

# SIR-модели в эпидемиологии

---

Сергей Николенко

Академия MADE — Mail.Ru

24 апреля 2021 г.

---

## *Random facts:*

- 24 апреля — День памяти жертв геноцида армян; 24 апреля 1915 г. турецкие власти начали массовые убийства, аресты и депортацию армян в Константинополе
- 24 апреля 1803 г. Александр I утвердил положение о Кавказских Минеральных Водах, а 24 апреля 1833 г. в США была запатентована газированная вода
- 24 апреля 1967 г. Владимир Комаров стал первым космонавтом, погибшим во время полёта — у аппарата «Союз-1» не открылся парашют при приземлении
- 24 апреля 1975 г. террористическая группа Баадер-Майнхоф устроила взрыв в посольстве ФРГ в Стокгольме, 24 апреля 1980 г. провалилась попытка США освободить заложников из американского посольства в Тегеране, а 24 апреля 1993 г. Ирландская республиканская армия устроила взрыв в районе Бишопсгейта (Лондон)
- 24 апреля 1990 г. остров Грюинард в Шотландии был официально объявлен свободным от чумы после 42-летнего карантина (с 1942 г. там проводились испытания биологического оружия)

# SIR-модели в эпидемиологии

---

- Прежде чем двигаться дальше — конкретный (и весьма актуальный) пример
- Давайте попробуем применить то, о чём мы говорили, к эпидемиологии
- В модели SIR есть:
  - объекты (люди)  $X = \{x_1, \dots, x_N\}$ ,
  - каждый эволюционирует между тремя состояниями  $\mathcal{S} = \{S, I, R\}^N$ ;
  - $S, I, R$  — ещё общее число объектов в соответствующих состояниях;
  - входные данные — число зарегистрированных случаев заболевания, изменяющееся во времени:  $y = \left(y^{(t)}\right)_{t=1}^T$ .

- Введём для каждого объекта *траекторию* (subject-path)  
 $\mathbf{x}_j = \left( x_j^{(t)} \right)_{t=1}^T, j = 1, \dots, N.$
- Тогда и общие статистики изменяются во времени:  $S^{(t)}, I^{(t)}, R^{(t)}.$
- Неизвестные параметры модели — это  $\boldsymbol{\theta} = \{\beta, \mu, \rho, \boldsymbol{\pi}\}$ :
  - $\boldsymbol{\pi}$  — начальное распределение заболевших,  $x_j^{(1)} \sim \boldsymbol{\pi}$ ;
  - $\rho$  — вероятность обнаружить инфицированного в общей популяции, то есть вероятность того, что человек  $x_j$  в момент  $t$ , когда  $x_j^{(t)} = I$ , будет обнаружен тестированием и зачислен в данные  $y^{(t)}$ ; тогда  $y_t \mid I^{(t)}, \rho \sim \text{Binom}(I^{(t)}, \rho)$ ;
  - $\mu$  — вероятность для заболевшего выздороветь, то есть вероятность перехода из состояния  $I$  в состояние  $R$ ;
  - $\beta$  — самый интересный параметр, вероятность заразиться за один отсчёт времени *от одного инфицированного человека*; будем предполагать самую простую модель, в которой вероятность заразиться от одного инфицированного равна  $\beta$  и все эти события независимы, а значит, вероятность остаться здоровым равна  $(1 - \beta)^{I^{(t)}}$ .

- Обозначим вектор состояний всех людей, кроме  $x_j$ , через  $\mathbf{x}_{-j}$  (и остальные величины так же).
- Вероятности перехода из  $x_j^{(t-1)}$  в  $x_j^{(t)}$ :

$$\left. \begin{aligned} p\left(x_j^{(t)} = S \mid x_j^{(t-1)} = S, \mathbf{x}_{-j}^{(t-1)}\right) &= (1 - \beta)^{I_{-j}^{(t-1)}}, \\ p\left(x_j^{(t)} = I \mid x_j^{(t-1)} = S, \mathbf{x}_{-j}^{(t-1)}\right) &= 1 - (1 - \beta)^{I_{-j}^{(t-1)}}, \\ p\left(x_j^{(t)} = R \mid x_j^{(t-1)} = I, \mathbf{x}_{-j}^{(t-1)}\right) &= \mu, \\ p\left(x_j^{(t)} = I \mid x_j^{(t-1)} = I, \mathbf{x}_{-j}^{(t-1)}\right) &= 1 - \mu, \\ p\left(x_j^{(t)} \mid x_j^{(t-1)}, \mathbf{x}_{-j}^{(t-1)}\right) &= 0 \quad \text{во всех остальных случаях.} \end{aligned} \right\}$$

- Скрытые переменные — те же самые траектории  $\mathbf{x}$  (не зря же мы их вводили).

