







Байесовские методы. Лекция 6.
Hamiltonian Monte-Carlo. Диагностика цепей.

Целищев М.А.

МГУ им. М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математической статистики

весна 2021

Список литературы

-  J. Kruschke. Doing Bayesian Data Analysis, Second Edition. A Tutorial with R, JAGS, and Stan. Academic Press, 2014.
-  R. McElreath. Statistical Rethinking. A Bayesian Course with Examples in R and Stan, Second Edition. Chapman and Hall CRC, 2020.
-  O. Martin. Bayesian Analysis with Python. Introduction to Statistical Modeling and Probabilistic Programming using PyMC3 and ArviZ, Second Edition. Packt, 2018.
-  K. P. Murphy. Machine Learning: A Probabilistic Perspective. The MIT Press, 2012.
-  A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, D. Rubin. Bayesian Data Analysis, Third Edition. CRC Press, 2013.
-  S. Brooks, A. Gelman, G. L. Jones and X.-L. Meng. Handbook of Markov Chain Monte Carlo. Chapman & Hall/CRC, 2011.

Напоминание: Metropolis-Hastings Algorithm

1. Пусть уже есть наблюдение $X_n = x_n$.
2. Генерируем X_* из некоторого $q(x_*|x_n) > 0$ (proposal), из которого умеем семплировать, например $\mathcal{N}(x_n, \sigma^2 I)$.
3.
$$X_{n+1} = \begin{cases} X_* & \text{с вер-тью } \alpha = \min \left(\frac{p(X_*)}{p(X_n)} \cdot \frac{q(X_n|X_*)}{q(X_*|X_n)} , 1 \right) \\ X_n & \text{иначе (т.е. отбрасывание } X_*, \text{ reject)} \end{cases}$$

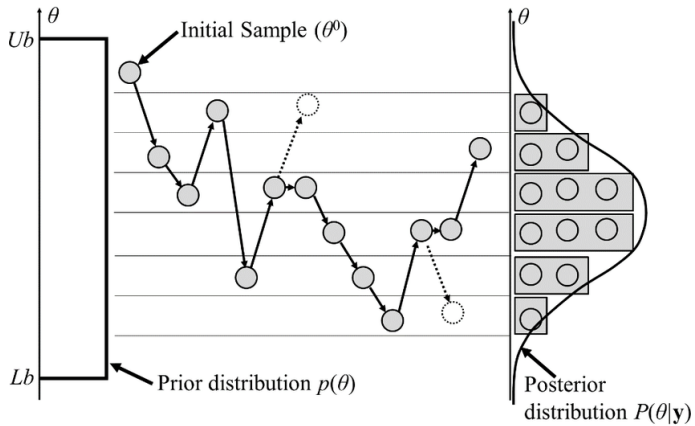
Для такой цепи распределение p удовлетворяет условию детального баланса и потому стационарно.

Замечание: для применения этого метода целевое распределение $p = \frac{\tilde{p}}{c}$ можно знать с точностью до нормировочной константы c , поскольку

$$\frac{\tilde{p}(y)}{\tilde{p}(x)} = \frac{p(y)}{p(x)}.$$

Это позволяет использовать метод Метрополиса-Гастингса для семплирования из апостериорного распределения в байесовском выводе!

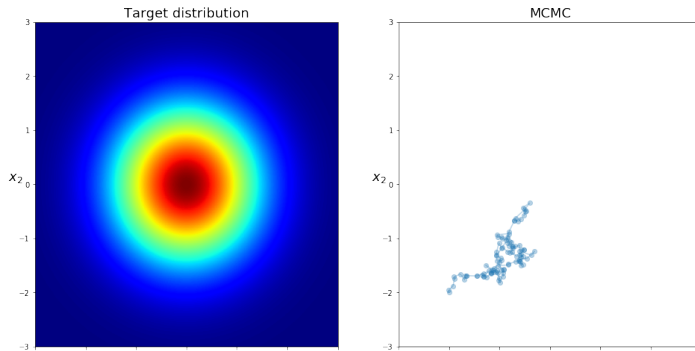
MCMC



Проблемы алгоритма М-Н

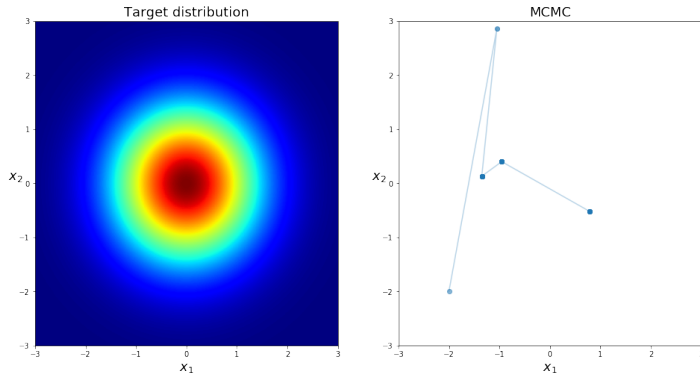
Если proposal-распределение $q(x_*|x_n)$ выбрано неудачно, то алгоритм Метрополиса-Гастингса будет работать неэффективно.

