Яндекс

SCHOOL OF DATA ANALYSIS

Skoltech
Skolkovo Institute of Science and Technology

Ресемплинг

Моделирование Монте-Карло. Бутстреп. Доверительные интервалы. Коррекция множественных сравнений. Бэггинг в машинном обучении.

ПМИ ФКН ВШЭ, 15 сентября 2018 г.

Максим Шараев[1]

[1] Сколтех

Содержание лекции

- Непараметрический бутстреп
- Параметрический бутстреп
- Оценка доверительных интервалов на основе бутстрепа
- Метод складного ножа
- Множественная проверка гипотез
- Ресемплинг в машинном обучении
- Примеры

Непараметрический бутстреп

Стандартная постановка задачи

- Модель:
  - Имеем конечную простую выборку $\{X_i\}_{i=1}^n \subset \mathbb{R}$, порожденную распределением вероятности F.
  - Задана некоторая статистика $T_n = T_n(X_1, \ldots, X_n)$.
- **Задача:** оценить дисперсию $\mathrm{V}_F(T_n)$, которая зависит от неизвестного распределения F.

Пример

Пусть $T_n = \overline{X}_n$.
Тогда $\mathrm{V}_F(T_n) = \sigma^2/n$, где $\sigma^2 = \int (x-\mu)^2 dF(x)$ и $\mu = \int x dF(x)$.
Таким образом, дисперсия T_n есть функция F.

Идея бутстрепа

Шаг 1. Оценить $\mathrm{V}_F(T_n)$ с помощью $\mathrm{V}_{\widehat{F}_n}(T_n)$.

Шаг 2. Приблизить $\mathrm{V}_{\widehat{F}_n}(T_n)$ при помощи моделирования.

Пример

- Для $T_n = \overline{X}_n$, $\mathrm{V}_{\widehat{F}_n}(T_n) = \widehat{\sigma}^2/n$, где $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$.
- В данном случае шага 1 достаточно.
- Однако зачастую не удаётся выписать явно $\mathrm{V}_{\widehat{F}_n}(T_n)$. В таком случае прибегают к шагу 2.

Оценка дисперсии на основе бутстрепа

Предположим, что $Y_1, \ldots, Y_B$ — реализации i. i. d.случайных величин с функцией распределения G. Согласно закону больших чисел,

$$\overline{Y}_n = \frac{1}{B}\sum_{j=1}^{B} Y_j \xrightarrow{P} \int y dG(y) = \mathrm{E}\ Y, \quad B \to \infty.$$

Таким образом, мы можем использовать $\overline{Y}_n$ при достаточно больших B для приближения $\mathrm{E}\ Y$. Более того, для любой функции h с конечным мат. ожиданием имеем:

$$\frac{1}{B}\sum_{j=1}^{B} h(Y_j) \xrightarrow{P} \int h(y) dG(y) = \mathrm{E}\,(h(Y)), \quad B \to \infty.$$

Оценка дисперсии на основе бутстрепа

В частности, это означает, что мы можем моделировать и дисперсию:

$$\frac{1}{B}\sum_{j=1}^{B}(Y_j - \overline{Y}_n)^2 = \frac{1}{B}\sum_{j=1}^{B}(Y_j)^2 - \left(\frac{1}{B}\sum_{j=1}^{B}Y_j\right)^2 \xrightarrow{\mathrm{P}}$$

$$\to \mathrm{P}\int y^2 dG(y) - \left(\int y dG(y)\right)^2 = \mathrm{V}\,(Y), \quad B \to \infty.$$

Это означает, что мы можем использовать выборку для оценки дисперсии. Данная процедура позволяет нам находить $\mathrm{V}_{\widehat{F}_n}(T_n)$ — «дисперсию T_n при данных, распределённых по $\widehat{F}_n$».

Оценка дисперсии на основе бутстрепа

Теперь ситуация выглядит следующим образом.

- С точки зрения реальности:

$$F \Rightarrow X_1, \ldots, X_n \Rightarrow T_n = g(X_1, \ldots, X_n)$$

- С точки зрения бутстрепа:

$$\widehat{F}_n \Rightarrow X_1^*, \ldots, X_n^* \Rightarrow T_n^* = g(X_1^*, \ldots, X_n^*)$$

- **Проблема:** как получить $X_1^*, \ldots, X_n^*$ из $\widehat{F}_n$?
- **Решение:** при подсчёте мат. ожидания с помощью $\widehat{F}_n$ мы использовали одинаковую массу $\frac{1}{n}$. Это значит, что получение наблюдения из $\widehat{F}_n$ эквивалентно выбору случайной точки из исходной выборки.

Алгоритм: оценка дисперсии

Приведём **алгоритм** оценки дисперсии с помощью бутстрепа:

1. Выбираем $X_1^*, \ldots, X_n^* \sim \widehat{F}_n$
2. Вычисляем $T_n^* = g(X_1^*, \ldots, X_n^*)$
3. Повторяем шаги 1 и 2 пока не получим $T_{n,1}^*, \ldots, T_{n,B}^*$
4. Положим

$$v_{boot} = \frac{1}{B}\sum_{b=1}^{B}(T_{n,b}^* - \frac{1}{B}\sum_{r=1}^{B}T_{n,r}^*)^2$$

В итоге получаем:

$$\mathrm{V}_F(T_n) \approx \mathrm{V}_{\widehat{F}}(T_n) \approx v_{boot}$$

При этом первое приближение оказывается точнее второго.

Параметрический бутстреп

Параметрический бутстреп

- Предположим, что $F(x) \in \{F(x, \theta) \ : \ \theta \in \Theta \subset \mathbb{R}^d\}$.
- Тогда с помощью максимизации правдоподобия найдем параметр θ, а именно:

$$\theta = \arg\max_{\theta \in \Theta} \mathcal{L}(\vec{X}, \theta)$$

- Вместо ОМП можно использовать метод моментов.
- Далее действуем по описанной схеме непараметрического бутстрепа.

(+) Непрерывная выборка, в маленькой выборке, как правило, недооценка разброса.

(−) Произвольная модель и оценка параметров.

- Обычно выборки с менее чем 10 элементами считаются ненадежными для непараметрического бутстрепа.

Доверительное оценивание на основе бутстрепа

Нормальный интервал

- Если предположить, что данные распределены нормально, то имеет смысл рассмотреть следующий доверительный интервал:

$$(T_n - z_{\alpha/2}\widehat{se}_{boot}, T_n + z_{\alpha/2}\widehat{se}_{boot}),$$

- При этом $z_\alpha : F_{N(0,1)}(z_\alpha) = 1 - \alpha$, $\widehat{se}_{boot} = \sqrt{v_{boot}}$.

Центральный интервал

- Пусть $\theta = T(F)$ и $\widehat{\theta}_n = T(\widehat{F}_n)$.
- Пусть $\widehat{\theta}^*_{n,1}, \ldots, \widehat{\theta}^*_{n,B}$ — получены итерированием шагов 1 и 2 алгоритма бутстрепа.
- Пусть θ^*_β — обозначает β-квантиль для $(\theta^*_{n,1}, \ldots, \theta^*_{n,B})$.
- Тогда центральный $(1-\alpha)$-доверительный интервал :

$$C_n = (2\widehat{\theta}_n - \widehat{\theta}^*_{1-\alpha/2}, 2\widehat{\theta}_n - \widehat{\theta}^*_{\alpha/2}).$$

Центральный интервал

Теорема

При некоторых несильных условия на $T(F)$,

$$\mathrm{P}_F(T(F) \in C_n) \to 1 - \alpha, \qquad n \to \infty,$$

$$C_n = (2\widehat{\theta}_n - \widehat{\theta}^*_{1-\alpha/2}, 2\widehat{\theta}_n - \widehat{\theta}^*_{\alpha/2})$$