- Тогда полное правдоподобие  $\mathcal{L}(X, Y \mid \boldsymbol{\theta})$  получается как

$$\begin{aligned}\mathcal{L}(X, Y \mid \boldsymbol{\theta}) &= p(Y \mid X, \rho) p\left(X^{(1)} \mid \pi\right) p\left(X \mid X^{(1)}, \beta, \mu\right) \\ &= \left[ \prod_{t=1}^T \binom{I^{(t)}}{Y^{(t)}} \rho^{Y^{(t)}} (1 - \rho)^{I^{(t)} - Y^{(t)}} \right] \times \\ &\quad \times \left[ \pi_S^{S^{(1)}} \pi_I^{I^{(1)}} \pi_R^{R^{(1)}} \right] \cdot \left[ \prod_{t=2}^T \prod_{j=1}^N p\left(X_j^t \mid \mathbf{x}_{-j}^{t-1}, \boldsymbol{\theta}\right) \right],\end{aligned}$$

где  $p\left(X_j^t \mid \mathbf{x}_{-j}^{t-1}, \boldsymbol{\theta}\right)$  определено матрицей вероятностей переходов.

- Апостериорное распределение, которое нам нужно:

$$p(\boldsymbol{\theta} \mid Y) \propto p(\boldsymbol{\theta}) p(Y \mid \boldsymbol{\theta}) = \int \mathcal{L}(Y \mid X, \boldsymbol{\theta}) p(X \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) dX,$$

и этот интеграл, конечно, никак не подсчитать. Что же делать?

- На помощь приходит алгоритм Метрополиса-Гастингса, точнее, сэмплирование по Гиббсу.
- Будем сэмплировать траектории  $\mathbf{x}_j$  последовательно, зафиксировав все остальные  $\mathbf{x}_{-j}$ , данные  $\mathbf{y}$  и параметры модели  $\boldsymbol{\theta}$ :

$$\mathbf{x}_j \sim p(\mathbf{x}_j | \mathbf{x}_{-j}, \mathbf{y}, \boldsymbol{\theta}) .$$

- Для этого нужно сначала понять, как выглядит распределение на траектории  $\mathbf{x}_j$ .
- Очевидно, её элементы  $x_j^{(t)}$  нельзя считать независимыми, ведь человек проходит цепочку состояний  $S \rightarrow I \rightarrow R$  только один раз и слева направо (если проходит вообще). Всё это на первый взгляд опять выглядит сложно...

- ...но здесь получается модель, которая нам уже хорошо знакома: последовательность случайных переменных  $x_j^{(t)}$  образует марковскую цепь, а если добавить ещё известные нам данные, то получится скрытая марковская модель.
- Выбросим  $x_j$  из множества траекторий, получив статистики по всей остальной популяции  $S_{-j}^{(t)}$ ,  $I_{-j}^{(t)}$  и  $R_{-j}^{(t)}$ . Тогда параметры скрытой марковской модели таковы:
  - скрытые состояния  $x_j^{(t)}$  с множеством возможных значений  $\{S, I, R\}$ ;
  - матрица вероятностей перехода  $p(x_j^t | x_{-j}^{t-1}, \theta)$ , определённая выше;
  - наблюдаемые  $y$ , вероятности получить которые зависят от того, заражён ли человек  $x_j$  в момент времени  $t$ :

$$p(y^{(t)} | x_j^{(t)}) = \text{Binom} \left( I_{-j}^{(t)} + \mathbb{I}[x_j^{(t)} = I], \rho \right).$$



- Чтобы сэмплировать одну траекторию  $\mathbf{x}_j$  при условии фиксированных остальных траекторий  $\mathbf{x}_{-j}$ , нужно сэмплировать траекторию вдоль скрытых состояний марковской модели.
- Здесь  $\mathbf{x}_j$  будет эволюционировать от состояния  $S$  к состоянию  $R$  последовательно, с вероятностями перехода  $\mathbf{x}_j$  на каждом шаге от  $S$  к  $R$

$$p\left(x_j^{(t)} = I \mid x_j^{(t-1)} = S, \mathbf{x}_{-j}\right) = 1 - (1 - \beta)^{I_{-j}^{(t-1)}},$$

а вероятность перехода от  $I$  к  $R$  фиксирована и равна  $\mu$ .

