Метод Монте-Карло по схеме марковской цепи для оценки вероятности редких событий в задачах биоинформатики

Небожатко Екатерина Павловна, гр. 16.М03-мм

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Статистическое моделирование

Научный руководитель: к.ф.-м.н., доцент Коробеников А. И.
Рецензент: программист-биостатистик Абрамова А. Н.

Санкт-Петербург
2018

Постановка задачи

Рассмотрим случайную величину ξ с распределением $\mathcal{P}$, определенную в $(\Omega, \mathcal{F}, \mathbb{P})$.

Определение

Задачей является вычисление вероятности

$$p^* = \mathbb{P}(\mathrm{Score}(\xi) \geq s),$$

где s — наперед заданное значение, Score — некоторая функция.

Интерес представляют случаи, когда p^* принимает очень маленькие значения.

Пусть $(x_1, \dots x_n)$ — реализации ξ.
Обозначим

$$\mathcal{S} = \{x \in \Omega : \text{Score}(x) \geq s\}.$$

Определение

Оценкой по методу Монте-Карло будем называть оценку вида

$$\widehat{p}_{MC} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\mathcal{S}}(x_i).$$

Дисперсия такой оценки $\mathbb{D}(\widehat{p}_{MC}) = \frac{p^*(1-p^*)}{n}$ стремится к 0 при $n \to \infty$, при этом относительная ошибка возрастает с уменьшением p^*

$$RE(\widehat{p}_{MC}) = \frac{\mathbb{D}(\widehat{p}_{MC})}{p^{*2}} = \frac{p^*(1-p^*)}{np^{*2}} = \frac{1}{np^*} - \frac{1}{n} \to \infty, \; p^* \to 0.$$

Метод Монте-Карло очень трудоемкий для оценки вероятности редких событий.

Будем использовать класс алгоритмов на основе Монте-Карло на марковских цепях. Такой подход позволяет эффективно моделировать выборку из хвостов распределений, при этом оценки имеют меньшую дисперсию.

Рассматриваются следующие алгоритмы:

- Алгоритм Ванга-Ландау (Wang et al, 2001)
- Replica Exchange (Geyer, 2005)
- Stochastic Approximation Monte Carlo (Liang et al., 2007)

Пусть $\mathcal{Q}$ — некоторое распределение с плотностью $q(x)$. Пусть $p(x)$ — плотность распределения $\mathcal{P}$.

Предположим, что существует производная Радона-Никодима $d\mathcal{P}/d\mathcal{Q}$.

Пусть $(y_1, \ldots, y_n) \sim \mathcal{Q}$.

Определение

Оценкой по методу существенной выборки будем называть оценку

$$\widehat{p}_{IS} = \frac{1}{n} \sum_{i=1}^{n} \frac{p(y_i)}{q(y_i)} \mathbb{I}_{\mathcal{S}}(y_i).$$

Если $q(x) \propto w(x)p(x)$, тогда:

$$\widehat{p}_{IS} = \frac{\sum_{i=1}^{n} \mathbb{I}_{\mathcal{S}}(y_i)/w(y_i)}{\sum_{i=1}^{n} 1/w(y_i)}.$$

При таком выборе $q(x)$ оценка зависит только от весов w, и не зависит от, вообще говоря, неизвестной плотности $p(x)$.

Для получения оценки $\widehat{p}_{IS}$ нужно уметь моделировать случайные величины из распределения $\mathcal{Q}$.
Алгоритм Метрополиса – Гастингса позволяет построить марковскую цепь со стационарным распределением $\mathcal{Q}$.

Описание алгоритма
Пусть $\zeta_i \sim \mathcal{Q}$. Выберем переходную плотность $f(\cdot|z)$ так, чтобы для нее было выполнено условие детального баланса $f(\cdot|z)q(z) = f(z|\cdot)q(\cdot)$. Для получения ζ_{i+1}:

1. Моделируем $z \sim f(\cdot|\zeta_i)$;
2. Полагаем $\zeta_{i+1} = z$ с вероятностью $\alpha(\zeta_i, z)$, иначе $\zeta_{i+1} = \zeta_i$.

