3.36pt

Algorytmy Statystyki Praktycznej

Błażej Miasojedow

MIMUW

11 kwietnia 2019

Podstawowe informacje.

- Błażej Miasojedow, email: bmiasojedow@mimuw.edu.pl
- Konsultacje po wykładach.
- www.mimuw.edu.pl/∼bmia/asp

Zasady zaliczenia przedmiotu.

- Laboratorium : 20 pkt.
- Egzamin : 20 pkt.
- Ocena końcowa : Laboratorium + Egzamin

Egzamin odbędzie się na ostatnim wykładzie.

O czym będzie wykład?

Większość metod statystycznych w praktyce wymag dodatkowych metod obliczeniowych, pozwalających na efektywne wyznaczanie estymatorów lub innych wielkości w realnych problemach.

1. W staystyce częstościowej, najpopularnijeszym o dobrych własnościach estymatorem jest ENW postaci:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \mathrm{L}(\theta).$$

2. W statystyce bayesowskiej estymatory są postaci:

$$\hat{\theta} = \mathbb{E}(\theta|\mathrm{X}),$$

gdzie typowo nie można wyznaczyć analitycznie rozkładu $\theta|\mathrm{X}$.

O czym będzie wykład?

Przedstawimy metody statystyki obliczeniowej ich własności i ograniczenia. Wykład będzie sie składał z 3 części:

1. Optymalizacja (wypukła).
2. Wnioskowanie dla ukrytych modeli Markowa.
3. Metody MCMC.

Spis treści

3.36pt

1. Optymalizacja
   - Wstęp
   - Metody gradientowe
   - Algorytm Newtona-Raphsona
   - Iterowane ważone najmniejsze kwadraty
   - Proximal gradient
   - ADMM
   - Expectation Maximization

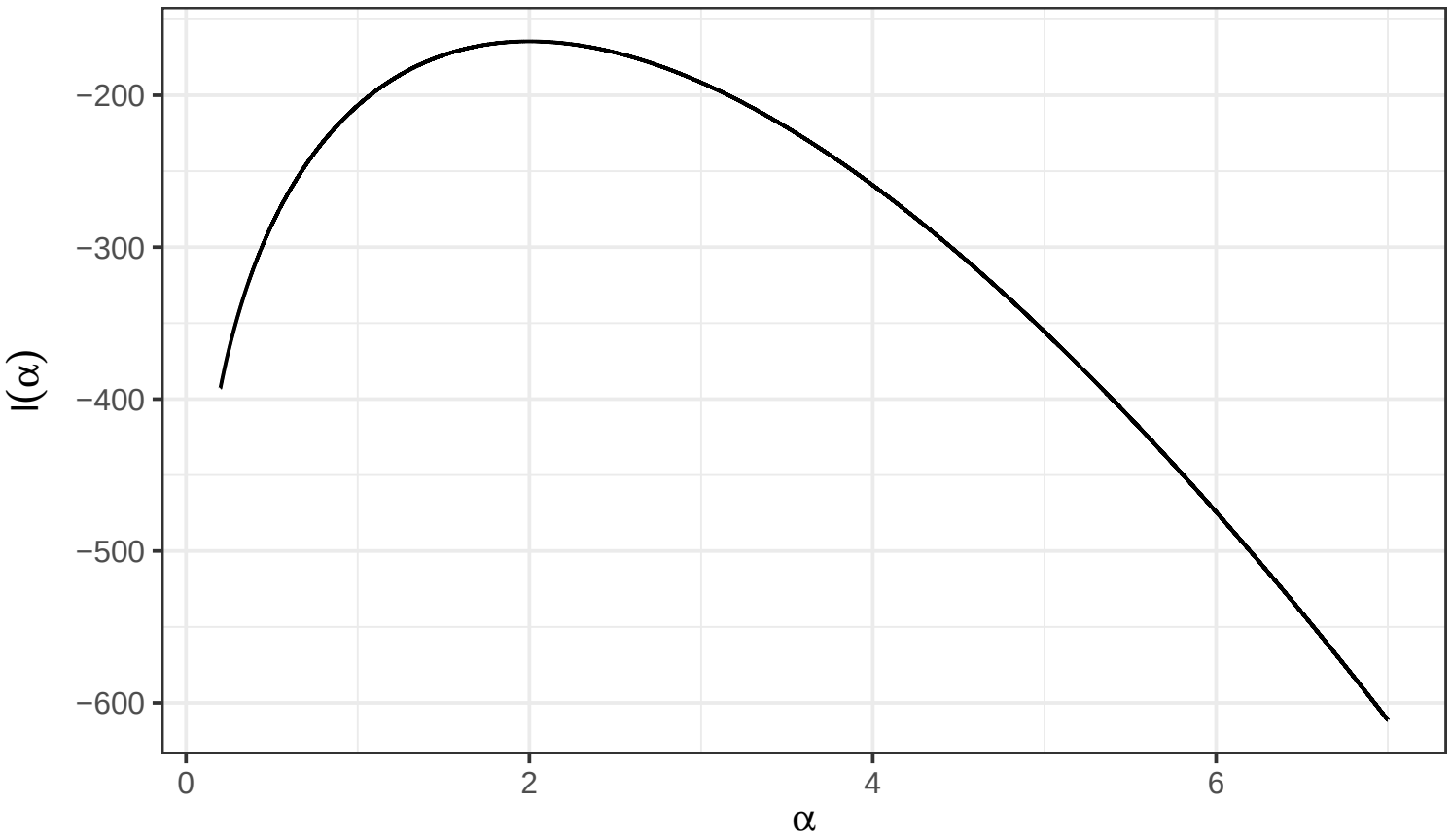
3.36pt

1. Optymalizacja
   - Wstęp
   - Metody gradientowe
   - Algorytm Newtona-Raphsona
   - Iterowane ważone najmniejsze kwadraty
   - Proximal gradient
   - ADMM
   - Expectation Maximization

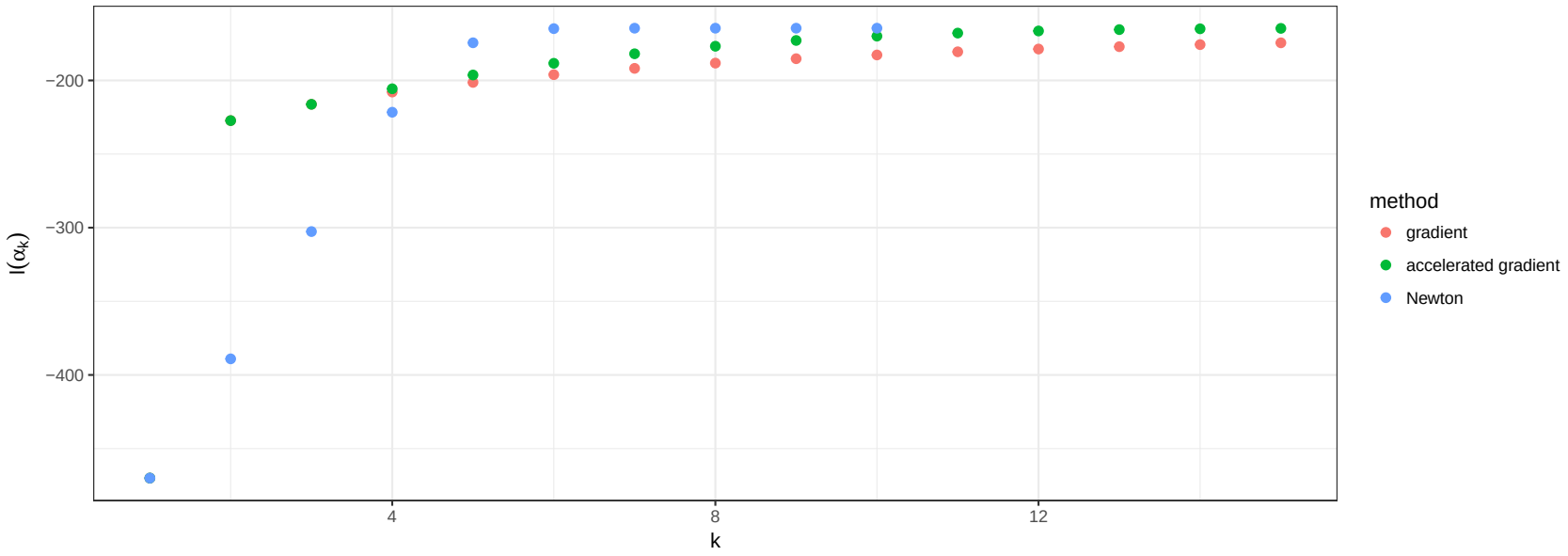
Problem 1.

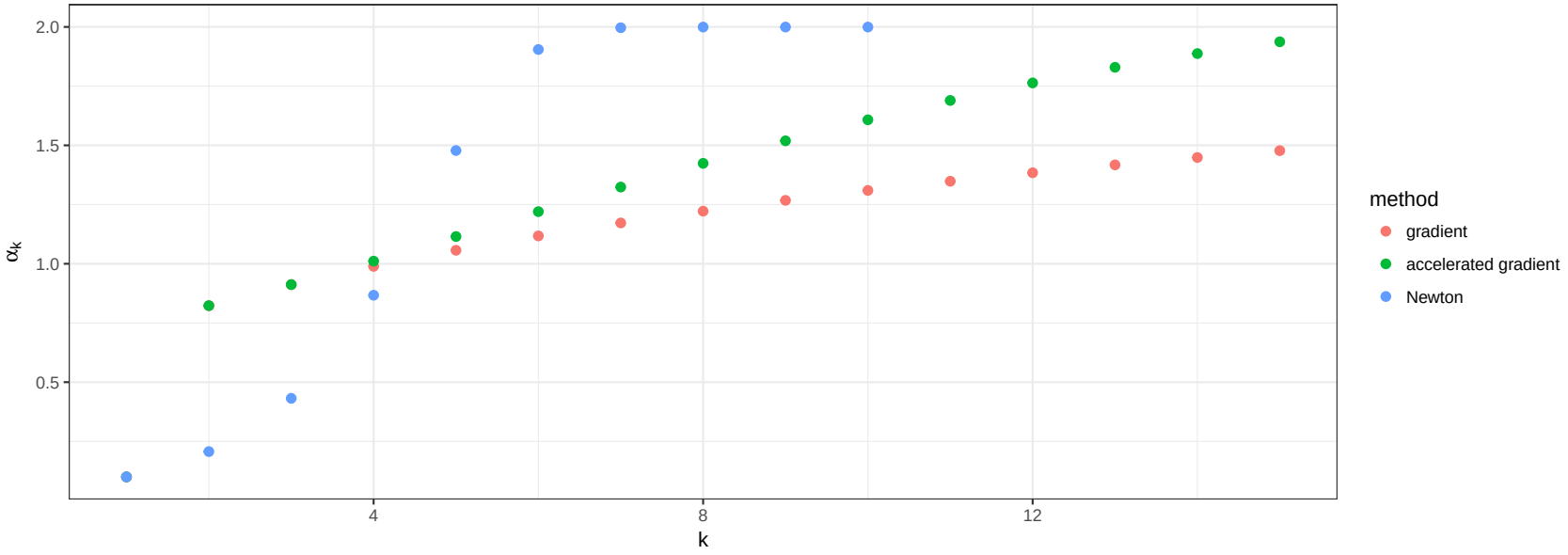
Niech $X_1, X_2, \ldots, X_n$ próbka z rozkładu $\Gamma(\alpha, 1)$. Jak wyznaczyć estymator największej wiarygodności α?

$$\hat{\alpha} = \arg\max_{\alpha} L(\alpha) = \Gamma(\alpha)^n \prod_{i=1}^{n} x_i^{\alpha-1}.$$



Problem 1. cd.


−200
−300
−400
l(αk)
4
8
12
k
method
gradient
accelerated gradient
Newton


2.0
1.5
1.0
0.5
αk
4
8
12
k
method
gradient
accelerated gradient
Newton

3.36pt

1. Optymalizacja
   - Wstęp
   - Metody gradientowe
   - Algorytm Newtona-Raphsona
   - Iterowane ważone najmniejsze kwadraty
   - Proximal gradient
   - ADMM
   - Expectation Maximization

Spadek po gradiencie

Problem:

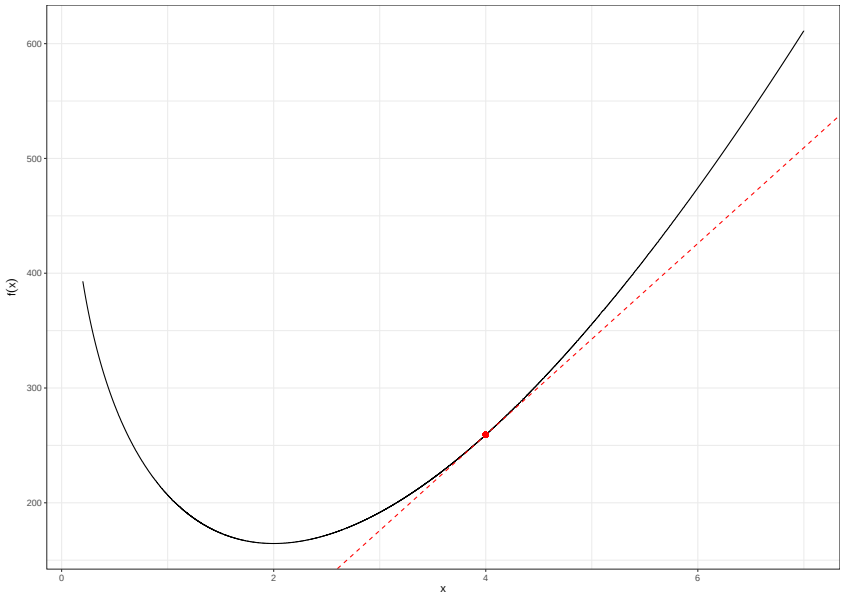
$$\arg\min_{x} f(x),$$

f -wypukła, klasy C^1.

Algorytm:

$$x_k = x_{k-1} - \gamma_k \nabla f(x_{k-1})$$

- Jakie długości kroków wybrać? (stałe, malejące)
- Czy i jak szybko zbiega ten algorytm?



Funkcje wypukłe

Definicja

Funkcja $f\colon X \subseteq \mathbb{R}^n \to \mathbb{R}$ jest wypukła wtedy i tylko wtedy gdy dla dowolnego $t \in [0,1]$ oraz dowolnych $x, y \in X$ zachodzi:

$$f(tx + (1-t)y) \leqslant tf(x) + (1-t)f(y).$$

Stwierdzenie

Różniczkowalna funkcja f jest wypukła na zbiorze wypukłym X wtedy i tylko wtedy gdy dla dowolnych $x, y \in X$ zachodzi:

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle.$$

Dowód

Niech f wypukła stąd

$$f(ty + (1-t)x) \leqslant tf(y) + (1-t)f(x).$$

Z definicji gradientu mamy

$$\begin{aligned}\langle \nabla f(x), y - x\rangle &= \lim_{t\to 0^+} \frac{f(x + t(y-x)) - f(x)}{t} \\ &\leqslant \lim_{t\to 0^+} \frac{tf(y) - tf(x)}{t} = f(y) - f(x)\end{aligned}$$

Dowód c.d.

Niech $z = tx + (1-t)y$

$$f(x) \geqslant f(z) + \langle \nabla f(z), x - z \rangle$$
$$f(y) \geqslant f(z) + \langle \nabla f(z), y - z \rangle$$

Mnożymy nierówności przez t i $1-t$ odpowiednio i dodajemy stronami. Otrzymujemy

$$tf(x) + (1-t)f(y) \geqslant f(z) + \langle \nabla f(z), tx + (1-t)y - z \rangle = f(z).$$

Silna wypukłość

Definicja

Funkcję f klasy C^1 nazywamy m-silnie wypukłą, dla $m \geqslant 0$ wtedy i tylko wtedy gdy

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2}\|x - y\|^2.$$

Uwagi:

- Równoważnie silną wypukłość można zdefiniować przez

$$f(tx + (1-t)y) \leqslant tf(x) + (1-t)f(y) - \frac{1}{2}mt(1-t)\|x - y\|^2$$

- f jest m silnie wypukła wtedy i tylko wtedy gdy $f(x) - \frac{m}{2}\|x\|^2$ jest wypukła.
- Gdy $m = 0$ silna wypukłość oznacza zwyczajną wypukłość.

L-gładkość

Definicja

Funkcje f nazywamy L-gładką jeżeli jej gradient ∇f jest Lipshitzowski ze stałą L:

$$\|\nabla f(x) - \nabla f(y)\| \leqslant L\|x - y\|.$$

Stwierdzenie

Jeżeli f jest L-gładka to dla dowolnego x, y zachodzi

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2. \tag{1}$$

Dowód

Zdefiniujmy $g(t) = f(x + t(y - x))$. Wówczas $g'(t) = \langle \nabla f(x + t(y - x)), y - x \rangle$ oraz

$$f(y) - f(x) = g(1) - g(0) = \int_0^1 g'(t)dt = \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle dt.$$

Stąd

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x), y - x \rangle &= \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle - \langle \nabla f(x), y - x \rangle dt \\ &= \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle dt \\ &\leqslant \|y - x\| \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\| dt \\ &\leqslant L\|y - x\|^2 \int_0^1 t dt = \frac{L}{2}\|y - x\|^2 \end{aligned}$$

Monotoniczność spadku po gradiencie

$$x_k = x_{k-1} - \gamma_k \nabla f(x_{k-1})$$

Lemat (1)

Jeśli nierówność (1) jest spełniona dla $y = x_k$ $x = x_{k-1}$ oraz $L = \frac{1}{\gamma_k}$ to

$$f(x_k) \leqslant f(x_{k-1})$$

Dowód

Zdefiniujmy $Q_\gamma(x,y) = f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2\gamma}\|y-x\|^2$.

- $Q_\gamma(x,x) = f(x)$
- $\arg\min_y Q_{\gamma_k}(x_{k-1}, y) = x_{k-1} - \gamma_k \nabla f(x_{k-1}) = x_k$
- Jeśli $\gamma < L^{-1}$ to

$$Q_\gamma(x,y) \geqslant f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y-x\|^2 \geqslant f(y)$$

Czyli

$$f(x_{k-1}) = Q(x_{k-1}, x_{k-1}) \geqslant Q(x_{k-1}, x_k) \geqslant f(x_k).$$

Zbieżność spadku po gradiencie

TWIERDZENIE

Jeśli f jest wypukła , L-gładka i osiąga minimum w x^* oraz dla pewnego $a > 0$ zachodzi $aL^{-1} \leqslant \gamma_k$ oraz spełnione są założenia Lematu (1) to dla dowolnego k

$$f(x_k) - f(x^*) \leqslant \frac{L\|x_0 - x^*\|^2}{2ak}.$$

Ponadto jeśli f jest m-silnie wypukła to

$$\|x_k - x^*\|^2 \leqslant \left(1 - \frac{ma}{L}\right)^k \|x_0 - x^*\|^2.$$

Dowód I

Lemat

Jeśli f jest m-silnie wypukła ,spełnione są założenia Lematu (1) oraz dla pewnego $a > 0$ zachodzi $aL^{-1} \leqslant \gamma_k$ to dla dowolnego k i dowolnego y

$$2\gamma_k(f(x_k) - f(y)) \leqslant \left(1 - \frac{ma}{L}\right) \|x_{k-1} - y\|^2 - \|x_k - y\|^2$$

Dowód twierdzenia

Niech x^* punkt w którym f osiąga minimum wtedy $f(x_k) - f(x^*) > 0$ i z lematu otrzymujemy

$$0 \leqslant \left(1 - \frac{ma}{L}\right) \|x_{k-1} - y\|^2 - \|x_k - y\|^2$$

Czyli

$$\|x_k - x^*\|^2 \leqslant \left(1 - \frac{ma}{L}\right) \|x_{k-1} - x^*\|^2$$

Dowód II

Iterując tą nierównośc otrzymujemy 2 część twierdzenia.