Метод складного ножа

Метод складного ножа

- Пусть $T_n = (X_1, \ldots, X_n)$.
- Рассмотрим n подвыборок: $T_{(-i)} = \frac{1}{n-1} \sum_{j \neq i} X_j$.
- Пусть $\overline{T}_n = n^{-1} \sum_{i=1}^{n} T_{(-i)}$.
- Построим следующую оценку $\mathrm{V}\,(T_n)$:

$$v_{jack} = \frac{n-1}{n} \sum_{i=1}^{n} (T_{(-i)} - \overline{T}_n)^2$$

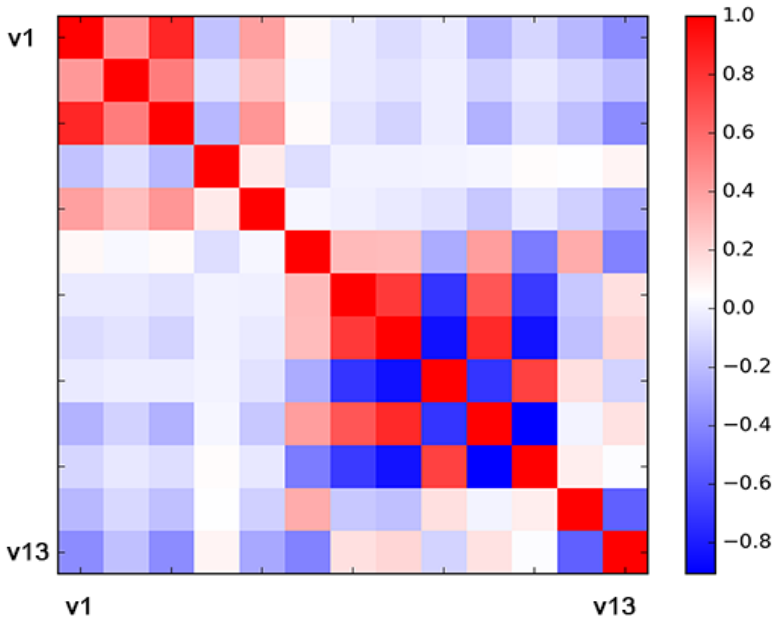
- Тогда оценка стандартной ошибки по методу складного ножа имеет вид $\widehat{se}_{jack} = \sqrt{v_{jack}}$.
- Может быть показано, что $v_{jack} / \mathrm{V}\,(T_n) \xrightarrow{\mathrm{P}} 1$.

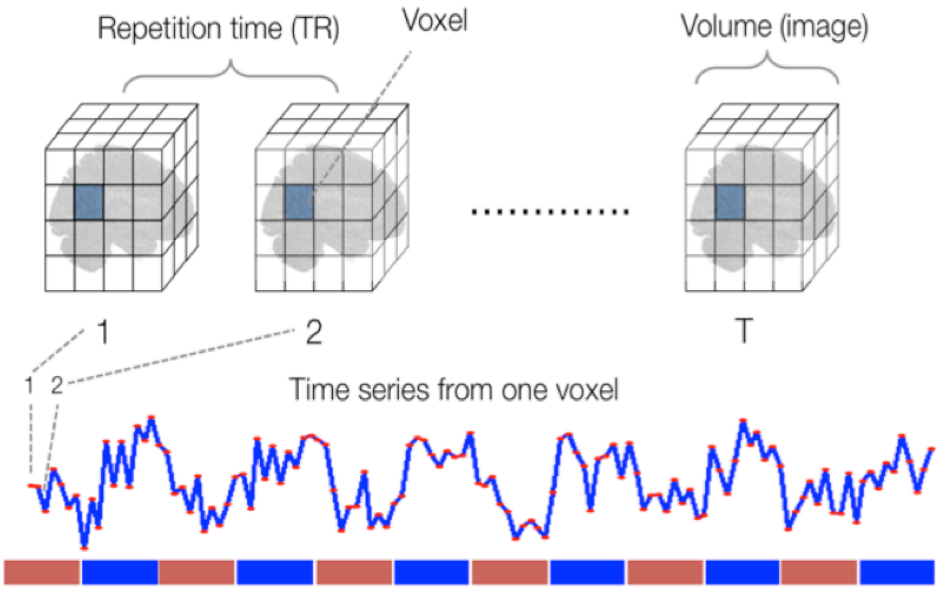
Метод складного ножа

1. Бутстреп — рандомизированная аппроксимация delete-m метода складного ножа.
2. Бутстреп — разные результаты, метод складного ножа — всегда одинаковые
3. Оценка дисперсии функционала vs. оценка всего распределения.
4. Условия гладкости функционала.
5. Метод складного ножа проще применять для сложных схем семплирования.

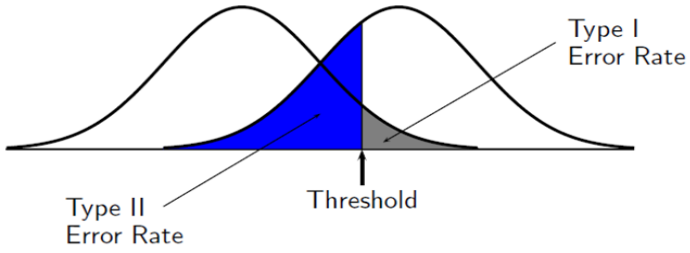
Множественная проверка гипотез

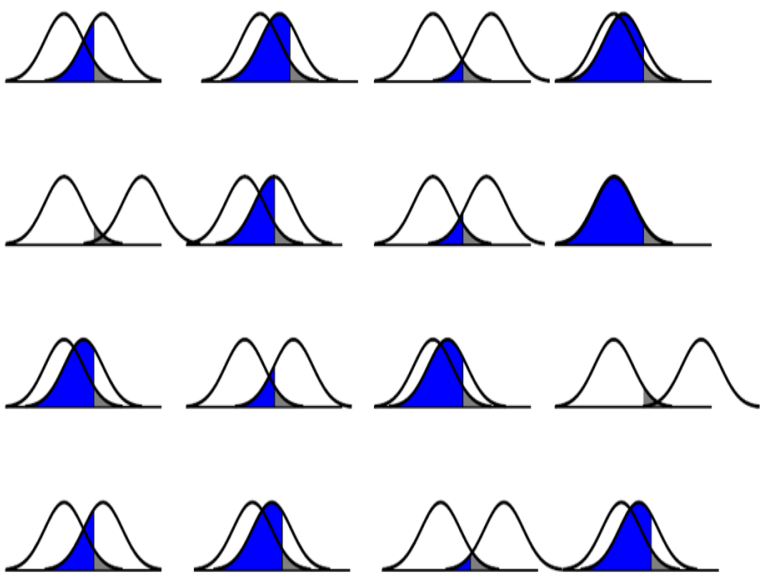
Множественная проверка гипотез

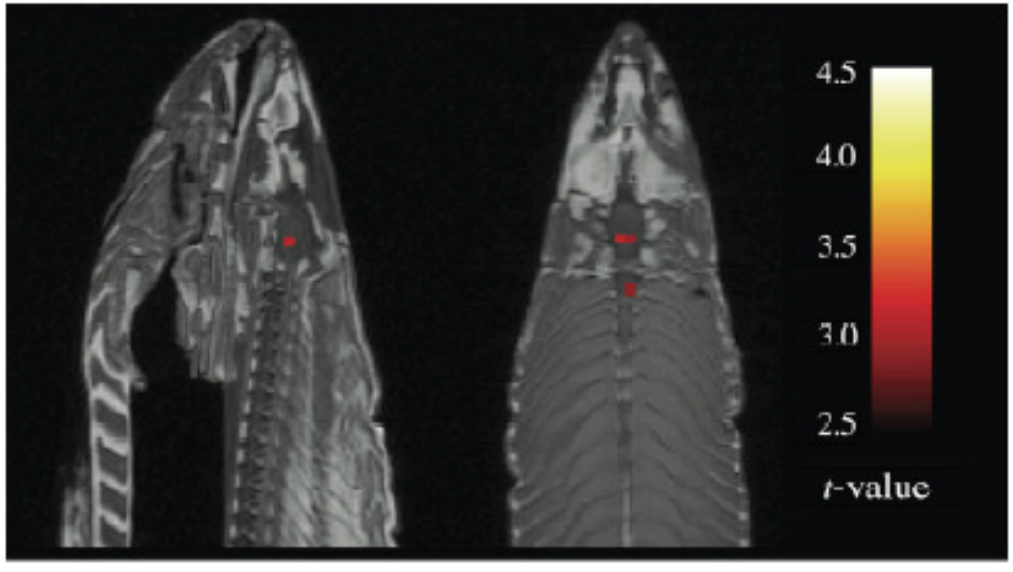




Проблема множественных сравнений







A *t*-contrast was used to test for regions with significant BOLD signal chang during the photo condition compared to rest. The parameters for th comparison were $t(131) > 3.15$, p(uncorrected) < 0.001, 3 voxel exte threshold.