При выборе $q(x) \propto w(x)p(x)$ и симметричной f вероятность $\alpha(\zeta_i, z)$ равна

$$\alpha(\zeta_i, z) = \min\left\{1, \frac{q(\zeta_i)}{q(z)}\frac{f(\zeta_i|z)}{f(z|\zeta_i)}\right\} = \min\left\{1, \frac{w(z)}{w(\zeta_i)}\right\}.$$

Алгоритм Ванга–Ландау

За счет выбора весов w можно уменьшить дисперсию оценки $\widehat{p}_{IS}$.

Если истинные значения вероятностей p известны и функция Score дискретна, оптимальные веса находятся из соотношения:

$$w(x) = w(\mathrm{Score}(x)) \propto \frac{1}{\mathbb{P}(\mathrm{Score}(x) = s)}.$$

Алгоритм Ванга–Ландау является модификацией алгоритма Метрополиса–Гастингса.

Описание алгоритма

1. Строится оценка $\widehat{w}$ методом Ванга–Ландау
2. Моделируется марковская цепь методом Метрополиса–Гастингса со стационарным распределением $q(x) \propto \widehat{w}(\mathrm{Score}(x))p(x)$

Определение

Оценкой по методу Ванга–Ландау будем называть оценку

$$\widehat{p}_{WL} = \frac{\sum_{i=1}^{n} \mathbb{I}_{\mathcal{S}}(x_i)/\widehat{w}(\mathrm{Score}(x_i))}{\sum_{i=1}^{n} 1/\widehat{w}(\mathrm{Score}(x_i))}.$$

Replica exchange выбирает веса $w_i = e^{\beta_i \cdot \mathrm{Score}(x)}$ из сетки значений.

Описание алгоритма

1. Параллельно моделируются k цепей с различными значениями параметров $(\beta_1 \dots \beta_k)$ и распределением $q_t(x) \propto e^{\beta_t \cdot \mathrm{Score}(x)} p(x)$;
2. Через r итераций выбирается пара цепей i и j для обмена состояниями. Обмен происходит с вероятностью

$$P_{swap} = \max\left\{1, \exp\left((\beta_i - \beta_j)(\mathrm{Score}(x_j) - \mathrm{Score}(x_i)\right)\right\};$$

3. Для каждой цепи строится оценка

$$\widehat{p}_{RE}^{(j)} = \frac{\sum_{i=1}^{n} h_s(x_i) \exp(-\beta_j \cdot \mathrm{Score}(x_i))}{\sum_{i=1}^{n} \exp(-\beta_j \cdot \mathrm{Score}(x_i))}.$$

Определение

Оценкой по методу replica exchange будем называть

$$\widehat{p}_{RE} = \frac{1}{k} \sum_{j=1}^{k} \widehat{p}_{RE}^{(j)}.$$

Разобьем выборочное пространство на l областей:
$E_1 = \{x : \text{Score}(x) \leq s_1\}, E_2 = \{x : s_1 < \text{Score}(x) \leq s_2\}, \ldots,$
$E_l = \{x : \text{Score}(x) > s_l\}.$

Пусть $\psi(x) \geq 0$ — некоторая функция и $g_i = \int_{E_i} \psi(x)dx.$

SAMC моделирует выборку из распределения

$$p_g(x) \propto \sum_{i=1}^{l} \frac{\pi_i \psi(x)}{g_i} \mathbb{I}(x \in E_i),$$

где $\pi_i > 0$ и $\sum_{i=1}^{l} \pi_i = 1.$

Так как истинные значения g_i не известны, алгоритм оценивает их итеративно при некотором заданном значении $\pi = (\pi_1, \ldots, \pi_l)$.

Пусть $\theta_t^{(i)}$ обозначает оценку $\log(g_i/\pi_i)$, полученную на итерации t.

Описание алгоритма
Пусть на шаге t получили значения x_t и $\theta_t = (\theta_t^{(1)}, \ldots, \theta_t^{(l)})$.

