблюдения $Y_i$ можно построить соответствующие значения факторов

$$\hat{W}_i = \left(\mathbb{F}^{\mathrm{T}}\mathbb{F}\right)^{-1} \mathbb{F}^{\mathrm{T}} Y_i.$$

Модель линейной регрессии работает в предположении о гомоскедастичности (одинаковой вариантивности) данных, т.е. что $\varepsilon_{i1}, \ldots, \varepsilon_{ip}$ — i.i.d. Модель фактороного анализа делает более слабое предположение: $\left(\varepsilon_{i1}, \ldots, \varepsilon_{ip}\right) \sim N(0, \operatorname{diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right))$, т.е. вариативность (степень уникальности в контексте факторного анализа) у разных наблюдений разная.

### 1.48.2. Weighted Least Squares (метод Бартлетта)

[30] Приведём наши данные к виду, в котором они бы удовлетворяли модели линейной регрессии. Так, если $\varepsilon_i \sim N(0, \Psi), \Psi = \operatorname{cov}\varepsilon_i = \operatorname{diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right), \Psi^{-1/2} = \operatorname{diag}\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_p}\right)$, то

$$\operatorname{cov}\left(\Psi^{-1/2}\varepsilon_i\right) = \Psi^{-1/2}\Psi\Psi^{-1/2} = \mathbb{I}_{p\times p} \implies \Psi^{-1/2}\varepsilon_i \sim N(0, \mathbb{I}_{p\times p})$$

Перепишем уравнение модели ФА так, чтобы уникальности в нём были вида $\Psi^{-1/2}\varepsilon_i$:

$$\Psi^{-1/2}Y_i = \Psi^{-1/2}\mathbb{F}W_i + \Psi^{-1/2}\varepsilon_i \iff \Psi^{-1/2}Y_i = \left(\mathbb{F}^{\mathrm{T}}\Psi^{-1/2}\right)^{\mathrm{T}} W_i + \Psi^{-1/2}\varepsilon_i, \tag{8}$$

тогда соответствующее выражение для оценки $W_i$ по МНК имеет вид

$$\hat{W}_i = \left(\mathbb{F}^{\mathrm{T}}\Psi^{-1/2}\Psi^{-1/2}\mathbb{F}\right)^{-1} \mathbb{F}^{\mathrm{T}}\Psi^{-1/2}\Psi^{-1/2}Y_i = \left(\mathbb{F}^{\mathrm{T}}\Psi^{-1}\mathbb{F}\right)^{-1} \mathbb{F}^{\mathrm{T}}\Psi^{-1}Y_i.$$

[30] Я не смогла найти публикации, где бы в явном виде был изложен именно этот «самый простой» вариант

## 1.49. Билет 48. Factor structure и factor pattern в случае ортогональных и неортогональных векторов

***Определение.*** Factor structure — *корреляции исходных признаков с факторами,* $\operatorname{cov}(\xi, \eta)$.

***Определение.*** Factor pattern — *коэффициенты выражения исходных признаков через факторы.*

Модель факторного анализа: $\xi = \mathbb{F}\eta + \varepsilon, \operatorname{cov}\eta = I_{r\times r}$. Любое вращение с матрицей вращения $\mathbb{W}$ выражается как $\xi = \mathbb{F}\mathbb{W}^{-1}\mathbb{W}\eta + \varepsilon = \mathbb{F}'\eta', \mathbb{W} \in M_r(\mathbb{R})$. Тогда

$$\operatorname{cov}\eta' = \operatorname{cov}\mathbb{W}\eta = \mathbb{W}\operatorname{cov}\eta\mathbb{W}^{\mathrm{T}} = \mathbb{W}\mathbb{W}^{\mathrm{T}}$$

$\mathbb{W}\mathbb{W}^{\mathrm{T}} \neq I_{r\times r}$, так как для неортогональных вращений $\mathbb{W}^{\mathrm{T}} \neq \mathbb{W}^{-1}$.

Посчитаем factor structure:

$$\begin{aligned}\operatorname{cov}(\xi,\eta') = \operatorname{cov}(\mathbb{F}'\eta' + \varepsilon, \eta') = \operatorname{cov}(\mathbb{F}'\eta', \eta') + \operatorname{cov}(\varepsilon, \mathbb{W}\eta) = \\ = \mathbb{F}'\operatorname{cov}\eta' + 0 = \mathbb{F}'\mathbb{W}\mathbb{W}^{\mathrm{T}} = \mathbb{F}\mathbb{W}^{\mathrm{T}}.\end{aligned}$$

Если вращение было ортогональным, то $\mathbb{W}\mathbb{W}^{\mathrm{T}} = \mathbb{W}\mathbb{W}^{-1} = I$ и $\operatorname{cov}(\xi,\eta') = \mathbb{F}$.

Factor structure — выражения исходных признаков через факторы, выписывается по определению модели:

$$\xi_i = f'_{i1}\eta'_{i1} + \cdots + f'_{ir}\eta'_{ir} + \varepsilon_i$$

Следовательно, матрица коэффициентов линейной комбинации выражения исходных признаков через факторы — это просто $\mathbb{F}'$.

# 2. Вторая часть

## 2.1. Билет 1. Распределение Уишарта, свойства

***Определение.*** *Пусть* $\xi_1, \dots, \xi_m \in \mathbb{R}^p$ *— независимые;* $\xi_i \in \mathcal{N}_p(0,\Sigma)$. *Тогда* $\eta = \eta_m = \sum_{i=1}^{m} \xi_i\xi_i^{\mathrm{T}}$ *имеет распределение Уишарта.*

***Обозначение.*** $\eta_m \sim \mathcal{W}_p(\Sigma, m)$

***Свойство*** (1). *Частный случай,* $p = 1$. *Случайная величина* $\eta_m$ *имеет распределение* $\chi^2$ *с точностью до нормировки:*

$$\frac{\eta_m}{\sigma^2} \sim \chi^2(m)$$

***Свойство*** (2). $\mathcal{E}\eta = m \cdot \Sigma$

***Свойство*** (3).

***Предложение.*** *Пусть* $w \sim \mathcal{W}_p(\Sigma, m)$, $\mathbb{C}^{q\times p}$ *— матрица,* $\operatorname{rk}\mathbb{C} = q$ *(матрица полного ранга). Тогда* $\mathbb{C}w\mathbb{C}^{\mathrm{T}} \sim \mathcal{W}_p(\Sigma, m)$

*Доказательство. Следует из свойства нормального распределения:* $\xi \sim \mathcal{N}(0,\Sigma) \implies \mathbb{C}\xi \sim \mathcal{N}(0,\mathbb{C}\Sigma\mathbb{C}^{\mathrm{T}})$ □

***Следствие.*** $w \sim \mathcal{W}_p(\Sigma, m)$. *Какое распределение имеет диагональный элемент?* ***Условно (так писать не надо, для понимания):*** $(\Sigma)_{ii}\chi^2(m)$. *Как это получить на самом деле? Берем вместо матрицы* $\mathbb{C}$ *вектор, где на нужном месте (элемент на диагонали, который нас интересует) стоит* $1$*:* $\mathbb{C} = (0, \dots, 0, \underset{(i)}{1}, 0, \dots 0)$. *В итоге получим распределение* $\chi^2$ *с соответствующей нормировкой.*

## 2.2. Билет 2. Pooled covariance matrix

## 2.3. Билет 3. Распределение Hotelling'a, свойства

***Определение***. *Пусть* $w \sim \mathcal{W}_p(\Sigma, m)$, $\xi \sim \mathcal{N}(\mu,\Sigma)$. *Тогда* $T^2 = (\xi - \mu)^{\mathrm{T}}\left(\frac{w}{m}\right)^{-1}(\xi - \mu)$ *имеет распределение Хотеллинга с* $m$ *степенями свободы.*

***Обозначение***. $T^2 \sim T_p^2(m)$

***Свойство*** (1). *Частный случай,* $p = 1$. $T_1^2(m) = (t(m)^2)$ *(обычный Стьюдент в квадрате).*

***Свойство*** (2). $\left(\frac{m-p+1}{p}\right)T_p^2(m) = F_{p,m-p+1}$. *Без доказательства.*

***Свойство*** (3). $T_p^2(m) \xrightarrow[m\to\infty]{\sim} \chi^2(p)$.

## 2.4. Билет 6. Проверка гипотезы о сравнении многомерных мат. ожиданий, независимые выборки

Пусть $\xi^{(1)}, \xi^{(2)} \in \mathbb{R}^p$ — независимые. При этом вектор $\xi^{(1)}$ имеет длину $n_1$, а $\xi^{(2)}$ имеет длину $n_2$. Проверяем гипотезу:

$H_0 : \mathcal{E}\xi^{(1)} = \mathcal{E}\xi^{(2)}$.

Пример: есть 2 группы людей. Снимаются показания по росту, весу, длине рук и т.д. Необходимо сравнить так называемые "средние размеры". Рассмотрим случаи:

- $\Sigma_1 = \Sigma_2 = \Sigma$ — известна. Тогда статистика критерия имеет вид:

  $$t = (\overline{x}^{(1)} - \overline{x}^{(2)})^{\mathrm{T}}\left(\Sigma(\frac{1}{n_1} + \frac{1}{n_2})\right)^{-1}(\overline{x}^{(1)} - \overline{x}^{(2)}) \sim \chi^2(p),$$

  где $\xi^{(i)} \sim \mathcal{N}_p(\mu,\Sigma)$ (иначе, асимптотически), $(\overline{x}^{(1)} - \overline{x}^{(2)})$ — расстояние Махаланобиса (от этой разности до 0), $\Sigma = \operatorname{cov}\xi^{(i)}$ — ковариационная матрица разности (но тут ковариационная матрица одинаковая для обеих выборок, поэтому в формуле такое выражение). Вообще говоря, $\operatorname{cov}\overline{x} = \frac{\Sigma}{n}$.

- $\Sigma_1 = \Sigma_2 = \Sigma$ — неизвестна. Тогда берем pooled covariance matrix $\widetilde{\mathbb{S}} = \frac{(n_1-1)\mathbb{S}^{(1)}+(n_2-1)\mathbb{S}^{(2)}}{n_1+n_2-2}$, где $\mathbb{S}^{(1)}, \mathbb{S}^{(2)}$ — выборочные ковариационные матрицы для векторов $\xi^{(1)}, \xi^{(2)}$, соответственно. Статистика критерия:

  $$t = (\overline{x}^{(1)} - \overline{x}^{(2)})^{\mathrm{T}}\left(\widetilde{\mathbb{S}}(\frac{1}{n_1} + \frac{1}{n_2})\right)^{-1}(\overline{x}^{(1)} - \overline{x}^{(2)}) \sim T_p^2(n_1 + n_2 - 2),$$

  где $\xi^{(i)} \sim \mathcal{N}_p(\mu,\Sigma)$ (иначе, асимптотически).