В частности, если это распределение слишком «узкое», то цепь будет долго «топтаться» вокруг одной точки, тем самым увеличивая автокорреляцию цепи. Кроме того, если начали из неудачной точки (с малой плотностью), то для сходимости к стационарному распределению потребуется очень много итераций.



Проблемы алгоритма М-Н

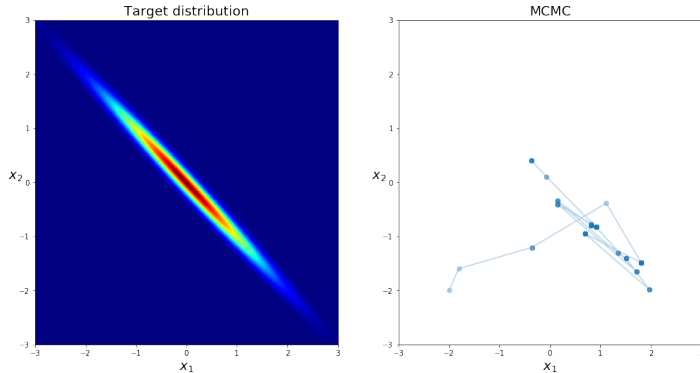
Если же proposal-распределение $q(x_*|x_n)$ слишком «широкое», то X_* будет хаотично прыгать по пространству \mathcal{X} , попадая чаще всего в точки с малой плотностью, которые будут отвергаться, поэтому цепь Маркова большую часть времени стоит на месте.



На картинке длина цепи равна 100, а rejection rate=0.96.

Проблемы алгоритма М-Н

Даже если «ширина» proposal-распределения $q(x_*|x_n)$ выбрана удачно, но само целевое распределение $p(x)$ имеет гребневую структуру (а в реальных задачах почти всегда так и бывает), то почти все X_* , которые идут не «по гребню», будут отклоняться.



На картинке длина цепи равна 100, а rejection rate=0.84.

Проблемы алгоритма М-Н

Идеально было бы адаптировать proposal-распределение $q(x_*|x_n)$, так чтобы оно по форме походило на целевое $p(x)$ (в каждой точке x_n !!!, так что это невозможно).

Хотелось бы, чтобы

- двигались по участку более-менее постоянства плотности $p(x)$ (плато) как прежде (с *постоянной скоростью*),
- двигались в направлении роста $p(x)$ с увеличивающейся скоростью,
- двигались против роста $p(x)$ с уменьшающейся скоростью.

При этом, конечно, нужно следить за сохранением условия детального баланса, чтобы обеспечить сходимость цепи к целевому распределению $p(x)$.

Оказывается, решение можно найти в классической механике.

Гамильтониан

Представим себе шайбу массы m , катающуюся по поверхности льда без воздействия неконсервативных сил. Динамика такой системы однозначно определяется координатами шайбы и её импульсом ($\phi = mv$) в начальный момент времени и задаётся полной энергией системы, называемой Гамильтонианом:

$$H(x, \phi) = U(x) + K(\phi),$$

где $U(x) = mgh(x)$ и $K(\phi) = \frac{|\phi|^2}{2m}$ — потенциальная и кинетическая энергия шайбы, соответственно.

Если шайба в какой-то момент катится вниз по склону, то её потенциальная энергия превращается в кинетическую, а если катится вверх по склону — наоборот, до тех пор, пока скорость не станет равной нулю.

Изменение координат и импульсов при этом с течением времени подчиняется системе дифференциальных уравнений:

$$\frac{dx_j}{dt} = \frac{\partial H}{\partial \phi_j}, \quad \frac{d\phi_j}{dt} = -\frac{\partial H}{\partial x_j}.$$

Hamiltonian MC

Такая аналогия с шайбой уместна и для нашей задачи семплирования из целевого *абсолютно непрерывного* распределения $p(x)$, $x \in \mathcal{X} \subset \mathbb{R}^d$.

Будем считать, что вектор x соответствует координатам шайбы и добавим вспомогательный вектор импульсов ϕ той же размерности d , с распределением $p(\phi) \sim \mathcal{N}(0, M)$, где $M = \text{diag}(m_1, \dots, m_d)$, $m_j > 0$, причём x и ϕ независимы, то есть $p(x, \phi) = p(x)p(\phi)$.

Определим Гамильтониан как

$$\begin{aligned} H(x, \phi) &= -\ln p(x, \phi) = -\ln p(x) - \ln p(\phi) = \\ &= -\ln p(x) + \frac{1}{2}\phi^T M^{-1}\phi \equiv U(x) + K(\phi) \end{aligned}$$

При этом дифф. уравнения Гамильтона принимают вид:

$$\begin{cases} \frac{dx_j}{dt} = \frac{\partial H}{\partial \phi_j} = \frac{\phi_j}{m_j} \\ \frac{d\phi_j}{dt} = -\frac{\partial H}{\partial x_j} = \frac{\partial}{\partial x_j} \ln p(x) \end{cases}, \quad j = 1, \dots, d.$$

Hamiltonian MC

Пусть текущие координаты шайбы $x = x^{(n)}$.