Коррекция Бонферрони

- Пусть $H_1, \ldots, H_m$ — семейство проверяемых гипотез, $p_1, \ldots, p_m$ — их p-values.
- m — общее число гипотез, m_0 — истинных гипотез.
- Тогда групповая вероятность ошибки (family-wise error rate, FWER) — вероятность отвергнуть хотя бы одну истинную H_i, то есть сделать хотя бы одну ошибку первого рода.
- При коррекции Бонферрони отвергается нулевая гипотеза для всех $p_i < \alpha/m$, тем самым удерживая FWER на уровне $\leq \alpha$:

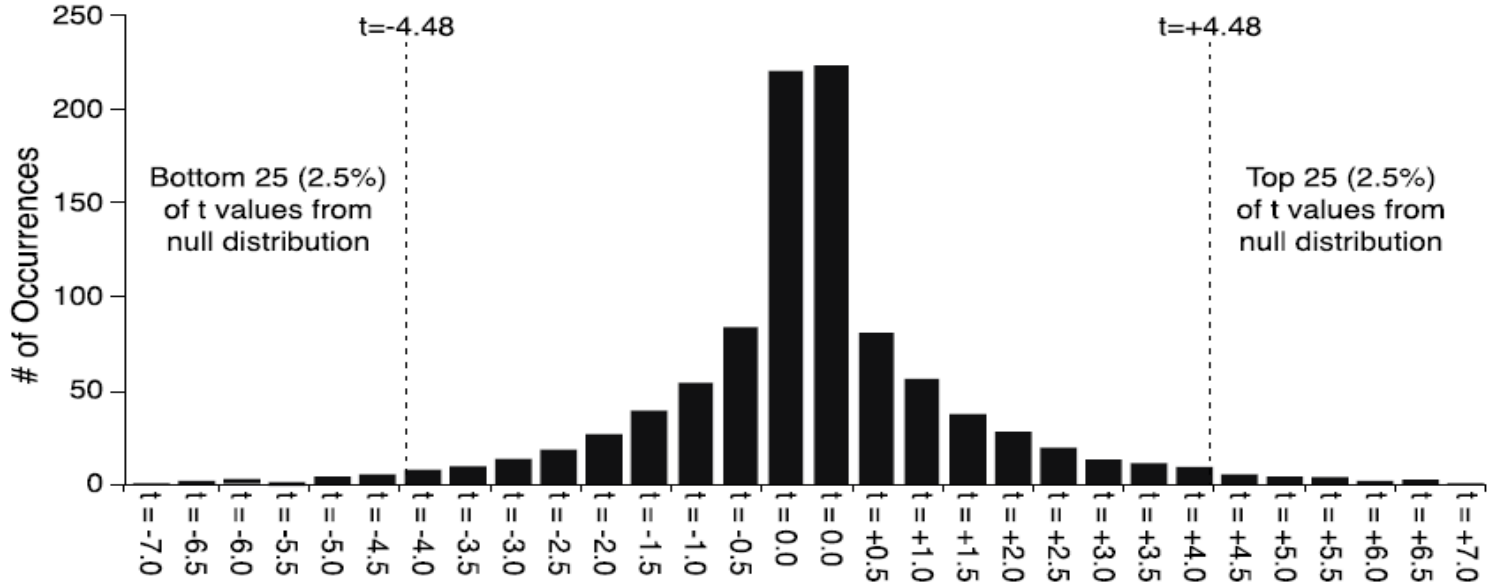
$$\begin{aligned} \text{FWER} &= \mathrm{P}\left\{\bigcup_{i=1}^{m_0}\left(p_i \leq \frac{\alpha}{m}\right)\right\} \leq \\ &\leq \sum_{i=1}^{m_0}\left\{\mathrm{P}\left(p_i \leq \frac{\alpha}{m}\right)\right\} = m_0\frac{\alpha}{m} \leq m\frac{\alpha}{m} = \alpha. \end{aligned}$$

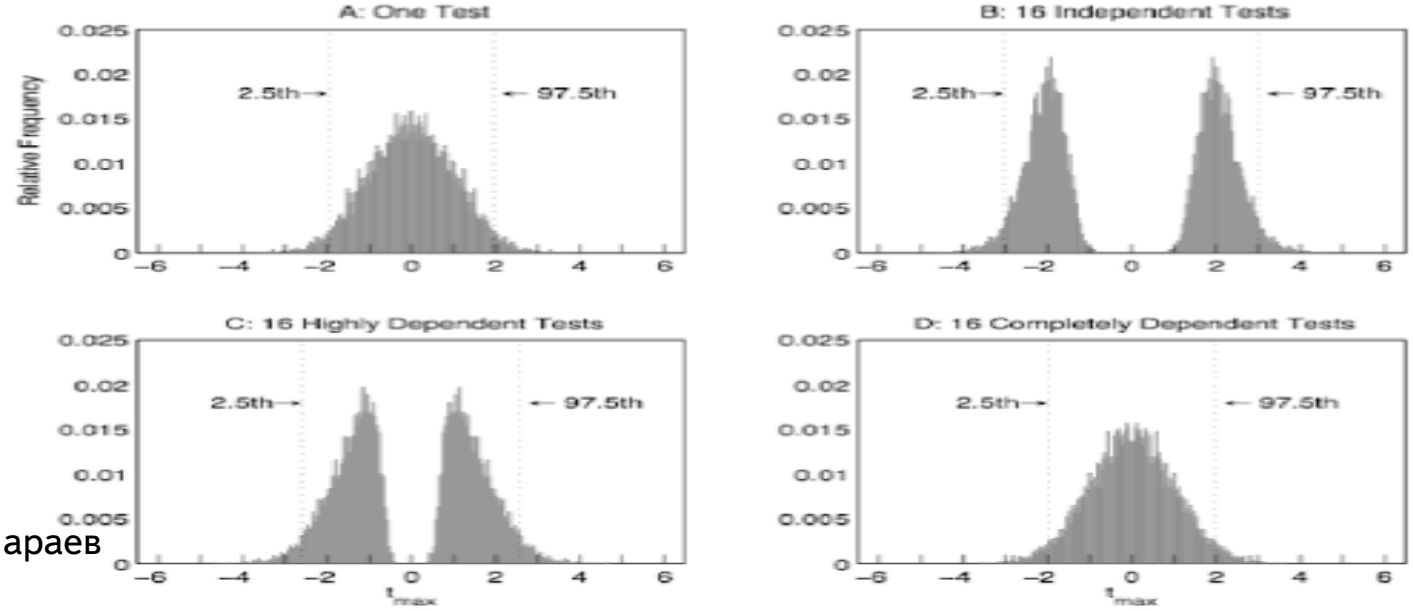
Пермутационный тест

Если знаем, как применить ресемплинг с учетом дизайна исследования и нулевой гипотезы. Простейший случай: две выборки (два условия), равенство средних.
Алгоритм:

1. Для каждой проверяемой гипотезы $H_1, \ldots, H_m$ считаем истинные $T_1, \ldots, T_m$.
2. Для каждой проверяемой гипотезы $H_1, \ldots, H_m$ случайным образом меняем метки N раз, считаем суррогатные $T_1^1, \ldots, T_1^N, T_m^1, \ldots, T_m^N$, находим $T_{\max}^1, \ldots T_{\max}^N$
3. Строим нулевое распределение $T_{\max}$, критическое значение статистики для истинных $T_1, \ldots T_m$ берется как квантиль (0.05, 0.01 и т.д.) этого нулевого распределения.