- Неизвестно, что $\Sigma_1 = \Sigma_2$. Статистика критерия:

  $$t = (\overline{x}^{(1)} - \overline{x}^{(2)})^{\mathrm{T}}\left(\frac{\widetilde{\mathbb{S}}_1}{n_1} + \frac{\widetilde{\mathbb{S}}_2}{n_2}\right)^{-1}(\overline{x}^{(1)} - \overline{x}^{(2)}) \xrightarrow[n_1,n_2\to\infty]{\sim} \chi^2(p),$$

  где $\mathbb{S}_1, \mathbb{S}_2$ — pooled covariance matrices для каждого вектора.

## 2.5. Билет 7. Для чего используется статистика Box's M?

**Гипотеза о равенстве ковариационных матриц** $\xi^{(i)} \in \mathbb{R}^p, \xi^{(i)} \sim \mathcal{N}_p(\mu_i, \Sigma_i), i = 1, \dots, k$. Проверяем гипотезу (о гомостохастичности): $H_0 : \Sigma_1 = \cdots = \Sigma_k$. Есть $k$ групп. Считаем несмещенные оценки ковариационных матриц, хотим проверить их равенство (критерий отношения дисперсий, одномерный случай).

$$M = \left(\frac{|\widetilde{\mathbb{S}}_1|}{\widetilde{\mathbb{S}}}\right)^{\nu_1/2} \dots \left(\frac{|\widetilde{\mathbb{S}}_k|}{\widetilde{\mathbb{S}}}\right)^{\nu_k/2},$$

где $\nu_i = n_i - 1$ — число степеней свободны, а $\widetilde{\mathbb{S}} = \frac{(n_1-1)\widetilde{\mathbb{S}}_1+\cdots+(n_k-1)\widetilde{\mathbb{S}}_k}{n_1+\cdots+n_k-k}$.

Считаем статистику критерия (Box's statistics):

$$t = -\log M$$

## 2.6. Билет 10. Единый подход к множественной регрессии и одномерному однофакторному дисперсионному анализу (ANOVA)

Отправной точкой для описания обоих статистических методов является известное Дисперсионное Тождество. В рамках ответа на вопрос сложно объяснить значимость этого равенства, однако следует иметь в виду, что корни уходят далеко за пределы статистики. Технически теорией, лежащей в основе простой идеи проекции случайных величин на другие случайные величины, является теория Условных Математических Ожиданий (в дальнейшем, УМО). В рамках этой теории, в частности, удается сформулировать вариант Теоремы Пифагора для случайных величин, который и называется Дисперсионным Тождеством.

***Теорема*** (Основное Дисперсионное Тождество, без доказательства). *Пусть $\eta, \xi$ — случайные величины, причем $\eta$, обладающая конечным вторым моментом. Тогда*

$$\mathbb{D}\eta = \mathbb{E}(\eta - \mathbb{E}(\eta \mid \xi))^2 + \mathbb{E}(\mathbb{E}(\eta \mid \xi) - \mathbb{E}\eta)^2, \tag{9}$$

*где $\mathbb{E}(\eta \mid \xi)$ — УМО $\eta$ «при условии» $\xi$.*

Стоит отметить несколько фактов:

1. Математическое ожидание УМО есть математическое ожидание исходной случайной величины:
$$\mathbb{E}\mathbb{E}(\eta \mid \xi) = \mathbb{E}\eta. \tag{10}$$
2. Математическое ожидание квадрата есть на самом деле ни что иное, как квадрат нормы в пространстве $L^2$, порожденным соответствующим распределением.
3. Как и в любой нормальной теореме Пифагора слагаемые в правой части (9) (точнее то, что стоит под нормой) оказываются ортогональными, то есть
$$\mathbb{E}\left[(\eta - \mathbb{E}(\eta \mid \xi))\mathbb{E}(\eta \mid \xi)\right] = 0. \tag{11}$$
4. УМО $\mathbb{E}(\eta \mid \xi)$ — есть результат ортогонального[31] проецирования $\eta$ на подпространство, порожденное случайной величиной $\xi$.

Рассмотрим теперь ANOVA, в которой используется частный случай Основного Тождества.

[31] именно это уточнено в предыдущем пункте

**ANOVA** Общая постановка задачи. Пусть есть $k$ групп и $p$ признаков, мы пытаемся проверить, что группы друг от друга не отличаются. Формально есть $k$ (одномерных) случайных величин $\eta_1, \dots, \eta_k$ и гипотеза формулируется так:

$$\mathrm{H}_0 \,:\, \mathcal{P}(\eta_1) = \mathcal{P}(\eta_2) = \dots = P(\eta_k). \tag{12}$$

Эту задачу можно переформулировать следующим образом. Пусть есть дискретный (м.б. качественный) признак $\xi$, принимающий ровно $k$ значений: $A_1, A_2, \dots, A_k$. Рассмотрим случайный вектор $(\eta, \xi)^{\mathrm{T}}$, такой что $\mathcal{P}(\eta \mid \xi = A_i) = \mathcal{P}(\eta_i)$ для всех $i \in 1 : k$. Тогда гипотеза (12) переписывается в виде:

$$\mathrm{H}_0^* \,:\, \mathcal{P}(\eta \mid \xi = A_i) = \mathcal{P}(\eta \mid \xi = A_j) \text{ для всех } i, j. \tag{13}$$

В ANOVA рассматривается частный случай (модель), когда $\mathcal{P}(\eta_i) = \mathcal{N}(\mu_i, \sigma^2)$. В рамках этой модели гипотезам (12) и (13) равносильны следующие две[32] гипотезы:

$$\mathrm{H}_0 \,:\, \mu_1 = \mu_2 = \dots = \mu_k. \tag{14}$$

$$\mathrm{H}_0^* \,:\, \mathbb{E}(\eta \mid \xi = A_1) = \dots = \mathbb{E}(\eta \mid \xi = A_k). \tag{15}$$

Заметим, что в рамках Основного Дисперсионного Тождества гипотезу можно сформулировать следующим образом:

$$\mathrm{H}_0 \,:\, \mathbb{E}(\mathbb{E}(\eta \mid \xi) - \mathbb{E}\eta)^2 = 0, \tag{16}$$

это запись по большому счету и означает, что средние внутри групп не отличаются от общего среднего.

Выборка задается следующим образом: имеется $n_i$ индивидов из $i$-той группы ($i \in 1 : k$). Обозначим $y_{ij}$ — $j$-того индивида из $i$-той группы ($i \in 1 : k$ и $j \in 1 : n_i$). Обозначим $\overline{y}$ — выборочное среднее по всем индивидам, а $\overline{y}_i$ — выборочное среднее индивидов $i$-той группы. Запишем (9) на выборочном языке:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2 + \sum_{i=1}^{k} n_i(\overline{y}_i - \overline{y})^2. \tag{17}$$

Убедитесь, что понимаете (хотя бы интуитивно), откуда это получается!

Далее факты из курса статистики:

$$\mathrm{SSTotal} = \mathrm{SST} \overset{\text{def}}{=} \widehat{\mathbb{D}\eta} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2 \sim \sigma^2\chi^2(n-1);$$

$$\mathrm{SSWithin} = \mathrm{SSW} \overset{\text{def}}{=} \widehat{\mathbb{E}(\eta - \mathbb{E}(\eta \mid \xi))^2} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2 \sim \sigma^2\chi^2(n-k);$$

$$\mathrm{SSBetween} = \mathrm{SSB} \overset{\text{def}}{=} \widehat{\mathbb{E}(\mathbb{E}(\eta \mid \xi) - \mathbb{E}\eta)^2} = \sum_{i=1}^{k} n_i(\overline{y}_i - \overline{y})^2 \sim \sigma^2\chi^2(k-1).$$

Сама статистика имеет вид

$$F = \frac{\mathrm{SSB}/(k-1)}{\mathrm{SSW}/(n-k)} \sim \mathrm{F}_{k-1,n-k}.$$

[32] равносильные

Убедитесь, что понимаете, что при условии знания распределений $\chi^2$, тут правда получается ФИШЕР! Откуда берется числитель должно быть ясно, из представления гипотезы в виде (16), знаменатель тоже можно объяснить, но это чуть более тонкое дело.

Также в ANOVA рассматривают следующее корреляционное отношение[33]:

$$r^2 = \hat{r}^2(\eta \mid \xi) = \frac{\mathrm{SSB}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSW}}{\mathrm{SST}} = 1 - \frac{\min_\varphi \widehat{\mathbb{E}}(\eta - \varphi(\xi))^2}{\widehat{\mathbb{D}}\eta}. \tag{18}$$

**Multiple Linear Regression** Здесь мы не будем касаться вопросов построения МНК-оценок, а сразу перейдем к проверке значимости регрессии. Уточним смысл сказанного. Во-первых, вспомним, что по определению регрессия — УМО. Во-вторых, какая бы модель у нас ни была, в стандартной ситуации нам нужно от этой модели только, чтобы УМО было линейным (это будет ясно чуть далее). Существует несколько моделей, в которых УМО — линейно. Приведем примеры двух из них.

$$\eta = \boldsymbol{B}^{\mathrm{T}}\overline{\xi} + \varepsilon,$$

где $\overline{\xi} \in \mathbb{R}^k$ — случайный вектор, $\eta \in \mathbb{R}$, $\varepsilon \in \mathbb{R}$ — случайные величины, $\boldsymbol{B} \in \mathbb{R}^k$ — вектор коэффициентов, причем $\overline{\xi}$, $\varepsilon$ — независимы и $\mathbb{E}\varepsilon = 0$.

В такой модели нетрудно посчитать, что $\mathbb{E}(\eta \mid \overline{\xi}) = \boldsymbol{B}^{\mathrm{T}}\overline{\xi}$, то есть регрессия является линейной, нас это устраивает.

Примером другой модели такого типа является следующая: $(\eta, \overline{\xi})^{\mathrm{T}} \sim \mathcal{N}$ с какими-то параметрами, которые нас сейчас не интересуют. Важно то, что в такой модели тоже найдется вектор $\boldsymbol{B}$: $\mathbb{E}(\eta \mid \overline{\xi}) = \boldsymbol{B}^{\mathrm{T}}\overline{\xi}$.

Гипотеза о незначимости регрессии в терминах MLR формулируется следующим образом:

$$\mathrm{H}_0 \,:\, \boldsymbol{B} = 0. \tag{19}$$

Именно здесь используется то, что мы требуем линейности регрессии (найдется такой вектор!).