Z lematu otrzymujemy

$$2\gamma_i(f(x_i) - f(x^*)) \leqslant \|x_{i-1} - x^*\|^2 - \|x_i - x^*\|^2$$

Sumujemy nierówności od $i = 1$ do k

$$\sum_{i=1}^{k} 2\gamma_i(f(x_i) - f(x^*)) \leqslant \|x_0 - x^*\|^2 - \|x_k - x^*\|^2$$

Z monotoniczności $f(x_i)$ oraz z $\gamma_i \geqslant aL^{-1}$ otrzymujemy

$$\frac{2ak}{L}(f(x_k) - f(x^*)) \leqslant \|x_0 - x^*\|^2 - \|x_k - x^*\|^2$$

Dowód lematu I

Z nierówności (1) mamy dla $L = \frac{1}{\gamma_k}$

$$
\begin{aligned}
f(x_k) &\leqslant f(x_{k-1}) + \langle \nabla f(x_{k-1}), x_k - x_{k-1} \rangle + \frac{L}{2}\|x_k - x_{k-1}\|^2 \\
&= f(x_{k-1}) - \frac{1}{\gamma_k}\|x_k - x_{k+1}\|^2 + \frac{L}{2}\|x_k - x_{k-1}\|^2 \\
&= f(x_{k-1}) - \frac{1}{\gamma_k}\left(1 - \frac{L\gamma_k}{2}\right)\|x_k - x_{k-1}\|^2 \\
&= f(x_{k-1}) - \frac{1}{2\gamma_k}\|x_k - x_{k-1}\|^2
\end{aligned}
$$

Z wypukłości

$$
\begin{aligned}
f(y) &\geqslant f(x_{k-1}) + \langle \nabla f(x_k), y - x_{k-1} \rangle + \frac{m}{2}\|x_{k-1} - y\|^2 \\
&= f(x_{k-1}) - \frac{1}{\gamma_k}\langle x_k - x_{k-1}, y - x_{k-1} \rangle + \frac{m}{2}\|x_{k-1} - y\|^2
\end{aligned}
$$

Dowód lematu II

Mnożąc nierówności przez $2\gamma_k$ i odejmując stronami otrzymujemy.

$$\begin{aligned}
2\gamma_k(f(x_k) - f(y)) &\leqslant 2\langle x_k - x_{k-1}, y - x_{k-1}\rangle - \|x_k - x_{k-1}\|^2 - \gamma_k m\|x_{k-1} - y\|^2 \\
&= (1 - m\gamma_k)\|x_{k-1} - y\|^2 - \|x_{k-1} - y\|^2 + 2\langle x_k - x_{k-1}, y - x_{k-1}\rangle - \|x_k - x_{k-1}\|^2 \\
&= (1 - m\gamma_k)\|x_{k-1} - y\|^2 - \|x_{k-1} - y - (x_{k-1} - x_k)\|^2 \\
&= (1 - m\gamma_k)\|x_{k-1} - y\|^2 - \|x_k - y\|^2
\end{aligned}$$

Jak wybrać γ_k

- Optymalne jest wybrać $\gamma_k = L$, jednak obliczenie L (dobre oszacowanie) może, być truden lub niemożliwe.
- Z dowodu wynika, że jeśli $\gamma_k \to 0$ oraz $\sum_{k=0}^{\infty} \gamma_k = \infty$ to tez otrzymamy zbieżność ale w tempie zależnym od tempa zbieżności γ_k.
- Innym rozwiązaniem jest backtracking. Ustalamy γ_0 i wybieramy $0 < \eta < 1$ I wykonujemy algorytm
  1. $\gamma_k = \gamma_{k-1}$
  2. $x_k = x_{k-1} - \gamma_k \nabla f(x_{k-1})$
  3. Jeśli

     $$f(x_k) \leqslant f(x_{k-1}) + \langle \nabla f(x_{k-1}), x_k - x_{k-1} \rangle + \frac{1}{2\gamma_k} \|x_k - x_{k-1}\|^2$$

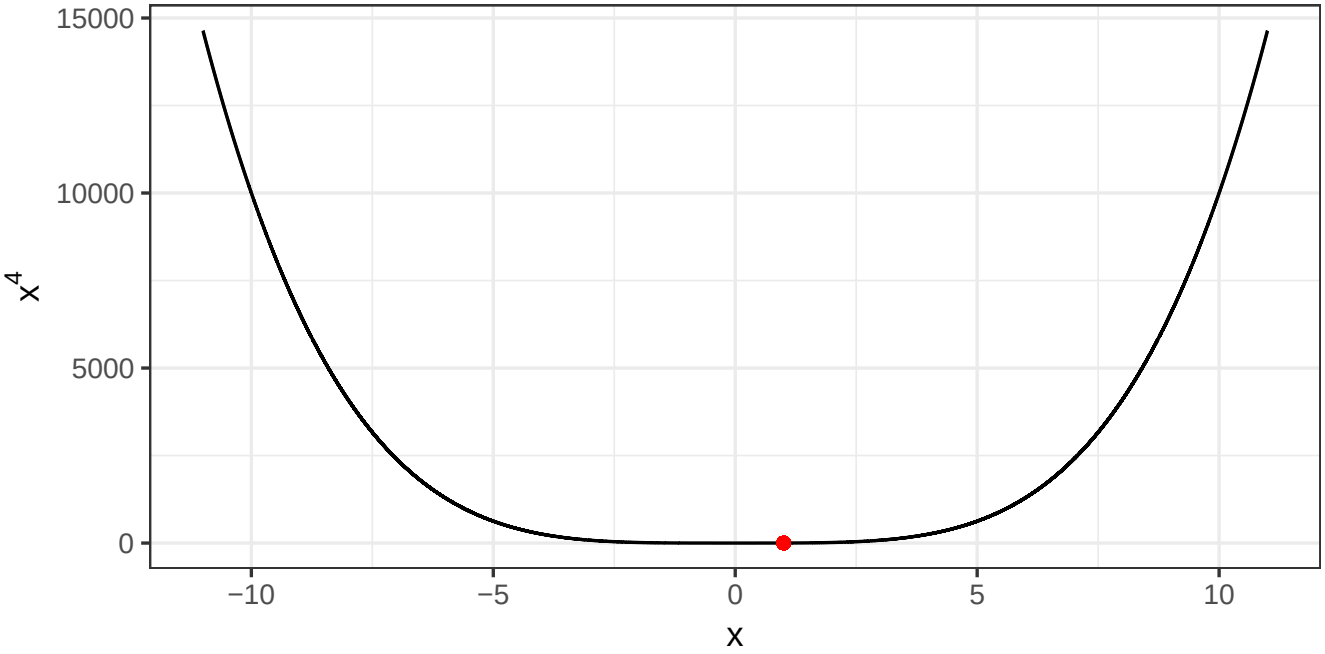
     to $k \to k+1$. W przeciwnym przypadku $\gamma_k = \eta \gamma_k$ i wracamy do kroku 2.

Istotność założenia o L-gładkości

Jeśli funkcja nie jest L gładka to algorytm gradintowy ze stałym krokiem, a nawet z γ_k dążacym do zera może uciekać do nieskończoności.

Przykład:

$$\arg\min_x x^4, \quad x_0 = 1 \quad \gamma_k = \frac{1}{k}$$

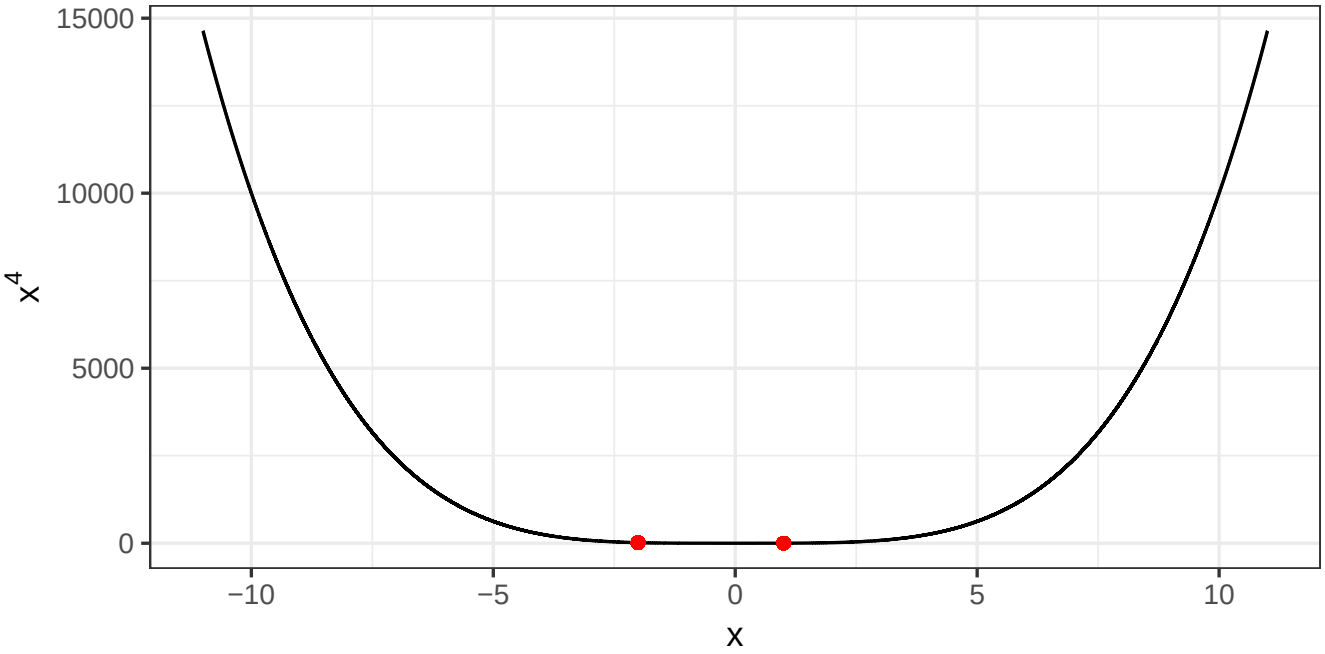


Istotność założenia o L-gładkości

Jeśli funkcja nie jest L gładka to algorytm gradintowy ze stałym krokiem, a nawet z γ_k dążacym do zera może uciekać do nieskończoności.

Przykład:

$$\arg\min_x x^4, \quad x_0 = 1 \quad \gamma_k = \frac{1}{k}$$

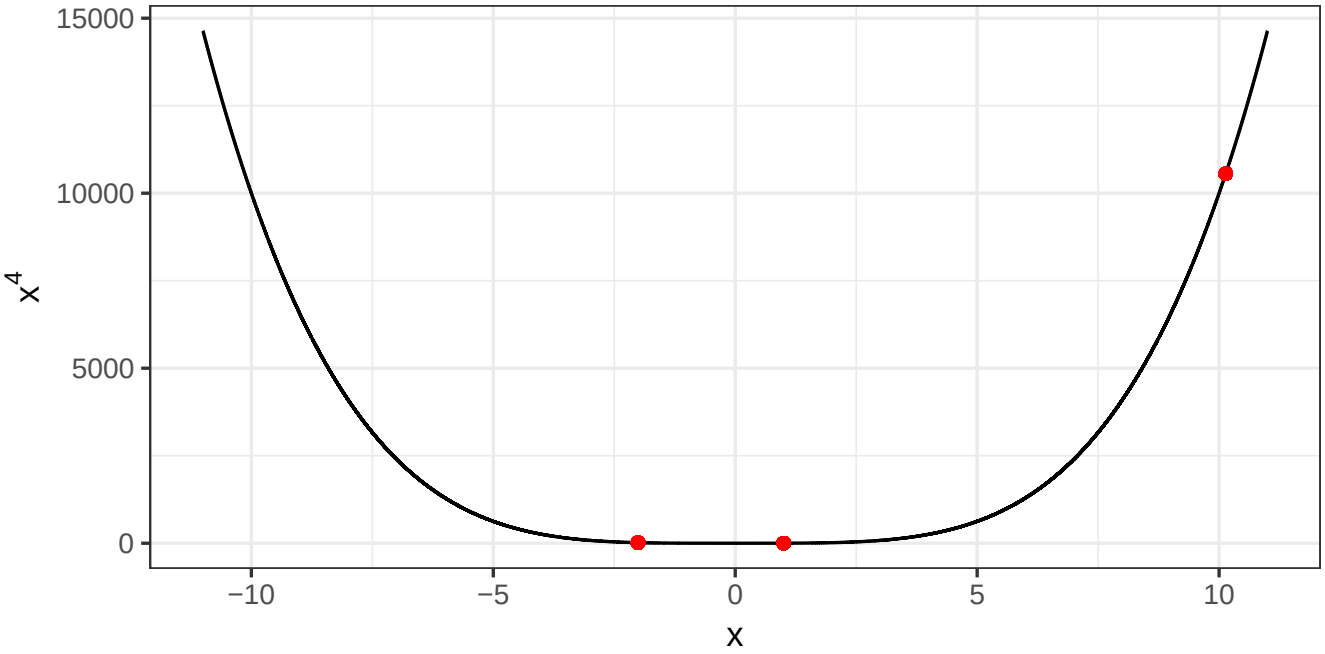


Istotność założenia o L-gładkości

Jeśli funkcja nie jest L gładka to algorytm gradintowy ze stałym krokiem, a nawet z γ_k dążacym do zera może uciekać do nieskończoności.

Przykład:

$$\arg\min_x x^4, \quad x_0 = 1 \quad \gamma_k = \frac{1}{k}$$



Przyspieszenie Nesterova

Można pokazać, że optymalne tempo zbieżności algorytmów pierwszego rzędu (opartych na gradiencie) jest $\mathcal{O}(1/k^2)$. Spadek po gradiencie zbiega w tempie $\mathcal{O}(1/k)$. Czy można go poprawić?

Nesterov 1983

Niech f wypukła

$$\arg\min_x f(x)$$

1. $y_0 = x_0, t_0 = 0$
2. $x_k = y_{k-1} - \gamma_k \nabla f(y_{k-1})$
3. $t_k = \frac{1+\sqrt{1+4t_{k-1}^2}}{2}$
4. $y_k = x_k + \frac{t_{k-1}-1}{t_k}(x_k - x_{k-1})$

Przyspieszenie Nesterowa

TWIERDZENIE

Niech f wypukła, L-gładka oraz osiąga minimum w x^*. Jeśli $\gamma_k = \gamma \leqslant L^{-1}$ to

$$f(x_k) - f(x^*) \leqslant \frac{2\|x_0 - x^*\|^2}{\gamma k^2}.$$

Dowód I

Tak jak w poprzednim dowodzie z wypukłości i L-gładkości

$$f(x_k) \leqslant f(y_{k-1}) - \frac{1}{2\gamma}\|x_k - y_{k-1}\|^2$$
$$f(z) \geqslant f(y_{k-1}) - \frac{1}{\gamma}\langle x_k - y_{k-1}, z - y_{k-1}\rangle$$

Odejmując stronami uzyskujemy

$$f(x_k) - f(z) \leqslant -\frac{1}{2\gamma}\|x_k - y_{k-1}\|^2 + \frac{1}{\gamma}\langle x_k - y_{k-1}, z - y_{k-1}\rangle$$

Uzywamy tej nierówności z $z = x_{k-1}$ i $z = x^*$

$$f(x_k) - f(x_{k-1}) \leqslant -\frac{1}{2\gamma}\|x_k - y_{k-1}\|^2 + \frac{1}{\gamma}\langle x_k - y_{k-1}, x_{k-1} - y_{k-1}\rangle$$
$$f(x_k) - f(x^*) \leqslant -\frac{1}{2\gamma}\|x_k - y_{k-1}\|^2 + \frac{1}{\gamma}\langle x_k - y_{k-1}, x^* - y_{k-1}\rangle$$

Dowód II

Mnożymy pierwszą nierównośc przez $t_{k-1} - 1$ i dodajemy do drugiej

$$t_{k-1}(f(x_k) - f(x^*)) - (t_{k-1} - 1)(f(x_{k-1}) - f(x^*)) \leqslant$$
$$-\frac{t_{k-1}}{2\gamma}\|x_k - y_{k-1}\|^2 + \frac{1}{\gamma}\langle x_k - y_{k-1}, (t_{k-1} - 1)x_{k-1} - t_{k-1}y_{k-1} + x^*\rangle$$

Oznaczmy przez $\Delta_k = f(x_k) - f(x^*)$ i pomnóżmy nierówmość przez t_{k-1} $(t_{k-2}^2 = t_{k-1}^2 - t_{k-1})$

$$t_{k-1}^2\Delta_k - t_{k-2}^2\Delta_{k-1} \leqslant \frac{-1}{2\gamma}\Big[\|(x_k - y_{k-1})t_{k-1}\|^2$$
$$+ 2t_{k-1}\langle x_k - y_{k-1}, t_{k-1}y_{k-1} - (t_{k-1} - 1)x_{k-1} - x^*\rangle\Big]$$

Dowód III

Po przekształceniu

$$t_{k-1}^2\Delta_k - t_{k-2}^2\Delta_{k-1} \leqslant \frac{-1}{2\gamma}$$

$$\left[\|x_k t_{k-1} - (t_{k-1}-1)x_{k-1} - x^*\|^2 - \|t_{k-1}y_{k-1} - (t_{k-1}-1)x_{k-1} - x^*\|^2\right]$$

Dodatkowo

$$\begin{aligned} y_k &= x_k + \frac{t_{k-1}-1}{t_k}(x_k - x_{k-1}) \\ t_k y_k &= t_k x_k + (t_{k-1}-1)(x_k - x_{k-1}) \\ t_k y_k - (t_k - 1)x_k &= t_{k-1}x_k - (t_{k-1}-1)x_{k-1} \end{aligned}$$

Oznaczając przez $u_k = t_{k-1}x_k - (t_{k-1}-1)x_{k-1} - x^*$
Stąd

$$t_{k-1}^2\Delta_k - t_{k-2}^2\Delta_{k-1} \leqslant \frac{1}{2\gamma}\left(\|u_{k-1}\|^2 - \|u_k\|^2\right)$$

Dowód IV

Sumując nierówność od $i = 2$ do k otrzymujemy

$$t_{k-1}^2 \Delta_k \leqslant \frac{1}{2\gamma}\left(\|u_1\|^2 - \|u_k\|^2\right) \leqslant \frac{1}{2\gamma}\|u_1\|^2$$

Gdzie

$$u_1 = t_0 x_k - (t_0 - 1)x_0 - x^* = x_0 - x^*$$

Oraz dla $k \geqslant 2$ (Indukcja)

$$t_{k-1} \geqslant \frac{k}{2}$$

Backtracking

1. $y_0 = x_0, t_0 = 0$, γ_0 i $0 < \eta < 1$
2. $\gamma_k = \gamma_{k-1}$
3. $x_k = y_{k-1} - \gamma_k \nabla f(y_{k-1})$
4. Jeśli
$$f(x_k) \leqslant f(y_{k-1}) + \langle \nabla f(y_{k-1}), x_k - y_{k-1} \rangle + \frac{1}{2\gamma_k} \|x_k - y_{k-1}\|^2$$
to krok 5. W przeciwnym przypadku $\gamma_k = \eta \gamma_k$ i wracamy do kroku 3.
5. $t_k = \frac{1+\sqrt{1+4t_{k-1}^2}}{2}$
6. $y_k = x_k + \frac{t_{k-1}-1}{t_k}(x_k - x_{k-1})$

TWIERDZENIE

Niech f wypukła, L-gładka oraz osiąga minimum w x^*. Jeśli γ_k jak powyżej to

$$f(x_k) - f(x^*) \leqslant \frac{2L\|x_0 - x^*\|^2}{\eta k^2}.$$

3.36pt

1. Optymalizacja
   - Wstęp
   - Metody gradientowe
   - Algorytm Newtona-Raphsona
   - Iterowane ważone najmniejsze kwadraty
   - Proximal gradient
   - ADMM
   - Expectation Maximization

Algorytm Newtona - Raphsona

Niech f klasy C^2, rozwiązanie problemu

$$\arg\min_{x} f(x)$$

sprowadza się do rozwiązania równania

$$\nabla f(x) = 0 \tag{2}$$

Ze wzoru Taylora

$$\nabla f(x) \approx \nabla f(x_0) + \nabla^2 f(x_0)(x - x_0)$$

Czyli możemy przybliżać rozwiązanie (2) poprzez

$$x = x_0 - (\nabla^2 f(x_0))^{-1} \nabla f(x_0)$$

Algorytm Newtona-Raphsona

1. Ustalamy punkt startowy x_0.
2. Kolejne przybliżenia rozwiązania uzyskujemy

$$x_k = x_{k-1} - (\nabla^2 f(x_{k-1}))^{-1} \nabla f(x_{k-1})$$

Kwadratowa zbieżnośc metody Newtona-Raphsona

Algorytm jest kwadratowo zbieżny do x^* jeżeli istnieje M takie, że

$$\lim_{k\to\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^2} < M$$

TWIERDZENIE

Niech f funkcja jest klasy C^3 i osiąga minimum w x^*. Jeśli istnieje $\epsilon > 0$, $M < \infty$, takie że

$$\sup_{\{x,y:\ \|x-x^*\|<\epsilon,\ \|y-x^*\|<\epsilon\}} \|(\nabla^2 f(x))^{-1}\nabla^3 f(y)\| < M$$

oraz $\|x_0 - x^*\| < \min(\epsilon, M^{-1})$, $\nabla^2 f$ odwracalna, to dla każdego k

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^2} < M$$

Dowód

Ponieważ w x^* jest minimum f ze wzoru Taylora dla pewnego ξ "pomiędzy" x_k i x^* otrzymujemy

$$0 = \nabla f(x^*) = \nabla f(x_k) + \nabla^2 f(x_k)(x^* - x_k) + \nabla^3 f(\xi)(x^* - x_k)^{\otimes 2}$$

Mnożąc, stronami przez $(\nabla^2 f(x_k))^{-1}$ otrzymujemy

$$0 = (\nabla^2 f(x_k))^{-1}\nabla f(x_k) + (x^* - x_k) + (\nabla^2 f(x_k))^{-1}\nabla^3 f(\xi)(x^* - x_k)^{\otimes 2}$$

Z definicji x_{k+1} otrzymujemy

$$0 = x^* - x_{k+1} + (\nabla^2 f(x_k))^{-1}\nabla^3 f(\xi)(x^* - x_k)^{\otimes 2}$$

Przenosząc na drugą strone i przykładając normę otrzymujemy

$$\|x^* - x_{k+1}\| \leqslant \|(\nabla^2 f(x_k))^{-1}\nabla^3 f(\xi)\| \|x^* - x_k\|^2$$

Własności metody Newtona-Raphsona

- Przy dodatkowych warunkach regularności można pokazać globalną kwadratową zbieżność.
- W ogólności obszar kwadratowej zbieżności może być mały i przy złym wyborze punktu startowego algorytm może wpaść w cykl lub nawet być rozbieżny.
- W przestrzeni wyosko wymiarowej, problemem może byc obliczanie macierzy drugiej pochodnej i jej odwracanie.

