Задание 2 по курсу "Байесовский выбор модели"

Грабовой Андрей, группа 574

Задача 1

Пусть $\mathbf{x} \in \mathbb{R}^n: \ \forall i \in \{0, ..., n\} \ x_i \in N(m, \sigma)$. Проверить гипотезу H_0 о том, что $m = 0$. Вычислить критическую область и сосчитать мощность критерия W от истиных m и σ.

1. Рассмотрим следующую статистику:

$$T(\mathbf{x}) = \frac{\bar{\mathbf{x}}}{\frac{\bar{\sigma}}{\sqrt{n+1}}} \sim t(n) \text{ — в условии истиност гипотезы}$$

2. Критическая область $|T(\mathbf{x})| \leq t_{\text{кр}}(\alpha, n)$, где $t_{\text{кр}}(\alpha, n) = |F^{-1}_{t(n)}(\frac{\alpha}{2})|$. Для $n = 100$ и $\alpha = 0.05$ показано на рис. 1:

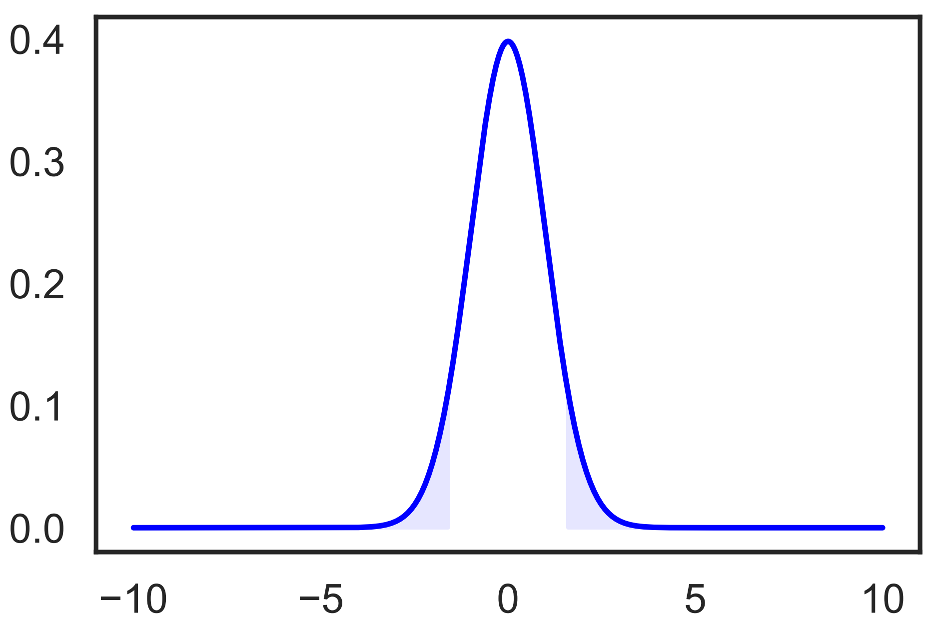

Рис. 1: График критической области

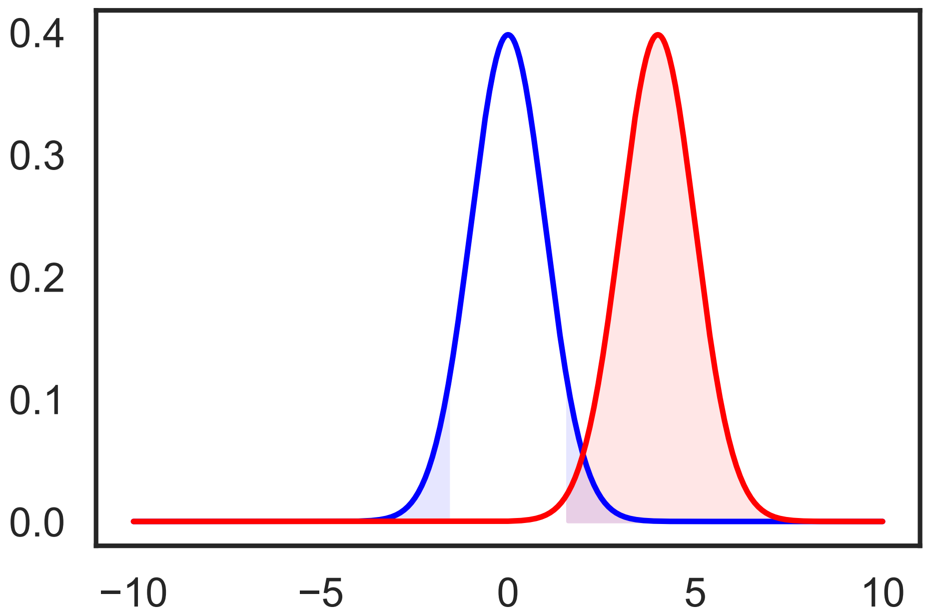

Рис. 2: График критической области

3. Мощность критерия это просто площадь под графиком(красная), как показано на рис. 2:

$$W = \left(\int_{-\infty}^{-t_{\text{кр}}(\alpha,n)} + \int_{t_{\text{кр}}(\alpha,n)}^{\infty} \right) p_{t(n)+\frac{m_0\sqrt{n+1}}{\bar{\sigma}}}(x)dx =$$

$$= \left(\int_{-\infty}^{-t_{\text{кр}}(\alpha,n)} - \int_{-\infty}^{t_{\text{кр}}(\alpha,n)} \right) p_{t(n)+\frac{m_0\sqrt{n+1}}{\bar{\sigma}}}(x)dx + 1 =$$

$$= 1 + F_{t(n)+\frac{m_0\sqrt{n+1}}{\bar{\sigma}}}(-t_{\text{кр}}(\alpha,n)) - F_{t(n)+\frac{m_0\sqrt{n+1}}{\bar{\sigma}}}(t_{\text{кр}}(\alpha,n))$$

Задача 2

Заметим:

$$p(\mathbf{y},\mathbf{w},\mathbf{X}|\alpha) = p(\mathbf{y},\mathbf{w}|\mathbf{X},\alpha)p(\mathbf{X}) \Rightarrow p(\mathbf{y},\mathbf{w}|\mathbf{X},\alpha) = p(\mathbf{w}|\alpha)p(\mathbf{y}|\mathbf{w},\mathbf{X})$$

1.

$$p(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}},\mathbf{X}_{\text{train}},\mathbf{y}_{\text{train}}) = \int p(\mathbf{y}_{\text{test}},\mathbf{w}|\mathbf{X}_{\text{test}},\mathbf{X}_{\text{train}},\mathbf{y}_{\text{train}})d\mathbf{w} =$$

$$= \int p(\mathbf{y}_{\text{test}}|\mathbf{w},\mathbf{X}_{\text{test}})p(\mathbf{w}|\mathbf{X}_{\text{train}},\mathbf{y}_{\text{train}})d\mathbf{w} = \int p(\mathbf{y}_{\text{test}}|\mathbf{w},\mathbf{X}_{\text{test}})\frac{p(\mathbf{w})p(\mathbf{X}_{\text{train}},\mathbf{y}_{\text{train}}|\mathbf{w})}{p(\mathbf{X}_{\text{train}},\mathbf{y}_{\text{train}})}d\mathbf{w} =$$

$$= \int p(\mathbf{w})p(\mathbf{y}_{\text{test}}|\mathbf{w}, \mathbf{X}_{\text{test}})p(\mathbf{y}_{\text{train}}|\mathbf{w}, \mathbf{X}_{\text{train}})\frac{p(\mathbf{X}_{\text{train}})}{p(\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}})}d\mathbf{w} =$$

$$= \int \frac{p(\mathbf{w})p(\mathbf{y}_{\text{train}}|\mathbf{w}, \mathbf{X}_{\text{train}})}{\int p(\mathbf{y}_{\text{train}}|\mathbf{w}, \mathbf{X}_{\text{train}})p(\mathbf{w})d\mathbf{w}}p(\mathbf{y}_{\text{test}}|\mathbf{w}, \mathbf{X}_{\text{test}})d\mathbf{w}$$

