Ответы к коллоквиуму по Многомерному Анализу Данных

Собрано 16 ноября 2015 г. в 20:10

Содержание

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

Нормальное распределение

Определение. *Говорят, что случайный вектор $\xi = (\xi_1, \xi_2, \ldots, \xi_p)^{\mathrm{T}}$ 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 матрица. Рассмотрим случайный вектор $\xi \sim \mathcal{N}(\mu, \Sigma)$. Тогда ξ 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 0.1 не требуется невырожденность ковариационной матрицы Σ. Если же Σ 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] $\square$

0.2 Билет 2. Оценки вектора средних и ковар.матрицы. Несмещенная оценка ковар. матрицы.

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

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

Введем $X_i^{(c)}$–i-тый центрированный признак и рассмотрим $\mathbb{X}^{(c)} = [X_1^{(c)}, \ldots, X_p^{(c)}]$ — 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.

0.3 Билет 3. Распределение вектора средних

Теорема. *Пусть дана выборка (на априорном языке) $\mathbf{x}_1, \ldots, \mathbf{x}_n \ldots$ с ковариационной матрицей Σ. 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 $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathcal{N}(\mu, \Sigma)$, 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?

0.4 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 ξ рассматривать как набор из p признаков, то $\eta = a^{\mathrm{T}}\xi$ — новый признак.

Рассмотрим же теперь матрицу $\mathbb{A} = [A_1 \colon \ldots \colon A_d] \in \mathrm{M}_{p,d}(\mathbb{R})$. 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 $\mathbb{Z} = [Z_1 \colon \ldots \colon Z_d] = \mathbb{X}\mathbb{A} \in \mathrm{M}_{n,d}(\mathbb{R})$ — для d признаков. Заметим, что в последней записи новыми признакими как раз будут являться столбцы $Z_1, \ldots, Z_d$.

Факторы и факторные нагрузки. Пусть задана матрица данных $\mathbb{X} = [X_1 \colon \ldots \colon X_p] \in \mathrm{M}_{n,p}(\mathbb{R})$. 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 $\mathbb{Z} = [Z_1 \colon \ldots \colon Z_d]$ 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 ортогональный: 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 $\mathbb{Q} = [Q_1 \colon \ldots \colon Q_d] \in \mathbb{M}_{n,d}(\mathbb{R})$.

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 $\mathbb{F} = [F_1 \colon \ldots \colon F_d]$, где $(F_j)_i = f_{ij}$. 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, $\mathbb{X} = \mathbb{Q}\mathbb{F}^{\mathrm{T}}$.

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

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

[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?

[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.

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 $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_d)$ и $\mathbb{P} = [P_1 \colon \ \ldots \colon P_d] \in \mathrm{M}_{n,d}(\mathbb{R})$. Тогда $\Sigma\mathbb{P}^{\mathrm{T}} = \mathbb{F}^{\mathrm{T}}$.[9] А значит $\mathbb{X} = \mathbb{Q}\Sigma\mathbb{F}^{\mathrm{T}}$.

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* $\mathbb{X} = [X_1 \colon X_2]$. *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 $\mathbb{Z} = [Z_1 \colon Z_2] = \mathbb{X}\mathbb{A}$, *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).*

0.5 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 $\mathbb{X} = \sum_{k=1}^{d} \mathbb{X}_k$.

0.6 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 $\|\mathbb{X}_i\|^2/\|\mathbb{X}\|^2$.

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

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

---

[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.

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

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.

0.7 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 λ_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* $\mathbb{Y}^{\mathrm{T}} U_i = \mathbb{0}$[19].
2. V_i *— собственный вектор матрицы* $\mathbb{Y}^{\mathrm{T}}\mathbb{Y}$*, соответствующий собственному числу* λ_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 $(V_i, V_j) = \left(\mathbb{Y}^{\mathrm{T}} U_i, \mathbb{Y}^{\mathrm{T}} U_j\right) / \sqrt{\lambda_i \lambda_j} = \\ = \left(U_i, \mathbb{Y}\mathbb{Y}^{\mathrm{T}} U_j\right) = \lambda_j (U_i, U_j) \sqrt{\lambda_i \lambda_j} = \delta_{i,j}$.

2. То, что V_i — собственный вектор, соответствующий λ_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] $\mathbb{Y}^{\mathrm{T}}(\mathbb{Y} V) = 0$.

[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?
[20] Здесь пока непонятно.