Шаг алгоритма НМС:

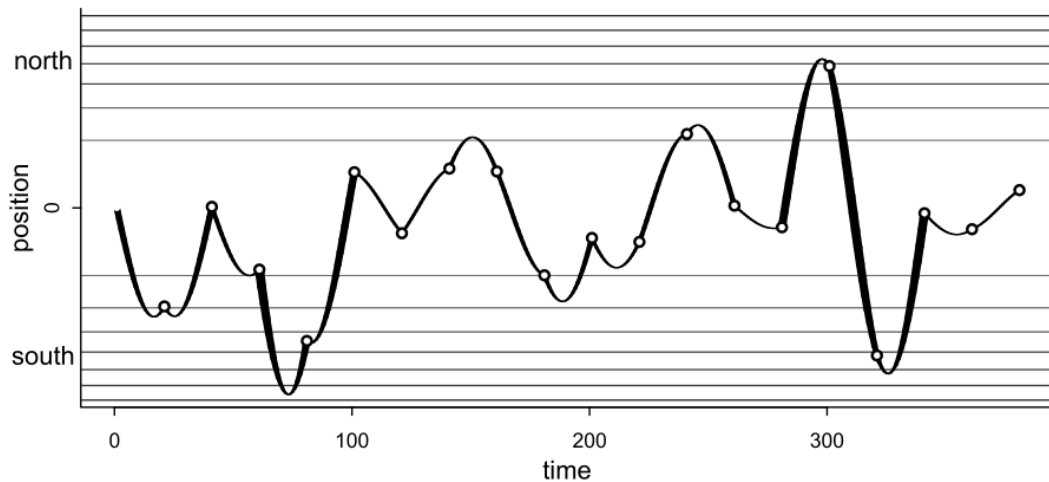
1. Семплируем $\phi^{(n)} \sim p(\phi)$ [придаём случайный импульс шайбе], $\phi \leftarrow \phi^{(n)}$.
2. Решая уравнения Гамильтона, находим положение и импульс шайбы через T секунд. Делать это приходится дискретизацией за L шагов ($T = \varepsilon L$), методом *Leapfrog*:
 - $\phi \leftarrow \phi + \frac{1}{2}\varepsilon \nabla \ln p(x)$
 - $x \leftarrow x + \varepsilon M^{-1}\phi$
 - $\phi \leftarrow \phi + \frac{1}{2}\varepsilon \nabla \ln p(x)$
3. Объявляем $x^* = x$, $\phi^* = \phi$ и смотрим по правилу Метрополиса, стоит ли принимать эту новую точку:

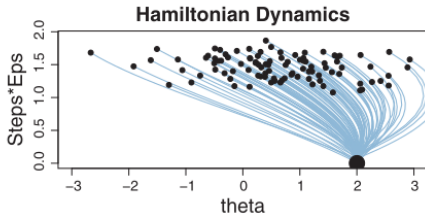
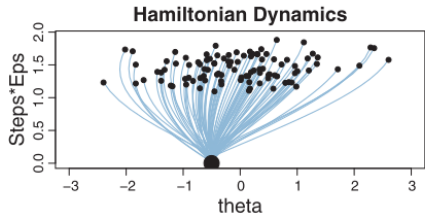
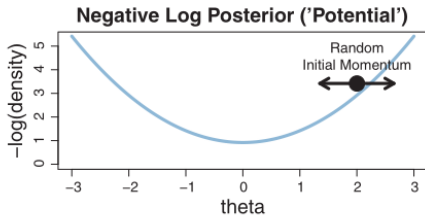
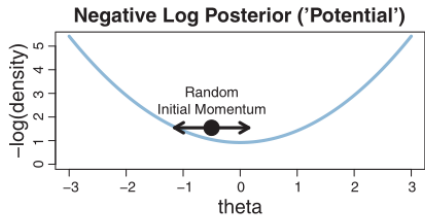
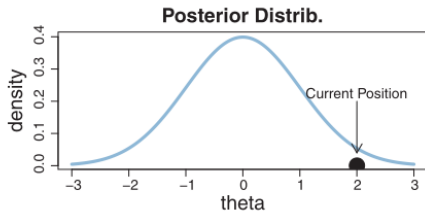
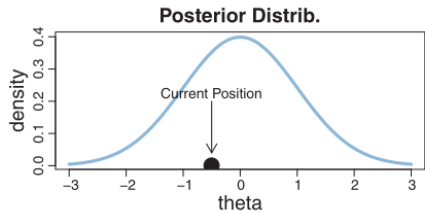
$$\alpha = \min \left(\frac{p(x^*)p(\phi^*)}{p(x^{(n)})p(\phi^{(n)})}, 1 \right).$$

С вер-тью α принимаем $x^{(n+1)} \leftarrow x^*$, иначе $x^{(n+1)} \leftarrow x^{(n)}$.