Пермутационный тест

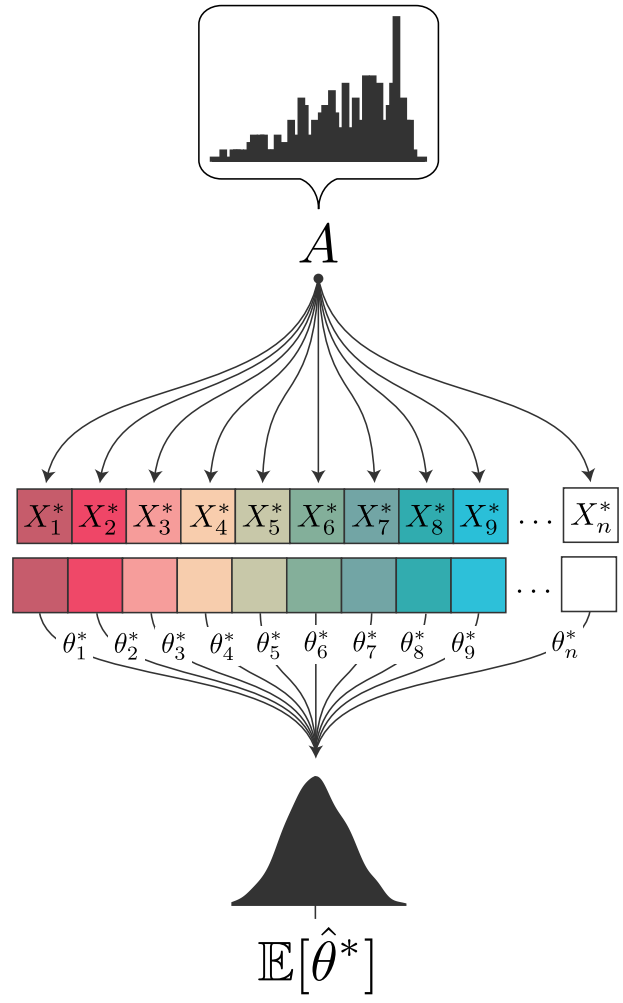




Ресемплинг
в машинном обучении

Бутстреп и слабые решающие правила

- Вход: выборка $X^\ell = \{(x_i, y_i)\}_{i=1}^\ell$
- Бутстреп: случайное семплирование новых элементов X_1^m вида (x_i, y_i) из X^ℓ с равными вероятностями и с возвращением (возможны повторы (x_i, y_i)!)
- Идея ансамблей гипотез:
  1. Сгенерировать B бутстрепных выборок $X_1^m, \ldots, X_B^m$
  2. Обучить B гипотез $h_1, \ldots, h_B$
  3. Усреднить прогнозы и получить $h(x) = \frac{1}{B}\sum_{i=1}^B h_i(x)$
  4. Profit!



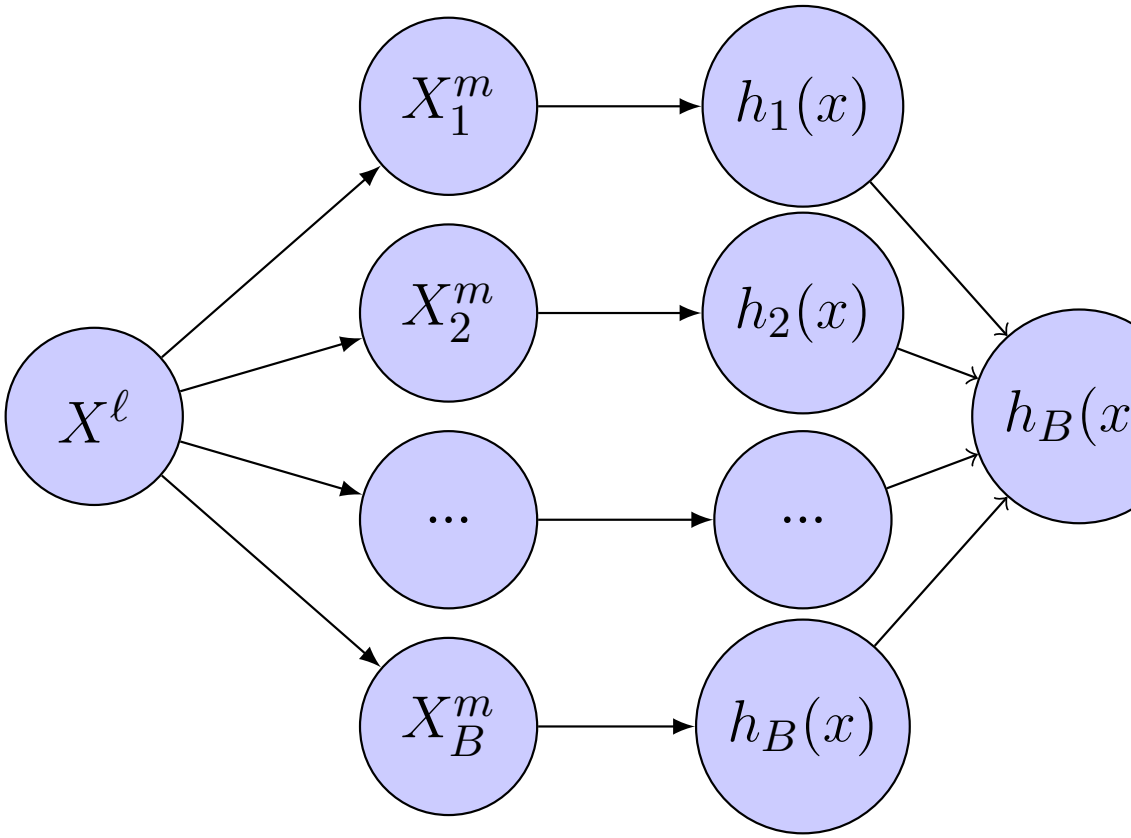
Picture credit: http://www.drbunsen.org/bootstrap-in-picture

Bagging: Bootstrap AGGregation

- Вход: выборка
  $X^\ell = \{(x_i, y_i)\}_{i=1}^\ell$
- Слабые гипотезы — из бутстрепа
  $\widetilde{\mu}(X^\ell) = \mu(\widetilde{X}^\ell)$
- Среднее по ансамблю

$$h_B(x) = \frac{1}{B}\sum_{i=1}^{B} h_i(x) =$$

$$= \frac{1}{B}\sum_{i=1}^{B} \tilde{\mu}(X^\ell)(x)$$



Пример: случайный лес

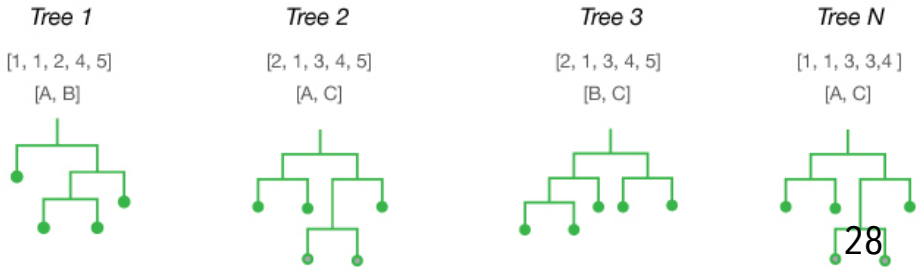
- Бэггинг над (слабыми) решающими деревьями
- Уменьшение ошибки с помощью усреднения по элементам выборки и признакам
- Вход: выборка $X^\ell = \{(\mathbf{x}_i, y_i)\}_{i=1}^\ell$, где $\mathbf{x}_i \in \mathbb{R}^d, y_i \in \mathbb{Y}$
- Алгоритм: повторять для $i = 1, \ldots, N$:
  1. Выбрать p случайных признаков из d
  2. Забутстрепить выборку $X_i^m = \{(\mathbf{x}_i, y_i)\}_{i=1}^\ell$ где $\mathbf{x}_i \in \mathbb{R}^p, y_i \in \mathbb{Y}$
  3. Обучить решающее дерево $h_i(\mathbf{x})$ по бутстрепной X_i^m
  4. Остановка: листья h_i содержат менее, чем $n_{\min}$ примеров