3. Подставьте 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 $\mathbb{Y} = \mathbb{E}_L \mathbb{Y}$.

$\square$

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 $\Lambda^{1/2} = \mathbb{U}^{\mathrm{T}}\mathbb{Y}\mathbb{V}^{\mathrm{T}}$ — 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!

0.8 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 $\mathbb{Y} = \sum\limits_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$.

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 $span(U_1,U_2)$.

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

Без доказательства.

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

[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.

0.9 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 $f^2 \in L_2 \leftrightarrow \int\limits_D |f|^2 d\mu$

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

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

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 ϕ 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 ψ соответствует нулевому собственному числу G^*G*

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

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

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

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

0.10.1 Сингулярное разложение

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

$D_1 = \{1,\dots,L\}$, μ_1 — считающая мера
$D_2 = \{1,\dots,K\}$, μ_2 — считающая мера

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

0.10.2 Выборочный анализ главных компонент

$\mathbb{X}, \mathbb{Y} = \mathbb{X}^{\mathrm{T}}$
$D_1 = \{1,\dots,L\}$, μ_1 — считающая мера
$D_2 = \{1,\dots,K\}$, μ_2 — вероятностная мера такая, что $\mu_2(\{i\}) = \frac{1}{K}$
$||\mathbb{Y}||_{1,2}^2 = \frac{||\mathbb{Y}||_F^2}{K}$
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:
$$\mathbb{Y} = \sum_i \sqrt{\lambda_i}U_iV_i^{\mathrm{T}}$$

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

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

$D_1 = \{1,\dots,L\}$, μ_1 — считающая мера
$(D_2, \mathfrak{A}_2, \mu_2)$ — вероятностное пространство
$g(x,s) \leftrightarrow \xi_i(\omega), g \leftrightarrow \xi, x \leftrightarrow i, s \leftrightarrow \omega$

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

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

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

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

В силу третьего свойства оптимальности SVD:

Пусть $Y_1, \ldots, Y_K \in \mathbb{R}^r$. $P \in \mathbb{R}^L$ 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* $U_i \forall j = 1, \ldots, k-1$, *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*

$Z_i = (< Y_1, U_i >, \ldots, < Y_k, U_i >)^{\mathrm{T}}$ — 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 $X_i \perp \mathbb{I}$.

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

Пусть

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

—- некоторое разложение SVD.

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

Предложение. *1. Аппроксимация матрицей меньшего ранга:*

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

2. Минимум достигается на первых r элементах сингулярного разложения:

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

Без доказательства.

0.14 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. ($span(U_1, \ldots, U_r)$).

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

0.15 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. $w = \arg\max\limits_{w\in\mathbb{R}^p} \sum\limits_{i=1}^{n} < y_i, w >^2$

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

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

Теорема. *В обозначениях, введенных выше:*

1. $\max\limits_{w} \sum\limits_{i=1}^{n} < y_i, w >^2 = \lambda_1$
2. $\arg\max\limits_{w} \sum\limits_{i=1}^{n} < y_i, w >^2 = u_1$

3. Для s-ого главного направления аналогично

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

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

Данное утверждение — переформулировка экстремальной задачи, определяющей SVD.

Заметим, что за счет центрирования $\sum\limits_{j=1}^{n} < y_j, u_i >^2$ с точностью до константы 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.

0.16 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:

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

0.17 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, $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_p \ge 0$

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.)

0.18 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 существует положительное собственной число единичной кратности и можно выбрать такой собственный вектор для данного с.ч., что все его компоненты будут строго положительны.*

Данная теорема, примененная для ковариационной матрицы, дает нам, что первое собственное направление будет взвешенной суммой исходных признаков. Таким образом, ее смысл — «среднее всех признаков».

0.19 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]:

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

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

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

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

[23]запись в векторном стиле: x_i — вектор-столбец для 1-ого признака

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

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.

0.20 Билет 20. Способы выбора числа главных компонент

Все методы во-многом эвристики. АГК - непараметрический, у нас нету модели, из-за чего никаких гипотез 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 $\lambda_i > \frac{tr(S)}{k}$, 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 $EL_i = \frac{1}{n}\sum\limits_{j=i}^{n}\frac{1}{j}$. 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.