1. Моделируем элемент марковской цепи x_{t+1} с помощью алгоритма Метрополиса – Гастингса со стационарным распределением $q_t(x) \propto \sum_{i=1}^{l} \psi(x)/\exp(\theta_t^{(i)})\mathbb{I}(x \in E_i)$ и равномерной переходной плотностью;
2. Обновляем параметры

$$\theta_{t+1}^{(i)} = \theta_t^{(i)} + \gamma_{t+1}(\mathbb{I}(x_{t+1} \in E_i) - \pi_i),\ \gamma_t - \text{параметр метода.}$$

Обозначим $\widehat{\nu}_t = \sum_{j \notin \mathcal{S}_t} \pi_j / |\mathcal{S}_t|$ и $\mathcal{S}_t$ обозначает множество областей, которые были посещены во время моделирования марковской цепи.

Теорема (Liang et al., 2007)

Если положить $\psi(x) \propto 1$, тогда g_i — мощность множества E_i, и оценка p^* определяется как

$$\widehat{p}_{SAMC_t} = \frac{\sum_{i=k+1}^{l} \exp(\theta_t^{(i)})(\pi_i + \widehat{\nu}_t)}{\sum_{j=1}^{l} \exp(\theta_t^{(j)})(\pi_j + \widehat{\nu}_t)},$$

и сходится (в среднем) к $p = \mathbb{P}(\mathrm{Score}(\xi) > s_k)$ при $t \to \infty$. Здесь $s_k = s$.

Рассмотрим конкретный случай.

Пусть теперь ξ — случайная строка длины m над Σ^m с равномерным распределением. $|\Sigma| = 4$. Рассмотрим также фиксированную последовательность $v \in \Sigma^m$.

Пусть $\mathrm{Score}(\xi, v) = \mathrm{Score}(\xi)$ — мера похожести двух строк.

Задача

Задачей является вычисление вероятности

$$p = \mathbb{P}(\mathrm{Score}(\xi) = m)$$

Такая постановка задачи позволяет вычислить истинные значения вероятностей:

$$p = 1/4^m.$$

Схема проведения экспериментов

Были построены:

- истинные значения оценок p^*,
- оценки по методу Монте-Карло $\widehat{p}_{MC}$,
- оценки по методу Ванга–Ланду $\widehat{p}_{WL}$,
- оценки по методу replica exchange $\widehat{p}_{RE}$,
- оценки по методу stochastic approximation monte carlo $\widehat{p}_{SAMC}$.

Все оценки были получены для сравнения строк длин 5, 6, 7 и 8. $n = 10^7$. Также для всех оценок были сосчитаны оценки дисперсий по методу batch means (Flegal et al, 2010) и построены 95% доверительные интервалы.