Тогда, прогнозируемые вероятности для $\mathbf{y}_{\text{test}}$ являються:

$$\mathcal{P}(\mathbf{y}_{\text{test}} = 1|\mathbf{X}_{\text{test}}) = \int \frac{p(\mathbf{w})p(\mathbf{y}_{\text{train}}|\mathbf{w}, \mathbf{X}_{\text{train}})}{\int p(\mathbf{y}_{\text{train}}|\mathbf{w}, \mathbf{X}_{\text{train}})p(\mathbf{w})d\mathbf{w}}p(\mathbf{y}_{\text{test}}|\mathbf{w}, \mathbf{X}_{\text{test}})d\mathbf{w}$$

Истиные вероятности для $\mathbf{y}_{\text{test}}$ являються:

$$\mathcal{P}(\mathbf{y}_{\text{test}} = 1) = \frac{1}{1 + \exp(-\mathbf{w}^{\mathrm{T}}\mathbf{X}_{\text{test}})}$$

Так-как нам известно, что $\mathbf{w} \sim N(0, \alpha^{-1}\mathbf{I})$, то мы можем вычислить этот интеграл.

Задача 3

a. Докажем, что Accuracy(ACC) является частным случаем ASY($\mathbf{P}$):

$$Acc = \frac{m_{11} + m_{00}}{m_2} = \frac{1}{m_2}\left(\sum_{i=1}^{m_{00}} 1 + \sum_{i=1}^{m_{11}} 1 + \sum_{i=1}^{m_{01}} 0 + \sum_{i=1}^{m_{10}} 0\right) =$$

$$= \left(\sum_{i=1}^{m_{00}} \frac{1}{m_2} + \sum_{i=1}^{m_{11}} \frac{1}{m_2} + \sum_{i=1}^{m_{01}} 0 + \sum_{i=1}^{m_{10}} 0\right) =$$

$$= \left(\sum_{i=1}^{m_{00}} \frac{1}{m_2} + \sum_{i=1}^{m_{11}} \frac{1}{m_2} + \sum_{i=1}^{m_{01}} 0 + \sum_{i=1}^{m_{10}} 0\right) = -\sum_{i=1}^{m_2} p_{y_i \hat{y}_i}$$

Знак минус появился в последнем равенстве из-за того, что ACC мы максимизируем, а ASY минимизируем. Получаем, что ACC частный случай ASY($\mathbf{P}$), когда $\mathbf{P} = \begin{bmatrix} -\frac{1}{m2} & 0 \\ 0 & -\frac{1}{m2} \end{bmatrix}$

b. Пусть класс объектов y_j не зависит от $\mathbf{x}_j$:

1. Построить наилучший прогноз $\hat{\mathbf{y}}_2$ в терминах ACC, если $\mathbb{P}(y_j = 1) = p$.

Наилучший прогноз достигается просто на константном классификаторе, причем за константу берем тот класс который имеет большую вероятность.

2. Построить наилучший прогноз $\hat{\mathbf{y}}_2$ в терминах ASY($\mathbf{P}$), если $\mathbb{P}(y_j = 1) = p$.

Наилучший прогноз достигается просто на константном классификаторе. Пусть $\mathbf{P} = \begin{bmatrix} w_{00} & w_{01} \\ w_{10} & w_{11} \end{bmatrix}$

Тогда нужно классифицировать константой $c = \arg\min_i \left((1-p)w_{i0} + pw_{i1}\right)$

3. Так-как y_j не зависит от $\mathbf{x}_j$, то просто считаем частоты каждого класса и это будет наша оценка вероятности.