$\mathbf{x}_i \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$

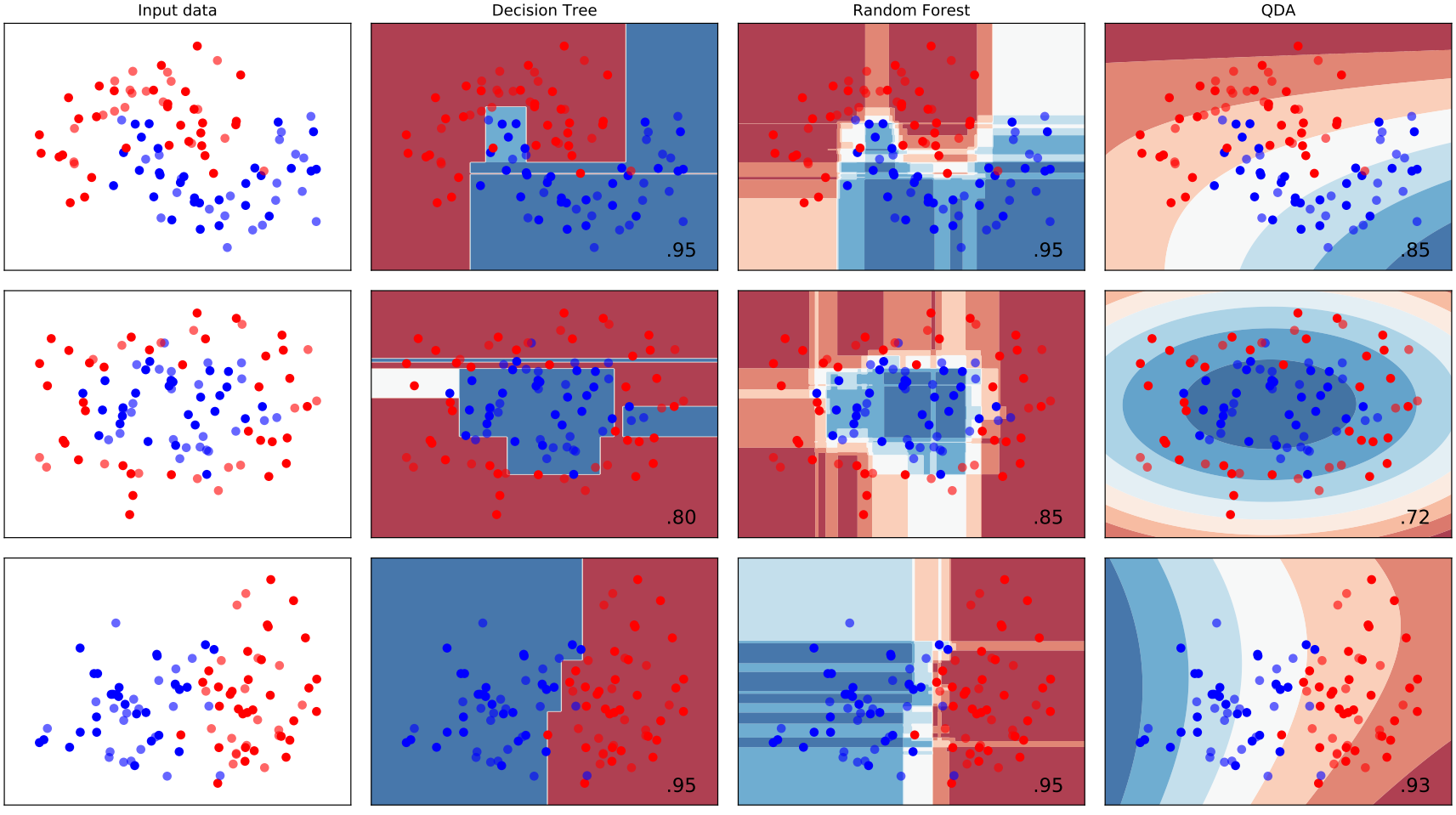
$X^\ell = \{(\mathbf{x}_i, y_i)\}_{i=1}^5$

Бутстреп $X_i^m, i \in \{1, 2, 3, 4\}$

Обучаем $_i(\mathbf{x})$ по X_i^m



Случайный лес: синтетическая выборка



Бэггинг: обсуждение эффективности

- $h_i(\mathbf{x})$ — гипотеза, построенная по бутстрепной выборке.
- Общий результат бэггинга: $h_B(\mathbf{x}) = \frac{1}{B}\sum_{i=1}^{B} h_i(\mathbf{x})$
- Пусть дана задача регрессии и на каждой бутстрепной выборке найдены функции $h_1(\mathbf{x}), \ldots, h_B(\mathbf{x})$.
- Ошибки на элементах выборки $\varepsilon_k(\mathbf{x}_i) = h_k(\mathbf{x}_i) - y_i$, $k = 1, \ldots, B$, $y_i = f(\mathbf{x}_i)$ — аппроксимируемая функция.
- Матожидание среднеквадратичной ошибки: $\mathcal{E}_1 = \mathrm{E}\,[h_k(\mathbf{x}) - y]^2 = \mathrm{E}\,[\varepsilon_k^2(\mathbf{x})]$.
- Средняя ошибка найденных гипотез

$$\mathcal{E}_B = \frac{1}{B}\,\mathrm{E}\,[\sum_{i=1}^{B} \varepsilon_k^2(\mathbf{x})]$$

Бэггинг: обсуждение эффективности

› Пусть ошибки несмещены и некоррелированы:

$$\mathrm{E}\left[\varepsilon_k(\mathbf{x})\right] = 0, \quad \mathrm{E}\left[\varepsilon_k(\mathbf{x})\varepsilon_j(\mathbf{x})\right] = 0, \quad k \neq j.$$

› Если возьмем функцию: $h(\mathbf{x}) = \frac{1}{B}\sum_{k=1}^{B} h_k(x)$, то

$$\begin{aligned}\mathcal{E}_B &= \mathrm{E}\left[\frac{1}{B}\sum_{k=1}^{B} h_k(\mathbf{x}) - f(\mathbf{x})\right]^2 = \mathrm{E}\left[\frac{1}{B}\sum_{k=1}^{B} \varepsilon_k(\mathbf{x})\right]^2 = \\ &= \frac{1}{B^2}\,\mathrm{E}\left[\sum_{k=1}^{B} \varepsilon_k^2(\mathbf{x}) + \sum_{k\neq j} \varepsilon_k(\mathbf{x})\varepsilon_j(\mathbf{x})\right] = \frac{1}{B}\mathcal{E}_1\end{aligned}$$

› Таким образом, снизили дисперсию в B раз.

› Общая ошибка модели («разложение ошибки на смещение и разброс») = смещение + разброс + неустранимая ошибка.