Оценки и их доверительные интервалы

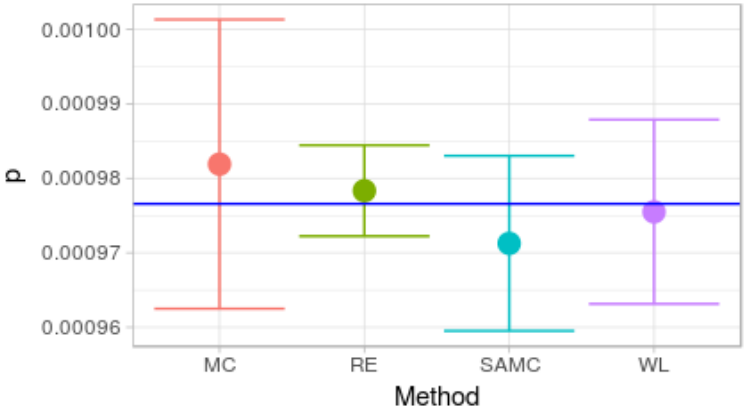

Рис. : Длина 5

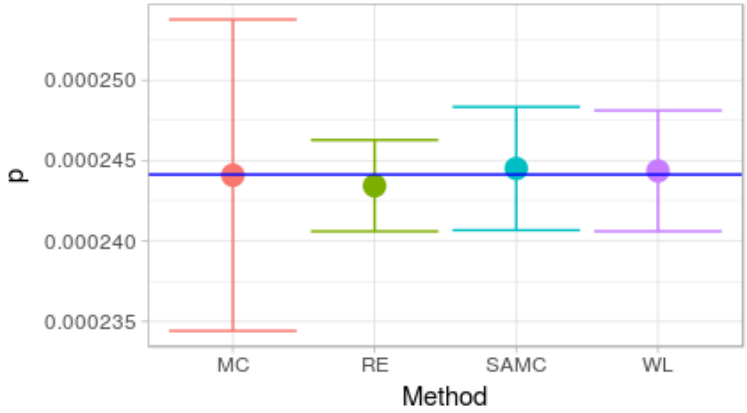

Рис. : Длина 6

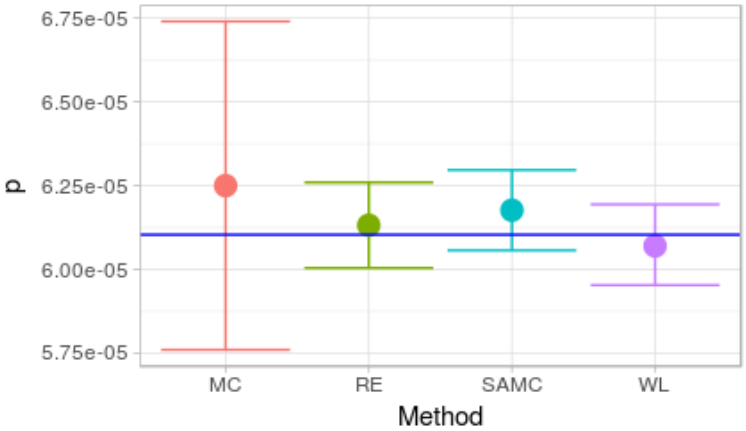

Рис. : Длина 7

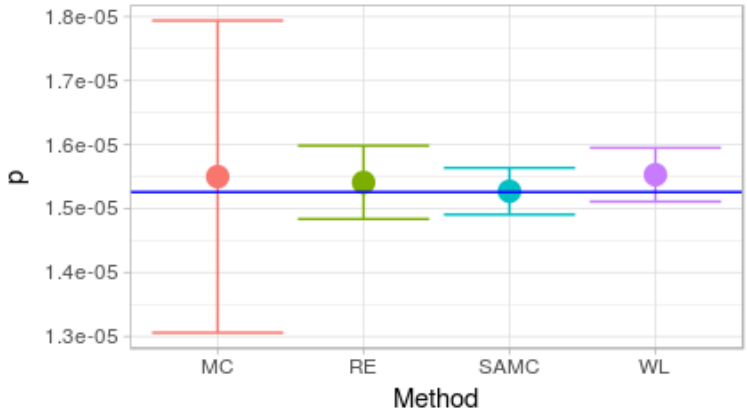

Рис. : Длина 8

Численные результаты: сравнение относительных ошибок

Score	p^*	$\mathrm{re}(\hat{p}_{MC})$	$\mathrm{re}(\hat{p}_{WL})$	$\mathrm{re}(\hat{p}_{RE})$	$\mathrm{re}(\hat{p}_{SAMC})$
5	$9.8 \cdot 10^{-4}$	0.32	0.13	0.03	0.12
6	$2.4 \cdot 10^{-4}$	1.29	0.19	0.11	0.20
7	$6.1 \cdot 10^{-5}$	4.73	0.32	0.35	0.31
8	$1.5 \cdot 10^{-5}$	20.4	0.60	1.13	0.47

- Относительная ошибка всех трех методов значительно меньше, чем у Монте-Карло. И эта разница тем больше, чем меньше значение оцениваемой вероятности.
- Эмпирически было показано, что все оценки лежат в доверительных границах оценок по методу Монте-Карло.