Учитывая то обстоятельность, что для независимых $\xi_1, \xi_2$ справедливо, что $\mathbb{E}(\xi_1 \mid \xi_2) = \mathbb{E}\xi_1$, гипотеза о незначимости регрессии для моделей можно свести к другой, сформулированной в рамках Основного Дисперсионного Тождества (9):

$$\mathrm{H}_0 \,:\, \mathbb{E}(\mathbb{E}(\eta \mid \overline{\xi}) - \mathbb{E}\eta)^2 = 0. \tag{20}$$

Несложно понять, что эта запись есть то же самое, что и (19).

Заметим, что вторая гипотеза является гораздо более общей, в ней не нужны предположения о линейности регрессии! Обратим внимание, что эта гипотеза полностью совпадает по своему виду с гипотезой (16), которая проверяется с помощью ANOVA.

Выборка задана набором $\{y_i\}_{i=1}^n$. Обозначим $\hat{y}_i$ — наилучшее[34] по МНК предсказание $y_i$. Запишем (9) на выборочном языке:

$$\sum_{i=1}^n (y_i - \overline{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \overline{y})^2. \tag{21}$$

Убедитесь, что понимаете (хотя бы интуитивно), откуда это получается!

[33] Это тоже термин из УМО.

[34] Все должны помнить в каком смысле.

Далее факты из курса статистики:

$$\text{SSTotal} = \text{SST} \stackrel{\text{def}}{=} \widehat{\mathbb{D}\eta} = \sum_{i=1}^{n} (y_i - \overline{y})^2 \sim \sigma^2 \chi^2(n-1);$$

$$\text{SSError} = \text{SSE} \stackrel{\text{def}}{=} \mathbb{E}(\eta - \widehat{\mathbb{E}(\eta \mid \xi)})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \sim \sigma^2 \chi^2(n-k-1);$$

$$\text{SSRegression} = \text{SSR} \stackrel{\text{def}}{=} \mathbb{E}(\widehat{\mathbb{E}(\eta \mid \xi)} - \mathbb{E}\eta)^2 = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 \sim \sigma^2 \chi^2(k).$$

Сама статистика имеет вид

$$F = \frac{\text{SSR}/k}{\text{SSE}/(n-k-1)} \sim \text{F}_{k,n-k-1}.$$

Убедитесь, что понимаете, что при условии знания распределений $\chi^2$, тут правда получается ФИШЕР!

Так же при регрессии рассматривают множественный коэффициент корреляции между $\eta$, $\xi$:

$$r^2 = \hat{\text{R}}^2(\eta; \{\xi_1, \xi_2, \dots, \xi_k\}) = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} = \\ = №1 - \frac{\min_{a_0, a_1, \dots, a_k} \hat{\mathbb{E}}(\eta - (a_0 + \sum_{i=1}^{k} a_i \xi_i))^2}{\widehat{\mathbb{D}}\eta}. \tag{22}$$

**Выводы** Я думаю, что даже тот, кто просматривал этот текст одним глазом заметил, что я написал два раза почти одно и то же. Теперь дадим некоторый общий (и на самом деле удобный) взгляд на то, что было раньше.

Начнем с гипотез. Заметим, что (16) и (20) вообще ничем не отличается, поэтому можно ввести следующую удобную универсальную терминологию (шляпа, как всегда обозначает оценки, смысл которых уточняется в частных случаях).

$$\text{Hypothesis} \stackrel{\text{def}}{=} \mathbb{E}(\mathbb{E}(\eta \mid \xi) - \mathbb{E}\xi)^2.$$

$$\text{Error} \stackrel{\text{def}}{=} \mathbb{E}(\eta - \mathbb{E}(\eta \mid \xi))^2;$$

$$\text{SSTotal} = \text{SST} \stackrel{\text{def}}{=} \widehat{\mathbb{D}\eta};$$

$$\text{SSError} = \text{SSE} \stackrel{\text{def}}{=} \mathbb{E}(\eta - \widehat{\mathbb{E}(\eta \mid \xi)})^2;$$

$$\text{SSHypothesis} = \text{SSH} \stackrel{\text{def}}{=} \mathbb{E}(\widehat{\mathbb{E}(\eta \mid \xi)} - \mathbb{E}\eta)^2.$$

В этих обозначениях общая гипотеза будет иметь вид:

$$\text{H}_0 : \text{Hypothesis} = 0.$$

А Дисперсионное Тождество на выборочном языке:

$$\text{SST} = \text{SSH} + \text{SSE}.$$

Статистика в общем виде будет выглядеть, как

$$F = \frac{\text{SSH}/\nu_H}{\text{SSE}/\nu_E} \sim \text{F}_{\nu_H, \nu_E},$$

где $\nu_\cdot$ — нормировочные числа, отвечающие за степени свободы.

На будущее вводим обозначения:

$$\lambda \stackrel{\text{def}}{=} \frac{\text{SSH}}{\text{SSE}};$$
$$\Lambda \stackrel{\text{def}}{=} \frac{1}{1+\lambda} \sim \Lambda_1(\nu_H, \nu_E);$$ [35]

Обобщающее (18) и (22) определение:

$$r^2 = \frac{\text{SSH}}{\text{SST}}.$$

## 2.7. Билет 11. Представление одномерного однофакторного дисперсионного анализа в виде множественной регрессии с фиктивными переменными.

Вот есть ANOVA и есть многомерная регрессия... У них много общего. Сейчас мы покажем, что на самом деле на ANOVA можно посмотреть, как на частный случай многомерной регрессии. Будем пользоваться обозначениями билета 10. Для $i \in 1 : k-1$ введем $\xi_i = \mathbb{I}(\xi = A_i)$ – фиктивные переменные.

***Теорема.*** $r^2(\eta \,|\, \xi) = \mathrm{R}^2(\eta, \{\xi_1, \xi_2, \dots, \xi_{k-1}\})$.

*Доказательство.* Распишем по определению обе части равенства.

$$r^2(\eta \,|\, \xi) = 1 - \frac{\min_\varphi \mathbb{E}(\eta - \varphi(\xi))^2}{\mathbb{D}\eta};$$
$$\mathrm{R}^2(\eta; \{\xi_1, \xi_2, \dots, \xi_k\}) = 1 - \frac{\min_{a_0, a_1, \dots, a_k} \mathbb{E}(\eta - (a_0 + \sum_{i=1}^k a_i \xi_i))^2}{\mathbb{D}\eta}.$$

Минимум в корреляционном отношении идет по всем измеримым функциям, каждая из которых однозначно определяется набором значений $\{\varphi(A_i)\}$. Пусть минимум в корреляционном отношении достигается на функции $\varphi^*$. Положим $a_i = \varphi^*(A_i)$ для $i \in 1 : (k-1)$ и $a_0 = \varphi^*(A_k)$. От противного легко доказать, что именно на этом наборе коэффициентов достигается минимум в коэффициенте корреляции. В обратную сторону – аналогично.[36] □

## 2.8. Билет 12. Корреляционное отношение с дискретным одномерным признаком и множественный коэффициент корреляции.

По-моему, это тот же 11ый билет.

## 2.9. Билет 13. Обобщенная задача на собственные значения

### Алгебраические факты

1. Пусть $\mathbb{C} \in \mathbb{R}^{q \times p}$ — неотрицательно определенная матрица. $\lambda_1 \geq \ldots \geq \lambda > 0$ — собственные числа.

   Тогда $\mathrm{tr}\mathbb{C} = \sum_{i=1}^p \lambda_i$; $|\mathbb{C}| = \prod_{i=1}^p \lambda_i$.

[35] Докажите это! НЭ будет это спрашивать. Удачи.

[36] Вроде бы тут все корректно. Проверьте, кто может.

При этом верно следующее: $|\mathbb{I} + \mathbb{C}| = \prod_{i=1}^{p}(1 + \lambda_i)$

2. Пусть $\mathbb{A}$ — симметричная, неотрицательно определенная матрица. $\mathbb{B}$ — асимметричная, положительно определенная матрица (т.е. существует обратная к ней). Тогда $\mathbb{B}^{-1}\mathbb{A}$ имеет неотрицательные собственные числа: $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ ($\mathbb{B}^{-1}\mathbb{A}$ не является симметричной).

   Имеет место задача:
   $$\mathbb{B}^{-1}\mathbb{A}\mathbb{U} = \lambda\mathbb{U}\mathbb{A}\mathbb{U} = \lambda\mathbb{B}\mathbb{U},$$
   при этом $\mathbb{U}$ — не являются ортогональными (так как матрица $\mathbb{B}^{-1}\mathbb{A}$). Это **обобщенная задача на собственные числа**. Обобщенность задачи состоит в том, что в правом равенстве появляется матрица $\mathbb{B}$.

Докажем, что собственные числа матрицы $\mathbb{B}^{-1}\mathbb{A}$ — неотрицательные. Пусть матрица имеет какое–то разложение: $\mathbb{B} = \mathbb{L}\mathbb{L}^{\mathrm{T}}$ (например, разложение Холецкого — произведение верхнедиагональной матрицы на нижнедиагональную), при этом важно помнить, что так как $\mathbb{B}$ — положительно определена, то существует $\mathbb{L}^{-1}$.

Получаем: $\mathbb{A}\mathbb{U} = \lambda\mathbb{L}\mathbb{L}^{\mathrm{T}}\mathbb{U}$. Добавим выражение $(\mathbb{L}\mathbb{L}^{-1})^{\mathrm{T}}$: $\mathbb{A}(\mathbb{L}\mathbb{L}^{-1})^{\mathrm{T}}\mathbb{U} = \lambda\mathbb{L}\mathbb{L}^{\mathrm{T}}\mathbb{U}$
$\mathbb{L}^{-1}\mathbb{A}(\mathbb{L}^{-1})^{\mathrm{T}}\mathbb{L}^{\mathrm{T}}\mathbb{U} = \lambda\mathbb{L}^{\mathrm{T}}\mathbb{U}$.
В итоге получили, что те самые $\lambda_i$ на самом деле являются собственными числами матрицы $\mathbb{A}$, а она — симметричная. А это значит, что собственные числа $\lambda_i$ — неотрицательные. Кроме того, получили, что $\mathbb{L}^{\mathrm{T}}\mathbb{U}$ — ортонормированные, то есть $\mathbb{U}^{\mathrm{T}}\mathbb{L}\mathbb{L}^{\mathrm{T}}\mathbb{U} = \mathbb{L}^{\mathrm{T}}\mathbb{B}\mathbb{L} = \mathbb{I}$. Таким образом, свели обобщенную задачу к обычной: в обычной задаче $\mathbb{L}^{\mathrm{T}}\mathbb{U}$ — ортонормированные, то есть по сути просто записали условие ортонормированности, а потом оказалось, что в середине находится исходная матрица $\mathbb{B}$.Таким образом, получаем условие “косоугольной” ортонормированности. (тут $\mathbb{B}$ в каком–то смысле регулирует ортогональность)

***Напоминание**. Рассмотрим задачу $\left(\sup||z|| \leq 1 Z^{\mathrm{T}}\mathbb{A}Z\right)$. Супремум будет достигаться на первом собственном числе, а $Z$, на котором достигается супремум — первый собственный вектор. Далее, если решаем эту же задачу для векторов, ортогональных тому, что было изначально, то получим второе собственное число и т.д.*

Эту задачу можно переписать так:

$$\left(\sup||Z|| \leq 1 Z^{\mathrm{T}}\mathbb{A}Z = \sup Z\frac{Z^{\mathrm{T}}\mathbb{A}Z}{Z^{\mathrm{T}}Z}\right)$$

А в обобщенной задаче возникает следующее:

$$\sup Z\frac{Z^{\mathrm{T}}\mathbb{A}Z}{Z^{\mathrm{T}}\mathbb{B}Z}\Big) = \lambda_1,$$

где $\lambda_1$ — максимальное собственное число матрицы $\mathbb{B}^{-1}\mathbb{A}$. Супремом достигается на соответствующем собственном векторе. (потом будет использоваться в дисперсионноа анализе)

$$\sup Z, Z\mathbb{B}U_i = 0, i = 1, \ldots, j\frac{Z^{\mathrm{T}}\mathbb{A}Z}{Z^{\mathrm{T}}\mathbb{B}Z}\Big) = \lambda_j,$$

где $\lambda_j$ — собственное число, а $A_j$ — собственный вектор матрицы $\mathbb{B}^{-1}\mathbb{A}$, на котором достигается максимум.