Примеры

Пример №1

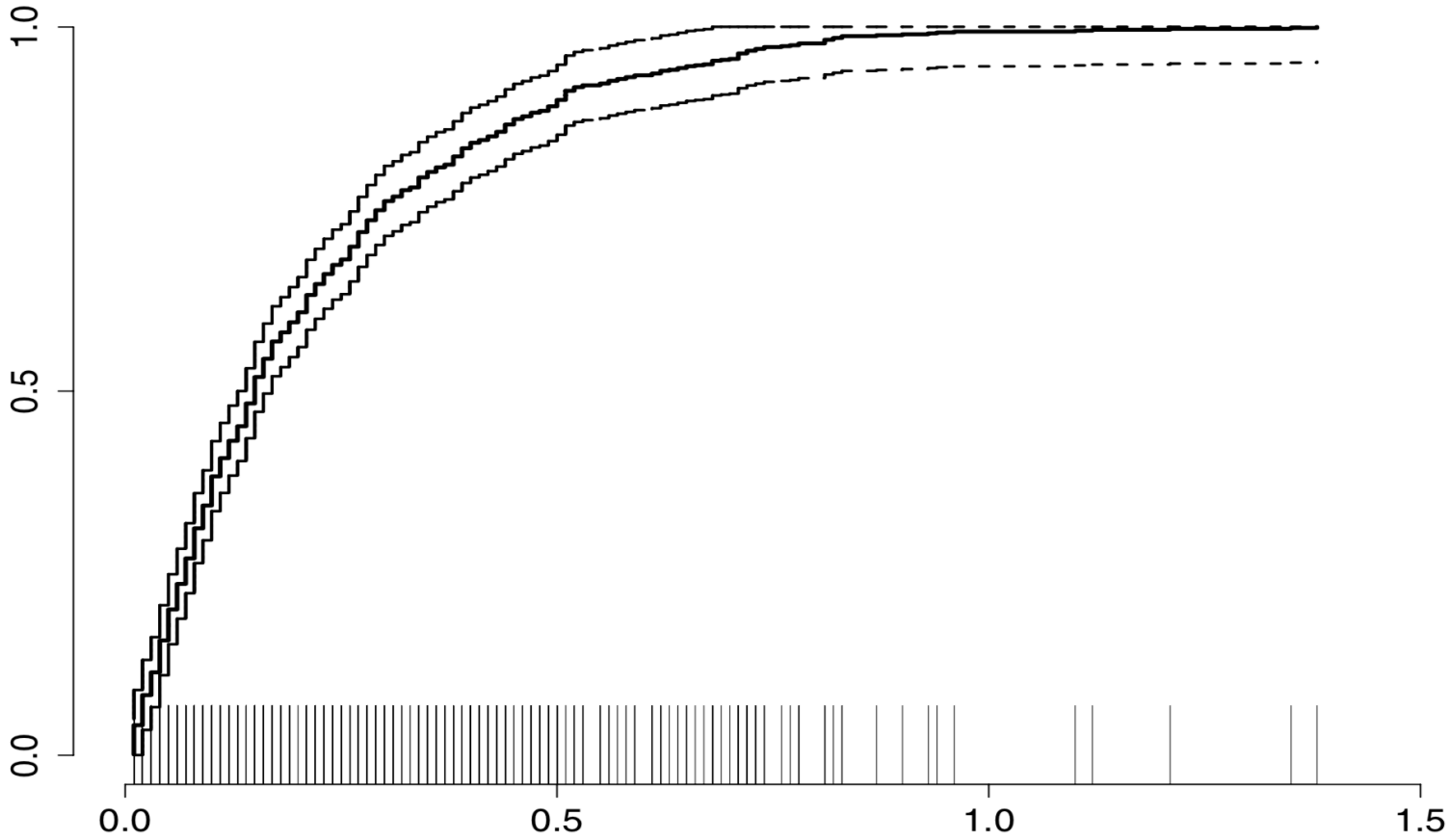


FIGURE 2.1. Nerve data. Each vertical line represents one data point. The solid line is the empirical distribution function. The lines above and below the middle line are a 95 percent confidence band.

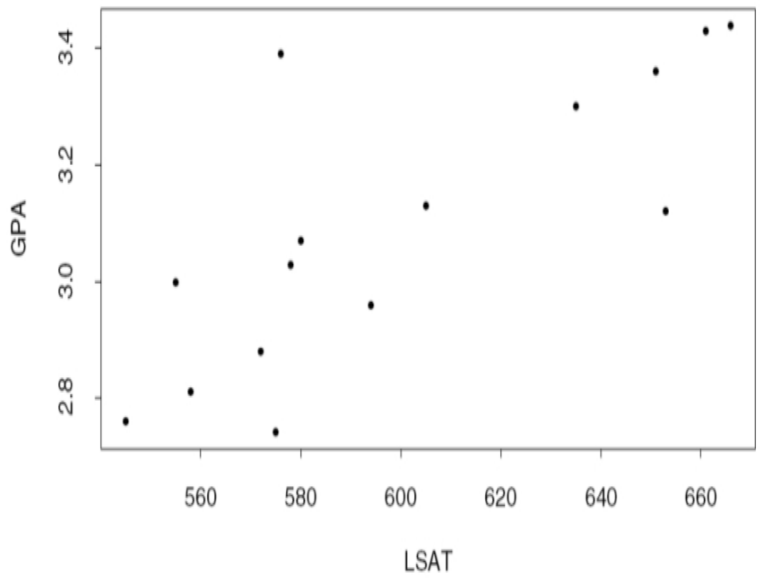
Picture credit: Wasserman, Larry. “All of Nonparametric Statistics, 268 pp.” (2007).

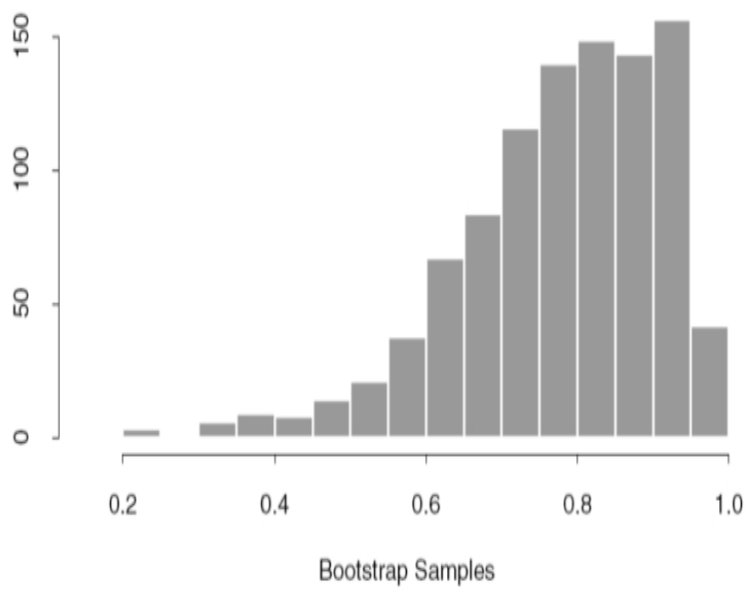
Пример №1

- Данные о моментах времени прохождения импульсов вдоль нервного волокна.
- $\theta = T(F) = \int \frac{(x-\mu)^3}{\sigma^3} dF(x)$ — коэффициент асимметрии.
- $\widehat{\theta} = T(\widehat{F}_n) = \frac{1}{\widehat{\sigma}^3} \cdot \left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^3\right] = 1.76.$
- Оценка дисперсии с помощью непараметрического бутстрепа: $\widehat{\mathrm{V}}_{\widehat{F}_n}^{\mathrm{boot}}(T_n) = (0.16)^2$, при $B = 1000$.
- 95% интервал для коэффициента асимметрии:
  - Нормальный интервал: (1.44, 2.09).
  - Центральный интервал: (1.48, 2.11).
  - Интервал на основе процентилей: (1.42, 2.03).

Пример №2

Данные о LSAT (Law School Admissible Test) и GPA (Grade Point Average).





Нас интересует корреляция между ними.