Сравнение пептидов

Пептид P с массой M состоит из k аминокислот. Обозначим

- $\mu = (\mu_1, \ldots, \mu_k)$ — вектор масс аминокислот
- $H_{r \times k}$ — *фрагментационная матрица*
- $\mathcal{M} = \{\mu = (\mu_1, \ldots, \mu_k) | \mu_i > 0, \sum_{i=1}^{k} \mu_i = M\}$

Можно описать множество пептидов с одинаковой массой и химической структурой множеством $\mathcal{M}$ и матрицей H.

Пептид P можно представить в виде вектора

$$TheoreticalSpectrum(P) = H\mu,$$

который называется *теоретическим спектром*.

Пусть S — некоторый фиксированный спектр пептида.

Введем функцию Score, которая измеряет «близость» двух спектров (расстояние между векторами)

$$\mathrm{Score}(\mu) = \mathrm{Score}(S, H\mu).$$

Задача

Нашей задачей является вычисление вероятности

$$p^* = \mathbb{P}(\mathrm{Score}(S, H\mu) \geq s)) = \mathbb{P}(\mathrm{Score}(\mu) \geq s),$$

где μ — равномерно распределенная случайная величина на множестве $\mathcal{M}$ и s — заранее заданный порог.

Схема проведения экспериментов

Были построены:

- оценки по методу Монте-Карло $\widehat{p}_{MC}$,
- оценки по методу Ванга–Ланду $\widehat{p}_{WL}$,
- оценки по методу replica exchange $\widehat{p}_{RE}$,
- оценки по методу stochastic approximation monte carlo $\widehat{p}_{SAMC}$.

Все оценки были получены для четырех пептидов. $n = 10^7$. Также для всех оценок были сосчитаны оценки дисперсий по методу batch means и построены 95% доверительные интервалы.