На случайном языке:

***Предложение***. *Пусть* $\mathbb{B}, \mathbb{A} \sim$ *Уишарта — независимые, с числом степеней свободы* $\nu_B, \nu_A$, *соответственно. При этом матрица* $\mathbb{B}$ *— положительно определенная. Пусть также* $\lambda_i$ *— с.ч.* $\mathbb{B}^{-1}\mathbb{A}$. *Тогда* $\prod_{i=1}^{p} \frac{1}{1+\lambda_i} \sim Wilks(\nu_A, \nu_B)$.

Верно по свойствам алгебры (см. выше). Кроме того

$$\frac{|\mathbb{B}|}{|\mathbb{B}+\mathbb{A}|} = \frac{1}{|\mathbb{I}+\mathbb{B}^{-1}\mathbb{A}|} = \prod_{i=1}^{p} \frac{1}{1+\lambda_i}.$$

## 2.10. Билет 14. Распределения Лямбда Уилкса. Частный случай $p = 1$

Пусть $\mathbb{A} \sim \mathcal{W}_p(\mathbb{I},\nu_A)$, $\mathbb{B} \sim \mathcal{W}_p(\mathbb{I}, \nu_B)$, при этом $\nu_B > p > \nu_A$. Тогда имеет место следующий теоретический факт:

$$\Lambda = \frac{|\mathbb{B}|}{|\mathbb{A}+\mathbb{B}|} = \frac{1}{|\mathbb{I}+\mathbb{B}^{-1}\mathbb{A}|} \sim \Lambda_p(\nu_A, \nu_B),$$

где $\nu_B$ будет стремиться к бесконечности (как обозначается в статистике критерия Фишера: вторая степень свободы стремится к бесконечности). Однако это не является обобщением критерия Фишера (там измеряется значимость отклонения одного к другому).

В частности логарифмированное распределение Уилска аппроскимируется распределением $\chi^2$ (говорим "аппроксимация", когда то, что стремится к бесконечности стоит и слева, и справа):

$$\left(\frac{p-\nu_A+1}{2} - \nu_B\right) \log \Lambda_p(\nu_A, \nu_B) \underset{\nu_B\to\infty}{\longrightarrow} \chi^2(\nu_A)$$

**Частный случай**, $p = 1$ $\quad \Lambda_1(\nu_A,\nu_B) = \frac{\chi^2(\nu_B)}{\chi^2(\nu_A)+\chi^2(\nu_B)}$ — отношение вспомогательных случайных величин. Кроме того,

$$\frac{(1-\Lambda_1)/\nu_A}{\Lambda_1/\nu_B} \sim F_{\nu_B,\nu_A}$$

## 2.11. Билет 15. MANOVA. Модель, запись через условные мат. ожидания и мат.ожидания Разложение ковариационной матрицы.

Этот билет во многом аналогичен части билета 10 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 $\overline{\eta}_1, \dots, \overline{\eta}_k \in \mathbb{R}^p$ и гипотеза формулируется так:

$$\mathrm{H}_0 \,:\, \mathcal{P}(\overline{\eta}_1) = \mathcal{P}(\overline{\eta}_2) = \dots = P(\overline{\eta}_k). \tag{23}$$

Эту задачу можно переформулировать следующим образом. Пусть есть дискретный (м.б. качественный) признак $\xi$, принимающий ровно $k$ значений: $A_1, A_2, \dots, A_k$. Рассмотрим случайный вектор $(\overline{\eta}, \xi)^{\mathrm{T}} \in \mathbb{R}^k \times \{A_1, A_2, \dots, A_k\}$, такой что $\mathcal{P}(\overline{\eta} \mid \xi = A_i) = \mathcal{P}(\eta_i)$ для всех $i \in 1 : k$. Тогда гипотеза (12) переписывается в виде:

$$\mathrm{H}_0^* \,:\, \mathcal{P}(\overline{\eta} \mid \xi = A_i) = \mathcal{P}(\overline{\eta} \mid \xi = A_j) \text{ для всех } i, j. \tag{24}$$

В MANOVA рассматривается частный случай (модель), когда $\mathcal{P}(\overline{\eta}_i) = \mathcal{N}(\overline{\mu}_i, \Sigma)$. В рамках этой модели гипотезам (23) и (24) равносильны следующие две[37] гипотезы:

$$\begin{gathered} \mathrm{H}_0 : \overline{\mu}_1 = \overline{\mu}_2 = \ldots = \overline{\mu}_k. \\ \mathrm{H}_0^* : \mathbb{E}(\overline{\eta} \mid \xi = A_1) = \ldots = \mathbb{E}(\overline{\eta} \mid \xi = A_k). \end{gathered} \tag{25}$$

Прежде чем сформулировать гипотезу в рамках Основного Дисперсионного Тождества (9) запишем его на генеральном языке в текущей модели.

$$\mathrm{Cov}\overline{\eta} = \mathbb{E}\left[(\overline{\eta} - \mathbb{E}\overline{\eta})(\overline{\eta} - \mathbb{E}\overline{\eta})^{\mathrm{T}}\right] = \tag{26}$$

$$= \mathbb{E}\left[(\overline{\eta} - \mathbb{E}(\overline{\eta} \mid \xi))(\overline{\eta} - \mathbb{E}(\overline{\eta} \mid \xi))^{\mathrm{T}}\right] + \mathbb{E}\left[(\mathbb{E}(\overline{\eta} \mid \xi) - \mathbb{E}\overline{\eta})(\mathbb{E}(\overline{\eta} \mid \xi) - \mathbb{E}\overline{\eta})^{\mathrm{T}}\right]. \tag{27}$$

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

$$\mathrm{H}_0 : \mathbb{E}\left[(\mathbb{E}(\overline{\eta} \mid \xi) - \mathbb{E}\overline{\eta})(\mathbb{E}(\overline{\eta} \mid \xi) - \mathbb{E}\overline{\eta})^{\mathrm{T}}\right] = 0. \tag{28}$$

Эта запись является многомерным обобщением (16) 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 $\mathbf{y}_{ij} \in \mathbb{R}^p$[38] — $j$-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 (26) на выборочном языке:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{y}_{ij} - \overline{\mathbf{y}})(\mathbf{y}_{ij} - \overline{\mathbf{y}})^{\mathrm{T}} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{y}_{ij} - \overline{\mathbf{y}_i})(\mathbf{y}_{ij} - \overline{\mathbf{y}_i})^{\mathrm{T}} + \sum_{i=1}^{k} n_i(\overline{\mathbf{y}_i} - \overline{\mathbf{y}})(\overline{\mathbf{y}_i} - \overline{\mathbf{y}})^{\mathrm{T}}. \tag{29}$$

Должно быть ясно, что это лишь многомерное обобщение (17).

Следуя обобщенным обозначениям, предложенным в вопросе 10, можно ввести следующие:

$$\mathbb{E} \overset{\text{def}}{=} \mathbb{E}\left[\widehat{(\overline{\eta} - \mathbb{E}(\overline{\eta} \mid \xi))}(\overline{\eta} - \mathbb{E}(\overline{\eta} \mid \xi))^{\mathrm{T}}\right] = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{y}_{ij} - \overline{\mathbf{y}_i})(\mathbf{y}_{ij} - \overline{\mathbf{y}_i})^{\mathrm{T}};$$

$$\mathbb{H} \overset{\text{def}}{=} \mathbb{E}\left[(\mathbb{E}(\overline{\eta} \mid \xi) - \widehat{\mathbb{E}\overline{\eta}})(\mathbb{E}(\overline{\eta} \mid \xi) - \mathbb{E}\overline{\eta})^{\mathrm{T}}\right] = \sum_{i=1}^{k} n_i(\overline{\mathbf{y}_i} - \overline{\mathbf{y}})(\overline{\mathbf{y}_i} - \overline{\mathbf{y}})^{\mathrm{T}}.$$

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 $(\mathbb{H}, \mathbb{E})$.

## 2.12. Билет 16. MANOVA для дискриминантного анализа и для многомерной множественной регрессии, общее и различие.

Будем использовать обозначения Билета 15.

[37] равносильные

[38] Шрифт тут неудачный, но пусть пока так будет

**Дискриминантный анализ** Дискриминатный анализ работает с $k$ группами и $p$ индивидами[39]. Есть несколько задач: проверка того, что $k$ групп различаются (иными словами проверка значимости), далее объяснение различия и далее классификация.

Использование MANOVA в первой задаче совершенно естественно (в принципе, можно использовать и другие критерии, например, непараметрические, но нас это сейчас не интересует).

Гипотеза будет иметь вид (28).

**Многомерная множественная регрессия** В случае многомерной множественной регрессии все аналогично одномерной множественной регрессии. Мы хотим проверить гипотезу, что регрессия незначима. Уточним смысл сказанного. Во-первых, вспомним, что по определению регрессия — УМО. Во-вторых, какая бы модель у нас ни была, в стандартной ситуации нам нужно от этой модели только, чтобы УМО было линейным (это будет ясно чуть далее). Существует несколько моделей, в которых УМО — линейно. Приведем примеры двух из них.

$$\overline{\eta} = \mathbb{B}\overline{\xi} + \overline{\varepsilon},$$

где $\overline{\xi} \in \mathbb{R}^k$, $\overline{\eta} \in \mathbb{R}^p$, $\overline{\varepsilon} \in \mathbb{R}^p$ — случайные вектора, $\mathbb{B} \in \mathrm{M}_{p,k}(\mathbb{R})$ — матрица коэффициентов, причем $\overline{\xi}$, $\overline{\varepsilon}$ — независимы и $\mathbb{E}\overline{\varepsilon}0$.