3.36pt

1. Optymalizacja
   - Wstęp
   - Metody gradientowe
   - Algorytm Newtona-Raphsona
   - Iterowane ważone najmniejsze kwadraty
   - Proximal gradient
   - ADMM
   - Expectation Maximization

Regresja logistyczna

$Y_1, Y_2, \ldots, Y_n$ - binarne zmienne objaśniane, $x_1, x_2, \ldots, x_n$ - zmienne objaśniające z $\mathbb{R}^p$. Oznaczmy przez $p_i = \mathbb{P}(Y_i = 1)$

Regresja logistyczna

$$\mathsf{logit}(p_i) = \mathsf{log}\left(\frac{p_i}{1 - p_i}\right) = x_i^T \beta$$

Log wiarygodność

$$\ell(\beta) = \sum_{i=1}^{n} Y_i x_i^T \beta - \sum_{i=1}^{n} \mathsf{log}(1 + \mathsf{exp}(x_i^T \beta))$$

Estymator największej wiarygodności

$$\hat{\beta} = \arg\max_{\beta} \ell(\beta) = \arg\min_{\beta} -\ell(\beta)$$

Jak obliczyć $\hat{\beta}$?

Uogólnione modele liniowe

Regresja logistyczna jest szczególnym przypadkiem uogólnionych modeli liniowych. Rodzina wykładnicza

$$\mathrm{f}_\theta(\mathrm{y}) = \mathsf{exp}\left[\frac{\theta \mathrm{y} - \mathrm{b}(\theta)}{\mathrm{a}(\phi)}\right] \mathrm{h}(\mathrm{y})$$

- θ - parametr kanoniczny
- ϕ - parmetr dyspersji
- $\mathbb{E}(\mathrm{Y}) = \mathrm{b}'(\theta) = \mu$
- $\mathrm{var}(\mathrm{Y}) = \mathrm{a}(\phi)\mathrm{b}''(\theta)$

Uogólniony model liniowy $\mathrm{Y}_1, \ldots, \mathrm{Y}_\mathrm{n} \sim \mathrm{f}_{\theta_\mathrm{i}}$

$$\mathrm{g}(\mu_\mathrm{i}) = \mathrm{x}_\mathrm{i}^\mathrm{T}\beta = \eta_\mathrm{i}$$

g- funkcja łącząca (jeśli $\mathrm{g}(\mu_\mathrm{i}) = \theta_\mathrm{i}$ to kanoniczna funkcja łącząca)

Estymator największej wiarygodności w GLM I

$$\ell(\beta) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{\theta_{\mathrm{i}}\mathrm{y} - \mathrm{b}(\theta_{\mathrm{i}})}{\mathrm{a}(\phi)}$$

Chcemy wyprowadzić algorytm Newtona-Raphsona (przybliżony).

$$\frac{\partial}{\partial \beta_{\mathrm{j}}}\ell = ? \quad \frac{\partial^2}{\partial \beta_{\mathrm{j}} \partial \beta_{\mathrm{k}}}\ell = ?$$

Dla uproszczenia zapisu przyjmijmy na chwilę $\mathrm{n} = 1$ i pomińmy indeks i

$$\frac{\partial}{\partial \beta_{\mathrm{j}}}\ell(\beta) = \frac{\partial \ell}{\partial \theta}\frac{\partial \theta}{\partial \mu}\frac{\partial \mu}{\partial \eta}\frac{\partial \eta}{\partial \beta_{\mathrm{j}}}$$

Estymator największej wiarygodności w GLM II

$$\frac{\partial \ell}{\partial \theta} = \frac{\mathrm{y} - \mathrm{b}'(\theta)}{\mathrm{a}(\phi)}$$
$$\frac{\partial \theta}{\partial \mu} = \frac{1}{\mathrm{b}''(\theta)} = \frac{\mathrm{a}(\phi)}{\mathrm{var}(\mathrm{Y})}$$
$$\frac{\partial \eta}{\partial \beta_{\mathrm{j}}} = \mathrm{x}_{\mathrm{ij}}$$

W przypadku kanonicznej funkcji łączącej $\eta = (\mathrm{b}')^{-1}(\mu)$

$$\frac{\partial \mu}{\partial \eta} = \mathrm{b}''(\theta)$$

Stąd

$$\frac{\partial \ell}{\partial \beta_{\mathrm{j}}} = \frac{\mathrm{y} - \mu}{\mathrm{var}(\mathrm{Y})} \mathrm{x}_{\mathrm{ij}} \frac{\partial \mu}{\partial \eta}$$

Estymator największej wiarygodności w GLM III

i w przypadku kanonicznej funnkcji łączącej

$$\frac{\partial \ell}{\partial \beta_{\mathrm{j}}} = \frac{\mathrm{y} - \mu}{\mathrm{a}(\phi)} \mathrm{x}_{\mathrm{ij}}$$

Fisher scoring

W algorytmie Newtona-Raphsona chcemy przybliżyć $\frac{\partial^2 \ell}{\partial \beta_{\mathrm{j}} \partial \beta_{\mathrm{k}}}$ przez $\mathbb{E} \frac{\partial^2}{\partial \beta_{\mathrm{j}} \partial \beta_{\mathrm{k}}} \ell$. Czyli

$$\nabla^2 \ell(\beta) \approx -\mathrm{I}(\beta)$$

$\mathrm{I}(\beta)$ - informacja Fishera

W przypadku kanonicznej funkcji łączacej

$$\nabla^2 \ell(\beta) = -\mathrm{I}(\beta)$$

Estymator największej wiarygodności w GLM IV

Korzystając z faktu

$$\mathbb{E}\frac{\partial^2 \ell}{\partial \beta_{\mathrm{j}} \partial \beta_{\mathrm{k}}} = -\mathbb{E}\frac{\partial \ell}{\partial \beta_{\mathrm{j}}}\frac{\partial \ell}{\partial \beta_{\mathrm{k}}}$$
$$= \mathbb{E}\frac{(\mathrm{y}-\mu)^2}{\mathrm{var}(\mathrm{Y})^2}\mathrm{x_{ij}x_{ik}}\left(\frac{\partial \mu}{\partial \eta}\right)^2 = \frac{\mathrm{x_{ij}x_{ik}}}{\mathrm{var}(\mathrm{Y})}\left(\frac{\partial \mu}{\partial \eta}\right)^2$$

Oznaczmy

$$\mathrm{A} = \mathrm{diag}\left(\frac{\frac{\partial \mu_{\mathrm{i}}}{\partial \eta_{\mathrm{i}}}}{\mathrm{var}(\mathrm{Y_i})}\right) \quad \mathrm{W} = \mathrm{diag}\left(\frac{\left(\frac{\partial \mu_{\mathrm{i}}}{\partial \eta_{\mathrm{i}}}\right)^2}{\mathrm{var}(\mathrm{Y_i})}\right)$$

Otrzymujemy

$$\nabla \ell(\beta) = \mathrm{X^T A}(\mathrm{y}-\mu), \quad \mathrm{I}(\beta) = \mathrm{X^T W X}$$

Estymator największej wiarygodności w GLM V

Oraz algorytm Fisher scoring przyjmuje postać

$$\begin{aligned}\beta^{k+1} &= \beta^{k} + (X^T W X)^{-1} X^T A (y - \mu) \\ &= (X^T W X)^{-1} [X^T W X \beta^{k} + X^T A (y - \mu)]\end{aligned}$$

$$A \operatorname{diag}\left(\frac{\partial \mu}{\partial \eta}\right) = W, \quad X^T \beta = \eta$$

Otrzymujemy iterowane ważone najmniejsze kwadraty

$$\beta^{k+1} = (X^T W^k X)^{-1} X^T W^k z^k,$$

gdzie

$$z = \eta - \operatorname{diag}\left(\frac{\partial \eta}{\partial \mu}\right)(y - \mu)$$

To jeszcze nie jest praktyczny algorytm, pozostaje pytanie jak numerycznie rozwiązać równanie liniowe

Powrót do regresji logistycznej

$Y_i \sim \mathrm{Bin}(1, p_i)$

$$\mathsf{logit}(p_i) = x_i^T \beta, \quad p_i = \frac{1}{1 + \mathsf{exp}(-x_i^t \beta)}$$

Gęstość można zapisać w postaci wykładniczej

$$f(y; p) = \mathsf{exp}(y\, \mathsf{logit}(p) + \mathsf{log}(1 - p)))$$

kanoniczny parametr $\theta = \mathsf{logit}(p)$

$$f(y; \theta) = \mathsf{exp}(y\theta - \mathsf{log}(1 + \mathsf{exp}(\theta)))$$

$$\mathbb{E}(Y) = p, \ \mathrm{var}(Y) = p(1 - p), \ \eta = \theta, \ \frac{\partial \eta}{\partial \mu} = \frac{1}{p(1 - p)}$$

Stąd W i z są postaci

$$W = \mathrm{diag}(p_i(1 - p_i)), \quad z_i = x_i^T \beta + \frac{y_i - p_i}{p_i(1 - p_i)}$$

Obliczanie rozwiązania MNK

Rozkład Choleskiego Dodatnio określoną macierz A można przedstawić w postaci

$$A = LL^T$$

gdzie L dolnotrójkątna.
Równanie

$$X^T X\beta = X^T y$$

Rozwiązujemy

1. $X^T X = LL^T$
2.

$$Lz = X^T Y$$
$$L^T\beta = z$$

Obliczanie rozwiązania MNK II spsób

Rozkład QR Macierz X -n $\times$ p, n $>$ p można przedstawić w postaci

$$X = QR$$

gdzie Q - n $\times$ p o ortogonalnych kolumnach $Q^T Q = I$ i R górnotrójkątna p $\times$ p. Wówczas

$$\beta = (X^T X)^{-1} X^T Y = (R^T Q^T Q R)^{-1} R^T Q^T Y = R^{-1} Q^T Y$$

I dostajemy równanie

$$R\beta = Q^T Y$$

Dodatkowo

$$\hat{Y} = X\hat{\beta} = QQ^T Y$$

3.36pt

1. Optymalizacja
   - Wstęp
   - Metody gradientowe
   - Algorytm Newtona-Raphsona
   - Iterowane ważone najmniejsze kwadraty
   - Proximal gradient
   - ADMM
   - Expectation Maximization

LASSO

Model liniowy:

$$\mathrm{Y} = \mathrm{X}\beta + \epsilon$$

Metoda najmniejszych kwadratów:

$$\arg\min_{\beta} \|\mathrm{Y} - \mathrm{X}\beta\|^2$$

LASSO:

$$\arg\min_{\beta} \{\|\mathrm{Y} - \mathrm{X}\beta\|^2 + \lambda \sum_{\mathrm{i}=1}^{\mathrm{p}} |\beta_{\mathrm{i}}|\}$$

Jak obliczyć rozwiązanie LASSO?

Subgradient

Jeśli f jest wypukła i gładka to

$$f(y) \geqslant f(x) + \nabla f(x)^T (y - x)$$

Definicja

Jeśli f jest wypukła to g nazywamy subgradientem f w punkcie x wtedy i tylko wtedy gdy dla każdego y

$$f(y) \geqslant f(x) + g^T (y - x)$$

Zbiór wszystkich subgradientów f w punkcie x nazywamy subróżniczką i będziemy oznaczać przez $\partial f(x)$

Własności subróżniczki

Jeśli f wypukła to:

- $\partial f(x)$ jest niepustym wypukłym zbiorem
- Jeśli $\partial f(x) = \{g\}$ wtedy i tylko wtedy gdy f różniczkowalna w x oraz $g = \nabla f(x)$
- f osiąga minimum w x^* wtedy i tylko wtedy gdy $0 \in \partial f(x^*)$
- $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$

Spadek po subgradinecie

$$x_k = x_{k-1} - \gamma_k g_{k-1}, \quad g_{k-1} \in \partial f(x_{k-1})$$

Własności: Jeśli f G-Lipshitzowska oraz $\|x_o - x^*\| < R$ to

- Dla $\gamma_k = \gamma$
$$\lim_{k\to\infty} f(x_k^N) \leqslant f(x^*) + G^2\gamma$$
- Jeśli $\gamma_k \to 0, \sum \gamma_k^2 < \infty$ oraz $\sum \gamma_k = \infty$ to
$$\lim_{k\to\infty} f(x_k^N) = f(x^*)$$
- Jeśli $\gamma_k = \frac{R}{G\sqrt{k}}$
$$f(x_k^N) - f(x^*) \leqslant \frac{RG}{\sqrt{k}}$$

Proximal operator

$$\mathsf{prox}_{\gamma,\mathrm{f}}(\mathrm{x}) = \arg\min_{\mathrm{y}}\{\mathrm{f}(\mathrm{y}) + \frac{1}{2\gamma}\|\mathrm{x}-\mathrm{y}\|^2\}$$

Dla dowolnej wypukłej funkcji f jeżeli

$$\mathrm{y} = \mathsf{prox}_{\gamma,\mathrm{f}}(\mathrm{x})$$

to

$$\mathrm{y} \in \mathrm{x} - \gamma\partial\mathrm{f}(\mathrm{y})$$

Przykład:

$$\mathsf{prox}_{\gamma,\lambda|\mathrm{x}|} = \mathrm{sign}(\mathrm{x})(|\mathrm{x}| - \lambda\gamma)_+$$

Proximal gradient

Problem:

$$\arg\min_{x}\{f(x) + g(x)\}$$

- f wypukła i gładka
- g wypukła

Algorytm

$$x_{k+1} = \mathsf{prox}_{\gamma_k, g}(x_k - \gamma_k \nabla f(x_k))$$

Stąd

$$x_{k+1} \in x_k - \gamma_k \nabla f(x_k) - \gamma_k \partial g(x_{k+1})$$

Własności proximal gradient

Jeśli f jest L-gładka oraz $\gamma \leqslant L^{-1}$ to

$$f(x)+g(x) \leqslant f(y)+\nabla f(y)^T(x-y)+\frac{1}{2\gamma}\|y-x\|^2+g(x) = Q_\gamma(x,y)$$

Przekształacając

$$Q_\gamma(x,y) = f(y) - \frac{\gamma}{2}\|\nabla f(y)\|^2 + g(x) + \frac{1}{2\gamma}\|x-(y-\gamma\nabla f(y))\|^2$$

Czyli

$$\mathsf{prox}_{\gamma_k,g}(x_k - \gamma_k \nabla f(x_k)) = \arg\min_x Q_{\gamma_k}(x, x_k)$$

Zatem

$$\begin{aligned} f(x_k) + g(x_k) = Q_{\gamma_k}(x_k, x_k) &\geqslant \arg\min_x Q_{\gamma_k}(x, x_k) \\ &= Q_{\gamma_k}(x_{k+1}, x_k) \geqslant f(x_{k+1}) + g(x_{k+1}) \end{aligned}$$

Zbieżność proximal gradient

Oznaczmy przez $F = f + g$

Lemat

Jeśli f jest wypukła i L-gładka, g jest wypukła oraz $\gamma_k \leqslant L^{-1}$ to dla dowolnego y

$$2\gamma_k(F(x_k) - F(y)) \leqslant \|x_{k-1} - y\|^2 - \|x_k - y\|^2$$

Dowód:

Niech $v \in \partial g(x_k)$ takie, że

$$x_k = x_{k-1} - \gamma_k \nabla f(x_{k-1}) - \gamma_k v$$

Mamy trzy nierówności

$$F(x_k) \leqslant g(x_k) + f(x_{k-1}) + \nabla f(x_{k-1})^T(x_k - x_{k-1}) + \frac{1}{2\gamma_k}\|x_k - x_{k-1}\|^2$$

$$f(y) \geqslant f(x_{k-1}) + \nabla f(x_{k-1})^T(y - x_{k-1})$$

$$g(y) \geqslant g(x_k) + v^T(y - x_k)$$

Dowód lematu c.d.