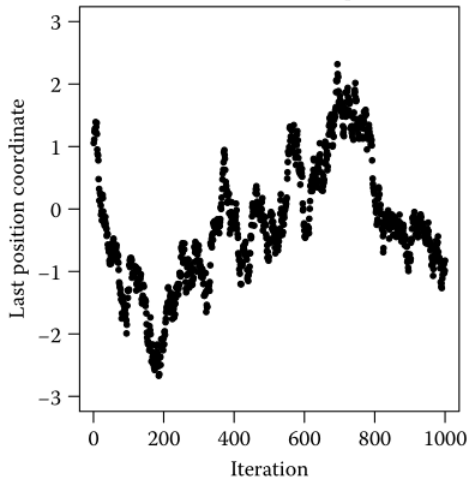
Можно показать, что при таком подходе proposal симметричен, и потому цепь сойдётся к стационарному распределению $p(x)p(\phi)$.

Hamiltonian MC Trace

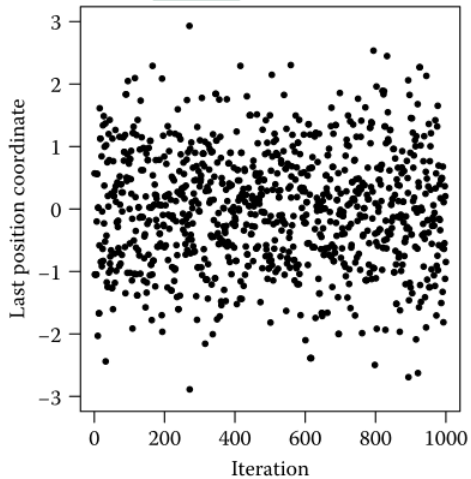


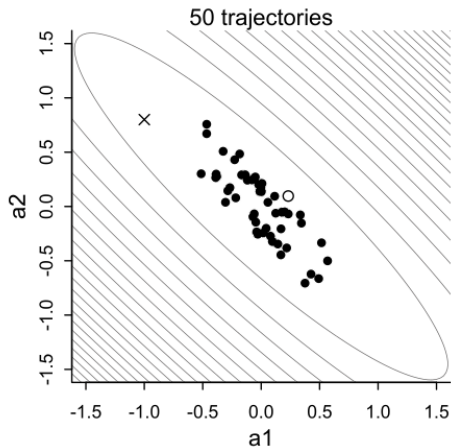
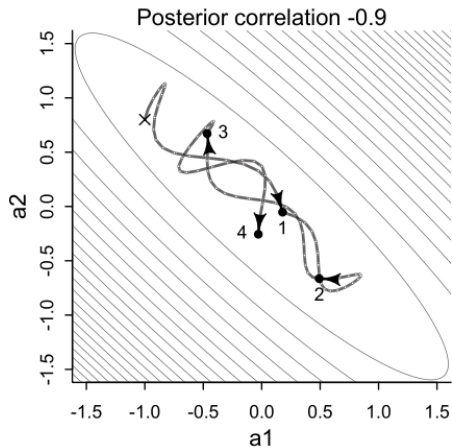


Random-walk Metropolis

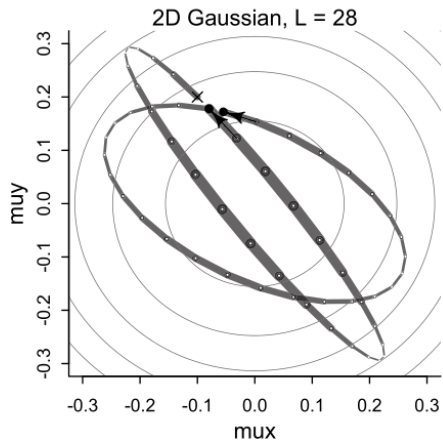
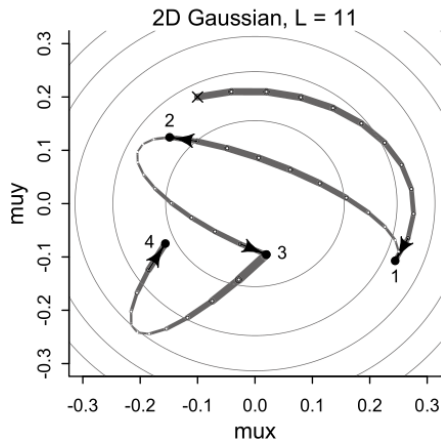


Hamiltonian Monte Carlo





Слева: динамика НМС в случае распределений гребневой структуры.
Справа: 50 сэмплов, всего один reject (пустой кружок).



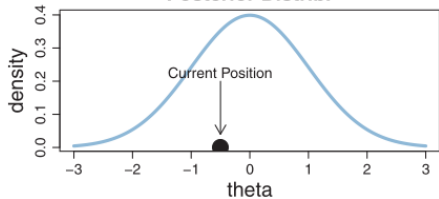
Слева: динамика НМС, когда удачно подобран параметр L .