В такой модели нетрудно посчитать, что $\mathbb{E}(\overline{\eta} \mid \overline{\xi}) = \mathbb{B}\overline{\xi}$, то есть регрессия является линейной, нас это устраивает.

Примером другой модели такого типа является следующая: $(\overline{\eta}, \overline{\xi})^{\mathrm{T}} \sim \mathcal{N}$ с какими-то параметрами, которые нас сейчас не интересуют. Важно то, что в такой модели тоже найдется матрица $\mathbb{B}$: $\mathbb{E}(\overline{\eta} \mid \overline{\xi}) = \mathbb{B}\overline{\xi}$.

В обеих моделях гипотеза о незначимости формулируется так: $\mathrm{H}_0 : \mathbb{B} = 0$. Именно здесь используется то, что мы требуем линейности регрессии (найдется такая матрица!).

Учитывая то обстоятельность, что для независимых $\xi_1, \xi_2$ справедливо, что $\mathbb{E}(\xi_1 \mid \xi_2) = \mathbb{E}\xi_1$, гипотеза о незначимости регрессии для моделей можно свести к другой, сформулированной в рамках Основного Дисперсионного Тождества (9):

$$\mathrm{H}_0 : \mathbb{E}\left[(\mathbb{E}(\overline{\eta} \mid \xi) - \mathbb{E}\overline{\eta})(\mathbb{E}(\overline{\eta} \mid \xi) - \mathbb{E}\overline{\eta})^{\mathrm{T}}\right] = 0.$$

То есть при независимости $\overline{\xi}$ и $\overline{\eta}$ обе гипотезы эквивалентны. Заметим, что вторая гипотеза является гораздо более общей, в ней не нужны предположения о линейности регрессии! Обратим внимание, что эта гипотеза полностью совпадает по своему виду с гипотезой (28), которая проверяется с помощью MANOVA.

Таким образом, проверку значимости дискриминантного анализа и многомерной множественной линейной регрессии можно проводить с помощью MANOVA.

**Построение критерия** Перейдем к построению критерия. Для этого перейдем на выборочный язык.

Оценкой $\mathbb{E}(\mathbb{E}(\eta \mid \xi) - \mathbb{E}\xi)^2$ является матрица $\mathbb{H}$. Сначала нам потребуется следующее утверждение.

***Теорема*** (Без доказательства). *Пусть* $\mathbb{C} \sim \mathcal{W}_p(\Sigma, \nu_{\mathbb{C}})$ *и* $\mathbb{D} \sim \mathcal{W}_p(\Sigma, \nu_{\mathbb{D}})$, *причем* $\Sigma$ *— невырождена и* $\mathbb{C}$ *и* $\mathbb{D}$ *— независимы. Потребуем так же, чтоб* $\nu_{\mathbb{C}} < \nu_{\mathbb{D}}$ *и* $\nu_{\mathbb{D}} > p$. *Обозначим* $\lambda_i$ *— $i$-тое по*

[39] Обозначения вводятся, чтобы потом сформулировать критерий.

*упорядоченности по невозрастанию собственное число матрицы* $\mathbb{D}^{-1}\mathbb{C}$. *Тогда*

$$\frac{|\mathbb{D}|}{|\mathbb{C}+\mathbb{D}|} = \frac{1}{|1+\mathbb{D}^{-1}\mathbb{C}|} = \prod_{i=1}^{p} \frac{1}{1+\lambda_i} \sim \Lambda_1(\nu_{\mathbb{C}}, \nu_{\mathbb{D}}).$$

Заметим, что оба равенства являются фактами из линейной алгебры, причем первое равенство легко интерпретируется тем, что определитель мультипликативен, а $\mathbb{D}$ является положительно определенной. Второе равенство легко понять, если подумать о Жордановой Нормальной Форме матрицы $\mathbb{D}^{-1}\mathbb{C}$.

Из этого утверждения подстановкой нужных букв следует то, что позволяет сформулировать критерий.

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

$$\frac{|\mathbb{E}|}{|\mathbb{H}+\mathbb{E}|} = \frac{1}{|1+\mathbb{E}^{-1}\mathbb{H}|} = \prod_{i=1}^{s} \frac{1}{1+\lambda_i} \sim \Lambda_1(\nu_{\mathbb{H}}, \nu_{\mathbb{E}}),$$

*где* $s = \min(\nu_{\mathbb{H}}, p)$.

*В случае дискриминантного анализа* $\nu_{\mathbb{H}} = k-1$, $\nu_{\mathbb{E}} = n-k$. *В случае многомерной множественной регрессии* $\nu_{\mathbb{H}} = k$, $\nu_{\mathbb{E}} = n-k-1$.

*Доказательство.* То, что $s$ выглядит так, объясняется в вопросе 20. То, что число степеней свободы такое, во-первых, можно увидеть непосредственно, во-вторых, по аналогии с одномерным случаем. □

Заметим, что построенный критерий для проверки соответствующей гипотезы является далеко не единственным. О построении других критериев речь идет в вопросе 19.

## 2.13. Билет 17. Какой смысл у канонических дискриминантных функций (коэффициентов) и переменных?

В вопросе 16 мы построили критерий для проверки значимости дискриминантного анализа. На самом деле это не единственный возможный критерий. Главными объектами, с которыми приходится работать, являются матрицы $\mathbb{H}$ и $\mathbb{E}$. На основе разных способов «совмещения» этих объектов могут получаться разные критерии. Подробнее об этом в билете 19.

В этом билете подробнее изучаются свойства этих двух матриц.

Напомним, что имеется $p$ признаков и $k$ групп: $\overline{\eta} \in \mathbb{R}^p$, причем $\mathcal{P}(\overline{\eta} \mid \xi = A_i) = \mathcal{N}(\mu_i, \Sigma)$. 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 $\zeta = A^{\mathrm{T}}\overline{\eta}$: $\mathcal{P}(\zeta \mid \xi = A_i) = \mathcal{N}(A^{\mathrm{T}}\mu_i, A^{\mathrm{T}}\Sigma A)$. На выборочном языке — $Z = \mathbb{Y}A$ и выборочная ковариационная матрица (с точностью до коэффициента имеет вид): $A^{\mathrm{T}}\mathbb{Y}^{\mathrm{T}}\mathbb{Y}A = A^{\mathrm{T}}(\mathbb{H}+\mathbb{E})A$. Из этого следуетчто «аналогом» $\mathbb{H}$ для нового признака является $A^{\mathrm{T}}\mathbb{H}A$, а аналогом $\mathbb{E}$ является $A^{\mathrm{T}}\mathbb{E}A$. [40]

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

Посмотрим теперь на $F$-статистику, следуя обобщенному определению статистики, данному в билете 10:

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

где $C = \nu_{\mathbb{E}}/\nu_{\mathbb{H}}$ — 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 статистики $F$ — это означает[41], что решается обобщенная задача на собственные числа и собственные вектора.

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

Собственные числа матрицы $\mathbb{E}^{-1}\mathbb{H}$ 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 $A_i^{\mathrm{T}}\mathbb{E}A_j = 0$ для $j < i$. Вектора $A_i$ — канонические коэффициенты (дискриминантные функции). Новые признаки $Z_i = \mathbb{Y}A_i$ — канонические переменные. В дальнейшем, оказывается, что новые признаки являются ортогональными.

Исходя из описания, смысл $\lambda_i$ — степень разброса по $i$-тому направлению, задаваемым $i$-той дискриминантной функцией. $A_i$ — коэффициенты с которыми нужно взять исходные признаки, чтобы получить новый признак, имеющий наибольший разброс и ортогональный предыдущим. А сами канонические переменные тем самым — ортогональные новые признаки, имеющие наибольший разброс, измеряемый $\lambda_i$.

## 2.14. Билет 18. Как вычисляются канонические дискриминантные функции (коэффициенты)?

Если посмотреть в вопрос 17, то ясно, что дискриминантные функции вычисляются как собственные вектора матрицы $\mathbb{E}^{-1}\mathbb{H}$.

## 2.15. Билет 19. Значимость LDA. Разные критерии, чем отличаются.

Будем использовать обозначения Билета 15. Как уже говорилось в вопросе 16, дискриминатный анализ работает с $k$ группами и $p$ индивидами Есть несколько задач: проверка того, что $k$ групп различаются (иными словами проверка значимости), далее объяснение различия и далее классификация. Использование MANOVA для проверки значимости дискриминантного анализа совершенно естественно. Один из критериев 2.12 уже был построен — этот критерий использует общее утверждение 2.12. 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 $\mathbb{H}$ и $\mathbb{E}$. На основе разных способов «совмещения» этих объектов могут получаться разные критерии. Рассмотрим несколько способов. Напомним, что $s = \min\{\nu_{\mathbb{H}}, p\}$.

Проверяется гипотеза о равенстве средних[42]

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

[41] Приглашаю убедиться.

[42] На самом деле о равенстве распределений в рамках нормальной модели при условии гомоскедастичности.

1. Lambda Wilks's

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

2. Roy's largest root

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

   Этот тест использует только первое собственное число матрицы $\mathbb{E}^{-1}\mathbb{H}$. Напомним, что $r_i^2$ — называют $i$-тым каноническим корнем. Как мы увидим в вопросе 42, на самом деле $r_i^2$ — канонические корреляции[43]. Потому они на самом деле имеют интерпретируемый смысл.

3. Pillai's:

$$V^{(s)} = \operatorname{tr}\left(\frac{\mathbb{H}}{\mathbb{E}+\mathbb{H}}\right) = \sum_{i=1}^{s} r_i^2.$$

4. Hotelling:

$$U^{(s)} = \operatorname{tr}\left(\mathbb{E}^{-1}\mathbb{H}\right) = \sum_{i=1}^{s} \lambda_i^2.$$

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 каждый[44] из предложенных критериев мощнее остальных?».

Посмотрим на это на примере первых двух критериев. Первый критерий включает в статистику все (!) направления, по которым разброс максимален (иными словами, все $\lambda_i$)[45]. 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 на лекции утверждалось, что они «где-то по середине» между первым и вторым.[46]

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.

---

[43] При повторном чтении спросите себя: «между чем и чем это корреляции?».

[44] Или хотя те из них, про которые это легко понять.

[45] Смысл раскрыт в вопросе 17.

[46] Более полные и точные объяснения приветствуются.

## 2.16. Максимальное число дискриминантных функций, почему такое?

Пользуемся обозначениями из билетов 15 и 16. Здесь мы рассматриваем именно случай применения MANOVA к дискриминантному анализу. Рассуждаем в предположении истинности гипотезы $H_0$.

