

Санкт-Петербургский государственный политехнический университет

Лабораторная работа № 2

по курсу «Стохастические модели»

«Оценка параметров модели случайного процесса нагрузки сервера,
обрабатывающего заявки»

Студент:	Руцкий В. В.
Группа:	5057/2
Преподаватель:	Иванков А. А.

Санкт-Петербург 2011

Содержание

1	Постановка задачи	2
2	Решение в случае бесконечного времени обработки заявки	2
2.1	Итеративный метод	2
2.1.1	Идентификация моментов времени поступления заявок T_c	2
2.1.2	Оценка интенсивности поступления заявок λ	3
2.1.3	Оценка параметров фоновой нагрузки m и σ	3
2.1.4	Оценка параметров увеличения уровня загрузки ресурсов сервера от заявок m_c и σ_c	3
2.2	Оценивание ЕМ-алгоритмом	4
2.2.1	ЕМ-алгоритм	4
2.2.2	Вычисление θ_0	5
2.2.3	Получение искомых оценок	5
3	Решение в случае конечного времени обработки заявки	5
3.1	Идентификация моментов времени поступления заявок	6
3.2	Оценка параметров увеличения уровня загрузки ресурсов сервера от заявок m_c и σ_c	6
3.3	Оценка параметров фоновой нагрузки m и σ	7
4	Результаты работы	7
4.1	Случай бесконечного времени обработки заявки	7

1 Постановка задачи

В данной работе производится анализ лога использования ресурсов сервера при поступающих заявках на обработку.

В отсутствие заявок уровень использования ресурсов сервера представляет собой сумму некоторой постоянной величины загрузки m и случайных отклонений:

$$B(t) = m + \sigma W(t),$$

где $W(t)$ — это винеровский процесс.

Заявки поступают в соответствии с законом распределения Пуассона $\mathcal{P}(\lambda)$.

При поступлении одной заявки использование ресурсов мгновенно возрастает, а затем экспоненциально снижается до прежнего уровня. Увеличение использования ресурсов сервера от одной заявки, поступившей в момент времени t_c , выражается следующим образом:

$$K_{t_c}(t) = \mathcal{N}(m_c, \sigma_c^2) \cdot I(t - t_c) \cdot e^{-\lambda_c(t - t_c)},$$

где $I(x)$ — функция Хевисайда.¹

В логе использования ресурсов сервера наблюдается общая загрузка сервера:

$$X(t) = B(t) + \sum_{t_c \in T_c} K_{t_c}(t),$$

где T_c — это моменты времени поступления заявок.

Необходимо по дискретным наблюдениям x_i случайного процесса $X(t)$ в моменты времени t_i , $i = 1, \dots, N$

1. идентифицировать моменты времени поступления заявок T_c ,
2. оценить параметры модели m , σ^2 , λ , m_c , σ_c^2 , λ_c .

Наблюдения производятся через равные промежутки времени $\Delta t = t_{i+1} - t_i$.

2 Решение в случае бесконечного времени обработки заявки

Рассмотрим случай, когда $\lambda_c \rightarrow 0$, т. е. при поступлении заявка увеличивает уровень загрузки сервера на постоянную величину и ресурсы, выделенные на обработку заявки никогда не освобождаются.

2.1 Итеративный метод

2.1.1 Идентификация моментов времени поступления заявок T_c

Предположим, что в отрезке времени $[t_k, t_{k+n}]$ не поступило ни одной заявки. Тогда $n + 1$ наблюдений x_k, \dots, x_{k+n} представляют собой наблюдения $B(t)$. Оценим по этим наблюдениям параметры $B(t)$.

Рассмотрим разности соседних наблюдений — они представляют собой наблюдения нормально распределённой случайной величины:

$$B(t_{i+1}) - B(t_i) = \sigma W(t_{i+1}) - \sigma W(t_i) = \sigma \mathcal{N}(0, \Delta t) = \mathcal{N}(0, \sigma^2 \Delta t).$$

Построим точечную оценку $\hat{\sigma}^2$ методом максимального правдоподобия:²

$$\hat{\sigma}^2 = \frac{1}{\Delta t} \cdot \frac{1}{n-1} \sum_{i=1}^n ((x_{k+i} - x_{k+i-1}) - 0)^2.$$

¹Функция Хевисайда: $I(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases}$.