Odejmując nierówności otrzymujemy

$$F(x_k) - F(y) \leqslant \frac{1}{2\gamma_k}\|x_k - x_{k-1}\|^2 + \nabla f(x_{k-1})^T(x_k - x_{k-1}) \\ - \nabla f(x_{k-1})^T(y - x_{k-1}) - v^T(y - x_k)$$

Przekształcając prawą stronę otrzymujemy

$$F(x_k) - F(y) \leqslant \frac{1}{2\gamma_k}\|x_k - x_{k-1}\|^2 + (\nabla f(x_{k-1}) + v)^T(x_k - y)$$

$$F(x_k) - F(y) \leqslant \frac{1}{2\gamma_k}\|x_k - x_{k-1}\|^2 - \frac{1}{\gamma_k}(x_k - x_{k-1})^T(x_k - y)$$

$$2\gamma_k(F(x_k) - F(y)) \leqslant \|x_k - x_{k-1} - x_k + y\|^2 - \|x_k - y\|^2$$

Zbieżność proximal gradient

TWIERDZENIE

Niech $F = f + g$, gdzi f jest wypukła i L-gładka, g jest wypukła oraz F osiąga minimum w x^*. Jeśli $\gamma_k = \gamma < L^{-1}$ to

$$F(x_k) - F(x^*) \leqslant \frac{L\|x_0 - x^*\|^2}{2k}.$$

Przyspieszenie Nesterova dla proximal gradient

1. $y_0 = x_0, t_0 = 0$, γ_0 i $0 < \eta < 1$
2. $\gamma_k = \gamma_{k-1}$
3. $x_k = \mathsf{prox}_{\gamma_k g}(y_{k-1} - \gamma_k \nabla f(y_{k-1}))$
4. Jeśli
$$f(x_k) \leqslant f(y_{k-1}) + \langle \nabla f(y_{k-1}), x_k - y_{k-1} \rangle + \frac{1}{2\gamma_k}\|x_k - y_{k-1}\|^2$$
to krok 5. W przeciwnym przypadku $\gamma_k = \eta\gamma_k$ i wracamy do kroku 3.
5. $t_k = \frac{1+\sqrt{1+4t_{k-1}^2}}{2}$
6. $y_k = x_k + \frac{t_{k-1}-1}{t_k}(x_k - x_{k-1})$

TWIERDZENIE

Niech f wypukła, L-gładka, g wypukła oraz F osiąga minimum w x^*. Jeśli γ_k jak powyżej to

$$F(x_k) - F(x^*) \leqslant \frac{2L\|x_0 - x^*\|^2}{\eta k^2}.$$

3.36pt

1. Optymalizacja
   - Wstęp
   - Metody gradientowe
   - Algorytm Newtona-Raphsona
   - Iterowane ważone najmniejsze kwadraty
   - Proximal gradient
   - ADMM
   - Expectation Maximization

Alternating Direction Method of Multipliers

Rozważmy problem

$$\min_{x:\ Ax=b} f(x)$$

Lagrangian dla tego problemu

$$\mathcal{L}(x, \mu) = f(x) + \mu^T(Ax - b)$$

Problem dualny

$$\max_{\mu} g(\mu), \quad g(\mu) = \inf_{x} \mathcal{L}(x, \mu)$$

Rozwiązanie

$$x^* = \arg\min_{x} \mathcal{L}(x, \mu^*)$$

Algorytm gradientowy dla problemu dualnego

$$\mu_{k+1} = \mu_k + \gamma_k \nabla g(\mu_k),$$

gdzie

$$\nabla g(\mu_k) = A\tilde{x} - b, \quad \tilde{x} = \arg\min \mathcal{L}(x, \mu_k).$$

Czyli

$$\begin{aligned} x_{k+1} &= \arg\min \mathcal{L}(x, \mu_k) \\ \mu_{k+1} &= \mu_k + \gamma_k (Ax_{k+1} - b) \end{aligned}$$

Algorytm jest zbieżny przy restrykcyjnych założeniach. Nie praktyczny

Metoda mnożników I

Warunki konieczne

$$\mathrm{Ax}^* - \mathrm{b} = 0, \quad \nabla\mathcal{L}(\mathrm{x}^*, \mu^*) = \nabla \mathrm{f}(\mathrm{x}^*) + \mathrm{A}^{\mathrm{T}}\mu^* = 0$$

Chcielibyśmy aby

$$\nabla\mathcal{L}(\mathrm{x}_\mathrm{k}, \mu_\mathrm{k}) = 0$$

Równoważnie, warunki konieczne można zapisać w postaci

$$\mathrm{Ax}^* - \mathrm{b} = 0, \quad \nabla \mathrm{f}(\mathrm{x}^*) + \mathrm{A}^{\mathrm{T}}\mu^* + \rho\mathrm{A}^{\mathrm{T}}(\mathrm{Ax}^* - \mathrm{b}) = 0$$

Co odpowiada

$$\mathrm{Ax}^* - \mathrm{b} = 0, \quad \nabla\mathcal{L}_\rho(\mathrm{x}^*, \mu^*) = 0,$$

gdzie

$$\mathcal{L}_\rho(\mathrm{x}, \mu) = \mathrm{f}(\mathrm{x}) + \mu^{\mathrm{T}}(\mathrm{Ax} - \mathrm{b}) + \frac{\rho}{2}\|\mathrm{Ax} - \mathrm{b}\|^2$$

Metoda mnożników II

Algorytm:

$$\begin{aligned} x_{k+1} &= \arg\min_x \mathcal{L}_\rho(x, \mu_k) \\ \mu_{k+1} &= \mu_k + \rho(Ax_{k+1} - b) \end{aligned}$$

Mamy

$$\begin{aligned} 0 &= \nabla \mathcal{L}_\rho(x_{k+1}, \mu_k) \\ &= \nabla f(x_{k+1}) + A^T \mu_k + \rho A^T (Ax_{k+1} - b) \\ &= \nabla f(x_{k+1}) + A^T (\mu_k + \rho(Ax_{k+1} - b)) \\ &= \nabla f(x_{k+1}) + A^T \mu_{k+1} \end{aligned}$$

II warunek osiągany jest asymptotycznie

$$Ax_k - b \to 0$$

ADMM

Rozważmy problem

$$\min_{\{x,z:\ Ax+Bz=c\}} f(x) + g(z)$$

f,g wypukłe

$$\mathcal{L}_\rho(x, z, \mu) = f(x) + g(z) + \mu^T(Ax + Bz - c) + \frac{\rho}{2}\|Ax + Bz - c\|^2$$

Algorytm

1. $x_{k+1} = \arg\min_x \mathcal{L}_\rho(x, z_k, \mu_k)$
2. $z_{k+1} = \arg\min_z \mathcal{L}_\rho(x_{k+1}, z, \mu_k)$
3. $\mu_{k+1} = \mu_k + \rho(Ax_{k+1} + Bz_{k+1} - c)$

Uwagi

- Kroki 2 i 3 można zastąpić przez rozwiązanie numeryczne.
- Algorytm jest zbieżny jeżeli $\mathcal{L}_0$ ma punkt przegięcia.

ADMM dla LASSO

Problem

$$\min_{\beta} \frac{1}{2}\|\mathrm{Y} - \mathrm{X}\beta\|^2 + \lambda\|\beta\|_1$$

można zapisać jako

$$\min_{\{\beta,\mathrm{z}:\ \beta=\mathrm{z}\}} \frac{1}{2}\|\mathrm{Y} - \mathrm{X}\beta\|^2 + \lambda\|\mathrm{z}\|_1$$

oraz

$$\mathcal{L}_\rho(\beta, \mathrm{z}, \mu) = \frac{1}{2}\|\mathrm{Y} - \mathrm{X}\beta\|^2 + \lambda\|\mathrm{z}\|_1 + \mu^{\mathrm{T}}(\beta - \mathrm{z}) + \frac{\rho}{2}\|\beta - \mathrm{z}\|^2$$

ADMM

1. $\beta_{\mathrm{k}+1} = (\mathrm{X}^{\mathrm{T}}\mathrm{X} + \rho\mathrm{I})^{-1}(\mathrm{X}^{\mathrm{T}}\mathrm{Y} - \mu_{\mathrm{k}} + \rho\mathrm{z}_{\mathrm{k}})$
2. $\mathrm{z}_{\mathrm{k}+1} = \mathsf{prox}_{\frac{\lambda}{\rho}\|\mathrm{z}\|_1}(\beta_{\mathrm{k}+1} + \frac{1}{\rho}\mu_{\mathrm{k}}) =$
   $\mathrm{sign}(\beta_{\mathrm{k}+1} + \frac{1}{\rho}\mu_{\mathrm{k}})(|\beta_{\mathrm{k}+1} + \frac{1}{\rho}\mu_{\mathrm{k}}| - \frac{\lambda}{\rho})_+$
3. $\mu_{\mathrm{k}+1} = \mu_{\mathrm{k}} + \rho(\beta_{\mathrm{k}+1} - \mathrm{z}_{\mathrm{k}+1})$

ADMM dla optymalizacji z ograniczeniami wypukłymi I

Problem:

$$\min_{x \in \mathcal{K}} f(x)$$

f- wypukła, $\mathcal{K}$ domknięty wypukły podzbiór $\mathbb{R}^d$.

$$\min_{x=z} f(x) + 1_{\mathcal{K}}(z)$$

gdzie

$$1_{\mathcal{K}}(z) = \begin{cases} 1 & z \in \mathcal{K} \\ +\infty & z \notin \mathcal{K} \end{cases}$$

$$\mathcal{L}_\rho(x, z, \mu) = f(x) + 1_{\mathcal{K}}(z) + \mu^T(x - z) + \frac{\rho}{2}\|x - z\|^2$$

ADMM

1. $x_{k+1} = \arg\min_x \mathcal{L}_\rho(x, z_k, \mu_k)$
2. $z_{k+1} = \Pi_{\mathcal{K}}(x_{k+1} + \frac{1}{\rho}\mu_k)$
3. $\mu_{k+1} = \mu_k + \rho(x_{k+1} - z_{k+1})$

ADMM dla optymalizacji z ograniczeniami wypukłymi II

Jeżeli f funkcja kwadratowa czyli

$$f(x) = \frac{1}{2}x^T P x + q^T x + r$$

to

$$\arg\min_x \mathcal{L}_\rho(x, z_k, \mu_k) = (P + \rho I)^{-1}(\rho z_k - q - \mu_k)$$

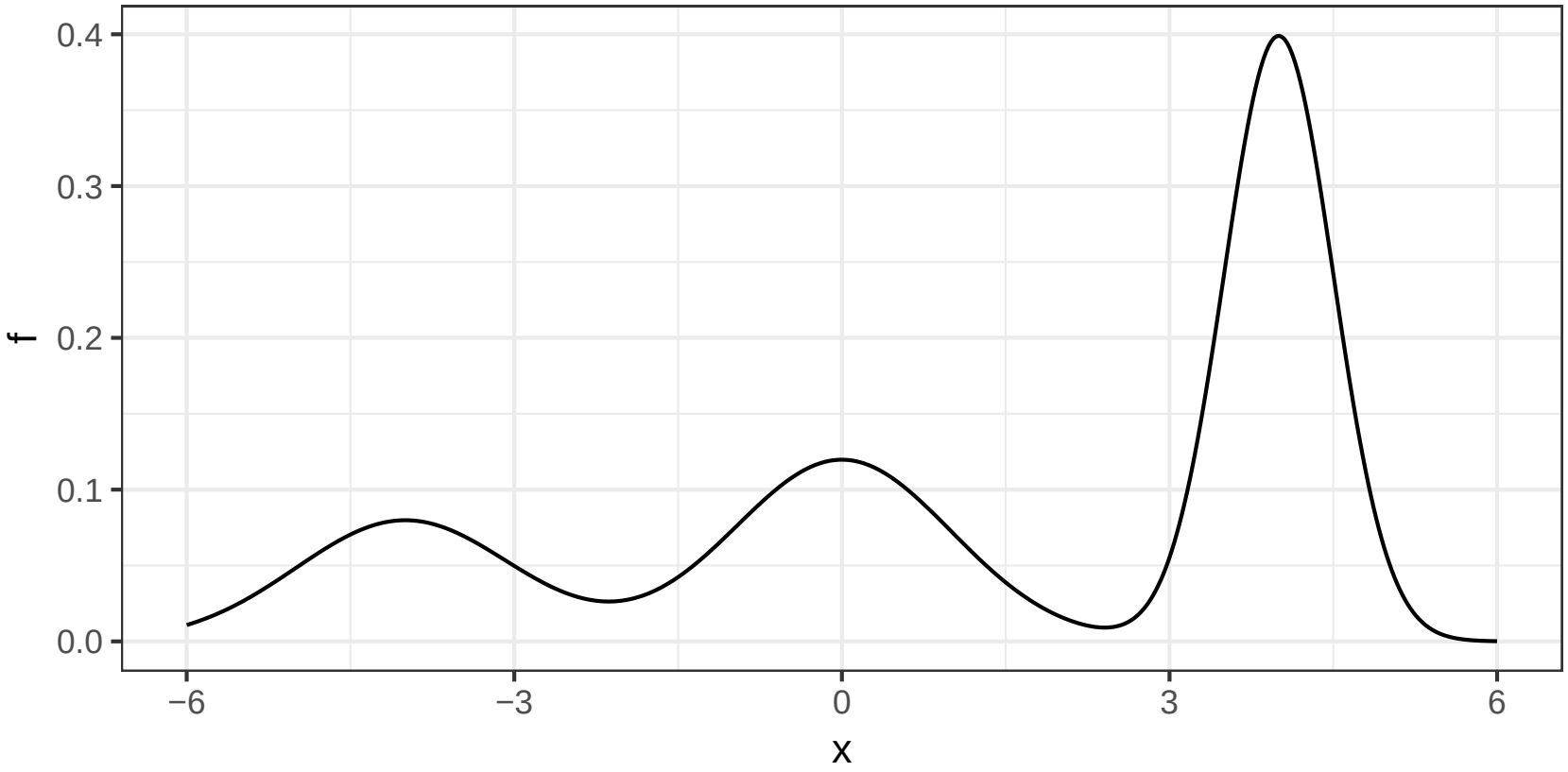
3.36pt

1. Optymalizacja
   - Wstęp
   - Metody gradientowe
   - Algorytm Newtona-Raphsona
   - Iterowane ważone najmniejsze kwadraty
   - Proximal gradient
   - ADMM
   - Expectation Maximization

Mieszanka rozkładów gaussowskich

Niech $Y_1, \ldots, Y_n$ będą zmiennymi losowymi o rozkładzie

$$f_\theta(y) = \sum_{i=1}^{k} \alpha_i \phi_{\mu_i,\sigma_i}(y), \quad \theta = (\alpha_i, \mu_i, \sigma_i)_{i=1}^{k}$$



Jak obliczyć estymator θ?

Mieszanka jako model z brakującymi danymi

Rozważmy parę (X, Y) o rozkładzie

$$\mathrm{P}(\mathrm{X} = \mathrm{i}) = \alpha_{\mathrm{i}}, \quad \mathrm{Y}|\mathrm{X} = \mathrm{i} \sim \mathcal{N}(\mu_{\mathrm{i}}, \sigma_{\mathrm{i}}^2)$$

Wówczas rozkład brzegowy Y ma gęstość

$$\mathrm{f}_\theta(\mathrm{y}) = \sum_{\mathrm{i}=1}^{\mathrm{k}} \alpha_{\mathrm{i}} \phi_{\mu_{\mathrm{i}}, \sigma_{\mathrm{i}}}(\mathrm{y})$$

Expectation Maximization I

Rozkład łączny

$$\mathrm{f}_\theta(\mathrm{x}, \mathrm{y})$$

Obserwujemy Y z rozkładu brzegowego

$$\mathrm{f}_\theta(\mathrm{y}) = \int \mathrm{f}_\theta(\mathrm{x}, \mathrm{y})\mathrm{dx}$$

Chcemy policzyć MLE

$$\arg\max_\theta \ell(\theta), \quad \ell(\theta) = \mathsf{log}(\mathrm{f}(\theta)(\mathrm{y}))$$

Wprowadźmy funkcję pomocniczą

$$\mathrm{Q}(\theta, \vartheta) = \int \mathsf{log}(\mathrm{f}_\theta(\mathrm{x}, \mathrm{y}))\mathrm{f}_\vartheta(\mathrm{x}|\mathrm{y})\mathrm{dx} = \mathbb{E}_\vartheta(\mathsf{log}(\mathrm{f}_\theta(\mathrm{X}, \mathrm{Y}))|\mathrm{Y})$$

Expectation Maximization II

Krok Expectation:

$$\mathrm{Q}(\theta, \theta_{\mathrm{k}}) = \mathbb{E}_{\theta_{\mathrm{k}}}(\mathsf{log}(\mathrm{f}_{\theta}(\mathrm{X}, \mathrm{Y}))|\mathrm{Y})$$

Krok Maximization:

$$\theta_{\mathrm{k}+1} = \mathsf{arg}\,\mathsf{max}_{\theta}\,\mathrm{Q}(\theta, \theta_{\mathrm{k}})$$

Własności funkcji Q

Oznaczmy przez

$$\mathrm{H}(\theta, \vartheta) = -\int \log(\mathrm{f}_\theta(\mathrm{x}|\mathrm{y}))\mathrm{f}_\vartheta(\mathrm{x}|\mathrm{y})\mathrm{dx}$$

1. $\mathrm{Q}(\theta, \vartheta) = \ell(\theta) - \mathrm{H}(\theta, \vartheta)$
2. $\ell(\theta) - \ell(\vartheta) \geqslant \mathrm{Q}(\theta, \vartheta) - \mathrm{Q}(\vartheta, \vartheta)$
3. $\nabla_\theta \mathrm{Q}(\theta, \vartheta)|_{\theta=\vartheta} = \nabla \ell(\vartheta)$

Powyższe własności wynikają z faktu, że

$$\begin{aligned}\mathrm{H}(\theta, \vartheta) - \mathrm{H}(\vartheta, \vartheta) &= -\int \log(\mathrm{f}_\theta(\mathrm{x}|\mathrm{y}))\mathrm{f}_\vartheta(\mathrm{x}|\mathrm{y}) + -\int \log(\mathrm{f}_\vartheta(\mathrm{x}|\mathrm{y}))\mathrm{f}_\vartheta(\mathrm{x}|\mathrm{y}) \\ &= -\int \log\left(\frac{\mathrm{f}_\theta(\mathrm{x}|\mathrm{y})}{\mathrm{f}_\vartheta(\mathrm{x}|\mathrm{y})}\right)\mathrm{f}_\vartheta(\mathrm{x}|\mathrm{y})\mathrm{dx} \geqslant 0\end{aligned}$$

EM dla rodziny wykladniczej

Niech

$$\mathrm{f}_\theta(\mathrm{x},\mathrm{y}) = \mathsf{exp}\{\theta^{\mathrm{T}}\mathrm{S}(\mathrm{x},\mathrm{y}) - \mathrm{b}(\theta)\}$$

Zatem

$$\mathrm{Q}(\theta,\vartheta) = \mathbb{E}_\vartheta(\mathsf{log}(\mathrm{f}_\theta(\mathrm{X},\mathrm{Y}))|\mathrm{Y}) = \theta^{\mathrm{T}}\mathbb{E}_\vartheta(\mathrm{S}(\mathrm{X},\mathrm{Y})|\mathrm{Y}) - \mathrm{b}(\theta)$$

EM dla mieszanek I

$Y_1, \ldots, Y_n$ z rozkładu

$$f_\theta(y) = \sum_{i=1}^{k} \alpha_k \phi_{\mu_i,\sigma_i}(y)$$

Łączny rozkład $P(X_j = i) = \alpha_i$, $f_\theta(y_j|x_j = i) = \phi_{\mu_i,\sigma_i}(y_j)$

$$f_\theta(y, x) = \prod_{j=1}^{n} \prod_{i=1}^{k} [\alpha_i \phi_{\mu_i,\sigma_i}(y_j)]^{1(x_j=i)}$$

Zatem

$$Q(\theta, \theta') = \sum_{j=1}^{n} \sum_{i=1}^{k} \mathbb{E}_{\theta'}(1(X_j = i)|Y_j = y_j)[\log(\alpha_i) + \log(\phi_{\mu_i,\sigma_i}(y_j))]$$

EM dla mieszanek II

Oznaczmy przez
$w_{ij} = \mathbb{E}_{\theta'}(1(X_j = i)|Y_j = y_j) = \mathbb{P}_{\theta'}(X_j = i|Y_j = y_j)$ Ze wzoru Bayesa mamy

$$w_{ij} = \frac{\alpha'_i \phi_{\mu'_i,\sigma'_i}(y_j)}{\sum_{i=1}^{k} \alpha'_i \phi_{\mu'_i,\sigma'_i}(y_j)}$$

Stąd

$$Q(\theta, \theta') = \sum_{j=1}^{n} \sum_{i=1}^{k} w_{ij}[\log(\alpha_i) + \log(\phi_{\mu_i,\sigma_i}(y_j))]$$

Maksymalizacja po α:
Chcemy zmaksymalizować $Q(\theta, \theta')$ względem α przy ograniczeniu $\sum \alpha_i = 1$. Lagrangian dla tego problemu to

$$\mathcal{L}(\alpha, \lambda) = \sum_{j=1}^{n} \sum_{i=1}^{k} w_{ij} \log(\alpha_j) - \lambda(\sum_{i=1}^{k} \alpha_j - 1)$$

EM dla mieszanek III

Stąd

$$\frac{\partial \mathcal{L}(\alpha, \lambda)}{\partial \alpha_j} = \frac{\sum_{j=1}^{n} w_{ij}}{\alpha_j} - \lambda = 0$$