0.21 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 λ_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 $\frac{\lambda_i}{\sum\limits_{i=1}^{p} \lambda_i}$ 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*

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

где

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

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

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

1. $\mathrm{D}_1 = \{1, \ldots, L\}$, $\mathrm{D}_2 = \{1, \ldots, K\}$,
2. $\mathfrak{A}_{1,2}$ *— множества всех подмножеств* $\mathrm{D}_{1,2}$,
3. μ_1 *— считающая мера,* μ_2 *— вероятностная,* $\mu_2(\{i\}) = 1/K$.

Если

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

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

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

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.

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

Главные компоненты — это $Z_i = \sqrt{\lambda_i} V_i$, $i = 1, \ldots, d$, 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 $\mathbb{X}^{\mathrm{T}}\mathbb{X}/K$. 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.*

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

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

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

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

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

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

С точки зрения построения базиса в пространстве индивидов:

$\{U_j\}_{j=1}^d$ — это ортонормированный базис в пространстве индивидов, $z_{ij} = Y_i^{\mathrm{T}} U_j$ — 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,

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

— это координаты i-го признака в этом базисе.

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

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

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

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}.

0.25 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 $B = A^{\mathrm{T}}\mathbb{U}$ — значения[27] главных компонент для этого индивида.

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

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

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

0.27 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 $\mathbb{U}^{\mathrm{T}} = \mathbb{U}^{-1}$.

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

Матрица факторных нагрузок — это $\mathbb{F} = [F_1 : \ldots : F_d] = \{f_{ij}\} \in \mathbb{R}^{p\times d}$, $f_{ij} = \langle X_i, V_j \rangle$.
Сумма квадратов по строкам — $\sum_{i=1}^p f_{ij}^2 = \|F_j\|^2 = \lambda_j$. Сумма квадратов по столбцам

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

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

$$\langle f_{i,j}, f_{k,j} \rangle = \langle \langle X_i, V_j \rangle, \langle X_k, X_j \rangle \rangle = \left\langle \frac{V^{\mathrm{T}} X_i}{\|X_i\|}, \frac{V^{\mathrm{T}} X_k}{\|X_k\|} \right\rangle = \left\langle \frac{X_i^{\mathrm{T}} V V^{\mathrm{T}} X_k}{\|X_i\|\|X_k\|} \right\rangle = \frac{X_i^{\mathrm{T}} X_k}{\|X_i\|\|X_k\|} = cos^2(X_i, X_k).$$

Если признаки нормированы и центрированы, то $cos^2(X_i, X_k) = cov(X_i, X_k)$.

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

Матрица $\mathbb{U}$ — ортогональная матрица, столбцы — координаты исходных признаков в новом базисе $\Rightarrow$ (U_{i1}, U_{i2}) — координаты i-го орта в плоскости первых двух главных компонент, поэтому в плоскости первых двух главных компонент исходные орты — это вектора с началом в полюсе и концом в (U_{i1}, U_{i2}).

[26]Здесь $L = p$ и обычно $d = p$

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

0.31 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.

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

Она равна корню из i-го собственного числа матрицы $\mathbb{Y}$.

0.33 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 μ_2: $\mu_2(i) = \omega_i$. В итоге получаем по прежнему биортогональное разложение, но с весами:$\langle V_i, V_j \rangle = \sum_{k=1}^{n} \omega_k$. Веса на признаках —- это масштаб (когда мы осознанно придаем больший вес индивиду, у которого больше разброс).

0.34 Билет 34. Какова модель в факторном анализе?

Модель факторного анализа

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

где ξ — случайный вектор размерности p, $\mathbb{F}$ — матрица размерности $p \times r$, η —случайный вектор размерности r, ε — случайный вектор размерности p.

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

Можно переписать все в виде:

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

Замечание. *F_i не может иметь вид $(0, \ldots, a, \ldots, 0)^{\mathrm{T}} \neq 0$, 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, η – факторными значениями.

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

0.35 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. ξ — *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. ε *и* η *некорр.,* ε_i *некорр. между собой.*
3. $\mathbb{E}\varepsilon = 0,\ \mathrm{cov}(\varepsilon) = \mathrm{diag}(\sigma_1^2,..,\sigma_p^2),\ \mathrm{cov}(\xi) = \Sigma,\ \widetilde{\xi} = \mathbb{F}\eta,\ \mathrm{cov}(\varepsilon) = \Psi,\ \mathrm{cov}(\widetilde{\xi}) = \mathbb{F}\mathbb{F}^{\mathrm{T}}$

   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* $\mathrm{cov}(\xi) = \mathrm{corr}(\xi)$.

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 ε, а мы интересуемся общей частью.

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 Σ — 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.

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.