²См. § 3.5 пункт 1 в [1].

Обозначим гипотезу о том, что в промежутке времени $[t_{k+n}, t_{k+n+1}]$ не поступило ни одной заявки, как H_0 . Тогда

$$(X(t_{k+n+1}) - X(t_{k+n}) | H_0) = \mathcal{N}(0, \sigma^2 \Delta t).$$

В качестве критерия принятия гипотезы H_0 с уровнем значимости α возьмём условие, что разность значений наблюдений $(x_{k+n+1} - x_{k+n})$ лежит в $(1 - \alpha)$ квантиле нормального распределения $\mathcal{N}(0, \hat{\sigma}^2 \Delta t)$, обозначенного как $\mathcal{N}_{1-\alpha}$:

$$H_0 \text{ принимается} \iff (x_{k+n+1} - x_{k+n}) < \mathcal{N}_{1-\alpha}$$

(рассматривается только правый квантиль нормального распределения, т.к. заявка может дать только положительное увеличение уровня нагрузки сервера).

Алгоритм нахождения моментов времени поступления заявок T_c состоит в следующем:

1. В предположении, что в первые $n + 1$ наблюдений не поступило ни одной заявки, оценим $\hat{\sigma}$ и построим критерий для принятия H_0 .
2. Будем добавлять к первым $n+1$ наблюдениям по одному наблюдению и проверять гипотезу H_0 . Если H_0 принимается, то $\hat{\sigma}$ и критерий для принятия H_0 пересчитываются для добавленного наблюдения.
3. Как только встретится наблюдение $n + 1 + l$, для которого гипотеза H_0 отвергается, то $t_{n+1+l} \in \hat{T}_c$. Все наблюдения до $t_{n+1+l+1}$ отбрасываются и алгоритм начинается с шага 1 для поиска следующего момента времени поступления заявки.

2.1.2 Оценка интенсивности поступления заявок λ

Зная оценку времени прибытия заявок \hat{T}_c интенсивность поступления заявок можно оценить методом максимального правдоподобия:³

$$\hat{\lambda} = \frac{1}{|\hat{T}_c|} \sum_{i=1}^{|\hat{T}_c|} (t_{c_{i+1}} - t_{c_i}).$$

2.1.3 Оценка параметров фоновой нагрузки m и σ

Оценку m и σ произведём по наблюдениям уровня загруженности сервера до поступления первой заявки t_{c_1} , т.к. дальше изменение уровня фоновой нагрузки сервера сравнимо с дисперсией уровня увеличения нагрузки от поступления заявки:

$$\sigma \approx \sigma_c.$$

Оценку произведём методом максимального правдоподобия для нормального распределения:⁴

$$\hat{m} = \frac{1}{K-1} \sum_{i=1}^{K-1} x_i, \quad \hat{\sigma}^2 = \frac{1}{K-2} \sum_{i=1}^{K-1} (x_i - \hat{m})^2,$$

где $K = t_{c_1} / \Delta t$ — номер наблюдения x_i , когда, согласно приведённой выше оценке, поступила первая заявка.

2.1.4 Оценка параметров увеличения уровня загрузки ресурсов сервера от заявок m_c и σ_c

Рассмотрим ненормированный разностный аналог производной случайного процесса $X(t)$:

$$dX(t) = X(t) - X(t - \Delta t).$$

³http://en.wikipedia.org/wiki/Poisson_distribution#Maximum_likelihood или в общем случае в § 3.5 пункт 1 в [1].

⁴http://en.wikipedia.org/wiki/Normal_distribution#Estimation_of_parameters.

$dX(t)$ в момент времени поступления заявки t_c выражается следующим образом:

$$\begin{aligned} dX(t_c) &= X(t_c) - X(t_c - \Delta t) = B(t_c) + K_{t_c}(t_c) - B(t_c - \Delta t) = \\ &= \sigma W(t_c) - \sigma W(t_c - \Delta t) + \mathcal{N}(m_c, \sigma_c^2) = \\ &= \sigma \mathcal{N}(0, \Delta t) + \mathcal{N}(m_c, \sigma_c^2) = \mathcal{N}(m_c, \sigma^2 \Delta t + \sigma_c^2), \end{aligned}$$

предполагая, что в момент времени $t_c - \Delta t$ заявки не было.