Справа: аналогично, но параметр L подобран неудачно (проблема U-turn).

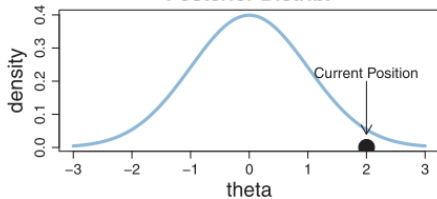
Проблемы НМС

- метод применим для семплирования только из абсолютно непрерывных распределений
- нужно уметь считать $\nabla \ln p(x)$
- нужно аккуратно подбирать параметры M, ε, L .

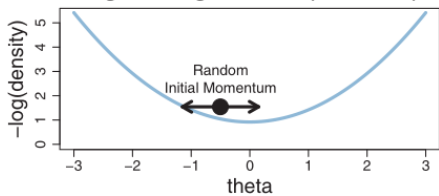
Posterior Distrib.



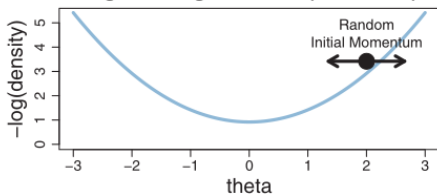
Posterior Distrib.



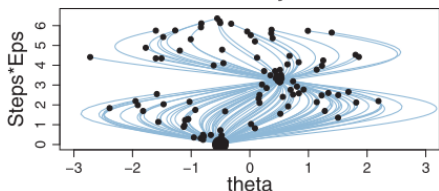
Negative Log Posterior ('Potential')



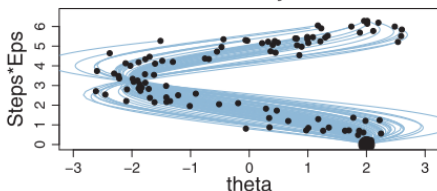
Negative Log Posterior ('Potential')



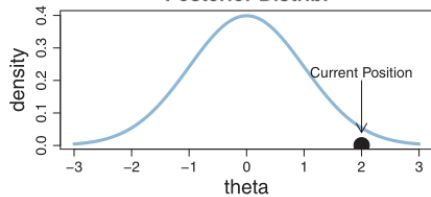
Hamiltonian Dynamics



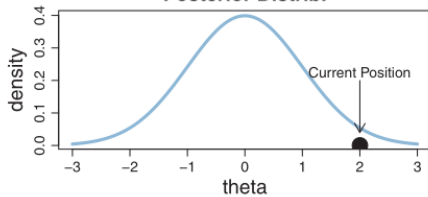
Hamiltonian Dynamics



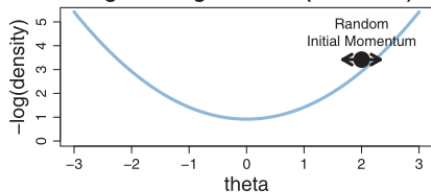
Posterior Distrib.



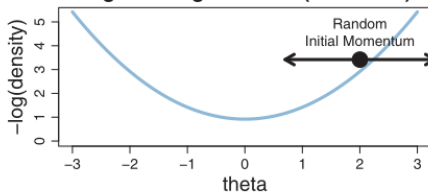
Posterior Distrib.



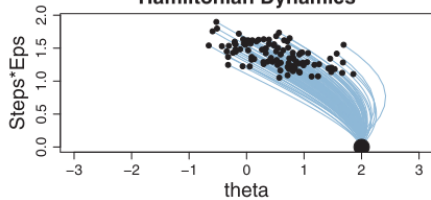
Negative Log Posterior ('Potential')



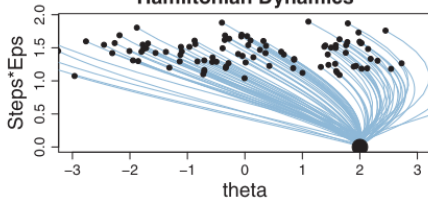
Negative Log Posterior ('Potential')



Hamiltonian Dynamics



Hamiltonian Dynamics



Диагностика цепей

Итак, марковская цепь позволяет генерировать наблюдения из стационарного распределения, при условии что цепь сошлась к этому распределению.

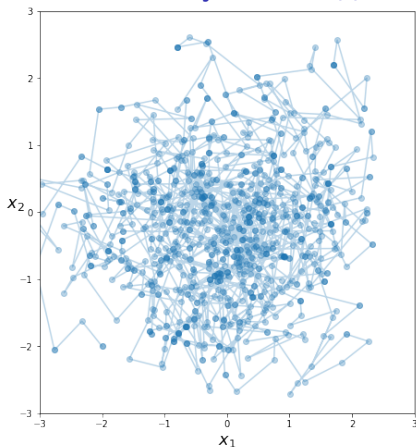
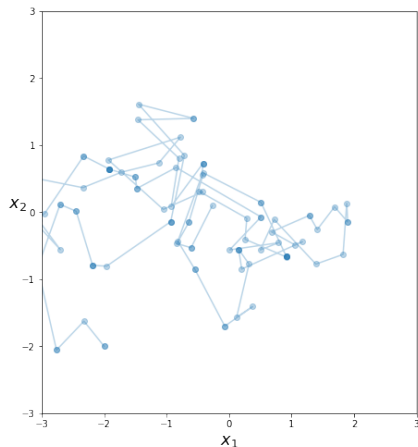
Как понять, что цепь сошлась к стационарному распределению?