Czyli

$$\hat{\alpha}_i = \frac{\sum_{j=1}^{n} w_{ij}}{n}$$

Maksymalizacja po μ i σ^2:

$$g(\mu, \sigma^2) = \sum_{j=1}^{n} \sum_{i=1}^{k} w_{ij} \left[\frac{-(y_j - \mu_i)^2}{2\sigma_i^2} - \frac{1}{2} \log(\sigma_i^2) \right]$$

$$\frac{\partial g}{\partial \mu_i} = \sum_{j=1}^{n} w_{ij} \frac{y_j - \mu_i}{\sigma_i^2} = 0$$

To

$$\hat{\mu}_i = \frac{\sum_{j=1}^{n} w_{ij} y_j}{\sum_{j=1}^{n} w_{ij}}$$

EM dla mieszanek IV

oraz $(\tau_i = \sigma_i^2)$

$$\frac{\partial g}{\partial \tau_i} = \sum_{j=1}^{n} w_{ij} \left[\frac{(y_j - \mu_i)^2}{\tau_i^2} - \frac{1}{2\tau_i} \right] = 0$$

Czyli

$$\hat{\sigma}_i^2 = \frac{\sum_{j=1}^{n} w_{ij} (y_j - \hat{\mu}_i)^2}{\sum_{j=1}^{n} w_{ij}}$$

Podsumowanie EM dla mieszanek gaussowskich

Krok E:

$$w_{ij}^{[t+1]} = \frac{\alpha_i^{[t]} \phi_{\mu_i^{[t]}, \sigma_i^{[t]}}(y_j)}{\sum_{i=1}^{k} \alpha_i^{[t]} \phi_{\mu_i^{[t]}, \sigma_i^{[t]}}(y_j)}$$

Krok M:

$$\alpha_i^{[t+1]} = \frac{\sum_{j=1}^{n} w_{ij}^{[t+1]}}{n}$$

$$\mu_i^{[t+1]} = \frac{\sum_{j=1}^{n} w_{ij}^{[t+1]} y_j}{\sum_{j=1}^{n} w_{ij}^{[t+1]}}$$

$$(\sigma_i^2)^{[t+1]} = \frac{\sum_{j=1}^{n} w_{ij}^{[t+1]} (y_j - \mu_i^{[t+1]})^2}{\sum_{j=1}^{n} w_{ij}^{[t+1]}}$$

EM dla wielowymiarowych gaussowskich

Krok E:

$$w_{ij}^{[t+1]} = \frac{\alpha_i^{[t]} \phi_{\mu_i^{[t]}, \Sigma_i^{[t]}}(y_j)}{\sum_{i=1}^{k} \alpha_i^{[t]} \phi_{\mu_i^{[t]}, \Sigma_i^{[t]}}(y_j)}$$

Krok M:

$$\alpha_i^{[t+1]} = \frac{\sum_{j=1}^{n} w_{ij}^{[t+1]}}{n}$$

$$\mu_i^{[t+1]} = \frac{\sum_{j=1}^{n} w_{ij}^{[t+1]} y_j}{\sum_{j=1}^{n} w_{ij}^{[t+1]}}$$

$$\Sigma_i^{[t+1]} = \frac{\sum_{j=1}^{n} w_{ij}^{[t+1]} (y_j - \mu_i^{[t+1]})(y_j - \mu_i^{[t+1]})^T}{\sum_{j=1}^{n} w_{ij}^{[t+1]}}$$

Uwagi: Estymowanie k macierzy kowariancji jest trudne (szczególnie dla większego wymiaru). W takiej sytuacji najczęściej zakłada się, że macierz kowariancji jest wspólna dla wszystkich rozkładów lub zakłada się szczególną postać macierzy kowariancji: $\Sigma_i = \sigma_i^2 I$, Σ_i diagonalna, itp.

EM jako proximal algorytm

Algorytm EM można zapisać w postaci

$$\theta_{\mathrm{k}+1} = \arg\max_{\theta} \mathrm{Q}(\theta, \theta_{\mathrm{k}})$$

Ponieważ

$$\mathrm{Q}(\theta, \theta_{\mathrm{k}}) + \mathrm{H}(\theta_{\mathrm{k}}, \theta_{\mathrm{k}}) = \ell(\theta) - \mathrm{KL}(\mathrm{f}_{\theta_{\mathrm{k}}}(\mathrm{x}|\mathrm{y})||\mathrm{f}_{\theta}(\mathrm{x}|\mathrm{y}))$$

Czyli EM równoważnie można zapisać jako

$$\theta_{\mathrm{k}+1} = \arg\min_{\theta}\{-\ell(\theta) + \mathrm{KL}(\mathrm{f}_{\theta_{\mathrm{k}}}(\mathrm{x}|\mathrm{y})||\mathrm{f}_{\theta}(\mathrm{x}|\mathrm{y}))\}$$

Podsumowanie I

1. Spadek po gradiencie
   1. Wymaga obliczania tylko gradinetu, koszt kroku $\mathcal{O}(\mathrm{dn})$
   2. Dla funkcji wypukłych i L gładkich tempo zbieżności $\mathcal{O}(\mathrm{k}^{-1})$
   3. Przyspieszenie Nesterova poprawia tempo zbieżności do $\mathcal{O}(\mathrm{k}^{-2})$
2. Algorytm Newtona-Raphsona
   1. Wymaga obliczania odwrotności hessianu, koszt kroku $\mathcal{O}(\mathrm{d}^2\mathrm{n})$.
   2. Lokalna kwadratowa zbieżność, wrażliwy na punkt startowy.
   3. W przypadku GLM, można zapisać jako iterowane ważone najmniejsze kwadraty.
3. Proximal gradient
   1. Pozwala obliczać funkcję niegładkie o ile obliczenie kroku jest możliwe.
   2. Takie same tempo zbieżności co spadek po gradiencie. Istnieje przypsieszenie Nesterova.

Podsumowanie II

4. ADMM
   1. Również pozwala na niegładkie funkcje.
   2. Często szybszy od proximal gradient w szczególności gdy poszczególne kroki można wyznaczyć analitycznie.
5. EM
   1. Pozwala efektywnie obliczać estymatory dla modeli z ukrytymi zmiennymi (brakującymi danymi).
   2. Krok E polega na imputacji ukrytych zmiennych.
   3. EM jest monotoniczny. Dla funckji wypukłych zbiega nie wolniej niż spadek po gradiencie.

3.36pt

2. Ukryte modele Markowa
   - Dyskretne modele Markowa
   - Obliczanie prawdopodobieństwa obserwacji
   - Forward Backward
   - Algorytm Viterbiego
   - Algorytm FFBS
   - Algorytm Bauma-Welcha
   - Modele gaussowskie i filtr Kalmana

3. Algorytmy MCMC
   - Podstawowe pojęcia statystyki bayesowskiej
   - Metody Monte Carlo
   - Algorytm Metropolisa-Hastingsa
   - Próbnik Gibbsa
   - Własności teoretyczne MCMC
   - Adaptacyjne algorytmy MCMC
   - Hamiltonian Monte Carlo

Łańcuchy Markowa

$\mathcal{S}$ - dyskretna przestrzeń stanów

Definicja

Ciąg zmiennych losowych $X_1, \ldots, X_n, \ldots$ jest łańcuchem Markowa wtedy i tylko wtedy gdy

$$\mathbb{P}(X_k = s_k | X_{k-1} = s_{k-1}, X_{k-2}, \ldots X_1) = \mathbb{P}(X_k = s_k | X_{k-1} = s_{k-1})$$

łańcuch Markowa jest jednorodny jeżeli

$$\mathbb{P}(X_k = s_k | X_{k-1} = s_{k-1}) = \mathbb{P}(X_2 = s_k | X_1 = s_{k-1}).$$

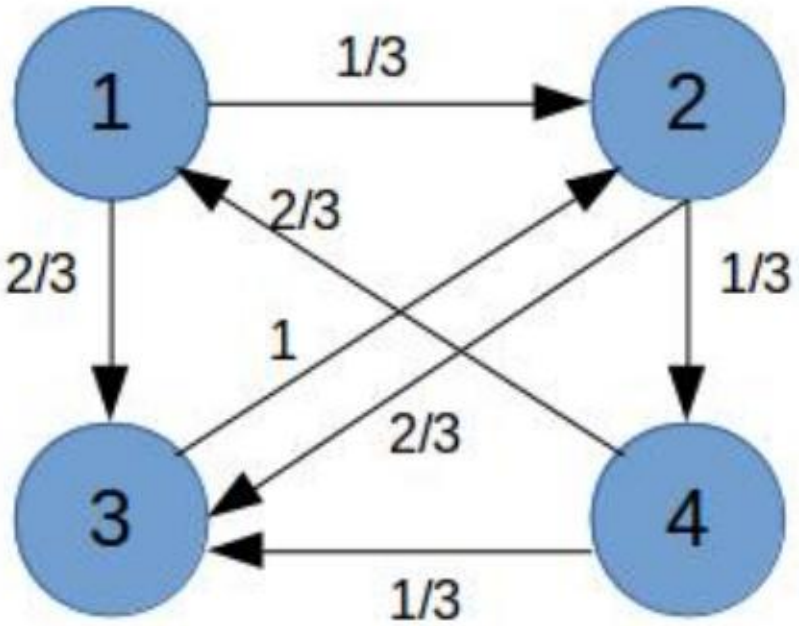
Macierz przejścia

$$\mathbb{P}(X_{k+1} = i | X_k = j) = P_k(j, i)$$

- Jeśli łańcuch jest jednorodny to dla każdego k $P_k = P$
- $\sum_{i \in \mathcal{S}} P_k(j, i) = 1$
- Jeżeli $X_k \sim v$ to
$$X_{k+1} \sim vP_k$$
- $$\mathbb{E}(f(X_{k+1} | X_k = s) = (P_k f)(s)$$
- Równanie Chapmana - Kołmogorowa

$$\mathbb{P}(X_{k+2} | X_k = s) = \sum_{s'} \mathbb{P}(X_{k+2} | X_{k+1} = s') \mathbb{P}(X_{k+1} = s' | X_k = s)$$

Przykład



$$P = \begin{bmatrix} 0 & 1/3 & 2/3 & 0 \\ 0 & 0 & 2/3 & 1/3 \\ 0 & 1 & 0 & 0 \\ 2/3 & 0 & 1/3 & 0 \end{bmatrix}$$

Łańcuchy Markowa na ciągłej przestrzeni stanów

Jądro przejścia

$$\mathbb{P}(X_{k+1} \in dy | X_k = x) = P_k(x, dy)$$

- $P(x, \cdot)$ - miara probabilistyczna
- $P(\cdot, A)$ - funkcja mierzalna
- $$\mu P(dy) = \int P(x, dy)\mu(dx)$$
- $$Pf(x) = \int f(y)P(x, dy) = \mathbb{E}(f(X_2)|X_1 = x)$$

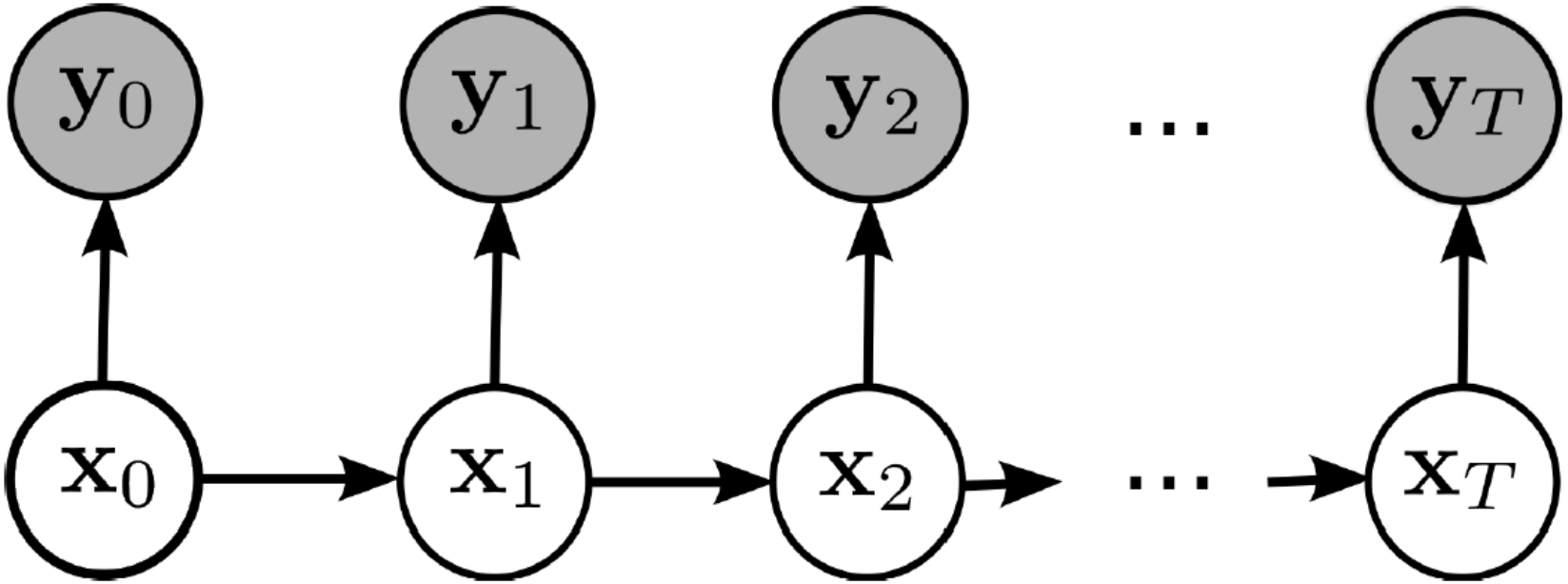
Przykłady

- Proces autoregresji

$$X_{k+1} = \rho X_k + Z_k$$

- Niech $Y = \rho X_k + Z_k$ i z prawdopodobieństwem $\alpha(X_k, Y)$ $X_{k+1} = Y$ z pozostałym prawdopodobieństwem $X_{k+1} = X_k$

Ukryte modele Markowa



1. Policzyć prawdopodobieństwo obserwacji
2. Jak znaleźć najbardziej prawdopodobną ukrytą trajektorię?
3. Jak generować ukrytą trajektorię?
4. Jak wyestymować parametry modelu?

Ukryte modele Markowa

$$\mathbb{P}(X_{k+1} = i | X_k = j) = p(j, i)$$
$$\mathbb{P}(Y_k = y | X_k = j) = g(j, y)$$
$$\mathbb{P}(X_1 = x) = q(x)$$

- Rozkład łączny

$$\mathbb{P}(X_{1:T} = x_{1:T}, Y_{1:T} = y_{1:T}) = q(x_1)g(x_1, y_1) \prod_{k=2}^{T} [g(x_k, y_k)p(x_{k-1}, x_k)]$$

- Rozkład wygładzania (smoohting)

$$\mathbb{P}(X_k = x_k | Y_{1:T} = y_{1:T}) = \frac{\mathbb{P}(X_k = x_k, Y_{1:T} = y_{1:T})}{\mathbb{P}Y_{1:T} = y_{1:T})}$$

- Rozkład filtru (filtering)

$$\mathbb{P}(X_T = x_T | Y_{1:T} = y_{1:T}) = \frac{\mathbb{P}(X_T = x_T, Y_{1:T} = y_{1:T})}{\mathbb{P}Y_{1:T} = y_{1:T})}$$

3.36pt

2. Ukryte modele Markowa
   - Dyskretne modele Markowa
   - Obliczanie prawdopodobieństwa obserwacji
   - Forward Backward
   - Algorytm Viterbiego
   - Algorytm FFBS
   - Algorytm Bauma-Welcha
   - Modele gaussowskie i filtr Kalmana

3. Algorytmy MCMC
   - Podstawowe pojęcia statystyki bayesowskiej
   - Metody Monte Carlo
   - Algorytm Metropolisa-Hastingsa
   - Próbnik Gibbsa
   - Własności teoretyczne MCMC
   - Adaptacyjne algorytmy MCMC
   - Hamiltonian Monte Carlo

Obliczanie prawdopodobieństwa obserwacji

$$\mathbb{P}(Y_{1:T} = y_{1:T}) = \sum_{x_{1:T}} \mathbb{P}(Y_{1:T} = y_{1:T}, X_{1:T} = x_{1:T})$$

Oznaczmy przez

$$\alpha_t(k) = \mathbb{P}(X_t = k, Y_{1:t} = y_{1:t})$$

Wówczas

$$\mathbb{P}(Y_{1:T} = y_{1:T}) = \sum_k \alpha_T(k)$$

oraz

$$\begin{aligned}
\alpha_1(k) &= q(k)g(k, y_1) \\
\alpha_t(k) &= \mathbb{P}(X_t = k, Y_{1:T} = y_{1:T}) \\
&= \sum_l \mathbb{P}(X_{t-1} = l, X_t = k, Y_{1:T} = y_{1:T}) \\
&= \sum_l \alpha_{t-1}(l)p(l, k)g(k, y_t)
\end{aligned}$$

Przykład

Mamy dwie monety S -symetryczną i N - niesymetryczną z prawdopodobieństwem O = 0.8. Rzucający może w sposób niezuważalny dla nas podmienić monetę, robi to z prawdopodobieństwem 0.3. Obserwujemy wyniki kolejnych rzutów.

$$\mathrm{O}, \mathrm{O}, \mathrm{O}, \mathrm{R}$$

Przykład

Mamy dwie monety S -symetryczną i N - niesymetryczną z prawdopodobieństwem O = 0.8. Rzucający może w sposób niezuważalny dla nas podmienić monetę, robi to z prawdopodobieństwem 0.3. Obserwujemy wyniki kolejnych rzutów.

$$\mathrm{O}, \mathrm{O}, \mathrm{O}, \mathrm{R}$$

	α_1	α_2	α_3	α_4
S	0,25	0,1475	0,094225	0,06214475
N	0,4	0,284	0,19444	0,0328751

$$\mathbb{P}(\mathrm{O}, \mathrm{O}, \mathrm{O}, \mathrm{R}) = 0,095$$

3.36pt

2. Ukryte modele Markowa
   - Dyskretne modele Markowa
   - Obliczanie prawdopodobieństwa obserwacji
   - Forward Backward
   - Algorytm Viterbiego
   - Algorytm FFBS
   - Algorytm Bauma-Welcha
   - Modele gaussowskie i filtr Kalmana

3. Algorytmy MCMC
   - Podstawowe pojęcia statystyki bayesowskiej
   - Metody Monte Carlo
   - Algorytm Metropolisa-Hastingsa
   - Próbnik Gibbsa
   - Własności teoretyczne MCMC
   - Adaptacyjne algorytmy MCMC
   - Hamiltonian Monte Carlo

Forward Backward I

$$x_t^* = \arg\max_{x_t} \mathbb{P}(X_t = x_t | Y_{1:T} = y_{1:T})$$

Równoważnie

$$x_t^* = \arg\max_{x_t} \mathbb{P}(X_t = x_t, Y_{1:T} = y_{1:T})$$

Oznaczmy przez

$$\beta_t(k) = \mathbb{P}(Y_{t+1:T} | X_t = k), \quad \beta_T(k) = 1$$

Wówczas

$$\begin{aligned}\mathbb{P}(X_t = x_t, Y_{1:T} = y_{1:T}) = \mathbb{P}(X_t = x_t, Y_{1:t} = y_{1:t})\mathbb{P}(Y_{t+1:T} | X_t = x) \\ = \alpha_t(x)\beta_t(x)\end{aligned}$$

Forward Backward II

Dodatkowo

$$
\begin{aligned}
\beta_t(k) &= \mathbb{P}(Y_{t+1:T}|X_t = k) \\
&= \sum_l \mathbb{P}(Y_{t+1:T}, X_{t+1} = l|X_t = k) \\
&= \sum_l \mathbb{P}(X_{t+1} = l|X_t = k)\mathbb{P}(Y_{t+2:T}, |X_{t+1} = l)\mathbb{P}(Y_{t+1}|X_{t+1} = l) \\
&= \sum_l p(k,l)g(l, y_{t+1})\beta_{t+1}(l)
\end{aligned}
$$