Наблюдения $dX(t)$ в моменты времени T_c соответствуют наблюдениям нормально распределённой случайной величины $\mathcal{N}(m_c, \sigma^2 \Delta t + \sigma_c^2)$ — оценим по этим наблюдениям параметры m_c и σ_c^2 методом максимального правдоподобия аналогично оценкам в пункте 2.1.3.

2.2 Оценивание ЕМ-алгоритмом

В пункте 2.1.4 было показано, что $dX(t)$ в момент времени поступления заявки t_c выражается как:

$$dX(t_c) = \mathcal{N}(m_c, \sigma^2 \Delta t + \sigma_c^2),$$

предполагая, что в момент времени $t_c - \Delta t$ заявки не было.

Во время отсутствия заявок $dX(t)$ выражается как:

$$dX(t) = X(t) - X(t - \Delta t) = B(t) - B(t - \Delta t) = \mathcal{N}(0, \sigma^2 \Delta t).$$

Значит в каждый отдельно взятый момент времени t случайная величина $dX(t)$ представляет собой смесь двух нормально распределённых случайных величин, причем параметры случайных величин со временем не меняются. Оценим их параметры ЕМ-алгоритмом (на основе примера из [2]).

Введём скрытые случайные величины Z_i , $i = 1, \dots, N$, принимающие значения 1 или 2, в зависимости от того, поступила ли заявка в момент времени t_i или нет соответственно, а z_i — наблюдения Z_i в момент времени t_i .

$$\begin{aligned} dX(t_i) | (Z_i = 1) &\sim \mathcal{N}(\mu_1, \sigma_1^2) = \mathcal{N}(m_c, \sigma^2 \Delta t + \sigma_c^2), & (\text{случай } t_i \in T_c), \\ dX(t_i) | (Z_i = 2) &\sim \mathcal{N}(\mu_2, \sigma_2^2) = \mathcal{N}(0, \sigma^2 \Delta t), & (\text{случай } t_i \notin T_c). \end{aligned}$$

Случайная величина Z_i распределена по закону Пуассона $\mathcal{P}(\lambda)$, её можно аппроксимировать Биномиальным законом $\mathcal{B}(N, p)$:

$$\mathcal{P}(\lambda) = \mathcal{P}(Np) = \mathcal{B}(N, p),$$

т. к. N велико, $\lambda = Np$ невелико, а значит p мало (см. §1.1 пункт 3 в [1]).

Пусть $\mathbf{P}(Z_i = 1) = \tau_1$ и $\mathbf{P}(Z_i = 2) = \tau_2 = 1 - \tau_1$.

Введём обозначения: $\theta = (\tau_1, \tau_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$, $\mathbf{x} = (x_1, \dots, x_N)$, $\mathbf{z} = (z_1, \dots, z_N)$.

Построим функцию правдоподобия:

$$L(\theta; \mathbf{x}, \mathbf{z}) = \mathbf{P}(\mathbf{x}, \mathbf{z} | \theta) = \prod_{i=1}^N \sum_{j=1}^2 \mathbb{I}(z_i = j) \tau_j f(x_i, \mu_j, \sigma_j^2),$$

где $\mathbb{I}(\text{expr})$ — функция индикатор,⁵ а $f(x, \mu, \sigma^2)$ — это функция плотности распределения, в данном случае нормального:

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Перепишем функцию правдоподобия в экспоненциальной форме:

$$L(\theta; \mathbf{x}, \mathbf{z}) = \exp \left\{ \sum_{i=1}^N \sum_{j=1}^2 \mathbb{I}(z_i = j) \left[\log \tau_j - \frac{1}{2} \log(2\pi) - \log(\sigma_j) - \frac{(x_i - \mu_j)^2}{2\sigma_j^2} \right] \right\}.$$

2.2.1 ЕМ-алгоритм

Пусть имеется начальная оценка параметров θ : $\theta^{(0)}$ (использованный способ вычисления $\theta^{(0)}$ описан в пункте 2.2.2). Последовательно выполняя Е- и М-шаги будем уточнять оценку $\theta^{(k)}$, пока она не сойдётся: $\theta^{(k)} \xrightarrow[k \rightarrow \infty]{} \hat{\theta}$. Вектор $\hat{\theta}$ примем за результат оценки.