Увы, никаких критериев нет... Но есть необходимые признаки. Если с ними наблюдаются проблемы, то цепи уж точно никуда не сошлись.

Эти признаки подразделяются на две категории:

- визуальные («по картинке»),
- численные (базируются на некоторых статистиках).

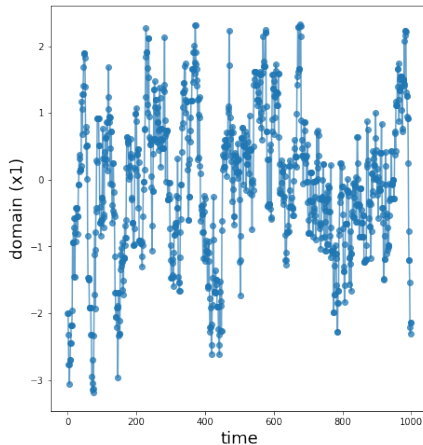
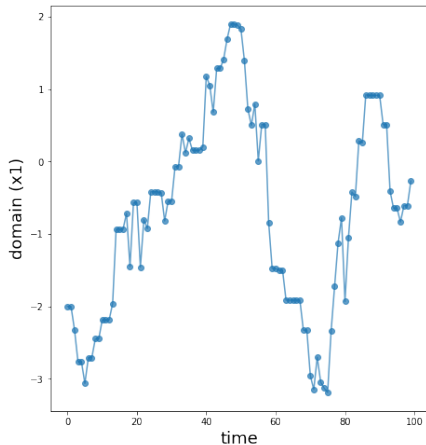
Визуальная диагностика



Двумерная цепь для целевого нормального распределения $\mathcal{N}(0, I)$. Слева — ещё не сошлась, справа — сошлась.

В многомерных пространствах используют одномерные проекции цепей (trace plots).

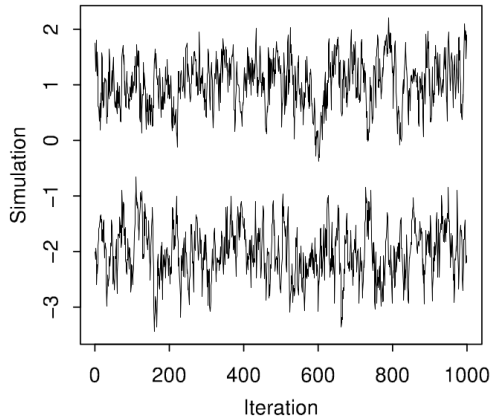
Trace plot



Здесь представлена динамика проекции цепи с прошлого слайда на первую компоненту. Слева — ещё не сошлась, справа — сошлась (цепь бродит вокруг общего центра, и её разброс с течением времени более-менее постоянен).

Мультимодальные распределения

Имеет смысл запускать несколько независимых цепей (из разных начальных точек) и накладывать их друг на друга.



На картинке обе цепи выглядят по отдельности стационарными, но первая из них застряла вблизи одной моды (точки локального максимума целевой плотности $p(x)$), а вторая — вблизи другой). Цепи не перемешались, значит не сошлись к $p(x)$.

Если же цепи статистически неотличимы (*перемешались*), то это признак того, что они сошлись к общему стационарному распределению $p(x)$.

Средние и дисперсии цепей

Пусть запущено m независимых цепей длины n , с начальным маргинальным распределением $p(x)$, совпадающим со стационарным.

X_{ij} — состояние j -ой цепи в i -ый момент, $i = 1 \dots n$, $j = 1 \dots m$.

Обозначим

$$\bar{X}_j = \frac{1}{n} \sum_{i=1}^n X_{ij} \text{ — среднее } j\text{-ой цепи,}$$

$$\bar{X} = \frac{1}{m} \sum_{j=1}^m \bar{X}_j \equiv \frac{1}{mn} \sum_{i,j=1}^{n,m} X_{ij} \text{ — общее среднее.}$$

В силу линейности мат. ожидания, это несмещённые оценки среднего $a = \int xp(x) dx$ распределения $p(x)$.

По эргодической теореме это ещё и состоятельные оценки среднего a (даже если начальное распределение отличается от $p(x)$).