Оценки и их доверительные интервалы

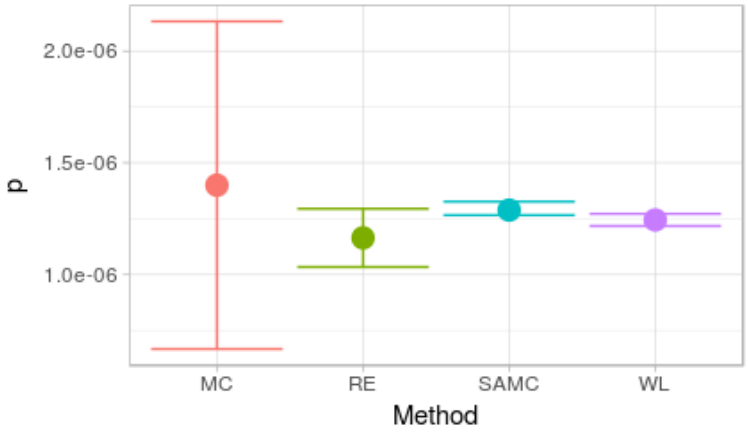

Рис. : PPAEDSQK

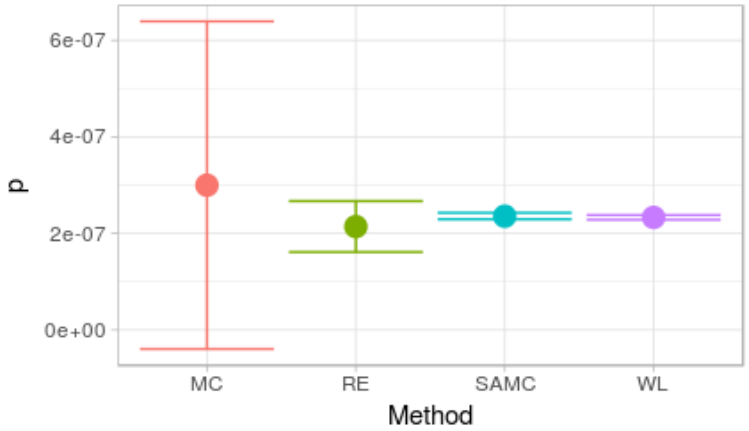

Рис. : ATAAGSEDAEK

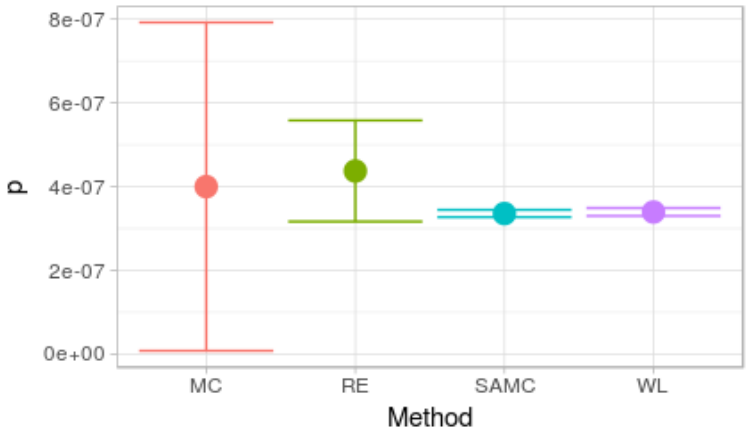

Рис. : GEEEPSQGQK

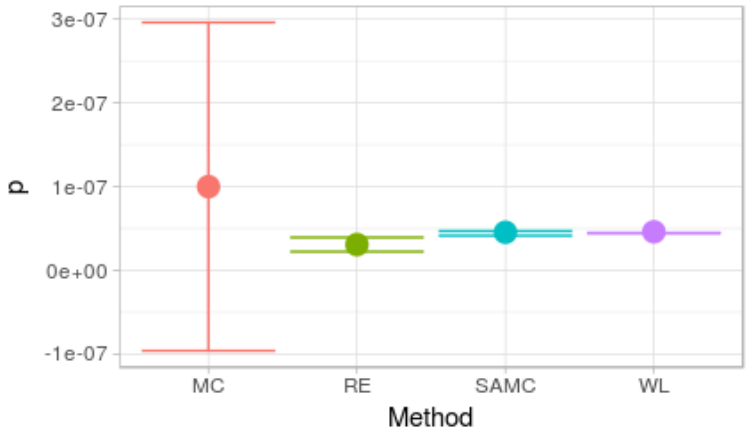

Рис. : PTTNPSAGK

Численные результаты: сравнение относительных ошибок

Пептид	$\widehat{p}_{WL}$	$\mathrm{re}(\hat{p}_{MC})$	$\mathrm{re}(\hat{p}_{WL})$	$\mathrm{re}(\hat{p}_{RE})$	$\mathrm{re}(\hat{p}_{SAMC})$
PPAEDSQK	$1.2 \cdot 10^{-6}$	225.87	0.39	10.32	0.49
ATAAGSEDAEK	$2.3 \cdot 10^{-7}$	1054.09	0.35	50.08	0.55
GEEEPSQGQK	$3.4 \cdot 10^{-7}$	790.5	0.63	62.7	0.54
PTTNPSAGK	$4.6 \cdot 10^{-8}$	3162.27	0.95	61.98	0.81

- Чем меньше вероятность, тем больше отношение относительных ошибок Монте-Карло и остальных методов.
- Относительная ошибка для RE значительно возрастает с уменьшением вероятности.
- Методы WL и SAMC показывают примерно одинаковые результаты.
- Эмпирически показано, что все оценки лежат в доверительных границах оценок по методу Монте-Карло.

Выводы

- Были рассмотрены способы оценки статистической значимости меры похожести строк и пептидных спектров.
- Эмпирически показано, что оценки для каждого примера, полученные с помощью рассмотренных алгоритмов, лежат в границах доверительных интервалов оценок Монте-Карло и имеют меньшую дисперсию.
- Наименьшая дисперсия была достигнута для оценок вероятностей по методам Ванга-Ландау и stochastic approximation Monte Carlo.