Задача 4

Пусть имеется выборка $\mathbf{X}^0 = \mathbf{x}_1^0, ..., \mathbf{x}_{m_0}^0$ объектов класса 0 размера m_0 и выборка $\mathbf{X}^1 = \mathbf{x}_1^1, ..., \mathbf{x}_{m_1}^1$ объектов класса 1 размера m_1. Пусть признаки независимы в совокупност в обеих выборках, а также признаки имеют нормальное распределение с дисперсиями σ_j^2, одинаковой для одного итого же признака в разных классах, и возможно разной между признаками. Пусть требуется проверить гипотезу о том, что мат. ожидание значения признака с номером j совпадают для обоих классов.

1. Пусть $\sigma_j = \sigma$. Проверить гипотезу о равенстве мат. ожиданий.

Будем проверять гипотезу о равенстве мат. ожиданий для каждого признака по отдельности.

$$\bar{\mathbf{x}}_j^0 \sim N(M_j^0, \frac{\sigma^2}{m_0}), \quad \bar{\mathbf{x}}_j^1 \sim N(M_j^1, \frac{\sigma^2}{m_1}),$$

где M_j^0 и M_j^1 — мат. ожидание j-го признака для класса 0 и класса 1 соответственно. Тогда построим статистику

$$T(\mathbf{X}) = \frac{\bar{\mathbf{x}}_j^0 - \bar{\mathbf{x}}_j^1}{\sigma\sqrt{\frac{m_0+m_1}{m_0 m_1}}} \sim N(0,1) \text{ — в условиях истиности гипотезы}$$

Критическая область будет $|T(\mathbf{X})| > t_{N(0,1)}(0.05) = [\text{по таблице}] = 1.960$.

2. Не могу придумать ничего лучше, чем следующее. Мы знаем, что дисперсия в классе 0 и в классе 1 равны, следовательно оценим σ как среднее значение между выборочной дисперсией в классе 0 и выборочной дисперсией в классе 1, а дальше применим идеи использованы в пункте 1.

3. На рис. 3 показан график зависимости p_{value} от размера выборки. Из графика видно, что при увеличении размера выборки p_{value} для статистики где разность мат. ожиданий равно 1 убывает, и при размере выборки 1000 меньше чем 0.05. На рис. 5 и рис. 4 показано, как меняется количество ложных положительных отклонений и настоящих положительных отклонений разности матожиданий от нуля. Как видно из графиков при увеличении размера выборки мы все верно отвергаем гипотезу или говорим, что данные не противоречат ей. В эксперименте были выбраны следующие m_1 и m_2 из табл. 1

4.

m_1	100	500	1000	5000	100	500	1000	500	5000	1000
m_2	100	500	1000	5000	500	100	500	1000	1000	5000
$m_1 + m_2$	200	1000	2000	10000	600	600	1500	1500	6000	6000