Средние и дисперсии цепей

«Выборочная дисперсия» j -ой цепи:

$$S_j^2 = \frac{1}{n-1} \sum_{i=1}^n (X_{ij} - \bar{X}_j)^2$$

является оценкой дисперсии $\sigma^2 = \int (x - a)^2 p(x) dx$.

$$\begin{aligned} \mathbf{E} S_j^2 &= \frac{1}{n-1} \sum_{i=1}^n \mathbf{E} (X_{ij} - a + a - \bar{X}_j)^2 = \\ &= \frac{1}{n-1} \sum_{i=1}^n \left[\mathbf{E} (X_{ij} - a)^2 + \mathbf{E} (\bar{X}_j - a)^2 - 2\mathbf{E} (X_{ij} - a)(\bar{X}_j - a) \right] = \\ &= \frac{n}{n-1} \sigma^2 + \frac{n}{n-1} \mathbf{D} \bar{X}_j - \frac{2n}{n-1} \mathbf{D} \bar{X}_j = \frac{n}{n-1} \sigma^2 - \frac{n}{n-1} \mathbf{D} \bar{X}_j \end{aligned}$$

— уже не является несмещённой

(но всё равно состоятельна по эргодической теореме).

Средние и дисперсии цепей

Усредним S_j^2 по m цепям:

$$W := \frac{1}{m} \sum_{j=1}^m S_j^2 \quad (\text{within-sequence variance}).$$

Посчитаем выборочную дисперсию между средними цепей:

$$B := \frac{1}{m-1} \sum_{j=1}^m (\bar{X}_j - \bar{X})^2 \quad (\text{between-sequence variance}).$$

$$\begin{aligned} \mathbf{E}B &= \frac{1}{m-1} \sum_{j=1}^m \mathbf{E}(\bar{X}_j - a + a - \bar{X})^2 = \\ &= \frac{1}{m-1} \sum_{j=1}^m \left[\mathbf{E}(\bar{X}_j - a)^2 + \mathbf{E}(\bar{X} - a)^2 - 2\mathbf{E}(\bar{X}_j - a)(\bar{X} - a) \right] = \\ &= \frac{m}{m-1} \left[\mathbf{D}\bar{X}_j - \mathbf{D}\bar{X} \right] = \frac{m}{m-1} \left[\mathbf{D}\bar{X}_j - \frac{1}{m} \mathbf{D}\bar{X}_j \right] = \mathbf{D}\bar{X}_j. \end{aligned}$$

Средние и дисперсии цепей

Theorem

$$\widehat{\text{var}}^+(X) := \frac{n-1}{n}W + B$$

есть несмещённая оценка σ^2 .

Док-во.

$$\mathbf{E} \widehat{\text{var}}^+(X) = \frac{n-1}{n} \mathbf{E}W + \mathbf{E}B = \sigma^2 - \mathbf{D}\bar{X}_j + \mathbf{D}\bar{X}_j = \sigma^2. \quad \square$$

Если же начальное распределение отличается от стационарного распределения $p(x)$, то $\widehat{\text{var}}^+(X)$ переоценивает дисперсию σ^2 , в то время как средняя дисперсия по цепям W недооценивает σ^2 , поскольку каждая индивидуальная цепь j могла не сойтись к стационарному распределению.

Shrink factor

Gelman & Rubin (1992), Brooks & Gelman (1998).

В этих работах было предложено рассмотреть отношение стандартного отклонения между цепями к стандартному отклонению внутри цепи:

$$\hat{R} := \sqrt{\frac{\widehat{\text{var}}^+}{W}},$$

называемое статистикой Гельмана, или Shrink factor.

Значения \hat{R} , сильно большие 1, говорят в пользу того, что цепи не смешались (не сошлись к общему стационарному распределению).

Практическое правило: если $\hat{R} > 1.1$, то хотя бы одна из цепей не сошлась к стационарному распределению.

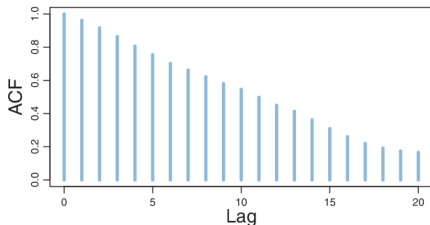
Размер выборки

Пусть цепи сошлись к стационарному распределению. Сколько нужно наблюдений, чтобы использовать насемплированные точки для оценки характеристик целевого распределения $p(x)$?

Definition

Автокорреляционной функцией цепи называют

$$\text{ACF}(t) \equiv \rho_t := \text{corr}(X_i, X_{i-t}), \quad t = 0, 1, 2, \dots$$



Ясно, что если автокорреляция слабо уменьшается с ростом t , то для статистического вывода нужно гораздо больше сэмплов, чем в случае i.i.d.