Przykład

Mamy dwie monety S -symetryczną i N - niesymetryczną z prawdopodobieństwem O = 0.8. Rzucający może w sposób niezuważalny dla nas podmienić monetę, robi to z prawdopodobieństwem 0.3. Obserwujemy wyniki kolejnych rzutów.

$$\mathrm{O}, \mathrm{O}, \mathrm{O}, \mathrm{R}$$

Przykład

Mamy dwie monety S -symetryczną i N - niesymetryczną z prawdopodobieństwem O = 0.8. Rzucający może w sposób niezuważalny dla nas podmienić monetę, robi to z prawdopodobieństwem 0.3. Obserwujemy wyniki kolejnych rzutów.

$$\mathrm{O}, \mathrm{O}, \mathrm{O}, \mathrm{R}$$

	α_1	α_2	α_3	α_4
S	0,25	0,1475	0,094225	0,06214475
N	0,4	0,284	0,19444	0,0328751

	β_1	β_2	β_3	β_4
S	0,128321	0,21315	0,41	1
N	0.157349	0,2239	0,29	1

3.36pt

2. Ukryte modele Markowa
   - Dyskretne modele Markowa
   - Obliczanie prawdopodobieństwa obserwacji
   - Forward Backward
   - Algorytm Viterbiego
   - Algorytm FFBS
   - Algorytm Bauma-Welcha
   - Modele gaussowskie i filtr Kalmana

3. Algorytmy MCMC
   - Podstawowe pojęcia statystyki bayesowskiej
   - Metody Monte Carlo
   - Algorytm Metropolisa-Hastingsa
   - Próbnik Gibbsa
   - Własności teoretyczne MCMC
   - Adaptacyjne algorytmy MCMC
   - Hamiltonian Monte Carlo

Algorytm Viterbiego

$$x^*_{1:T} = \arg\max_{x_{1:T}} \mathbb{P}(X_{1:T} = x_{1:T} | Y_{1:T} = y_{1:T})$$

Wystarczy

$$x^*_{1:T} = \arg\max_{x_{1:T}} \mathbb{P}(X_{1:T} = x_{1:T}, Y_{1:T} = y_{1:T})$$

Zdefiniujmy:

$$v_1(j) = \mathbb{P}(X_1 = j, Y_1 = y_1) = q(j)g(j, y_1)$$
$$v_t(j) = \max_{x_{1:(t-1)}} \mathbb{P}(X_{1:(t-1)} = x_{1:(t-1)}, x_t = j, Y_{1:t} = y_{1:t})$$
$$\Psi_t(j) = \arg\max_{x_{1:(t-1)}} \mathbb{P}(X_{1:(t-1)} = x_{1:(t-1)}, x_t = j, Y_{1:t} = y_{1:t})$$

Przy tych oznaczeniach jeśli $j^* = \arg\max_j v_T(j)$ to

$$x^*_{1:T} = [\Psi_T(j^*), j^*]$$

Obliczanie v_t

$$v_t(j) = g(j, y_t) \max_k [v_{t-1}(k)p(k, j)]$$

Przykład

Mamy dwie monety S -symetryczną i N - niesymetryczną z prawdopodobieństwem O = 0.8. Rzucający może w sposób niezuważalny dla nas podmienić monetę, robi to z prawdopodobieństwem 0.3. Obserwujemy wyniki kolejnych rzutów.

$$\mathrm{O}, \mathrm{O}, \mathrm{O}, \mathrm{R}$$

Przykład

Mamy dwie monety S -symetryczną i N - niesymetryczną z prawdopodobieństwem O = 0.8. Rzucający może w sposób niezuważalny dla nas podmienić monetę, robi to z prawdopodobieństwem 0.3. Obserwujemy wyniki kolejnych rzutów.

$$\mathrm{O}, \mathrm{O}, \mathrm{O}, \mathrm{R}$$

$$\mathrm{N}, \mathrm{N}, \mathrm{N}, \mathrm{S}$$

3.36pt

2. Ukryte modele Markowa
   - Dyskretne modele Markowa
   - Obliczanie prawdopodobieństwa obserwacji
   - Forward Backward
   - Algorytm Viterbiego
   - Algorytm FFBS
   - Algorytm Bauma-Welcha
   - Modele gaussowskie i filtr Kalmana

3. Algorytmy MCMC
   - Podstawowe pojęcia statystyki bayesowskiej
   - Metody Monte Carlo
   - Algorytm Metropolisa-Hastingsa
   - Próbnik Gibbsa
   - Własności teoretyczne MCMC
   - Adaptacyjne algorytmy MCMC
   - Hamiltonian Monte Carlo

Forward Filtering Backward Sampling

Cel: Generowanie $X_{1:T}|Y_{1:T}$
Przypomnijmy, że

$$\alpha_t(j) = \mathbb{P}(X_t = j, Y_{1:t} = y_{1:t})$$

Stąd

$$\mathbb{P}(X_T = x|Y_{1:T}) = \frac{\alpha_T(x)}{\sum_j \alpha_T(j)}$$

oraz

$$\mathbb{P}(X_{t-1}|X_t, Y_{1:t}) \propto \mathbb{P}(X_{t-1}, X_t, Y_{1:t})$$

gdzie

$$\begin{aligned}
&\mathbb{P}(X_{t-1}, X_t, Y_{1:t}) \\
&= \alpha_{t-1}(X_{t-1})\mathbb{P}(X_t|X_{t-1}, Y_{1:t-1})\mathbb{P}(Y_{t:T}|X_t, X_{t-1}, Y_{1:t-1}) \\
&\propto \alpha_{t-1} p(X_{t-1}, X_t)
\end{aligned}$$

3.36pt

2. Ukryte modele Markowa
   - Dyskretne modele Markowa
   - Obliczanie prawdopodobieństwa obserwacji
   - Forward Backward
   - Algorytm Viterbiego
   - Algorytm FFBS
   - Algorytm Bauma-Welcha
   - Modele gaussowskie i filtr Kalmana

3. Algorytmy MCMC
   - Podstawowe pojęcia statystyki bayesowskiej
   - Metody Monte Carlo
   - Algorytm Metropolisa-Hastingsa
   - Próbnik Gibbsa
   - Własności teoretyczne MCMC
   - Adaptacyjne algorytmy MCMC
   - Hamiltonian Monte Carlo

Algorytm Bauma-Welcha I

$$\theta = [\mathrm{q(j), p(i,j), g(j,y)}]$$

I chcemy znalezc

$$\arg\max_{\theta} \mathbb{P}(\mathrm{Y}_{1:\mathrm{T}})$$

Algorytm EM

Excpectation:

$$\begin{aligned}
\mathbb{P}(\mathrm{X}_{1:\mathrm{T}}, \mathrm{Y}_{1:\mathrm{T}}) &= \mathrm{q}(\mathrm{x}_1)\mathrm{g}(\mathrm{x}_1, \mathrm{y}_1) \prod_{\mathrm{t}=2}^{\mathrm{T}} \mathrm{p}(\mathrm{x}_{\mathrm{t}-1}, \mathrm{x}_{\mathrm{t}})\mathrm{g}(\mathrm{x}_{\mathrm{t}}, \mathrm{y}_{\mathrm{t}}) \\
&= \prod_{\mathrm{j}} \mathrm{q}(\mathrm{j})^{1(\mathrm{x}_1=\mathrm{j})} \prod_{\mathrm{i}} \prod_{\mathrm{j}} \mathrm{p}(\mathrm{i},\mathrm{j})^{\sum_{\mathrm{t}=2}^{\mathrm{T}} 1(\mathrm{x}_{\mathrm{t}-1}=\mathrm{i}, \mathrm{x}_{\mathrm{t}}=\mathrm{j})} \\
&\quad \times \prod_{\mathrm{j}} \prod_{\mathrm{y}} \mathrm{g}(\mathrm{j},\mathrm{y})^{\sum_{\mathrm{t}=1}^{\mathrm{T}} 1(\mathrm{x}_{\mathrm{t}}=\mathrm{j})1(\mathrm{y}_{\mathrm{t}}=\mathrm{y})}
\end{aligned}$$

Algorytm Bauma-Welcha II

Stąd

$$
\begin{aligned}
Q(\theta, \theta') = &\sum_{j} \mathbb{P}_{\theta'}(X_1 = j | Y_{1:T}) \log(q(j)) \\
&+ \sum_{i} \sum_{j} \left[\sum_{t=2}^{T} \mathbb{P}_{\theta'}(X_{t-1} = i, X_t = j | Y_{1:T}) \right] \log(p(i,j)) \\
&+ \sum_{y} \sum_{j} \left[\sum_{t=1}^{T} 1(y_t = y) \mathbb{P}_{\theta'}(X_t = j | Y_{1:T}) \right] \log(g(j,y))
\end{aligned}
$$

Ponadto

$$
\gamma_t(j) = \mathbb{P}(X_t = j | Y_{1:T}) = \frac{\alpha_t(j)\beta_t(j)}{\sum_l \alpha_t(l)\beta_t(l)}
$$

Algorytm Bauma-Welcha III

oraz

$$
\begin{aligned}
\xi_t(i,j) &= \mathbb{P}(X_{t-1} = i, X_t = j | Y_{1:T}) \\
&= \mathbb{P}(X_{t-1} = i | Y_{1:T}) \mathbb{P}(X_t = j | X_{t-1} = i, Y_{t:T}) \\
&= \frac{\gamma_{t-1}(i) \mathbb{P}(X_t = j, Y_{t:T} | X_{t-1} = i)}{\mathbb{P}(Y_{t:T} | X_{t-1} = i)} \\
&= \frac{\gamma_{t-1}(i) \mathbb{P}(X_t = j | X_{t-1} = i) \mathbb{P}(Y_t | X_t = j) \mathbb{P}(Y_{t+1:T} | X_t = j)}{\beta_{t-1}(i)} \\
&= \frac{\gamma_{t-1}(i) p(i,j) g(j, y_t) \beta_t(j)}{\beta_{t-1}(i)}
\end{aligned}
$$

Algorytm Bauma-Welcha IV

Maximization:

$$\max_{q} \sum_{j} \gamma_1(j) \log(q(j))), \qquad \sum_{j} q(j) = 1$$

$$\max_{p} \sum_{i} \sum_{j} \sum_{t=2}^{T} \xi_t(i,j) \log(p(i,j)), \qquad \sum_{j} p(i,j) = 1$$

$$\max_{g} \sum_{j} \sum_{y} \sum_{t=1}^{T} 1(y_t = y)\gamma_t(j) \log(g(j,y)), \qquad \sum_{y} g(j,y) = 1$$

Algorytm Bauma-Welcha V

Czyli

$$\hat{q}(j) = \frac{\gamma_1(j)}{\sum_j \gamma_1(j)}$$

$$\hat{p}(i,j) = \frac{\sum_{t=2}^{T} \xi_t(i,j)}{\sum_j \sum_{t=2}^{T} \xi_t(i,j)}$$

$$\hat{g}(j,y) = \frac{\sum_{t=1}^{T} 1(y_t = y)\gamma_t(j)}{\sum_{t=1}^{T} \gamma_j(t)}$$

3.36pt

2. Ukryte modele Markowa
   - Dyskretne modele Markowa
   - Obliczanie prawdopodobieństwa obserwacji
   - Forward Backward
   - Algorytm Viterbiego
   - Algorytm FFBS
   - Algorytm Bauma-Welcha
   - Modele gaussowskie i filtr Kalmana

3. Algorytmy MCMC
   - Podstawowe pojęcia statystyki bayesowskiej
   - Metody Monte Carlo
   - Algorytm Metropolisa-Hastingsa
   - Próbnik Gibbsa
   - Własności teoretyczne MCMC
   - Adaptacyjne algorytmy MCMC
   - Hamiltonian Monte Carlo

Linear state space model

$$\begin{aligned} X_k &= F_k X_{k-1} + B_k u_k + w_k \\ Y_k &= H_k X_k + v_k \end{aligned}$$

$$w_k \sim N(0, Q_k), \quad v_k \sim N(0, R_k), \; X_1 \sim N(\mu_0, \Sigma_0)$$

Rozkład filtracji i predykcji

$$p(x_{k+1}|Y_{1:k}) \sim N(\hat{X}_{k+1|k}, \Sigma_{k+1|k}), \quad p(x_k|Y_{1:k}) \sim N(\hat{X}_{k|k}, \Sigma_{k|k})$$

Stwierdzenie

Niech $X \sim N(\mu_x, \Sigma_x)$, $V \sim N(0, V)$ niezależne oraz

$$Y = HX + V$$

Wówczas

$$\begin{aligned}\mathbb{E}[X|Y] &= \mathbb{E}[X] + \mathrm{Cov}(X, Y)\mathrm{Cov}(Y)^{-1}(Y - E[Y]) \\ &= \mu_x + \Sigma_X H^T \left[H\Sigma_x H^T + \Sigma_V\right]^{-1} (Y - H\mu_X)\end{aligned}$$

oraz

$$\begin{aligned}\mathrm{Cov}(X|Y) &= \mathrm{Cov}(X - \mathbb{E}[X|Y]) = \mathbb{E}[(X - \mathbb{E}[X|Y])X^T] \\ &= \Sigma_X - \Sigma_X H^T \left[H\Sigma_X H^T + \Sigma_V\right]^{-1} H\Sigma_X\end{aligned}$$

Dowód:

Oznaczmy prawą stronę przez $\hat{X}$ i mamy

$$X - \hat{X} = X - \mathbb{E}[X] + \mathrm{Cov}(X, Y)\mathrm{Cov}(Y)^{-1}(Y - E[Y])$$

Stąd

$$\mathrm{Cov}(X - \hat{X}, Y) = \mathrm{Cov}(X, Y) - \mathrm{Cov}(X, Y)\mathrm{Cov}(Y)^{-1}\mathrm{Cov}(Y) = 0$$

Czyli $X - \hat{X}$ są niezależne oraz $\hat{X}$ jest mierzalny wzgledem $\sigma(Y)$ czyli

$$\mathbb{E}(X|Y) = \mathbb{E}(X - \hat{X} + \hat{X}|Y) = \mathbb{E}(X - \hat{X}) + \hat{X} = \hat{X}$$

Z definicji warunkowej wariancji

$$\mathrm{Cov}(X|Y) = \mathbb{E}((X-\hat{X})(X-\hat{X})^T|Y) = \mathbb{E}((X-\hat{X})(X-\hat{X})^T) = \mathrm{Cov}(X-\hat{X})$$

oraz

$$\mathbb{E}((X - \hat{X})(X - \hat{X})^T) = \mathbb{E}((X - \hat{X})X^T)$$

Filtr Kalmana

Inicjalizacja:

$$\hat{X}_{1|0} = \mu_0$$
$$\Sigma_{1|0} = \Sigma_0$$

Predykcja:

$$\hat{X}_{k|k-1} = F_k \hat{X}_{k-1|k-1} + B_k u_k$$
$$\Sigma_{k|k-1} = F_k \Sigma_{k-1|k-1} F_k^T + Q_k$$

Filtracja:

Inowacja	$\hat{Y}_k = Y_k - H_k \hat{X}_{k\|k-1}$
Wariancja inowacji	$S_k = R_k + H_k \Sigma_{k\|k-1} H_k^T$
Przyrost Kalmana	$K_k = \Sigma_{k\|k-1} H_k^T S_k^{-1}$
Filtr	$\hat{X}_{k\|k} = \hat{X}_{k\|k-1} + K_k \hat{Y}_k$
Wariancja filtru	$\Sigma_{k\|k} = \Sigma_{k\|k-1} - K_k H_k \Sigma_{k\|k-1}$

Wygładzanie I

Chcemy policzyć rozkład $X_{k|T}$ wiedząc, że $X_{k+1|T} \sim N(\hat{X}_{k+1|T}, \Sigma_{k+1|T})$ Zauważmy, że

$$\begin{aligned}
\mathbb{E}(X_k|Y_{1:T}) &= \mathbb{E}[\mathbb{E}(X_k|X_{k+1}, Y_{1:k})|Y_{1:T}] \\
\mathrm{Cov}(X_k|Y_{1:T}) &= \mathbb{E}[\mathrm{Cov}((X_k|X_{k+1}, Y_{1:k}))|Y_{1:T}] \\
&\qquad + \mathrm{Cov}[\mathbb{E}((X_k|X_{k+1}, Y_{1:k}))|Y_{1:T}]
\end{aligned}$$

Czyli trzeba policzyć rozkład $X_k|X_{k+1}, Y_{1:k}$

$$\mathbb{E}[X_k, X_{k+1}|Y_{1:k}]^T = [\hat{X}_{k|k}, \hat{X}_{k+1|k}]^T$$

$$\mathrm{Cov}([X_k, X_{k+1}|Y_{1:k}]^T) = \begin{bmatrix} \Sigma_{k|k} & \Sigma_{k|k}F_{k+1}^T \\ \Sigma_{k|k}F_{k+1}^T & \Sigma_{k+1|k} \end{bmatrix}$$

Wygładzanie II

Lemat

Jeśli

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{21} \\ \Sigma_{21}^T & \Sigma_{22} \end{pmatrix}\right)$$

to

$$Z_1|Z_2 \sim \mathcal{N}(\mu_1 + \Sigma_{21}\Sigma_{22}^{-1}(Z_2 - \mu_2), \Sigma_{11} - \Sigma_{21}\Sigma_{22}^{-1}\Sigma_{21}^T)$$

Stąd

$$X_k|X_{k+1}, Y_{1:k} \sim N(\hat{X}_{k|k} + L_k(X_{k+1} - \hat{X}_{k+1|k}), \Sigma_{k|k} - L_k\Sigma_{k+1|k}^{-1}L_k^T)$$

$$L_k = \Sigma_{k|k}F_{k+1}^T\Sigma_{k+1|k}^{-1}$$

Wygładzanie III

Czyli

$$
\begin{aligned}
\hat{X}_{k|T} &= \hat{X}_{k|k} + L_k(\hat{X}_{k+1|T} - \hat{X}_{k+1|k}) \\
\Sigma_{k|T} &= \Sigma_{k|k} - L_k \Sigma_{k+1|k}^{-1} L_k^T + L_k \Sigma_{k+1|T} L_k^T \\
&= \Sigma_{k|k} + L_k(\Sigma_{k+1|T} - \Sigma_{k+1|k}) L_k^T
\end{aligned}
$$

Rauch–Tung–Striebel

$$
\begin{aligned}
L_k &= \Sigma_{k|k} F_{k+1}^T \Sigma_{k+1|k}^{-1} \\
\hat{X}_{k|T} &= \hat{X}_{k|k} + L_k(\hat{X}_{k+1|T} - \hat{X}_{k+1|k}) \\
\Sigma_{k|T} &= \Sigma_{k|k} + L_k(\Sigma_{k+1|T} - \Sigma_{k+1|k})L_k^T
\end{aligned}
$$

EM dla liniowych modeli

$$\mathrm{X_k} = \mathrm{F}\mathrm{X_{k-1}} + \mathrm{w_k}$$
$$\mathrm{Y_k} = \mathrm{H}\mathrm{X_k} + \mathrm{v_k}$$

$$\mathrm{w_k} \sim \mathrm{N}(0, \mathrm{Q}), \quad \mathrm{v_k} \sim \mathrm{N}(0, \mathrm{R}),\ \mathrm{X_1} \sim \mathrm{N}(\mu_0, \Sigma_0)$$