Ясно, что $s \leqslant \mathrm{rk}(\mathbb{E}^{-1}\mathbb{H})$ (за $s$-тым собственным числом, если и «идут», то точно нулевые собственные числа, нам такие не интересны). Далее, размеры обеих матриц $p \times p$, но $\mathbb{E}$, с вероятностью 1, полного ранга, поэтому[47] $s = \mathrm{rk}\,\mathbb{H}$. Далее, на матрицу $\mathbb{H}$ можно смотреть (с точностью до коэффициентов), как на оценку ковариационной матрицы из центров групп (выборка размера $k$). То есть можно сказать, что есть исходная матрица данных центров групп $\widetilde{Y} \in \mathrm{M}_{p,k}(\mathbb{R})$ и $\mathrm{rk}(\mathbb{H}) = \mathrm{rk}(\widetilde{\mathbb{Y}} - \overline{\mathbb{Y}}) \leqslant \min\{p, k\}$, где $\overline{\mathbb{Y}}$ — усредненная матрица. Но на самом деле $\mathrm{rk}(\widetilde{\mathbb{Y}} - \overline{\mathbb{Y}}) \leqslant \min\{p, k-1\}$, из-за усреднения. Докажем это. По определению

$$\mathbb{H} = \sum_{i=1}^{k} n_i (\overline{\mathbf{y}}_i - \overline{\mathbf{y}})(\overline{\mathbf{y}}_i - \overline{\mathbf{y}})^{\mathrm{T}}.$$

При этом известно, что $\sum_{i=1}^{k} n_i \overline{\mathbf{y}}_i = n\overline{\mathbf{y}}$, где $n = \sum_{i=1}^{k} n_i$. То есть $\sum_{i=1}^{k} n_i(\overline{\mathbf{y}}_i - \overline{\mathbf{y}}) = 0$, тем самым вектора $\{\overline{\mathbf{y}}_i - \overline{\mathbf{y}}\}_{i=1}^{k}$ линейнозависимы, Но ведь матрица $\mathbb{H}$ — сумма матриц ранга 1, из которых одна представляется линейной комбинацией других, а значит ее ранг не превосходит $k-1$.

Таким образом, можно положить $s = \min\{p, k-1\}$.

Обращаю внимание, что $s$ — это не максимальное количество собственных чисел матрицы $\mathbb{E}^{-1}\mathbb{H}$, и даже не максимальное число ненулевых с.ч. Реально ненулевых собственных чисел может оказаться и меньше, чем $s$[48] , но главное, что никогда не окажется, что ненулевых собственных чисел больше, чем $s$.

## 2.17. Билет 21. С чем совпадают дискриминантные функции и переменные, если ошибки сферические?

Пусть $\mathbb{Y} \in M_{n,p}(\mathbb{R})$ — матрица данных (строки $\mathbf{y}_i, i \in 1:n$ – наблюдения, столбцы $Y_j, j \in 1p$ – признаки), наблюдения принадлежат к одной из $k$ групп, в каждой группе $n_i$ наблюдений, $n = \sum_{i=1}^{k} n_i$. Матрица межгрупповых отклонений $\mathbb{H} = \sum_{i=1}^{k} n_i \left(\overline{\mathbf{y}}_i - \overline{\mathbf{y}}\right)\left(\overline{\mathbf{y}}_i - \overline{\mathbf{y}}\right)^{\mathrm{T}}$, матрица внутригрупповых отклонений $\mathbb{E} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\mathbf{y}_{ij} - \overline{\mathbf{y}}_i\right) \mathbf{y}_{ij} - \overline{\mathbf{y}}_i^{\mathrm{T}}$.

В случае, если $\mathbb{E} = \sigma^2\mathbb{I}$ (все группы сферические и одинакового размера), дискриминантные функции данных, т.е. $A_1, \ldots, A_s$ – собственные вектора $\mathbb{E}^{-1}\mathbb{H}$, становятся собственными векторами матрицы $\sigma^2\mathbb{H}$. 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 $\mathbb{Y}A_i$ при такой постановке являются главными компонентами.

## 2.18. Билет 22. Как определить значимое число дискриминантных функций (размерность пространства, где группы различаются).

Модель дискриминантного анализа (на генеральном языке): $\xi$ – дискретная с.в. с носителем $\left\{A_i\right\}_{i=1}^{k}$, $\eta = \eta_i$, если $\xi = A_i$. Определены матрицы

$$\mathbb{H} = \mathrm{E}\left(\left(\mathrm{E}\left(\eta|\xi\right) - \mathrm{E}\eta\right)\left(\mathrm{E}\left(\eta|\xi\right) - \mathrm{E}\eta\right)^{\mathrm{T}}\right) \in M_p\left(\mathbb{R}\right),$$

[47] Это какое-то стандартное свойство матриц: домножение на обратимую не меняет ранга матрицы.

[48] Однако, такое событие имеет вероятность 0 при справедливости нулевой гипотезы.

$$\mathbb{E} = \mathrm{E}\left((\eta - \mathrm{E}(\eta|\xi))(\eta - \mathrm{E}(\eta|\xi))^{\mathrm{T}}\right) \in M_p(\mathbb{R}).$$

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

Для того, чтобы узнать, какие из функций являются значимыми, мы последовательно проверяем гипотезу $H_0$: «дискриминантные функции $A_m, \ldots, A_s$ не описывают различия в данных» для $m \in 1 : s$. 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 (см. билет 2.13).

Если $H_0$ верна, то $\mathrm{rk}\mathbb{H} = m - 1$, и (почему?)

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

С помощью статистики $\Lambda'_m$ (lambda prime) мы для каждого $m$ можем проверить, незначимы ли эта и все последующие дискриминантные функции.

## 2.19. Билет 23. Почему канонические дискриминантные переменные получаются ортогональными.

Пусть $\mathbb{Y} \in M_{n,p}(\mathbb{R})$ — матрица данных (строки $\mathbf{y}_i, i \in 1 : n$ – наблюдения, столбцы $Y_j, j \in 1p$ – признаки), наблюдения принадлежат к одной из $k$ групп, в каждой группе $n_i$ наблюдений, $n = \sum_{i=1}^{k} n_i$. Матрица межгрупповых отклонений $\mathbb{H} = \sum_{i=1}^{k} n_i \left(\overline{\mathbf{y}}_i - \overline{\mathbf{y}}\right)\left(\overline{\mathbf{y}}_i - \overline{\mathbf{y}}\right)^{\mathrm{T}}$, матрица внутригрупповых отклонений $\mathbb{E} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\mathbf{y}_{ij} - \overline{\mathbf{y}}_i\right)\mathbf{y}_{ij} - \overline{\mathbf{y}}_i^{\mathrm{T}}$.

$A_1, \ldots, A_s$ — собственные вектора $\mathbb{E}^{-1}\mathbb{H}$. Дискриминантные переменные — вектора данных в новых координатах $Z_i = \mathbb{Y}A_i$.

***Теорема.***

$$\forall i \neq j \in 1 : s \; A_i \perp A_j$$

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

## 2.20. Билет 24. Интерпретация разделения: стандартизованные дискриминантные функции и факторная структура

Пусть $\mathbb{Y} \in M_{n,p}(\mathbb{R})$ — матрица данных (строки $\mathbf{y}_i, i \in 1 : n$ – наблюдения, столбцы $Y_j, j \in 1p$ – признаки), наблюдения принадлежат к одной из $k$ групп, в каждой группе $n_i$ наблюдений, $n = \sum_{i=1}^{k} n_i$. Матрица межгрупповых отклонений $\mathbb{H} = \sum_{i=1}^{k} n_i \left(\overline{\mathbf{y}}_i - \overline{\mathbf{y}}\right)\left(\overline{\mathbf{y}}_i - \overline{\mathbf{y}}\right)^{\mathrm{T}}$, матрица внутригрупповых отклонений $\mathbb{E} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\mathbf{y}_{ij} - \overline{\mathbf{y}}_i\right)\mathbf{y}_{ij} - \overline{\mathbf{y}}_i^{\mathrm{T}}$.

$A_1, \ldots, A_s$ — собственные вектора $\mathbb{E}^{-1}\mathbb{H}$. Дискриминантные переменные — вектора данных в новых координатах $Z_i = \mathbb{Y}A_i$.

Задача: проанализировать разложение по дискриминантным переменным.

### 2.20.1. 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 $\mathbb{S}$ – матрица взвешенных ковариаций, $s_1^2, \dots, s_p^2$ – элементы на её диагонали (взвешенные дисперсии признаков), $A_i = \begin{pmatrix} a_{i1} \\ \vdots \\ a_{ip} \end{pmatrix}$. Тогда

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

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

Важно учитывать, что

1. стд. д. ф. могут существенно меняться при удалении/добавлении новых признаков,
2. на малых выборках (при малом отношении $N/p$) стд. д. ф. нестабильны от выборки к выборке.

### 2.20.2. Факторная структура

Факторная структура — матрица корреляций между исходными и каноническими переменными. У Ренчера указано, что им же показано, что корреляции показывают вклад исходных признаков в дискриминантные переменные независимо от вклада других признаков, устойчивы относительно удаления/добавления новых признаков и этим плохи, так как цель дискриминантного анализа – выявить совместные влияния признаков.

## 2.21. Билет 25. Свойства исходных признаков, по которым можно понять, какие признаки лишние.

«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 $H_0$: «добавление признака $i$ не влияет на качество разделения». Статистика:

$$(\text{Partial } \Lambda)_i = \frac{\Lambda\left(Y_1, \dots, Y_p\right)}{\Lambda\left(Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_p\right)} = \frac{1}{1+\lambda_i} \sim \Lambda_1\left(\nu_H, \nu_E - p + 1\right) \tag{30}$$

## 2.22. Билет 26. Пошаговый дискриминантный анализ

Пошаговый дискриминантный анализ подбирает тот набор признаков, который лучше всего будет разделять переменные (аналогично пошаговой множественной регрессии). На каждом шаге добавляется одна переменная, которая максимально увеличивает качество разделения групп (качество меряется статистикой partial lambda из (30), точнее, из-за наличия точного преобразования из $\Lambda_1$ в $F$, 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.

## 2.23. Билет 27. Что уменьшается с помощью lambda prime и что с помощью partial lambda?

### 2.23.1. Lambda prime

$H_0$: «дискриминантные функции $A_m, \dots, A_s$ не описывают различия в данных»

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

Если $H_0$ не верна, то $m$-я *дискриминантная функция* значима при описании различий. Т. е. $\Lambda'_m$ уменьшает число дискриминантных функций.