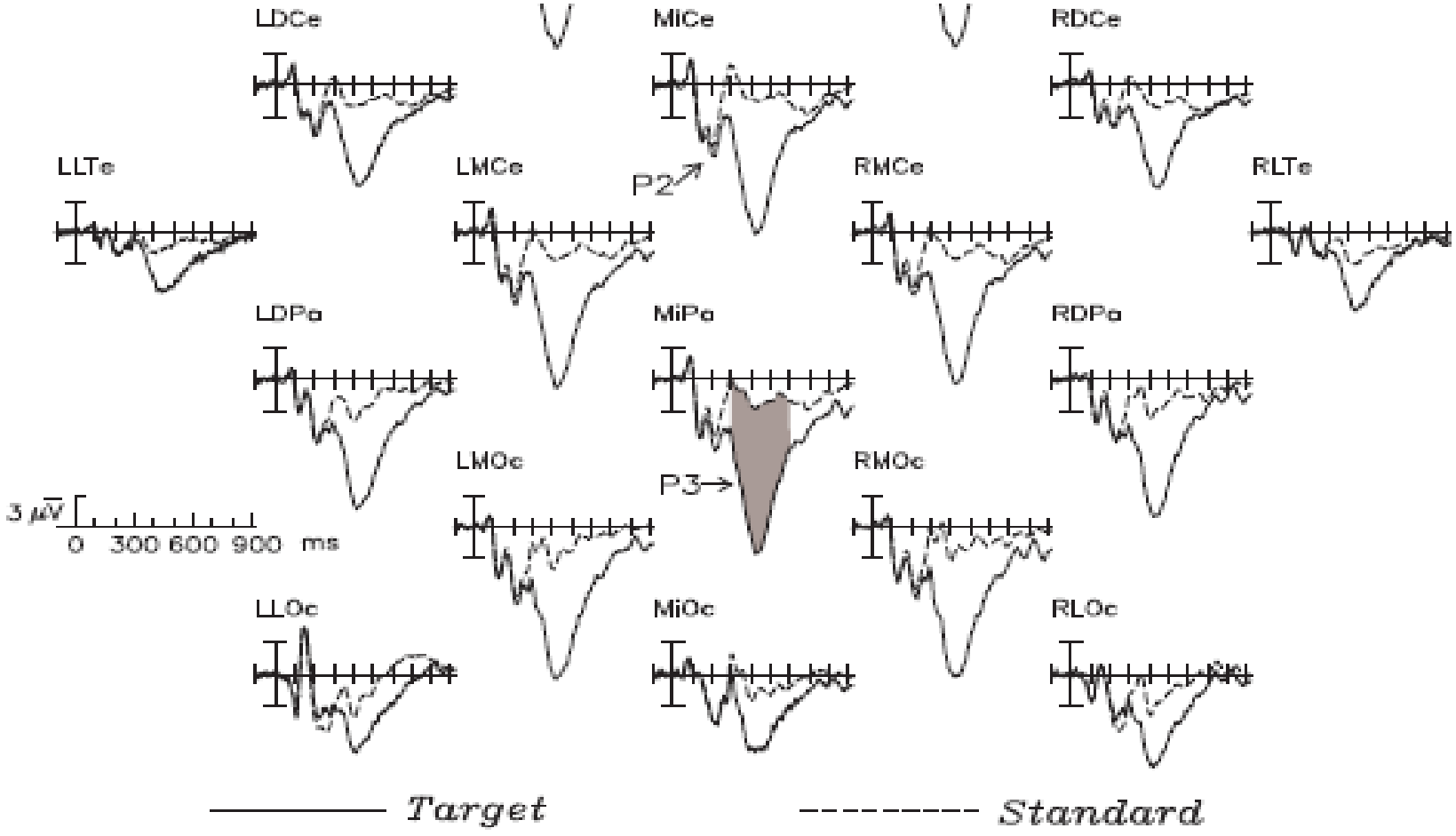
Пример №2

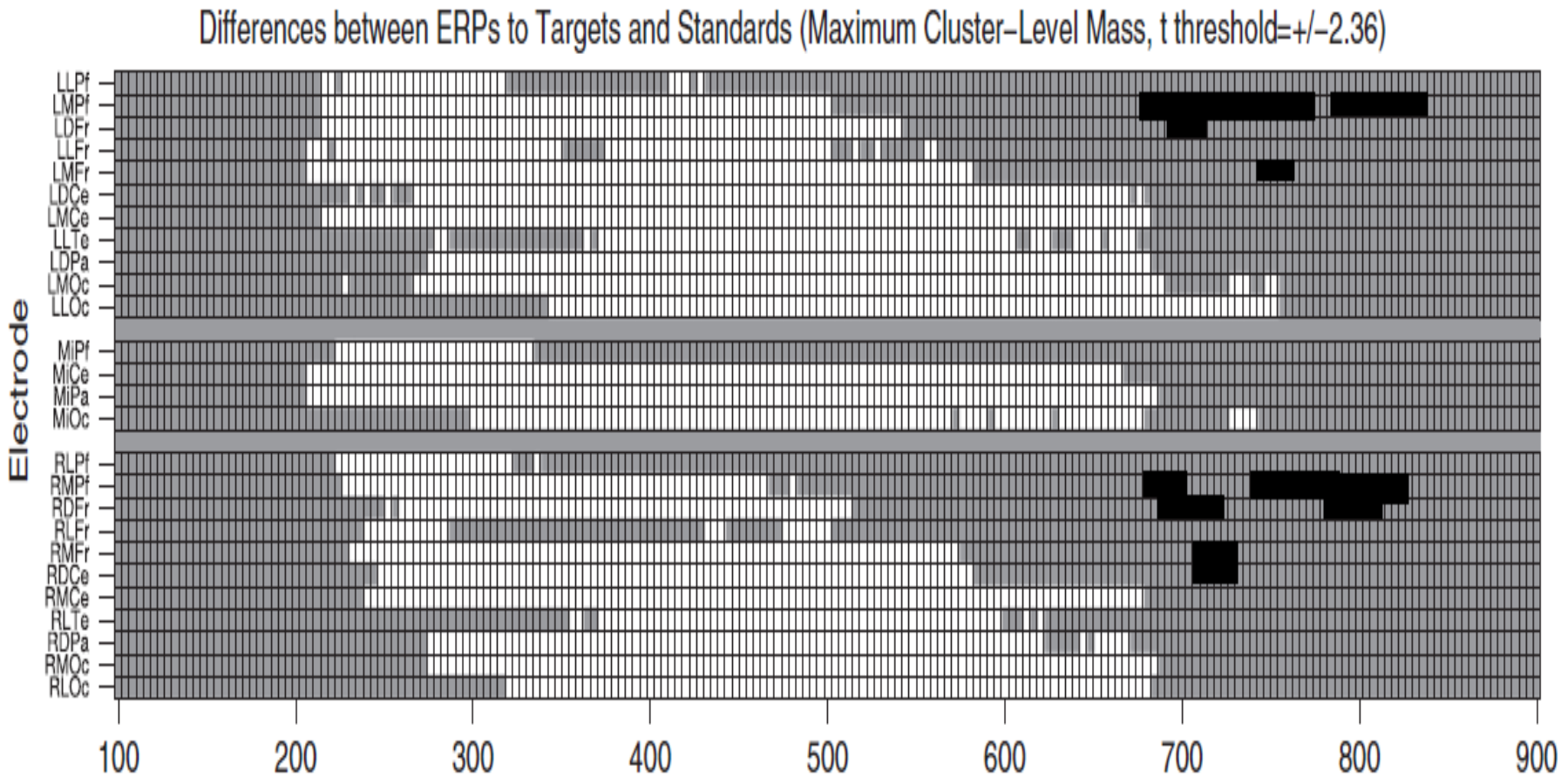
1. Подсчитаем выборочную корреляцию

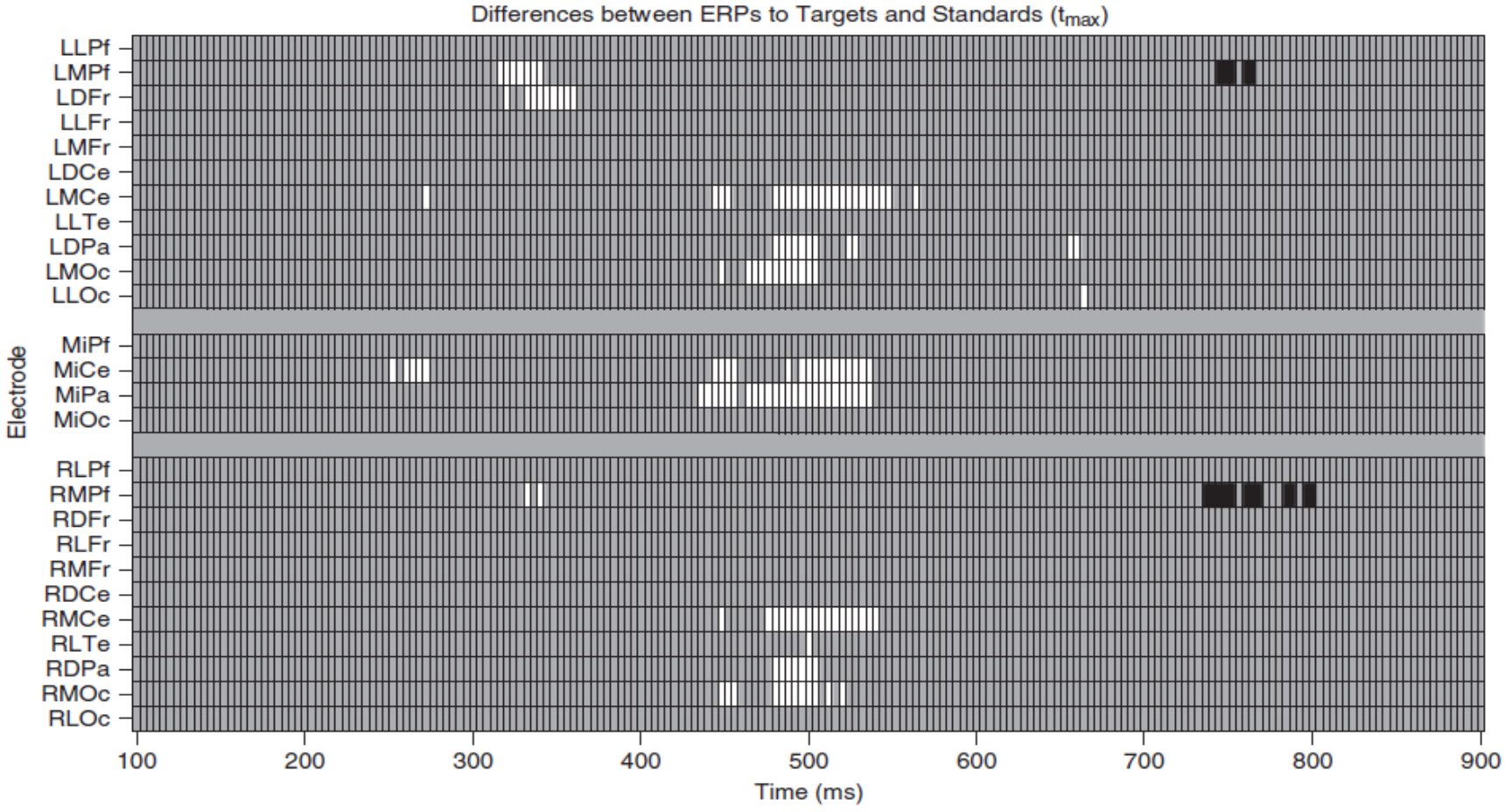
$$\widehat{r}(LSAT, GPA) = \frac{\sum_i (LSAT_i - \overline{LSAT})(GPA_i - \overline{GPA})}{\sqrt{[\sum_i (LSAT_i - \overline{LSAT})^2][\sum_i (GPA_i - \overline{GPA})^2]}} = 0.776$$

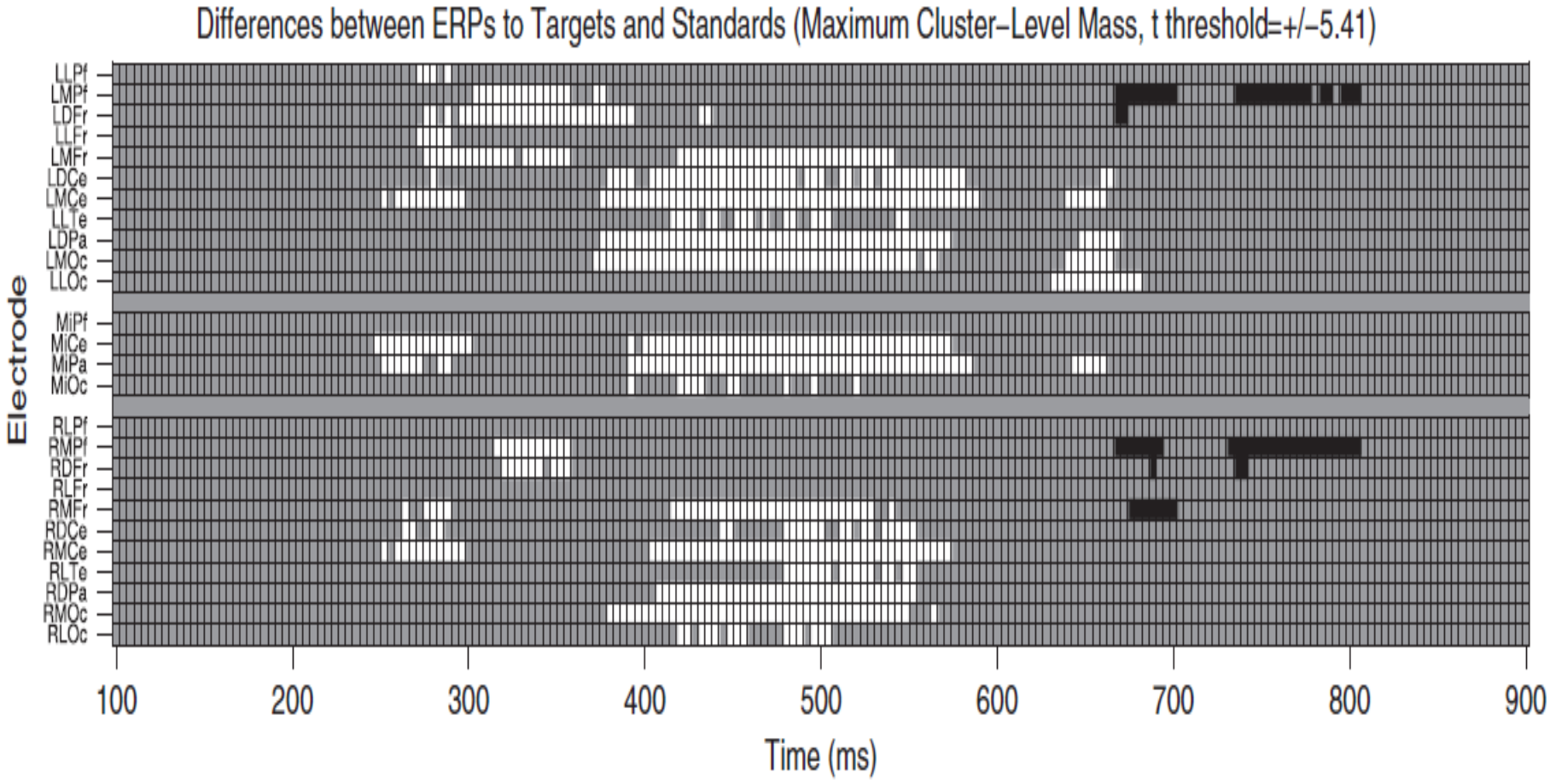
2. $\widehat{\mathrm{V}}\left(\widehat{r}(LSAT, GPA)\right) = 0.137^2$, при $N = 1000$.
3. 95% интервал для коэффициента асимметрии:
   - Нормальный интервал: (0.51, 1)
   - Интервал на основе процентилей: (0.46, 0.96)

Пример №3


LDCe
MiCe
RDCe
LLTe
LMCe
P2
RMCe
RLTe
LDPa
MiPa
RDPa
LMOc
P3
RMOc
3 μV
0 300 600 900 ms
LLOc
MiOc
RLOc
Target
Standard