⁵Функция индикатор: $\mathbb{I}(\text{expr}) = \begin{cases} 0, & \text{expr} = \text{False} \\ 1, & \text{expr} = \text{True} \end{cases}$.

Е-шаг Имея текущую оценку параметров $\theta^{(k)}$, вычислим по теореме Байеса условную вероятность принадлежности i -го наблюдения j -му нормальному распределению:

$$T_{j,i}^{(k)} = \mathbf{P} \left(Z_i = j \mid dX(t_i) = x_i; \theta^{(k)} \right) = \frac{\tau_j^{(k)} f(x_i; \mu_j^{(k)}, \sigma_j^{(k)})}{\tau_1^{(k)} f(x_i; \mu_1^{(k)}, \sigma_1^{(k)}) + \tau_2^{(k)} f(x_i; \mu_2^{(k)}, \sigma_2^{(k)})}.$$

Построим функцию — математическое ожидание логарифма функции правдоподобия:

$$Q \left(\theta \mid \theta^{(k)} \right) = \mathbf{E} [\log L(\theta; \mathbf{x}, \mathbf{z})] = \sum_{i=1}^N \sum_{j=1}^2 T_{j,i}^{(k)} \left[\log \tau_j - \frac{1}{2} \log(2\pi) - \log(\sigma_j) - \frac{(x_i - \mu_j)^2}{2\sigma_j^2} \right].$$

М-шаг Теперь найдём параметры $\theta^{(k+1)}$ максимизирующие $Q \left(\theta \mid \theta^{(k)} \right)$:

$$\theta^{(k+1)} = \underset{\theta}{\operatorname{argmax}} Q \left(\theta \mid \theta^{(k)} \right).$$

В соответствии с вычислениями в [2]:

$$\tau_j^{(k+1)} = \frac{1}{n} \sum_{i=1}^N T_{j,i}^{(k)}, \quad \mu_j^{(k+1)} = \frac{\sum_{i=1}^N T_{j,i}^{(k)} x_i}{\sum_{i=1}^N T_{j,i}^{(k)}}, \quad \sigma_j^{(k+1)} = \frac{\sum_{i=1}^N T_{j,i}^{(k)} (x_i - \mu_j^{(k+1)})^2}{\sum_{i=1}^N T_{j,i}^{(k)}}.$$

2.2.2 Вычисление θ_0

В качестве начальных значений τ возьмём равные вероятности:

$$\tau_j^{(0)} = 0.5, \quad j = 1, 2.$$

Для вычисления $\mu_j^{(0)}$ построим полигон частот⁶ dx_i : в качестве $\mu_1^{(0)}$ возьмём последний локальный максимум частот, а в качестве $\mu_2^{(0)}$ — первый (т.к. $\mathbf{E} [B(t) - B(t - \Delta t)] = 0$, а $\mathbf{E} [B(t_c) + K_{t_c}(t_c) - B(t - \Delta t)] = m_c > 0$).

В качестве $\sigma_j^{(0)}$ возьмём $\frac{1}{3} \left(\mu_1^{(0)} - \mu_2^{(0)} \right)$, $j = 1, 2$.

2.2.3 Получение искомых оценок

Из оценки θ несложно найти искомые оценки σ^2 , m_c , σ_c :

$$\widehat{m}_c = \mu_1, \quad \widehat{\sigma}^2 = \frac{\sigma_2^2}{\Delta t}, \quad \widehat{\sigma}_c^2 = \sigma_1^2 - \sigma_2^2.$$

Из $T_{j,i}^{(k)}$, полученного на последнем шаге ЕМ-алгоритма, можно идентифицировать моменты времени поступления заявок:

$$\widehat{T}_c = \left\{ t_i \mid i = 1, \dots, N \mid T_{1,i}^{(k)} > T_{2,i}^{(k)} \right\}.$$

Оценку m можно провести по наблюдениям $X(t)$ до момента времени поступления первой заявки аналогично тому, как это было сделано в пункте 2.1.3.