### 2.23.2. Partial lambda

$H_0$: «добавление признака $i$ не влияет на качество разделения»

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

Если $H_0$ не верна, то $i$-й *исходный признак* значим при описании различий исходных данных. Т. е. $(\text{Partial } \Lambda)_i$ уменьшает число исходных признаков.

## 2.24. Билет 29. Почему линейный дискриминантный анализ называется линейным, а квадратичный — квадратичным?

### 2.24.1. LDA

Модель: $\xi$ – дискретная с.в. с носителем $\left\{A_i\right\}_{i=1}^{k}$, $\eta \sim \mathcal{N}\left(\mu_i, \Sigma\right)$, если $\xi = A_i$. Тогда плотность $x$

$$p\left(x \middle| \xi = A_i\right) = \frac{1}{(2\pi)^{p/2}\left|\Sigma\right|^{1/2}} \exp\left(-\frac{1}{2}\left(x-\mu_i\right)^{\mathrm{T}} \Sigma^{-1}\left(x-\mu_i\right)\right),$$

и классифицирующая функция $f_i(x) = \pi_i p\left(x \middle| \xi = A_i\right)$, где $\pi_i$ — априорная вероятность наблюдения попасть в $i$-ю группу. Для упрощения вычислений можно переписать классифицирующую функцию как

$$g_i(x) = \log f_i(x) = \log \pi_i - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\left(x-\mu_i\right)^{\mathrm{T}} \Sigma^{-1}\left(x-\mu_i\right) \longrightarrow -\frac{1}{2}\mu_i^{\mathrm{T}}\Sigma^{-1}\mu_i + \mu_i\Sigma^{-1} + \log \pi_i,$$

то есть классифицирующая функция является линейной по $x$.

### 2.24.2. QDA

Модель: $\xi$ – дискретная с.в. с носителем $\left\{A_i\right\}_{i=1}^{k}$, $\eta \sim \mathcal{N}\left(\mu_i, \Sigma_i\right)$, если $\xi = A_i$. Тогда плотность $x$

$$p\left(x \middle| \xi = A_i\right) = \frac{1}{(2\pi)^{p/2}\left|\Sigma_i\right|^{1/2}} \exp\left(-\frac{1}{2}\left(x-\mu_i\right)^{\mathrm{T}} \Sigma_i^{-1}\left(x-\mu_i\right)\right),$$

и классифицирующая функция $f_i(x) = \pi_i p\left(x \middle| \xi = A_i\right)$. Оставляем в классифицирующей функции только монотонность и члены, отличающиеся в разных группах:

$$g_i(x) = \log f_i(x) = \log \pi_i - \frac{1}{2}\log\left|\Sigma_i\right| - \frac{1}{2}\left(x - \mu_i\right)^{\mathrm{T}} \Sigma_i^{-1}\left(x - \mu_i\right),$$

получаем квардатично зависящую от $x$ классифицирующую функцию.

## 2.25. Билет 30: Общий подход к классификации через апостериорные вероятности

Общая задача классификации: $\xi$ – дискретная с.в. с носителем $\left\{A_i\right\}_{i=1}^k$, $\eta \sim \mathcal{P}_i$, если $\xi = A_i$. Класс $C_i = \left\{\xi = A_i\right\}$. Чтобы классифицировать наблюдение $x$, необходимо найти

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

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

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

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

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.)

## 2.26. 34. Что такое канонические корреляции, сколько их

Пусть у нас имеется два набора случайных величин:

$$\eta = (\eta_1, \ldots, \eta_q)^{\mathrm{T}} \tag{31}$$

$$\xi = (\xi_1, \ldots, \xi_p)^{\mathrm{T}} \tag{32}$$

Не умаляя общности будем считать, что они центрированы. Тогда

***Определение***. *Первой канонической корреляцией называется*

$$r_1^2 = \max_{A,B} \rho^2(A^{\mathrm{T}}\xi, B^{\mathrm{T}}\eta), \tag{33}$$

*где $A \in \mathbb{R}^p, B \in \mathbb{R}^q$. $i$-ой каноническая корреляция определяется аналогично с условием, что максимум берется по некоррелированным с предыдущими случайным величинам:*

$$r_i^2 = \max_{\substack{A,B \\ \rho(A\xi, A_j\xi)=0 \\ \rho(B\eta, B_j\eta)=0 \\ 1 \le j < i}} \rho^2(A^{\mathrm{T}}\xi, B^{\mathrm{T}}\eta) \tag{34}$$

Канонических корреляций будет $s = \min(p, q)$.

## 2.27. 37. Канонические переменные, как находятся?

тут оригинальный порядок вопросов уплыл от порядка, в котором все удобно определять Смотрим на (34). У нас каждая каноническая корреляция определяется с помощью двух векторов — они определяют новый базис. Новые признаки $\tilde{\eta}_i = B_i^{\mathrm{T}}\eta$ и $\tilde{\xi}_i = A_i^{\mathrm{T}}\xi$ называются правыми и левыми каноническими переменными. Находятся, как обычно, с помощью лин. алгебры:

$$\rho^2(A^{\mathrm{T}}\xi,\, B^{\mathrm{T}}\eta) = \frac{(A^{\mathrm{T}}\Sigma_{xy}B)^2}{\left(A^{\mathrm{T}}\Sigma_{xx}A\right)\left(B^{\mathrm{T}}\Sigma_{yy}B\right)}, \text{ где} \tag{35}$$

$$\Sigma_{xx} = E\xi\xi^{\mathrm{T}} \tag{36}$$

$$\Sigma_{xy} = E\xi\eta^{\mathrm{T}} \tag{37}$$

$$\Sigma_{yy} = E\eta\eta^{\mathrm{T}} \tag{38}$$

Отсюда сразу получаем, что задачу максимизации из (34) можно переформулировать как

$$(A,\, B) = \underset{\substack{A,\, B \\ A^{\mathrm{T}}\Sigma_{xx}A=1 \\ B^{\mathrm{T}}\Sigma_{yy}B=1}}{\arg\max}\, A^{\mathrm{T}}\Sigma_{xy}B \tag{39}$$

Если требуется найти больше, чем одну каноническую корреляцию/переменную/etc, то вводим дополнительное ограничение

$$\rho(A_i\xi, A_j\xi) = 0 \Leftrightarrow A_i^{\mathrm{T}}\Sigma_{xx}A_j = 0 \tag{40}$$

Для $\eta$ аналогично.

Задача (39) эквивалентна (является) обобщенным задачам на собственные числа:

$$\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}A = r^2\Sigma_{xx}A \tag{41}$$

$$\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}B = r^2\Sigma_{yy}B \tag{42}$$

Собственные числа у обоих задач совпадают, соответствующие вектора задают выражения канонических переменных через оригинальные признаки.

## 2.28. 40. Корреляции внутри множества канонических переменных, левых и правых

Как и в предыдущих обозначениях, $A_i$, $B_j$ — линейные функции для выражения новых признаков. $1 \le i, j \le s = min(p, q)$. Факт без док-ва:

$$\langle\, A_i\xi, B_j\eta \,\rangle = 0 \; i \neq j$$

## 2.29. 36. Множественная корреляция как каноническая корреляция, если число признаков с одной стороны равно 1

Как все помнят, множественный коэффициент корреляции случайной величины $\eta$ относительно $\xi = (\xi_1, \ldots, \xi_p)$ определяется как

$$R^2(\eta, \{\xi\}) = 1 - \frac{\min\limits_{\hat{\eta}\in\{a_0+\sum a_i\xi_i\}} E\,(\eta - \hat{\eta})^2}{D\eta} \tag{43}$$

Не умаляя общности считаем, что $\eta$ и $\xi$ центрированы. Также предположим, что $E\eta^2 = 1$. Тогда в (39) условие на $\eta$ автоматически выполнено и $B = 1$. Далее

$$\min_{\hat{\eta}\in\{a_0+\sum a_i\xi_i\}} E\,(\eta-\hat{\eta})^2 = \min_{\substack{\hat{\eta}\\ E\hat{\eta}^2=1\\ k\geq 0}} E\,(\eta-k\hat{\eta})^2 = \tag{44}$$

$$\min\left(1-2k\rho(\eta,\hat{\eta})+k^2\right) = \min\left((k-\rho(\eta,\hat{\eta}))^2+1-\rho^2(\eta,\hat{\eta})\right)$$

Заметим, что нахождение минимума данного выражения тоже самое, что и нахождение максимума $\rho^2(\eta,\hat{\eta})$. Кроме того, видно, что $R^2$ равен максимуму $\rho^2$, а следовательно множественный коэффициент корреляции и каноническая корреляция (при одномерном $\eta$) совпадают.

## 2.30. 35. Значимость корреляции между множествами признаков и значимость многомерной множеств. регрессии.

В многомерной множественной регрессии все ясно. Пусть $\Xi, Y \in \mathbb{R}^{n\times q}$, $X \in \mathbb{R}^{n\times p}$, $B \in \mathbb{R}^{p\times q}$.

$$Y = XB + \Xi \tag{45}$$

Тогда MSE оценка $B$:

$$\hat{B} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y \tag{46}$$

$$\hat{Y} = X\hat{B} \tag{47}$$

Как и раньше считаем, что признаки центрированы[49]. Тогда проверка значимости аналогична MANOVA:

$$E = (\hat{Y}-Y)^{\mathrm{T}}(\hat{Y}-Y) \tag{48}$$

$$H = \hat{Y}^{\mathrm{T}}\hat{Y} \tag{49}$$

Проверяется гипотеза о том, что $B = 0$. Статистика критерия остается той же, что и была:

$$\Lambda = \frac{|E|}{|H+E|} \sim \Lambda_p(\nu_H,\nu_E),\ \text{где} \tag{50}$$

$$\nu_H = p \tag{51}$$

$$\nu_E = n-\nu_H-1 \tag{52}$$

И как мы узнали до этого, $\Lambda$ вычисляется через собственные числа:

$$\Lambda = \prod_{i=1}^{s}\frac{1}{\lambda_i+1}$$

В каноническом анализе значимость проверяется аналогично. С помощью $\Lambda_{prime}$ можно проверять, сколько корреляций значимы. Для проверки требуется нормальность данных. Кажется 39 должен быть частным случаем этого вопроса

## 2.31. 39. Как найти число значимых корреляционных переменных

Смотрим предыдущий (35) вопрос + вспоминаем проверку с $\Lambda_{prime}$ из LDA.