Chcemy estymować $\mathrm{F}, \mathrm{H}, \mathrm{Q}, \mathrm{R}$

Log wiarygodność

$$
\begin{aligned}
&\ell(F, H, Q, R) \\
&= \sum_{k=2}^{T} \left[-\frac{1}{2} \log(|Q|) - \frac{1}{2}(x_k - Fx_{k-1})^T Q^{-1} (x_k - Fx_{k-1}) \right] \\
&\quad + \sum_{k=1}^{T} \left[-\frac{1}{2} \log(|R|) - \frac{1}{2}(y_k - Hx_k)^T R^{-1} (y_k - Hx_k) \right] \\
&= \frac{T-1}{2} \log(|Q|^{-1}) - \frac{1}{2} \mathrm{tr} \left(Q^{-1} \sum_{k=2}^{T} (x_k - Fx_{k-1})(x_k - Fx_{k-1})^T \right) \\
&\quad + \frac{T}{2} \log(|R|^{-1}) - \frac{1}{2} \mathrm{tr} \left(R^{-1} \sum_{k=1}^{T} (y_k - Hx_k)(y_k - Hx_k)^T \right) \\
&= \frac{T-1}{2} \log(|Q|^{-1}) - \frac{1}{2} \mathrm{tr} \left(Q^{-1} \sum_{k=2}^{T} \left[x_k x_k^T + Fx_{k-1} x_{k-1}^T F^T - x_k x_{k-1}^T F^T - Fx_{k-1} x_k^T \right] \right) \\
&\quad + \frac{T}{2} \log(|R|^{-1}) - \frac{1}{2} \mathrm{tr} \left(R^{-1} \sum_{k=1}^{T} \left[y_k y_k^T + Hx_k x_k^T H^T - y_k x_k^T H^T - Hx_k y_k^T \right] \right)
\end{aligned}
$$

Expectation

$$
\begin{aligned}
\mathbb{E}(x_k|Y_{1:T}) &= \hat{X}_{k|T} \\
\mathbb{E}(x_k x_k^T|Y_{1:T}) &= \hat{X}_{k|T}\hat{X}_{k|T}^T + \Sigma_{k|T} \\
\mathbb{E}(x_k x_{k+1}^T|Y_{1:T}) &= \mathbb{E}(x_k x_{k+1}^T|x_{k+1}, Y_{1:k})|Y_{1:T}) \\
&= \mathbb{E}[(\hat{X}_{k|k} + L_k(x_{k+1} - \hat{X}_{k+1|k}))x_{k+1}^T|Y_{1:T}] \\
&= \hat{X}_{k|k}\hat{X}_{k+1|T}^T + L_k\left(\Sigma_{k+1|T} + (\hat{X}_{k+1|T} - \hat{X}_{k+1|K})\hat{X}_{k+1|T}^T\right)
\end{aligned}
$$

Maximization I

Potrzebne fakty:

- $$\frac{\partial x^T A x}{\partial x} = x^T (A^T + A)$$
- $$\frac{\partial B^T A B}{\partial B} = B^T (A^T + A)$$
- $$\frac{\partial \mathrm{tr}(AB)}{\partial A} = \frac{\partial \mathrm{tr}(BA)}{\partial A} = \frac{\partial \mathrm{tr}(B^T A^T)}{\partial A} = B^T$$
- $$\frac{\partial \mathrm{tr}(ABA^T)}{\partial A} = 2AB$$
- $$\frac{\partial \log(|A|)}{\partial A} = A^{-T}$$

Maximization II

Maksymalizacja po F

$$\frac{\partial \ell}{\partial F} = \sum_{k=2}^{T} x_k x_{k-1}^T Q^{-1} - F \sum_{k=2}^{T} x_{k-1} x_{k-1}^T Q^{-1} = 0$$

Czyli

$$F = \sum_{k=2}^{T} x_k x_{k-1}^T \left(\sum_{k=2}^{T} x_{k-1} x_{k-1}^T \right)^{-1}$$

Maximization III

Maksymalizacja po H

$$\frac{\partial \ell}{\partial H} = \sum_{k=1}^{T} x_k y_k^T R^{-1} - H \sum_{k=1}^{T} x_k x_k^T R^{-1} = 0$$

Czyli

$$H = \sum_{k=1}^{T} x_k y_k^t \left(\sum_{k=1}^{T} x_k x_k^T \right)^{-1}$$

Maximization IV

Maksymalizacja po Q

$$\frac{\partial \ell}{\partial Q^{-1}} = \frac{T-1}{2} Q$$
$$- \frac{1}{2} \left\{ \sum_{k=2}^{T} \left[x_k x_k^T + F x_{k-1} x_{k-1}^T F^T - x_k x_{k-1}^T F^T - F x_{k-1} x_k^T \right] \right\}^T$$

Czyli

$$Q = \frac{1}{T-1} \left\{ \sum_{k=2}^{T} \left[x_k x_k^T + F x_{k-1} x_{k-1}^T F^T - x_k x_{k-1}^T F^T - F x_{k-1} x_k^T \right] \right\}$$

Maximization V

Maksymalizacja po R

$$\frac{\partial \ell}{\partial R^{-1}} = \frac{T}{2} R - \frac{1}{2} \left\{ \sum_{k=1}^{T} \left[y_k y_k^T + H x_k x_k^T H^T - y_k x_k^T H^T - H x_k y_k^T \right] \right\}^T$$

Czyli

$$R = \frac{1}{T} \left\{ \sum_{k=1}^{T} \left[y_k y_k^T + H x_k x_k^T H^T - y_k x_k^T H^T - H x_k y_k^T \right] \right\}$$

3.36pt

2 Ukryte modele Markowa
- Dyskretne modele Markowa
- Obliczanie prawdopodobieństwa obserwacji
- Forward Backward
- Algorytm Viterbiego
- Algorytm FFBS
- Algorytm Bauma-Welcha
- Modele gaussowskie i filtr Kalmana

3 Algorytmy MCMC
- Podstawowe pojęcia statystyki bayesowskiej
- Metody Monte Carlo
- Algorytm Metropolisa-Hastingsa
- Próbnik Gibbsa
- Własności teoretyczne MCMC
- Adaptacyjne algorytmy MCMC
- Hamiltonian Monte Carlo

Model bayesowski

W statystyce bayesowskiej zakłada się, że parametry rozkładu są zmiennymi losowymi

$$\theta \sim \pi(\theta) \quad \text{rozkład a'priori.}$$

Obserwacje przy znanym θ maja rozkład

$$\mathrm{X} \sim \mathrm{p}(\mathrm{X}|\theta) \quad \text{wiarygodność.}$$

Wnioskujemy o parametrach na podstawaie rozkładu θ pod warunkiem zaobserwowanego X

$$\pi(\theta|\mathrm{X}) = \frac{\pi(\theta)\mathrm{p}(\mathrm{X}|\theta)}{\int_{\mathrm{X}} \pi(\theta)\mathrm{p}(\mathrm{x}|\theta)\mathrm{dx}} \propto \pi(\theta)\mathrm{p}(\mathrm{X}|\theta) \quad \text{rozkład a'posteriori}$$

Przykład 1

S liczba wyrzuconych orłów w N rzutach monetą o prawdopodobieństwie wyrzucenia orła θ.

$$\mathrm{p}(\mathrm{S}|\theta) \propto \theta^{\mathrm{S}}(1-\theta)^{\mathrm{N}-\mathrm{S}}$$

Zakładamy, że rozkład a'priori jest rozkładem $\mathrm{Beta}(\alpha, \beta)$ czyli

$$\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

Stąd

$$\pi(\theta|\mathrm{S}) \propto \theta^{\mathrm{S}+\alpha-1}(1-\theta)^{\mathrm{N}-\mathrm{S}+\beta-1}$$

czyli $\pi(\theta|\mathrm{S}) \sim \mathrm{Beta}(\mathrm{S}+\alpha, \mathrm{N}-\mathrm{S}+\beta)$.

Przykład 2

$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, μ - nieznane, σ^2 znane

$$\pi(\mu) \sim N(m, v^2)$$

Rozkład a'posteriori

$$\begin{aligned}
\pi(\mu|X_1 \ldots, X_n) &\propto \exp\left\{-\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2\sigma^2} - \frac{(\mu-m)^2}{2v^2}\right\} \\
&\propto \exp\left\{-\frac{nv^2+\sigma^2}{2v^2\sigma^2}\mu^2 + 2\frac{n\bar{X}v^2+m\sigma^2}{2v^2\sigma^2}\mu\right\} \\
&\propto \exp\left\{-\frac{1}{2}\frac{nv^2+\sigma^2}{v^2\sigma^2}\left(\mu - \frac{n\bar{X}v^2+m\sigma^2}{nv^2+\sigma^2}\right)^2\right\}
\end{aligned}$$

Czyli

$$\pi(\mu|X_1 \ldots, X_n) \sim N\left(\frac{n\bar{X}v^2+m\sigma^2}{nv^2+\sigma^2}, \frac{v^2\sigma^2}{nv^2+\sigma^2}\right)$$

Przykład 3

$X_1, \ldots, X_n \sim N(\mu, 1/\tau)$, μ - znane, τ- nieznane

$$\pi(\tau) \sim \text{Gamma}(a, b)$$

Rozkład a'posteriori

$$\begin{aligned}\pi(\tau|X_1 \ldots, X_n) &\propto \tau^{a-1}\tau^{n/2} \exp\left\{-\tau\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2} - \tau b\right\} \\ &\propto \tau^{a+n/2-1} \exp\left\{-\tau\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2} + b\right)\right\}\end{aligned}$$

Czyli

$$\pi(\tau|X_1 \ldots, X_n) \sim \text{Gamma}\left(a + n/2, b + \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2}\right)$$

Estymatory bayesowskie I

Funkcja straty:

$$\mathrm{L}(\theta, \mathrm{a})$$

Strata gdy pradziwa wartość jest θ a my przewidujemy a. L wypukła, równa 0 wtedy i tylko wtedy gdy $\theta = \mathrm{a}$. Standardowo używa się kwadratowej funkcji straty $(\theta - \mathrm{a})^2$.

Ryzyko estymatora:

$\delta(\mathrm{X})$ -estymator

$$\mathrm{R}(\delta) = \int_{\Theta} \int_{\mathrm{X}} \mathrm{L}(\theta, \delta(\mathrm{X}))\mathrm{p}(\mathrm{X}|\theta)\pi(\theta)\mathrm{d}\mathrm{X}\mathrm{d}\theta.$$

Ryzyko bayesowskie:

$$\mathrm{R}_{\mathrm{B}}(\delta(\mathrm{X})) = \int_{\Theta} \mathrm{L}(\theta, \delta(\mathrm{X}))\pi(\theta|\mathrm{X})\mathrm{d}\theta.$$

Estymatory bayesowskie II

Jeżeli zminimalizujemy ryzyko bayesowskie to minimalizujemy również zwykłe ryzyzko

$$
\begin{aligned}
\mathrm{R}(\delta) &= \int_{\Theta}\int_{\mathrm{X}} \mathrm{L}(\theta,\delta(\mathrm{X}))\mathrm{p}(\mathrm{X}|\theta)\pi(\theta)\mathrm{d}\mathrm{X}\mathrm{d}\theta \\
&= \int_{\mathrm{X}}\int_{\Theta} \mathrm{L}(\theta,\delta(\mathrm{X}))\pi(\theta|\mathrm{X})\mathrm{d}\theta\mathrm{p}(\mathrm{X})\mathrm{d}\mathrm{X} \\
&= \int_{\mathrm{X}} \mathrm{R}_{\mathrm{B}}(\delta(\mathrm{X}))\mathrm{d}\mathrm{X}
\end{aligned}
$$

Estymator bayesowski:

$$\hat{\theta}_{\mathrm{B}} = \arg\min_{\delta} \mathrm{R}_{\mathrm{B}}(\delta(\mathrm{X}))$$

W przypadku kwadratowej funkcji straty

$$\hat{\theta}_{\mathrm{B}} = \mathrm{E}(\theta|\mathrm{X})$$

Przykłady estymatorów bayesowskich

Model Dwumianowy-Beta:

$$\pi(\theta|\mathrm{S}) \sim \mathrm{Beta}(\mathrm{S}+\alpha, \mathrm{N}-\mathrm{S}+\beta)$$

$$\hat{\theta}_{\mathrm{B}} = \frac{\mathrm{S}+\alpha}{\mathrm{N}+\alpha+\beta} = \frac{\mathrm{S}}{\mathrm{N}}\frac{\mathrm{N}}{\mathrm{N}+\alpha+\beta} + \frac{\alpha}{\alpha+\beta}\frac{\alpha+\beta}{\mathrm{N}+\alpha+\beta}$$

Model Normalny-Normalny:

$$\pi(\mu|\mathrm{X}_1\ldots,\mathrm{X}_\mathrm{n}) \sim \mathrm{N}\left(\frac{\mathrm{n}\bar{\mathrm{X}}\mathrm{v}^2+\mathrm{m}\sigma^2}{\mathrm{n}\mathrm{v}^2+\sigma^2}, \frac{\mathrm{v}^2\sigma^2}{\mathrm{n}\mathrm{v}^2+\sigma^2}\right)$$

$$\hat{\mu}_{\mathrm{B}} = \frac{\mathrm{n}\bar{\mathrm{X}}\mathrm{v}^2+\mathrm{m}\sigma^2}{\mathrm{n}\mathrm{v}^2+\sigma^2} = \bar{\mathrm{X}}\frac{\mathrm{n}\mathrm{v}^2}{\mathrm{n}\mathrm{v}^2+\sigma^2} + \mathrm{m}\frac{\sigma^2}{\mathrm{n}\mathrm{v}^2+\sigma^2}$$

Model Normalny-Gammy:

$$\pi(\tau|\mathrm{X}_1\ldots,\mathrm{X}_\mathrm{n}) \sim \mathrm{Gamma}\left(\mathrm{a}+\mathrm{n}/2, \mathrm{b}+\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{X}_\mathrm{i}-\mu)^2}{2}\right)$$

$$\hat{\tau}_{\mathrm{B}} = \frac{\mathrm{a}+\mathrm{n}/2}{\mathrm{b}+\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{X}_\mathrm{i}-\mu)^2/2}$$

Obrobine bardziej złożony model

$X_1, \ldots, X_n \sim N(\mu, 1/\tau)$, μ - nieznane, τ -nieznane

$$\pi(\mu) \sim N(m, v^2), \quad \pi(\tau) \sim \mathrm{Gamma}(a, b)$$

μ i τ niezależne Rozkład a'posteriori

$$\pi(\mu, \tau | X_1 \ldots, X_n) \propto \tau^{a+n/2-1} \exp\left\{-\tau \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{2} - \frac{(\mu - m)^2}{2v^2} - \tau b\right\}$$

Czyli

$$\pi(\mu, \tau | X_1 \ldots, X_n) \sim ???, \quad \hat{\tau}_B = ??? \quad \hat{\mu}_B = ???$$

Znamy $\pi(\mu, \tau | X_1 \ldots, X_n)$ z dokładnością do stałej normującej

Hierarchiczny model bayesowski

Mamy k grup (np powiatów), w j-tej grupie mamy N_j obserwacji

$$X_{1j}, \ldots, X_{N_j,j} \sim N(\mu_j, 1/\tau_j)$$

Następnie

$$\mu_1, \ldots, \mu_k \sim N(\mu, 1/\tau)$$
$$\tau_1, \ldots, \tau_k \sim \mathrm{Gamma}(a, b)$$

μ i τ odpowiadają wartości oczekiwanej i wariancji badanej cechy w całym kraju. Są to z reguły wartości nieznane i na nie również nakładamy rozkłady a'priori

$$\mu \sim N(m, v^2)$$
$$\tau \sim \mathrm{Gamma}(\alpha, \lambda)$$

Rozkład a'posteriori

$$\pi(\mu_1, \ldots, \mu_k, \tau_1, \ldots, \tau_k, \mu, \tau | X, a, b, \alpha, \lambda, m, v^2) \propto$$
$$\prod_{j=1}^{k} \left\{ \prod_{i=1}^{N_j} [p(X_{ij}|\mu_j, \tau_j)]\, \pi(\mu_j|\mu, \tau)\pi(\tau_j|a, b) \right\} \pi(\mu|m, v^2)\pi(\tau|\alpha, \lambda)$$

Inne elementy wnioskowania bayesowskiego

Przedziały ufności:
Szukamy najkrótszego przedziału $[\underline{\theta}, \overline{\theta}]$ takiego, że

$$\mathrm{P}(\theta \in [\underline{\theta}, \overline{\theta}]|\mathrm{X}) = 1 - \alpha$$

Prognozowanie:

$$\pi(\mathrm{X}^*|\mathrm{X}) = \int_{\Theta} \mathrm{p}(\mathrm{X}^*|\theta)\pi(\theta|\mathrm{X})\mathrm{d}\theta$$

Dla kwadratowej funkcji straty predykcja bayesowska

$$\mathrm{E}(\mathrm{X}^*|\mathrm{X})$$

Problemy obliczeniowe

Potrzebujemy metod pozwalających:

- Aproksymować/ generować z rozkładu $\pi(\theta|\mathrm{X})$ znanego z dokładnością do stałej.
- Obliczać całki względem rozkładu jak powyżej.

3.36pt

2 Ukryte modele Markowa
- Dyskretne modele Markowa
- Obliczanie prawdopodobieństwa obserwacji
- Forward Backward
- Algorytm Viterbiego
- Algorytm FFBS
- Algorytm Bauma-Welcha
- Modele gaussowskie i filtr Kalmana

3 Algorytmy MCMC
- Podstawowe pojęcia statystyki bayesowskiej
- Metody Monte Carlo
- Algorytm Metropolisa-Hastingsa
- Próbnik Gibbsa
- Własności teoretyczne MCMC
- Adaptacyjne algorytmy MCMC
- Hamiltonian Monte Carlo

Proste Monte Carlo

Chcemy obliczyć

$$I = \int f(x)\pi(x)dx,$$

gdzie π gęstość prawdopodobieństwa. Czyli z MPWL

$$I = Ef(X) \approx \frac{1}{n}\sum_{i=1}^{n} f(X_i) =: \hat{I}_n,$$

gdzie $X_1, \ldots, X_n \sim \text{i.i.d}, \pi$

Własności esytomatora MC:

- $E\hat{I}_n = I$
- Jeśli $\mathrm{Var}(f(X_i)) = \sigma^2$ to

$$E(I - \hat{I}_n)^2 = \frac{\sigma^2}{n}$$

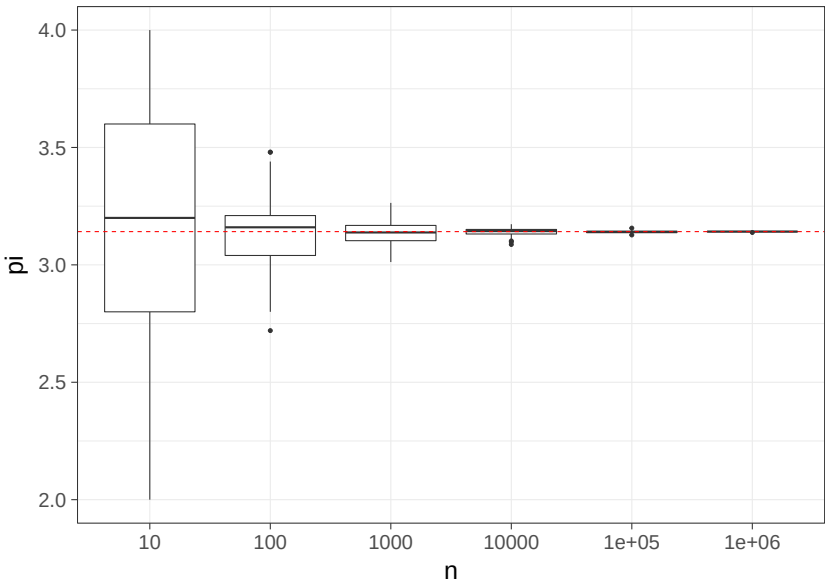
Przykład oblicznie liczby π

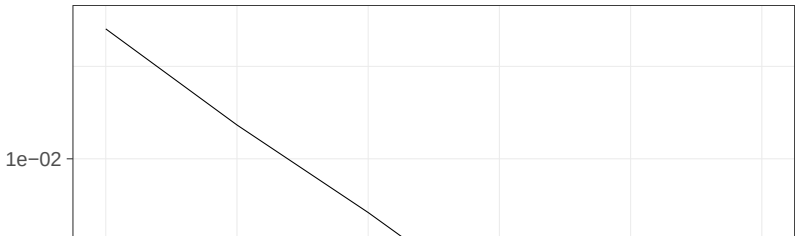
Jeśli $X \sim \mathrm{Unif}([0,1] \times [0,1])$ to

$$\pi = 4E1(X_1^2 + X_2^2 \leqslant 1)$$

Jeśli $X^1, \ldots X^n \sim \mathrm{Unif}([0,1] \times [0,1])$ to

$$\pi \approx \frac{1}{n} \sum_{i=1}^{n} 1((X_1^i)^2 + (X_2^i)^2 \leqslant 1)$$





Rejection Sampling I

Chcemy wylosować zmienną losową o gęstości f, ale potrafimy generować z g oraz wiemy, że

$$f(x) \leqslant Mg(x)$$

Algorytm:

1. $Y \sim g$
2. $U \sim \mathrm{Unif}([0,1])$
3. Jeśli $U \leqslant \frac{f(Y)}{Mg(Y)}$ to zwracamy Y w.p.p wracamy do kroku 1

Poprawność: Niech $Z_i = (U_i, Y_i)$, $A = \{(y,u) \colon u \leqslant \frac{f(y)}{Mg(y)}\}$,

$$N = \mathsf{min}\{n \colon Z_i \in A\}$$

Wówczas N i Z_N są niezależne oraz

$$P(Z_N \in B) = P(Z \in B | Z \in A)$$

Rejection Sampling II

Obliczmy najpierw

$$P(N = n) = P(Z_1 \notin A, \ldots, Z_{n-1} \notin A, Z_n \in A) = q^{n-1}p,$$

gdzie $p = P(Z_1 \in A)$ oraz $q = 1 - p$ Teraz

$$\begin{aligned} P(N = n, Z_N \in B) &= P(Z_1 \notin A, \ldots, Z_{n-1} \notin A, Z_n \in A \cap B) \\ &= P(Z_1 \notin A)^{n-1} P(Z_n \in B) \frac{P(Z_n \in A \cap B)}{P(Z_n \in B)} \end{aligned}$$

Teraz wystarczy obliczyć

$$P(Y \in B | Z \in A) = \frac{\int_B \int_0^{f(y)/Mg(y)} g(y) du dy}{\int \int_0^{f(y)/Mg(y)} g(y) du dy} = \frac{\int_B f(y) dy}{\int f(y) dy}$$

Oczekiwana liczba kroków algorytmu:

$$p = P(Z_1 \in A) = \int \frac{f(y)}{Mg(y)} g(y) dy = \frac{1}{M}$$

Stąd

$$EN = M$$

Przykład

Chcemy generować z X rozkładu gaussowskiego $N(0,1)$ uciętego do odcinka $[3,4]$ czyli

$$f(x) = \frac{\exp(-x^2/2)}{\int_3^4 \exp(-x^2/2)} 1(x \in [3,4])$$

I korzystamy z dwóch gęstości pomocniczych g_1 gęstość $N(0,1)$ i $g_2 \propto x \exp(-x^2/2) 1(x \in [3,4])$ W pierszym przypadku do wygenerowania próbki długości 100 potrzeba 73936 zmiennych z rozkładu g_1 natomiast dla g_2 wystarczyło 303 zmiennych.