Differences between ERPs to Targets and Standards (Maximum Cluster-Level Mass, t threshold=+/-2.36)
Electrode
LLPf
LMPf
LDFr
LLFr
LMFr
LDCe
LMCe
LLTe
LDPa
LMOc
LLOc
MiPf
MiCe
MiPa
MiOc
RLPf
RMPf
RDFr
RLFr
RMFr
RDCe
RMCe
RLTe
RDPa
RMOc
RLOc
100
200
300
400
500
600
700
800
900


Differences between ERPs to Targets and Standards (t_{max})
LLPf
LMPf
LDFr
LLFr
LMFr
LDCe
LMCe
LLTe
LDPa
LMOc
LLOc
MiPf
MiCe
MiPa
MiOc
RLPf
RMPf
RDFr
RLFr
RMFr
RDCe
RMCe
RLTe
RDPa
RMOc
RLOc
Electrode
100
200
300
400
500
600
700
800
900
Time (ms)


Differences between ERPs to Targets and Standards (Maximum Cluster-Level Mass, t threshold=+/-5.41)
Electrode
LLPf
LMPf
LDFr
LLFr
LMFr
LDCe
LMCe
LLTe
LDPa
LMOc
LLOc
MiPf
MiCe
MiPa
MiOc
RLPf
RMPf
RDFr
RLFr
RMFr
RDCe
RMCe
RLTe
RDPa
RMOc
RLOc
100
200
300
400
500
600
700
800
900
Time (ms)