[49] под признаками понимаем как элементы матрицы $X$, так и элементы матрицы $Y$

## 2.32. 38. Интерпретация канонических переменных через стандартизованные канонические функции и через факторную структуру

Определения такие же, как и в PCA:

***Определение***. *Пусть старые признаки $\xi$ и $\eta$ стандартизованы, т.е. для $1 \leq i \leq p$ $E\xi_i = 0, E\xi_i^2 = 1$ и $1 \leq i \leq q$ $E\eta_i = 0, E\eta_i^2 = 1$. Пусть для $1 \leq i \leq s$ $A_i$ и $B_i$ — функции (т.е. вектора, т.к. у нас все линейно), выражающие новые признаки через старые. Тогда $A_i$, $B_i$ — стандартизованные канонические функции.*

***Определение***. *Факторная структура — корреляции старых признаков с новыми.*

Как считать факторную структуру:

$$\rho(\xi, A_i^{\mathrm{T}}\xi) = E\xi\xi^{\mathrm{T}}A_i = \Sigma_{\xi\xi}A_i \quad (53)$$

$$\rho(\xi, B_i^{\mathrm{T}}\eta) = E\xi\eta^{\mathrm{T}}B_i = \Sigma_{\xi\eta}B_i \quad (54)$$

$$\rho(\eta, B_i^{\mathrm{T}}\eta) = E\eta\eta^{\mathrm{T}}B_i = \Sigma_{\eta\eta}B_i \quad (55)$$

Интерпретация — аналогично PCA и регрессии.

## 2.33. Билет 41 Объясненные каноническими переменными доли дисперсии. Избыточность.

Пусть $\overline{\xi} \in \mathbb{R}^p$, $\overline{\eta} \in \mathbb{R}^q$ — исходные признаки. $A_i^T\overline{\xi}$, $B_i^T\overline{\eta}$ — $i$-ые канонические направления. Утверждается, что, если $\mathbb{E}\xi_i = 0, \mathbb{D}\xi_i = 1$, $\mathbb{E}\eta_i = 0, \mathbb{D}\eta_i = 1$ $\forall i$, то

$$\sum_{k=1}^{s} \rho^2(\eta_i, B_k^T\overline{\eta}) \leq 1 = \mathbb{D}\eta_i, \ \sum_{k=1}^{s} \rho^2(\xi_i, A_k^T\overline{\xi}) \leq 1 = \mathbb{D}\xi_i, \ s = \min(p, q).$$

При этом, если $s = q$, то в первом случае достигается равенство, а если $s = p$, то во втором. Поэтому $\sum_{k=1}^{s} \rho^2(\eta_i, B_k^T\overline{\eta})$ и $\sum_{k=1}^{s} \rho^2(\xi_i, A_k^T\overline{\xi})$ называется объясненной долей дисперсии.

**Избыточность**: Избыточностью $i$-ого левого признака $\xi_i$ называется

$$\sum_{k=1}^{s} \rho^2(\xi_i, B_k^T\overline{\eta}).$$

Аналогично для правых. На консультации мы не разобрались, как ее интерпретировать, поэтому достаточно только определения.

## 2.34. Билет 42.Что общего между дискриминантным анализом в многомерной множественной регрессией?

Окей, давайте по порядку. Задача MANOVA (буду сразу писать в нормальной модели с равными ковариационными матрицами, а не начинать с общей задачи на равенство распределений):

Пусть $\eta_i \sim \mathcal{N}(\mu_i, \Sigma)$, $i = 1, \dots k$.
Необходимо проверить гипотезу

$$H_0 : \mu_1 = \dots = \mu_k.$$

Введем одномерный качественный признак $\xi$, принимающий значения $A_1, \dots A_k$. Тогда можно рассмотреть пару $(\eta, \xi)$, такую что $\mathcal{P}_{\eta|\xi=A_i} = \mathcal{P}_{\eta_i}$ и гипотеза принимает вид

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

Что эквивалентно независимости $\eta$ от $\xi$ Знаем, что

$$\mathbb{E}(\eta - \mathbb{E}\eta)(\eta - \mathbb{E}\eta)^T = \mathbb{E}(\mathbb{E}(\eta \mid \xi) - \mathbb{E}\eta)(\mathbb{E}(\eta \mid \xi) - \mathbb{E}\eta)^T + \mathbb{E}(\eta - \mathbb{E}(\eta \mid \xi))(\eta - \mathbb{E}(\eta \mid \xi))^T$$

То есть, на самом деле мы можем проверять гипотезу о том, что

$$H_0 : \mathbb{E}(\mathbb{E}(\eta \mid \xi) - \mathbb{E}\eta)(\mathbb{E}(\eta \mid \xi) - \mathbb{E}\eta)^T = 0$$

Теперь перейдем на выборочный язык и напишем основное дисперсионное тождество ($n_i$ – длина выборки для $i$-ой группы.

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2 = \sum_{i=1}^{k} n_i(y_{ij} - \overline{y}_{i.})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{y}_{i.} - \overline{y})^2 = \mathbb{H} + \mathbb{E}$$

Статистика критерия, проверяющая нашу гипотезу имеет вид

$$\Lambda = \frac{|\mathbb{E}|}{|\mathbb{E} + \mathbb{H}|} = \prod_{i=1}^{s}(\frac{1}{1 + \lambda_i}), \tag{56}$$

где $\lambda_i$ – собственные числа матрицы $\mathbb{E}^{-1}\mathbb{H}$, $s = \min(p, k - 1)$.

Теперь введем переменные $\zeta_1, \dots, \zeta_{k-1}$ как dummy variables для переменной $\xi$ и рассмотрим многомерную множественную линейную регрессию вектора $\overline{\zeta}$ на $\eta$. А именно рассмотрим

$$\mathbb{Y} = \mathbb{X}\mathbb{B} + \Xi$$

Предположим, что $\mathbb{Y}$ и $\mathbb{X}$ — центрированы. Решение данной задачи $\hat{\mathbb{B}} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$. Значимость регрессии проверяется гипотезой

$$H_0 : \mathbb{B} = 0.$$

Положим $\mathbb{E} = (\hat{\mathbb{Y}} - \mathbb{Y})(\hat{\mathbb{Y}} - \mathbb{Y})^T$, $\mathbb{H} = \hat{\mathbb{Y}}\hat{\mathbb{Y}}^T$ и статистика критерия имеет вид

$$\Lambda = \frac{|\mathbb{E}|}{|\mathbb{E} + \mathbb{H}|} = \prod_{i=1}^{s}\left(\frac{1}{1 + \lambda_i}\right) = \prod(1 - r_i^2), \tag{57}$$

где $r_i$ — это $i$-ая каноническая корреляция. Так вот, оказывается, что $\Lambda$ из (56) совпадает с $\Lambda$ из (57). Таким образом получается, что $i$-ая каноническая корреляция между наборами $\zeta_1, \dots, \zeta_{k-1}$ и $\eta_1, \dots, \eta_p$ выражается через собственные числа $\lambda_i$, полученные в результате дискриминантного анализа как

$$r_i^2 = \frac{\lambda_i}{1 + \lambda_i}$$

## 2.35. Билет 43.Две группы, использование множественной линейной регрессии для классификации.

Все в тех же обозначениях положим $k = 2$. Таким образом у нас получается только одна dummy-переменная и мы можем рассмотреть регрессию $\eta_1, \dots, \eta_p$ на $\zeta_1$. (Мы все перевернули, но это не смертельно ибо в канонических корреляциях все симметрично). В таком случае первая каноническая корреляция $r_1^2 = R(\zeta, \eta_1, \dots, \eta_p)$ — это выборочный коэффициент корреляции.

***Теорема.*** *Канонические коэффициенты $A_1$ пропорциональны $B^{(c)}$, которые получены при решении регрессионной задачи и пропорциональны $\Sigma^{-1}(\mu_1 - \mu_2)$.*

Вообще это частично показывалось, но я так поняла, что это не нужно..

## 2.36. Билет 44.Кластерный анализ, пример model-based подхода

Предположим, что многомерная выборка — неоднородная. Но в отличие от дискриминантного анализа у нас нет признака, объясняющего эту неоднородность и задачей является ее выявить. Тип классификации, когда есть модель называется model-based clustering. Например, пусть наша выборка из смеси $k$ нормальных распределений. Таким образом ее плотность имеет вид

$$p(x) = \pi_1 p(x, \mu_1, \Sigma_1) + \dots + \pi_k p(x, \mu_k, \Sigma_k), \tag{58}$$

где

$$p(x, \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{p/2}\sqrt{|\Sigma_i|}} \exp\left(-\frac{1}{2}(x - \mu_i)\Sigma_i^{-1}(x - \mu_i)^T\right) \tag{59}$$

Эта задача решается методом максимального правдоподобия с помощью EM-алгоритма.

## 2.37. Билет 45.Кластерный анализ:k-means, k-means++

Хотим искать кластеры $C_1, \dots, C_k$ минимизируя следующий функционал

$$\sum_{i=1}^{k} \sum_{j \in C_i} ||x_j - \mu_i|| \tag{60}$$

по разбиению всего пространства на $C_j$ и по всем $\mu_i$. Можно делать это по следующему алгоритму:

1. Выбираем случайно $\mu_1, \dots, \mu_k$.
2. $C_j$ – кластер, содержащий точки, которые лежат к $\mu_j$ ближе, чем к остальным $\mu_i$.
3. Для каждого $C_j$ пересчитываем центр как выборочное среднее элементов из этого кластера.

Проблема метода в том, что у такого функционала имеет много локальных минимумов и алгоритм может сойтись в значение, далекое от истинного. Метод $k - means++$ повторяет алгоритм, приведенный выше, но начальные значения выбираются не случайно, а следующим образом

1. Выбираем случайным образом первый центр $\mu_1$.
2. Считаем расстояние от всех точек до ближайшего центра $\{\rho_i\}$. После чего выбираем $x_i$ как новый центр с вероятностью, пропорциональной $\rho_i$.

3. Пока количество центров меньше, чем $k$, повторяем процедуру.

Результат полученной процедуры запишем как $J(\{C_j\}, \{\mu_j\})$. Известно, что

$$\frac{\mathbb{E}(J(\{C_j\}, \{\mu_j\}))}{J_{min}} = O(\ln k).$$

## 2.38. Билет 46. Запись задачи, решаемой k-means, как задачи low-rank approximatio с ограничениями. Использование PCA

Функционал вида (60) можно переписать в виде

$$||\mathbb{X} - \mathbb{GM}||_F^2, \tag{61}$$

где $M = (\mathbb{m}_1, \dots, \mathbb{m}_k)$, $\mathbb{G}$ — матрица, у которой в каждой строке стоит ровно одна единица и остальныеа нули. При этом, ранг матрицы $\mathbb{GM}$ не превосходит $k$, что приводит нас к задачи аппроксимации заданной матрицы матрицей меньшего ранга. Заметим, что задача похоже на задачу сингулярного разложения с ограничениями[50] Есть результаты, что если к данным применить анализ главных компонент, то пространство, натянутое на первые $k-1$ главных векторов будет близко к пространству, проходящему, через центры кластеров.

[50]Матрица $\mathbb{G}$ должна «вытаскивать» вектор среднего для соответсвующей строчки $\mathbb{X}$