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-т о зафиксировать$\mathcal{L}(\mathbb{S}$

$\max\limits_{\mathbb{F},\Psi}$

40. Проверка значимости модели ФА.

Не уверена, что это тот самый вопрос.

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

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

41. Критерий сферичности Бартлетта, для чего нужен

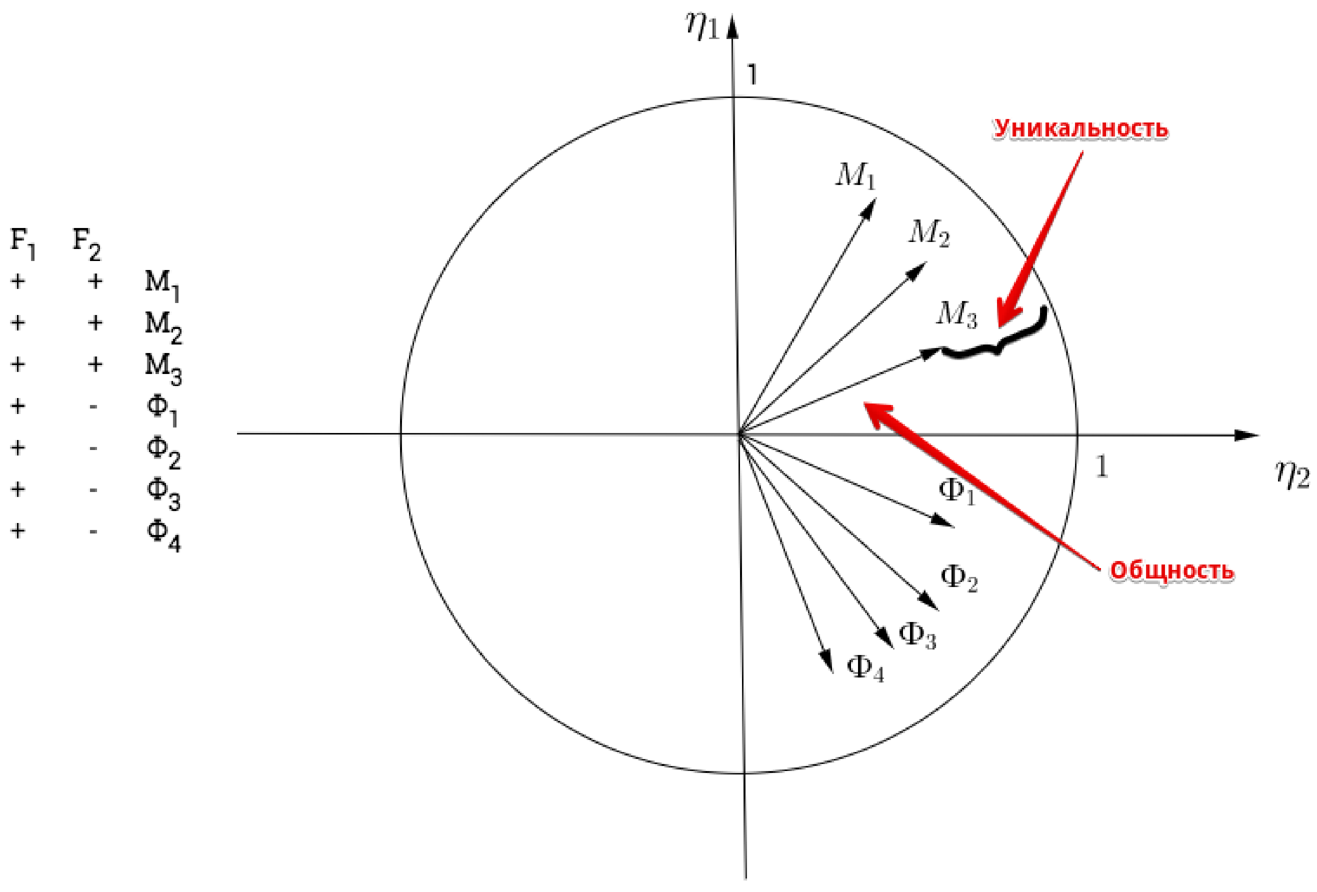
Частный случай критерия (Вопрос 40) при $r = 0$, имеет смысл проверять перед поиском факторов, вдруг общих факторов совсем нет.

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

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

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° получим скрытые факторы — это способности по математике и по физике.*

Не нашла материла про то, какие факторы не находит факторый анализ.