Таблица 1: Таблица размера выборок

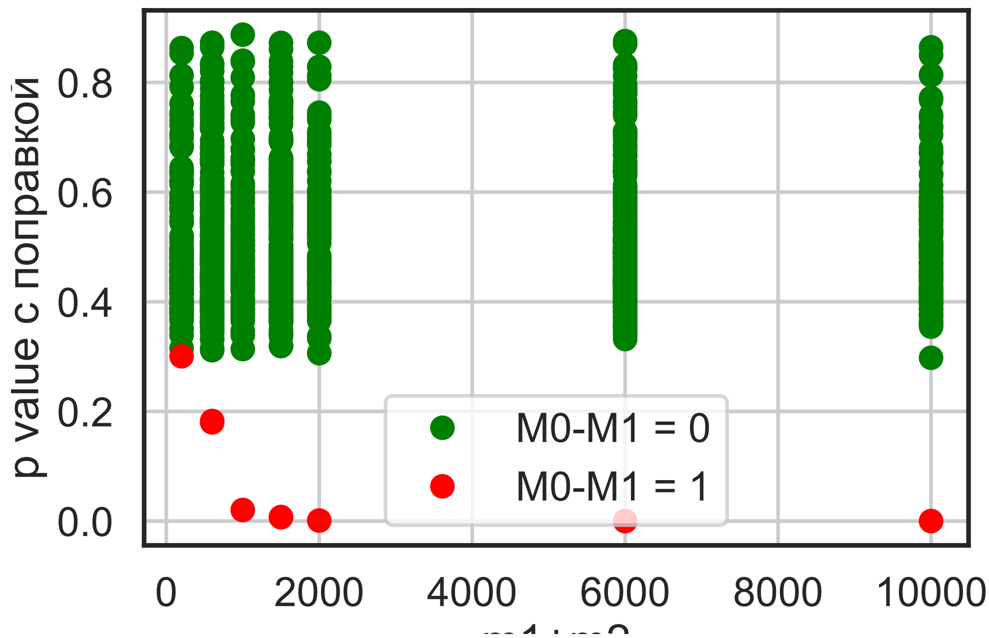

Рис. 3: График зависимости p_{value} от размера выборки$(m_1 + m_2)$

Задача 5

1. PCA - это один из способов уменьшение размерности данных, потеряв наименьшее количество информации. По сути это просто проекция на подпространство где важные признак выбираться из svd разложения(а именно оставляем те признаки которые имеют максимальные сингулярные числа). Формально решает задачу поиска подпространства меньшей размерности, в ортогональной проекции на которые среднеквадратиченое расстояние между точками максимально. В вероятностной постановке он говорит, что нужно найти такой базис в котором матрица ковариациии диагональна
2. Будем предполагать, что m»n. Тогда в силу того, что выборка независима, воспользовавшись определением, что PCA это «Поиск ортогональных проекций с наибольшим рассеянием», получаем что все рассеяния одинаковые, следовательно все компоненты важны и следовательно все компоненты нужны, причем std по каждой компоненте это σ
3. Пусть $\mathbf{X}$ состоит из n-1 зашумленной копии некоторого признака $\chi_{\mathbf{1}}$, а также из шкалированного признака $\chi_{\mathbf{2}}$, то есть $\mathbf{X} = [\chi_{\mathbf{1}} + \varepsilon_{\mathbf{1}}, \chi_{\mathbf{1}} + \varepsilon_{\mathbf{2}}, ..., \chi_{\mathbf{1}} + \varepsilon_{\mathbf{n-1}}, \xi\chi_{\mathbf{2}}]$, где все элементы независимым и одинаково нормально распределены. Найти в зависимости от ξ ожидаемую первую главную компоненту матрицы $\mathbf{X}$.

3.1. Аналитически:

Из определения следует, что первая главная компонента, эта компонента которая имеет максимальное среднеквадратическое отклоенние. Первые $n-1$ компоненты имеют среднеквад-