- Стохастический алгоритм Витерби: два прохода по НММ слева направо и справа налево.
- На прямом проходе подсчитываем матрицы совместных вероятностей пар последовательных состояний

$$Q_j^{(t)} = \left( q_{j,s',s}^{(t)} \right)_{s',s \in \{S,I,R\}}, \quad \text{где}$$

$$q_{j,s',s}^{(t)} = p \left( x_j^{(t)} = s, x_j^{(t-1)} = s' | Y, \mathbf{x}_{-j}, \boldsymbol{\theta} \right).$$

- Фактически в нашей модели возможных пар таких состояний всего шесть (остальные переходы запрещены), и все матрицы  $Q$  выглядят как

$$Q_j^{(t)} = \begin{pmatrix} q_{j,S,S}^{(t)} & q_{j,S,I}^{(t)} & 0 \\ 0 & q_{j,I,I}^{(t)} & q_{j,I,R}^{(t)} \\ 0 & 0 & q_{j,R,R}^{(t)} \end{pmatrix}.$$

- Чтобы вычислить  $q_{j,s',s}^{(t)}$ , нужно подсчитать

$$\begin{aligned}
 q_{j,s',s}^{(t)} &= p\left(x_j^{(t)} = s, x_j^{(t-1)} = s' | \mathbf{y}, \mathbf{x}_{-j}, \boldsymbol{\theta}\right) \\
 &\propto p\left(x_j^{(t-1)} = s' | \mathbf{y}, \mathbf{x}_{-j}, \boldsymbol{\theta}\right) p\left(x_j^{(t)} = s | x_j^{(t-1)} = \right. \\
 &\quad \left. = s', \mathbf{y}, \mathbf{x}_{-j}, \boldsymbol{\theta} \mid y_t | x_j^{(t)} = s, \mathbf{y}, \mathbf{x}_{-j}, \boldsymbol{\theta}\right) = \\
 &= \left[ \sum_{s''} q_{j,s'',s'}^{(t-1)} \right] \cdot p\left(x_j^{(t)} = s | x_j^{(t-1)} = s', \mathbf{x}_{-j}, \boldsymbol{\theta}\right) \times \\
 &\quad \times p_{\text{Binom}}\left(y^{(t)} \mid l_{-j}^{(t)} + \mathbb{I}[x_j^{(t)} = l], \rho\right),
 \end{aligned}$$

где  $p\left(x_j^{(t)} = s | x_j^{(t-1)} = s', \mathbf{x}_{-j}, \boldsymbol{\theta}\right)$  — это те самые вероятности перехода в нашей модели, подсчитанные по статистикам  $S_{-j}^{(t-1)}$ ,  $I_{-j}^{(t-1)}$  и  $R_{-j}^{(t-1)}$ , а  $p_{\text{Binom}}$  — вероятность по биномиальному распределению.

- Потом нужно нормировать, учитывая, что  $\sum_{s,s'} q_{j,s',s}^{(t)} = 1$ .

- Когда все матрицы  $Q_j^{(t)}$  подсчитаны, их можно использовать для того, чтобы сэмплировать целые последовательности скрытых состояний. Для этого нужно разложить  $p(\mathbf{x}_j \mid \mathbf{x}_{-j}, \mathbf{y}, \boldsymbol{\theta})$  не с начала времён, а с конца:

$$p(\mathbf{x}_j \mid \mathbf{x}_{-j}, \mathbf{y}, \boldsymbol{\theta}) = p(x_j^{(T)} \mid \mathbf{x}_{-j}, \mathbf{y}, \boldsymbol{\theta}) p(x_j^{(T-1)} \mid x_j^{(T)}, \mathbf{x}_{-j}, \mathbf{y}, \boldsymbol{\theta}) \times \dots \\ \dots \times p(x_j^{(2)} \mid x_j^{(3)}, \dots, x_j^{(T)}, \mathbf{x}_{-j}, \mathbf{y}, \boldsymbol{\theta}) p(x_j^{(1)} \mid x_j^{(2)}, \dots, x_j^{(T)}, \mathbf{x}_{-j}, \mathbf{y}, \boldsymbol{\theta}).$$

- И можно сэмплировать справа налево по матрицам  $Q$ .