Оценку λ можно получить так же, как было сделано в пункте 2.1.2.

3 Решение в случае конечного времени обработки заявки

Рассмотрим случай, когда интенсивность освобождения ресурсов λ_c существенно больше нуля и, сервер успевает обрабатывать заявки: $\exists \sup \{X(t)\}$.

⁶См. § 2.1 пункт 4 в [1].

$dX(t)$ в момент времени, когда не поступило ни одной заявки $t \notin T_c$, выражается следующим образом:

$$\begin{aligned} dX(t) &= X(t) - X(t - \Delta t) = \\ &= \left(B(t) + \sum_{t_c \in T_c, t_c \leq t} K_{t_c}(t) \right) - \left(B(t - \Delta t) + \sum_{t_c \in T_c, t_c \leq (t - \Delta t)} K_{t_c}(t - \Delta t) \right). \end{aligned}$$

Будем считать поступление заявок достаточно редким: таким, что на уровень загрузки ресурсов сервера существенно влияет лишь последняя пришедшая заявка в момент времени $t_c < t$:

$$\begin{aligned} dX(t) &= X(t) - X(t - \Delta t) = \\ &= (B(t) + K_{t_c}(t)) - (B(t - \Delta t) + K_{t_c}(t - \Delta t)) = \\ &= (\sigma W(t) - \sigma W(t - \Delta t)) + \left(\mathcal{N}(m_c, \sigma_c^2) \cdot e^{-\lambda_c(t-t_c)} - \mathcal{N}(m_c, \sigma_c^2) \cdot e^{-\lambda_c(t-\Delta t-t_c)} \right) = \\ &= \mathcal{N}(0, \sigma^2 \Delta t) + \mathcal{N}(m_c, \sigma_c^2) \cdot \left(e^{-\lambda_c(t-t_c)} - e^{-\lambda_c(t-\Delta t-t_c)} \right). \end{aligned}$$

В условиях предположения о влиянии только последней заявки, в момент поступления заявки t_c :

$$\begin{aligned} dX(t_c) &= X(t_c) - X(t_c - \Delta t) = \\ &= (B(t_c) + K_{t_c}(t_c)) - (B(t_c - \Delta t) + 0) = \\ &= \mathcal{N}(0, \sigma^2 \Delta t) + \mathcal{N}(m_c, \sigma_c^2) = \\ &= \mathcal{N}(m_c, \sigma^2 \Delta t + \sigma_c^2). \end{aligned}$$

3.1 Идентификация моментов времени поступления заявок

Построим вариационный ряд наблюдений $dX(t_i)$:

$$dx_{(1)} \leq dx_{(2)} \leq \dots \leq dx_{(N)}.$$

Поступающие заявки вносят существенно большее изменение уровня загрузки ресурсов сервера, чем фоновая нагрузка, поэтому в правой части этого ряда будут находиться наблюдения $dX(t_c)$. Воспользуемся итеративным методом из пункта 2.1.1 для выделения наблюдений поступления заявок из вариационного ряда — получим оценку множества моментов времени поступления заявок \hat{T}_c .

3.2 Оценка параметров увеличения уровня загрузки ресурсов сервера от заявок m_c и σ_c

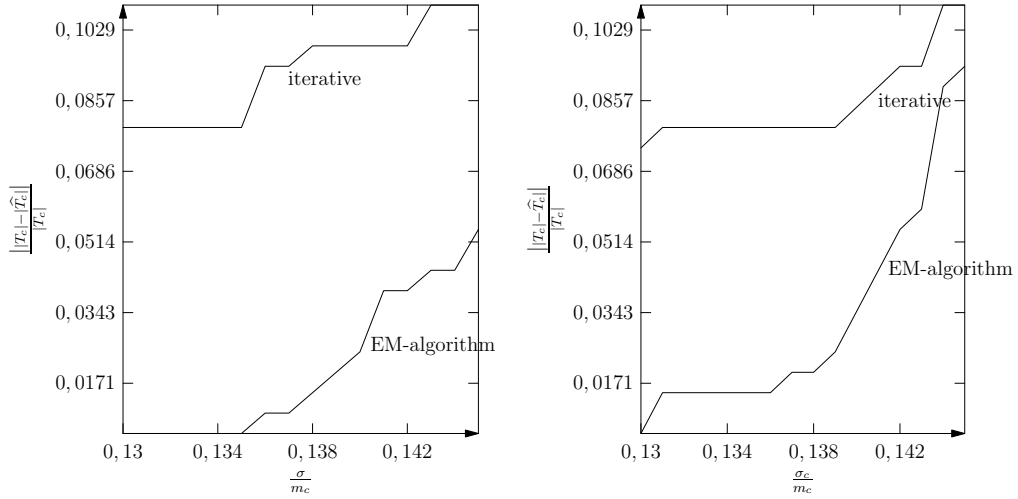
Рассмотрим промежуток времени между поступлениями заявок $[t_k, t_{k+n}]$. На нём $dX(t)$ выражается следующим образом:

$$\begin{aligned} dX(t) &= \mathcal{N}(0, \sigma^2 \Delta t) + \mathcal{N}(m_c, \sigma_c^2) \cdot \left(e^{-\lambda_c(t-t_c)} - e^{-\lambda_c(t-\Delta t-t_c)} \right) \\ &= \left(\mathcal{N}(0, \sigma^2 \Delta t) + \mathcal{N}(0, \sigma_c^2) \cdot \left(e^{-\lambda_c(t-t_c)} - e^{-\lambda_c(t-\Delta t-t_c)} \right) \right) + m_c \cdot \left(e^{-\lambda_c(t-t_c)} - e^{-\lambda_c(t-\Delta t-t_c)} \right). \end{aligned}$$

Оценим m_c и λ_c методом наименьших квадратов. Для этого необходимо минимизировать сумму квадратов отклонений:

$$\begin{aligned} (m_c, \lambda_c) &= \underset{m_c, \lambda_c}{\operatorname{argmin}} S(m_c, \lambda_c) = \\ &= \underset{m_c, \lambda_c}{\operatorname{argmin}} \sum_{i=1}^n \left(dx_i - m_c \cdot \left(e^{-\lambda_c(t_i-t_k)} - e^{-\lambda_c(t_i-\Delta t-t_k)} \right) \right)^2. \end{aligned}$$

$S(m_c, \lambda_c)$ выпукла вниз на интересующей для анализа области, поэтому для нахождения аргументов, обращающих её в минимум, можно воспользоваться численными методами минимизации, в данной работе был использован метод покоординатного спуска.



(a) Зависимость относительной погрешности оценивания $|T_c|$ от σ . (b) Зависимость относительной погрешности оценивания $|T_c|$ от σ_c .

Рис. 1: Погрешности оценивания итеративным методом и EM-алгоритмом.

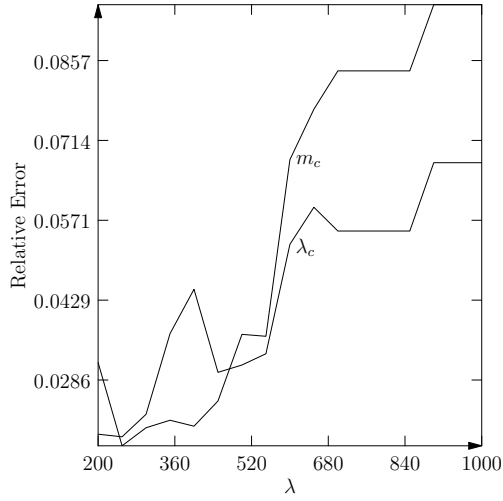
3.3 Оценка параметров фоновой нагрузки m и σ

Оценку m и σ , при найденной оценке \hat{T}_c , можно получить так же, как было сделано в пункте 2.1.3.