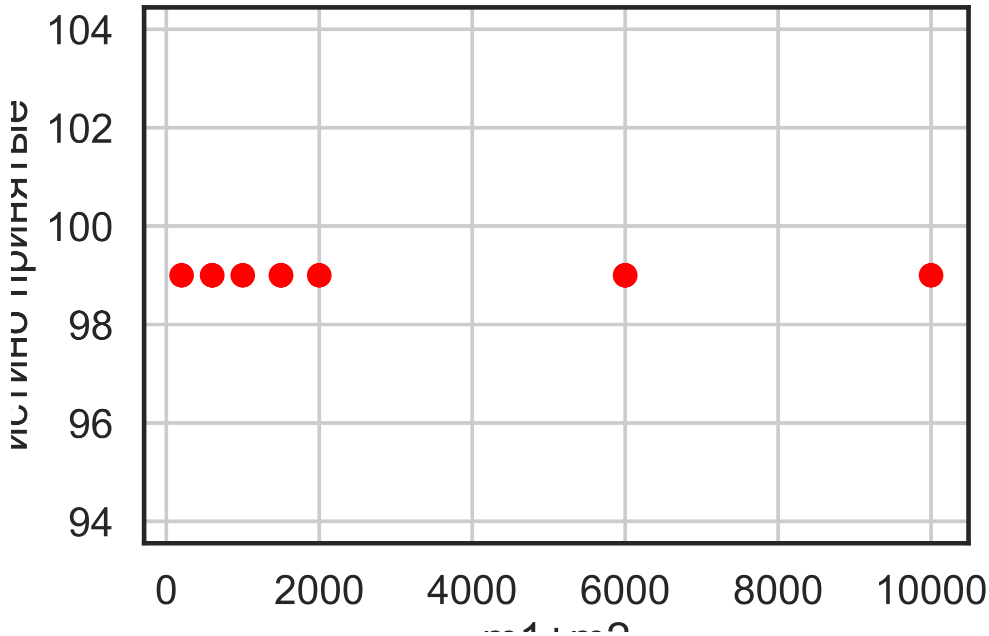

Рис. 4: График зависимости настоящих положительных отклонений от размера выборки($m_1 + m_2$)

ратические отклонения равные 2, последняя компонента имеет ξ. Тоесть ожидание первой главной компоненты при $\xi < 2$ это любая из $1, ...n-1$, а при $\xi >= 2$ это n-я компонента.

3.2. Сэмплирование:

Рассмотрим выборку из 5 признаков порожденной по тому как описано в условии задачи. Посмотрим какая будет главная компонента для k=1 — рис. 6 и для k=3 — рис. 7. Сэмплированый и аналитический результат совпадают.

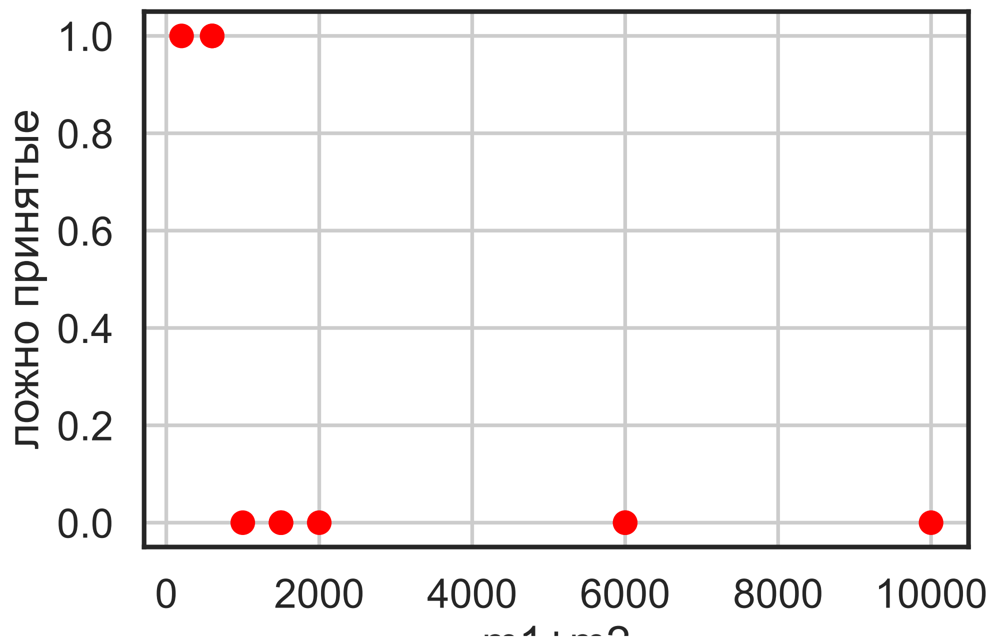

Рис. 5: График зависимости ложных положительных отклонений от размера выборки($m_1 + m_2$)