- Последнее состояние сэмплируется из сумм по строкам последней матрицы  $Q_j^{(T)}$ :

$$\begin{aligned}x_j^{(T)} \sim p\left(x_j^{(T)} = s | \mathbf{x}_{-j}, \mathbf{y}, \boldsymbol{\theta}\right) &= \sum_{s'} p\left(x_j^{(T)} = s, x_j^{(T-1)} = s' | \mathbf{x}_{-j}, \mathbf{y}, \boldsymbol{\theta}\right) = \\&= \sum_{s'} q_{j,s',s}^{(T)}.\end{aligned}$$

- А дальше достаточно, по марковскому свойству последовательности  $\mathbf{x}_j$ , сэмплировать при условии следующего состояния, то есть использовать распределение

$$\begin{aligned}x_j^{(t)} \sim p\left(x_j^{(t)} = s | x_j^{(t+1)}, \mathbf{x}_{-j}, \mathbf{y}, \boldsymbol{\theta}\right) &\propto \\&\propto p\left(x_j^{(t)} = s, x_j^{(t+1)} = s' | \mathbf{x}_{-j}, \mathbf{y}, \boldsymbol{\theta}\right) = q_{j,s,s'}^{(t+1)}.\end{aligned}$$

- Так мы получим новую траекторию  $\mathbf{x}_j$ , и её можно подставить в  $X$  на место старой траектории и продолжать процесс сэмплирования: выбрать новый индекс  $j$  и повторить всё заново.
- В какой-то момент надо будет остановиться и обновить значения параметров.
- Теоретически можно даже сделать полноценный байесовский вывод, пересчитав параметры сопряжённых априорных распределений.
- Три основных параметра  $\beta$ ,  $\rho$  и  $\mu$  — это три монетки, а оставшийся параметр  $\pi$  — кубик с тремя гранями. Поэтому сопряжёнными априорными распределениями будут

$$\begin{aligned} p(\beta) &= \text{Beta}(a_\beta, b_\beta), & p(\mu) &= \text{Beta}(a_\mu, b_\mu), \\ p(\rho) &= \text{Beta}(a_\rho, b_\rho), & p(\pi) &= \text{Dir}(\mathbf{a}_\pi). \end{aligned}$$

- Чтобы пересчитать их апостериорные значения, нужно аналогично обычным HMM подсчитать «статистику» того, сколько раз соответствующие монетки и кубики «бросали» и чем они «выпадали» в текущем наборе скрытых переменных (траекторий)  $X$ :
  - к параметрам  $\mathbf{a}_{\pi}$  добавляются статистики того, в каких состояниях начинаются траектории:

$$a_{\pi,s} := a_{\pi,s} + \sum_{j=1}^N \mathbb{I}[X_j^{(1)} = s];$$

- Чтобы пересчитать их апостериорные значения, нужно аналогично обычным НММ подсчитать «статистику» того, сколько раз соответствующие монетки и кубики «бросали» и чем они «выпадали» в текущем наборе скрытых переменных (траекторий)  $X$ :
  - параметры  $a_\mu$  и  $b_\mu$  обновляются в зависимости от того, каково было ожидаемое число переходов из состояния  $I$  в состояние  $R$  (выздоровлений) и сколько всего времени люди провели в состоянии  $I$  (проболели):

$$a_\mu := a_\mu + \sum_{t=1}^{T-1} \sum_{j=1}^N \mathbb{I}[x_j^{(t)} = I, x_j^{(t+1)} = R],$$

$$b_\mu := b_\mu + \sum_{t=1}^T I^{(t)} - \sum_{t=1}^{T-1} \sum_{j=1}^N \mathbb{I}[x_j^{(t)} = I, x_j^{(t+1)} = R].$$



- Чтобы пересчитать их апостериорные значения, нужно аналогично обычным HMM подсчитать «статистику» того, сколько раз соответствующие монетки и кубики «бросали» и чем они «выпадали» в текущем наборе скрытых переменных (траекторий)  $X$ :
  - аналогично, параметры  $a_\rho$  и  $b_\rho$  получаются из статистики выявленных случаев, попавших в  $y$ , по сравнению со случаями, которые оказались только в  $I^{(t)}$ :

$$a_\rho := a_\rho + \sum_{t=1}^T y^{(t)}, \quad b_\rho := b_\rho + \sum_{t=1}^T \left( I^{(t)} - y^{(t)} \right);$$

# SIR-модели

- Параметры  $a_\beta$  и  $b_\beta$  самые интересные: нужно подсчитать ожидаемое число «возможностей заразиться», которые реализовались и не реализовались для всех людей в популяции:

$$p(x_j \text{ заразился при одном контакте} | x_j \text{ заразился}) = \frac{\beta}{1 - (1 - \beta)^{I^{(t)}}},$$