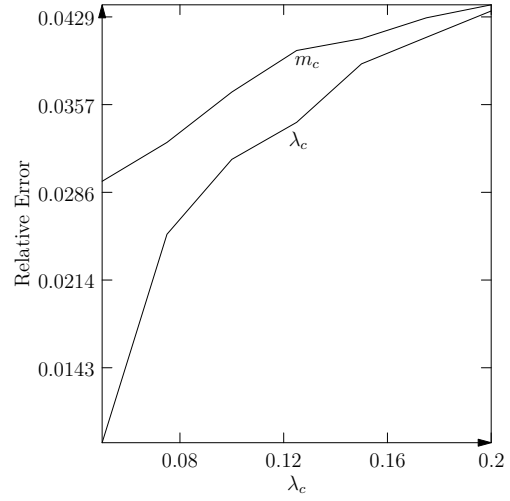
4 Результаты работы

4.1 Случай бесконечного времени обработки заявки

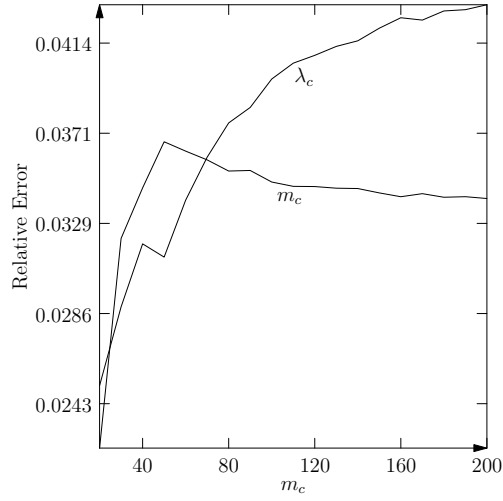
Графики зависимости относительной погрешности полученных оценок числа заявок $|T_c|$ в пунктах 2.1 и 2.2 от среднеквадратичных отклонений данных σ и σ_c приведены на рис. 1. На рис. 2 приведены графики зависимости относительной погрешности полученных оценок методом наименьших квадратов для случая $\lambda_c \gg 0$.



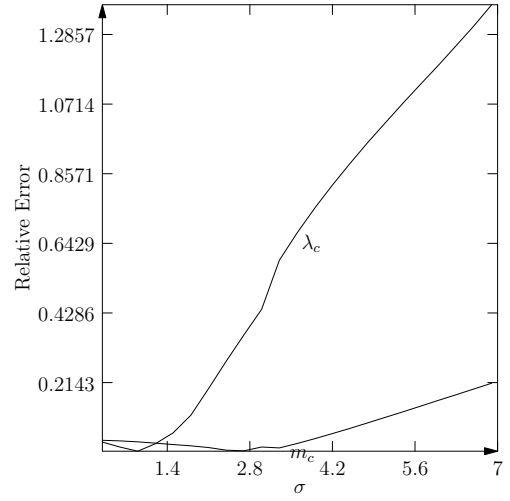
(a) Зависимость относительной погрешности оценивания m_c и λ_c от λ .



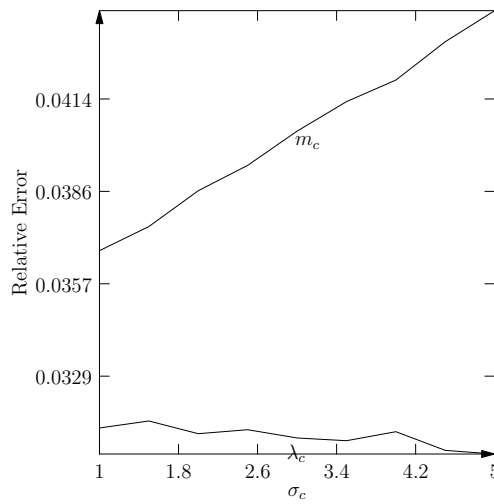
(b) Зависимость относительной погрешности оценивания m_c и λ_c от λ_c .



(c) Зависимость относительной погрешности оценивания m_c и λ_c от m_c .



(d) Зависимость относительной погрешности оценивания m_c и λ_c от σ .



(e) Зависимость относительной погрешности оценивания m_c и λ_c от σ_c .

Рис. 2: Погрешности оценивания в случае $\lambda_c \gg 0$.

Список литературы

- [1] Г.И. Ивченко and Ю.И. Медведев. *Введение в математическую статистику*. М: Издательство ЛКИ, 2010.
- [2] Wikipedia: Expectation-maximization algorithm. http://en.wikipedia.org/w/index.php?title=Expectation-maximization_algorithm&oldid=423422317.