Эффективный размер выборки

Рассмотрим дисперсию выборочного среднего \bar{X}_j одной цепи:

$$\begin{aligned}\mathbf{D}\bar{X}_j &= \mathbf{D} \left[\frac{1}{n} \sum_{i=1}^n X_{ij} \right] = \frac{1}{n^2} \sum_{i,k=1}^n \text{cov}(X_{ij}, X_{kj}) = \\ &= \frac{1}{n^2} \sigma^2 (n \cdot 1 + 2(n-1)\rho_1 + 2(n-2)\rho_2 + \dots)\end{aligned}$$

Иными словами,

$$\lim_{n \rightarrow \infty} n \mathbf{D}\bar{X}_j = \left(1 + 2 \sum_{t=1}^{\infty} \rho_t \right) \sigma^2.$$

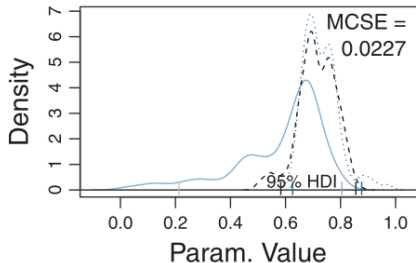
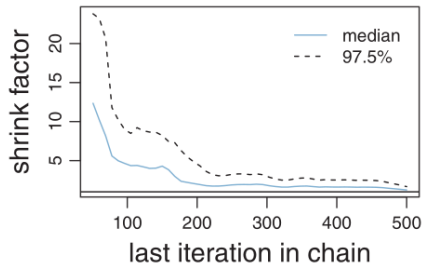
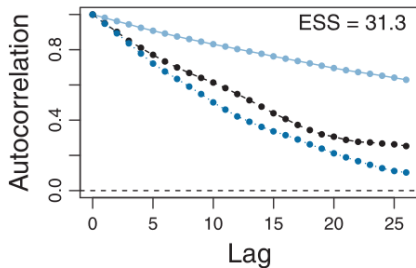
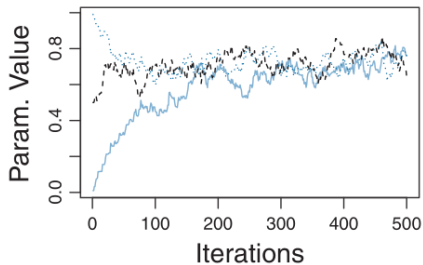
Если бы выборка была i.i.d., то $n \mathbf{D}\bar{X}_j = \sigma^2$.

Definition (Эффективный размер выборки)

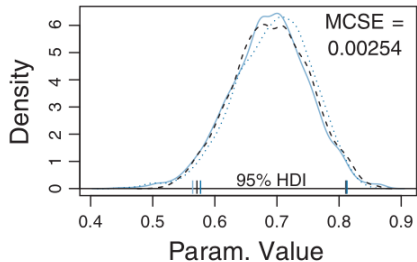
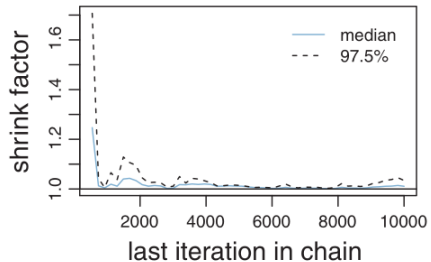
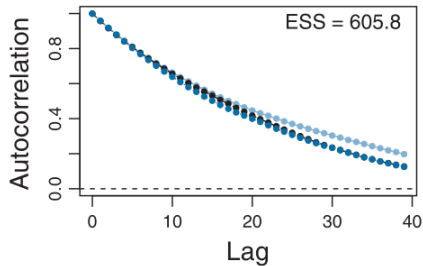
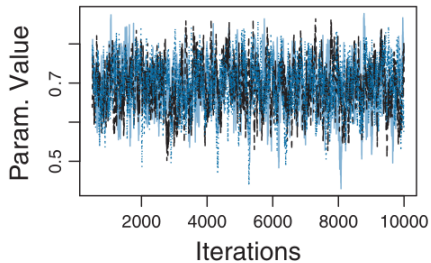
$$n_{\text{eff}} := \frac{n}{1 + 2 \sum_{t=1}^{\infty} \rho_t}$$

n_{eff} показывает, какому объёму i.i.d. выборки соответствует выборка, сгенерированная цепью.

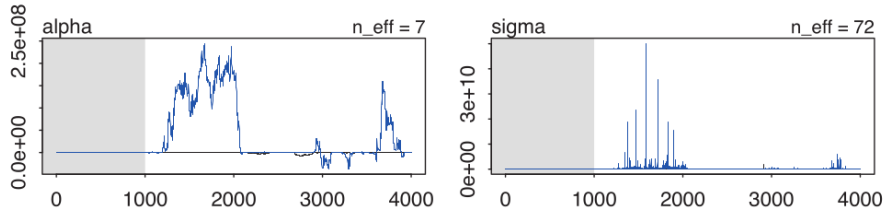
Пример диагностики



Пример диагностики



Напоследок



Если цепь выглядит нездоровой даже после длительного прогона, то, скорее всего, проблема с целевым распределением $p(x)$. Возможно, что это и вовсе не плотность распределения (интеграл от неё не сходится).

В байесовском выводе, когда в качестве целевого распределения используется апостериорное распределение параметров выбранной модели, в такой ситуации может помочь использование более информативных (т.е. более «узких») априорных распределений.