а значит,

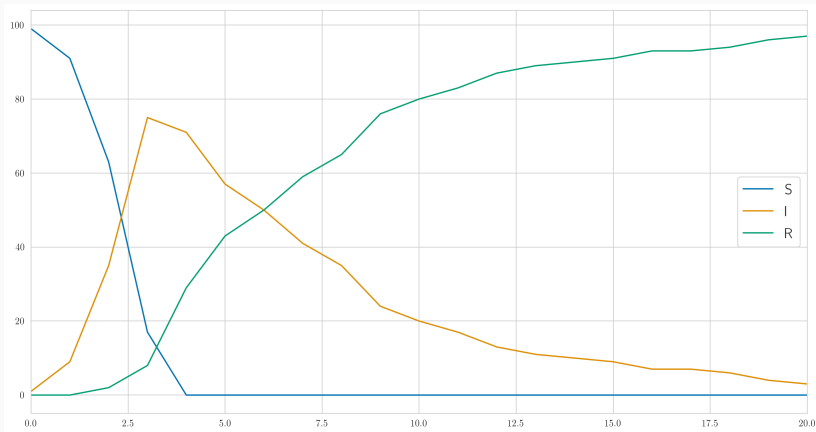
$$a_\beta := a_\beta + \sum_{t,j: x_j^{(t)}=S, x_j^{(t+1)}=I} \frac{\beta I^{(t)}}{1 - (1 - \beta)^{I^{(t)}}},$$

$$b_\beta := b_\beta + \sum_{t,j: x_j^{(t)}=S, x_j^{(t+1)}=S} I^{(t)} + \sum_{t,j: x_j^{(t)}=S, x_j^{(t+1)}=I} \left( I^{(t)} - \frac{\beta I^{(t)}}{1 - (1 - \beta)^{I^{(t)}}} \right).$$

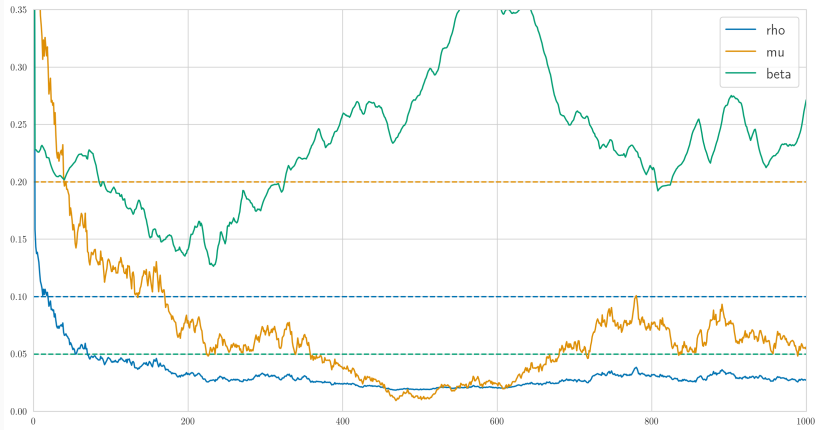
- Итого получили все компоненты нашей (сильно упрощённой!) SIR-модели: скрытые переменные в виде траекторий элементов популяции, алгоритм для сэмплирования по Гиббсу, который сэмплирует одну траекторию при условии всех остальных, и правила обновления параметров, которыми можно воспользоваться после того, как марковская цепь сэмплирования достаточно долго поработала.
- Давайте теперь посмотрим на практику...

# SIR-модели

- Пример визуализации статистик заражения при параметрах  $N = 100$ ,  $T = 20$ ,  $\rho = 0.1$ ,  $\beta = 0.05$ ,  $\mu = 0.1$ :

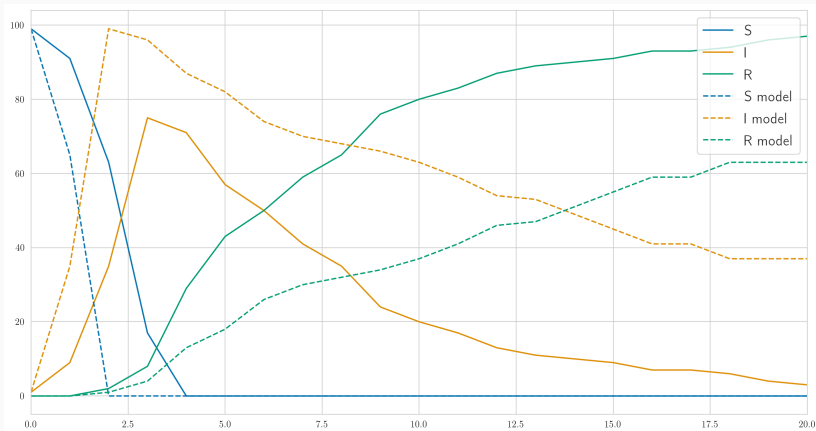


- Пример обучения параметров модели SIR:



# SIR-модели

- И если посэмплировать популяции из полученных параметров и из настоящих, получится совсем одно и то же:



- Какие выводы? Как это использовать на практике?

Спасибо!

Спасибо за внимание!