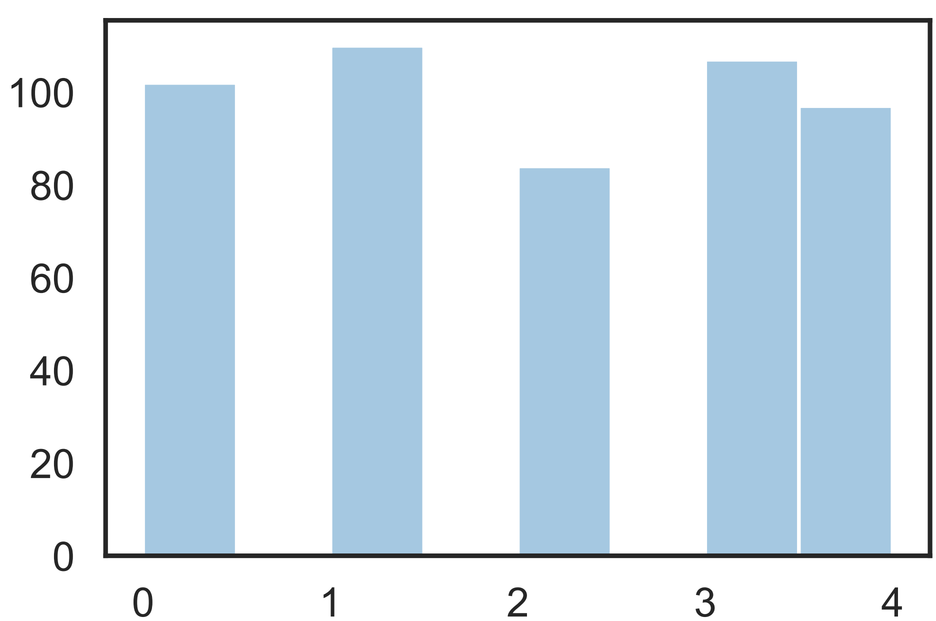

Рис. 6: Гистограмма номеров главных компонент

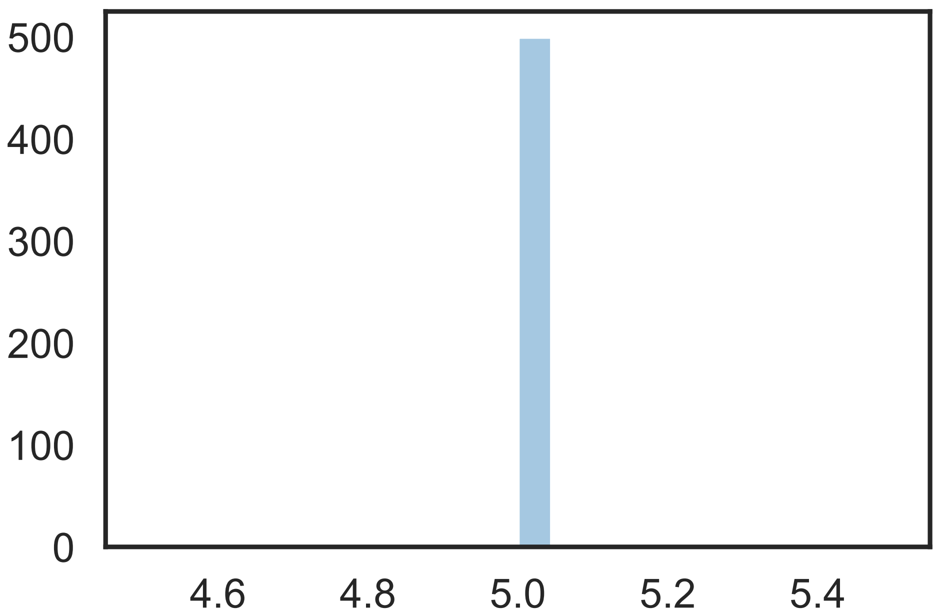

Рис. 7: Гистограмма номеров главных компонент