Własności rejection sampling

- Jeśli potrafimy dobrze, przybliżyć globalnie gęstość to algorytm jest efektywny.
- Generuje próbkę i.i.d
- W przypadku użycia do obliczania całek "marnujemy" znaczną część wygenerowanych zmiennych losowych.

Importance sampling

Chcemy obliczyć

$$I = \int f(x)p(x)dx$$

p gęstość rozkładu. Dysponujemy próbką $X_1, \ldots, X_n$ i.i.d z rozkładu q. Zauważmy, że

$$I = \int f(x)p(x)dx = \int f(x)\frac{p(x)}{q(x)}q(x) = E_q f(X)w(X),$$

gdzie $w = \frac{p}{q}$.

$$I_{IS} = \frac{1}{n}\sum_{i=1}^{n} f(X_i)w(X_i)$$

Własności IS

- $E(I_{IS}) = I$
- $Var(I_{IS}) = \frac{Var_q(f(X)w(X))}{n}$
- Wymaga znajomości stałej normującej f

Samonormujący IS

Zauważmy, że

$$E_q(w(X)) = 1$$

Zastąpmy n przez estymator $\hat{n} = \sum_{i=1}^{n} w(X_i)$. Czyli

$$\bar{I}_{IS} = \frac{\sum_i f(X_i)w(X_i)}{\sum_{i=1}^{n} w(X_i)}$$

Teraz w występuje w liczniku i mianowniku.

- $E(\bar{I}_{IS}) \neq I$ ale $\lim_{n\to\infty} E\bar{I}_{IS} = I$
- $Var(\bar{I}_{IS}) = \mathcal{O}(1/n)$
- Można stosować do gęstości znanej z dokładnością do stałej normującej.

Efektywność IF

Proste MC

$$\mathrm{MSE}(\hat{\mathrm{I}}) = \frac{\sigma^2}{\mathrm{n}}$$

Dla IS

$$\mathrm{MSE}(\bar{\mathrm{I}}_{\mathrm{IS}}) = \frac{\mathrm{c}}{\mathrm{n}}$$

Effective Sample Size

Dla jakiej długości próbki i.i.d. dostaniemy ten sam błąd co nasz estymator. Dla IS ESS przybliża się następująco: Traktujemy wagi $\mathrm{w_i} = \mathrm{w}(\mathrm{X_i})$ jako deterministyczne wówczas wariancja estymatora IS jest postaci

$$\frac{\sum \mathrm{w_i^2}}{(\sum \mathrm{w_i})^2} \mathrm{Var_q}(\mathrm{f}(\mathrm{X_i}))$$

I ESS definiujemy jako

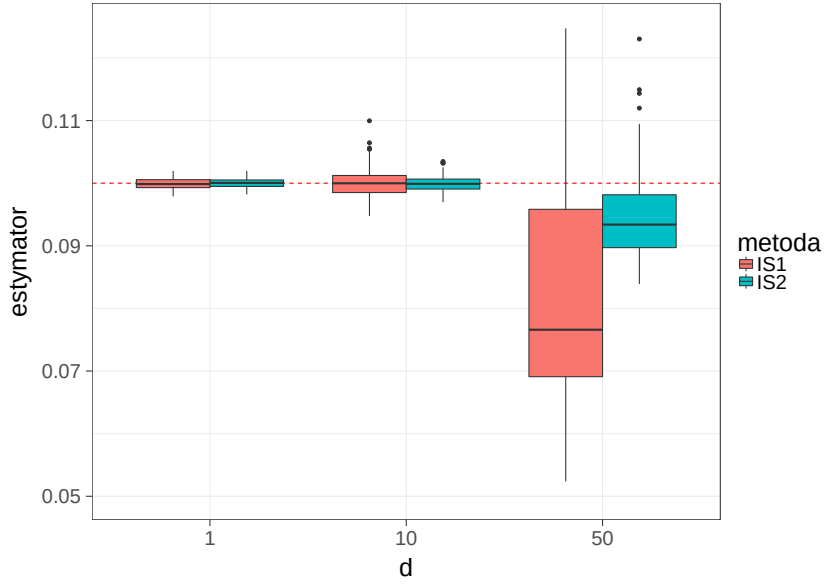
$$\mathrm{ESS} = \frac{(\sum \mathrm{w_i})^2}{\sum \mathrm{w_i^2}}$$

Przykład

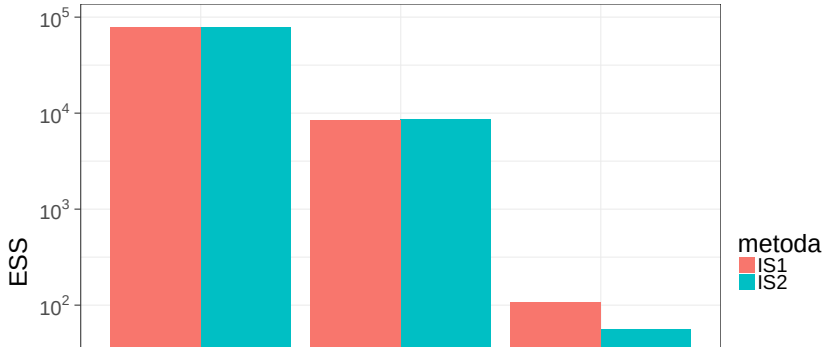
$Y \sim N(0.1, 0.2^2 I_d)$ i chcemy obliczyć

$$I = \frac{1}{d}\sum_{i=1}^{d} EY_i = 0.1$$

Bedziemy korzystać z (samonormujący) IS z rozkładem



propozycji $N(0, 0.2^2 I_d)$.



Własności IS

- Wykorzystuje całą próbkę do estymacji.
- Wrażliwa na dobór rozkładu pomocniczego. Trzeba dobrze przybliżyć rozkład globalnie.
- Nie wymaga znajomości stałej normującej

MCMC

Wadą standardowych MC jest to, że kolejne iteracje niezależą od poprzednich. Co powoduje, że każdy krok musi być dopasowany globalnie do rozkładu. Chcielibyśmy eksplorować rozkład lokalnie, np. w okolicy wyniku z poprzedniego kroku.

MCMC:

Chcemy wygenerować ergodyczny łańcuch Markowa $X_1, \ldots, X_n$ o rozkładzie stacjonarnym π. Wówczas na podstawie MPWL dla łańcuchów Markowa otrzymujemy

$$I = \int f(x)\pi(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

Z uwagi na to, że asymptotycznie mamy właściwy rozkład to stosuje sie "burn-in" time

$$\hat{I}_{n_0,n} = \frac{1}{n}\sum_{i=n_0+1}^{n+n_0} f(X_i)$$

Rozkład stacjonarny i odwracalność I

Łańcuh Markowa z jądrem przejścia $P(x, dy)$ ma rozkład stacjonarny π wtedy i tylko wtedy gdy

$$\pi(A) = \int_X P(x, A)\pi(x)dx = \pi P.$$

Łańcuch Markowa z jądrem przejścia $P(x, dy)$ jest odwracalny względem π wtedy i tylko wtedy gdy

$$\int_{A\times B} P(x, dy)\pi(x)dx = \int_{A\times B} P(y, dx)\pi(y)dy$$

lub w skrócie

$$P(x, dy)\pi(dx) = P(y, dx)\pi(dy)$$

Rozkład stacjonarny i odwracalność II

Łańcuch Markowa jest odwracalny względem $\pi \Longrightarrow \pi$ jest rozkładem stacjonarnym

$$\pi P(A) = \int P(x, A)\pi(x)dx = \int \int_A \pi(x)P(x, dy)dx$$
$$= \int_A \int P(y, dx)\pi(y)dy = \int_A \pi(y)$$

3.36pt

2 Ukryte modele Markowa
- Dyskretne modele Markowa
- Obliczanie prawdopodobieństwa obserwacji
- Forward Backward
- Algorytm Viterbiego
- Algorytm FFBS
- Algorytm Bauma-Welcha
- Modele gaussowskie i filtr Kalmana

3 Algorytmy MCMC
- Podstawowe pojęcia statystyki bayesowskiej
- Metody Monte Carlo
- Algorytm Metropolisa-Hastingsa
- Próbnik Gibbsa
- Własności teoretyczne MCMC
- Adaptacyjne algorytmy MCMC
- Hamiltonian Monte Carlo

Algorytm Metropolisa-Hastingsa

- $Q(x, dy)$ jądro przejścia o gęstości $q(x, y)$
- π rozkład docelowy

Algorytm MH

1. Generujemy $Y \sim Q(X_n, \cdot)$
2. Obliczamy prawdopodobieństwo akceptacji $\alpha(X_n, Y)$ gdzie

$$\alpha(x, y) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)} \wedge 1$$

3.

$$X_{n+1} = \begin{cases} Y & \text{z prawdopodobieństwem } \alpha(X_n, Y) \\ X_n & \text{z prawdopodobieństwem } 1 - \alpha(X_n, Y) \end{cases}$$

Algorytm Metropolisa-Hastingsa

Algorytm MH generuje łańcuch Markowa odwracalny względem π Czyli

$$\mathrm{P_{MH}(x,y)\pi(x) = P_{MH}(y,x)\pi(y)}$$

Jeśli $\mathrm{x} \neq \mathrm{y}$ to

$$\mathrm{P_{MH}(x,y) = \alpha(x,y)q(x,y)}$$

Załóżmy, że $\alpha(\mathrm{x},\mathrm{y}) < 1$ czyli $\alpha(\mathrm{x},\mathrm{y}) = \frac{\pi(\mathrm{y})\mathrm{q}(\mathrm{y},\mathrm{x})}{\pi(\mathrm{x})\mathrm{q}(\mathrm{x},\mathrm{y})}$ to $\alpha(\mathrm{y},\mathrm{x}) = 1$ (drugi przypadek symetryczny)

$$\begin{aligned}\mathrm{P_{MH}(x,y)\pi(x)} &= \alpha(\mathrm{x},\mathrm{y})\mathrm{q}(\mathrm{x},\mathrm{y})\pi(\mathrm{x}) = \pi(\mathrm{y})\mathrm{q}(\mathrm{y},\mathrm{x}) \\ &= \pi(\mathrm{y})\mathrm{q}(\mathrm{y},\mathrm{x})\alpha(\mathrm{y},\mathrm{x}) = \mathrm{P_{MH}(y,x)}\pi(\mathrm{y})\end{aligned}$$

Przykłady MH I

- Random Walk Metropolis $q(x, y) = q(|x - y|)$ np. $Q(x, dy) \sim N(x, \Sigma)$ wówczas

$$\alpha(x, y) = \frac{\pi(y)}{\pi(x)} \wedge 1$$

- Independent MH $q(x, y) = q(y)$

$$\alpha(x, y) = \frac{\pi(y)q(x)}{\pi(x)q(y)} \wedge 1$$

Przykłady MH II

- Metropolis Adjusted Langevin Algorithm

$$dX_t = \nabla \log(\pi(X_t))dt + \sqrt{2}dW_t$$

Dyskretyzacja Eulera- Maruyamy

$$X_{k+1} = X_k + \tau \nabla \log(\pi(X_k)) + \sqrt{2\tau} Z_k,$$

gdzie $Z_k \sim N(0, I)$.
MALA

$$Y = X_k + \tau \nabla \log(\pi(X_k)) + \sqrt{2\tau} Z_k$$

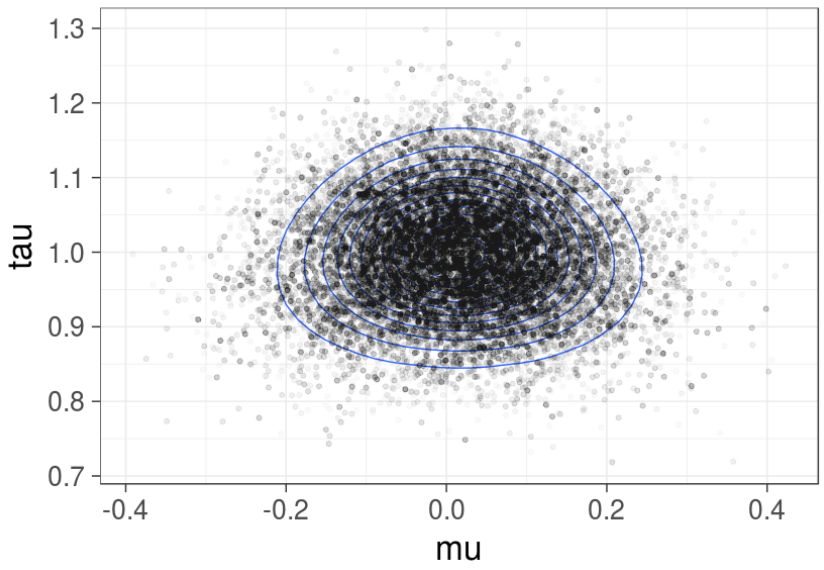
oraz

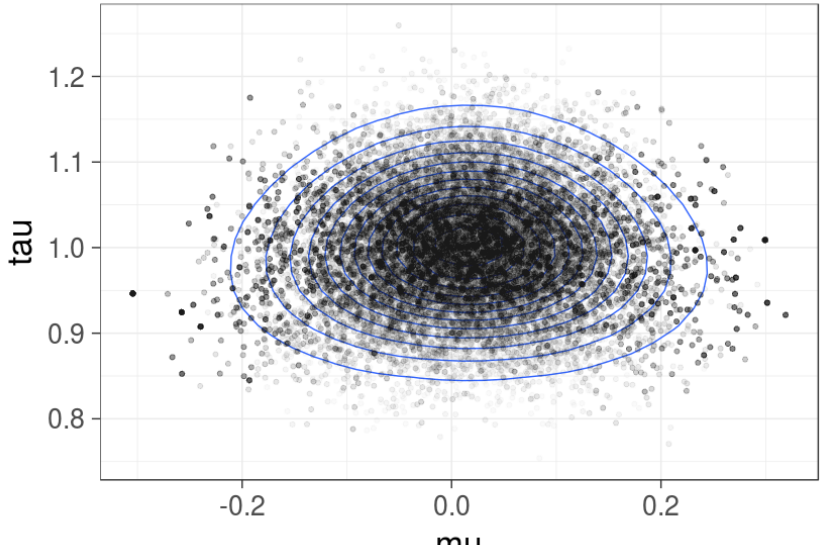
$$\alpha(x, y) = \frac{\pi(y)\phi(x; y + \tau \nabla \log(\pi(y)); 2\tau I)}{\pi(x)\phi(y; x + \tau \nabla \log(\pi(x)); 2\tau I)} \wedge 1$$

Przykład

$$X_1, \ldots, X_n \text{ i.i.d } N(\mu, 1/\tau),$$

$$\pi(\mu) \sim N(0, 10), \quad \pi(\tau) \sim \text{Gamma}(1, 1)$$





3.36pt

2 Ukryte modele Markowa
- Dyskretne modele Markowa
- Obliczanie prawdopodobieństwa obserwacji
- Forward Backward
- Algorytm Viterbiego
- Algorytm FFBS
- Algorytm Bauma-Welcha
- Modele gaussowskie i filtr Kalmana

3 Algorytmy MCMC
- Podstawowe pojęcia statystyki bayesowskiej
- Metody Monte Carlo
- Algorytm Metropolisa-Hastingsa
- Próbnik Gibbsa
- Własności teoretyczne MCMC
- Adaptacyjne algorytmy MCMC
- Hamiltonian Monte Carlo

Próbnik Gibbsa I

Mały krok próbnika Gibbsa:

$$X_i \sim \pi(X_i | X_{-i})$$

Oznaczmy przez $P_i(x, dy)$ jego jądro przejścia, jego gęstość

$$p_i(x, y) = \begin{cases} \pi(y_i | x_{-i}) & \text{jeśli } x_{-i} = y_{-i} \\ 0 & \text{w p.p.} \end{cases}$$

Systematic Scan

Wykonujemy kroki P_i w ustalonej kolejności czyli np.

$$\begin{aligned} X_1^n &\sim \pi(\cdot | X_{-1}^{n-1})) \\ X_2^n &\sim \pi(\cdot | X_1^n, X_2^{n-1}, \dots X_d^{n-1}) \\ &\vdots \\ X_k^n &\sim \pi(\cdot | X_1^n, \dots, X_{k-1}^n, X_{k+1}^{n-1}, \dots X_d^{n-1}) \\ &\vdots \end{aligned}$$

Próbnik Gibbsa II

Random Scan

W każdym kroku losujemy współrzędna która ma być zmieniaona

1. Losujemy $j \sim \mathrm{Unif}\{1, \ldots, d\}$
2. $$\begin{aligned} X_j^n &\sim \pi(\cdot | X_{-j}^{n-1}) \\ X_k^n &= X_k^{n-1}, \quad \text{dla } k \neq j \end{aligned}$$

Poprawność próbnika Gibbsa I

Mały krok próbnika Gibbsa jest odwracalny. Dla $x_{-i} = y_{-i}$

$$
\begin{aligned}
P_i(x, y)\pi(x) &= \pi(y_i|x_{-i})\pi(x_i, x_{-i}) \\
&= \frac{\pi(y_i, x_{-i})\pi(x_i|x_{-i})}{\pi(x_{-i})} \\
&= \frac{\pi(y_i, y_{-i})\pi(x_i|y_{-i})}{\pi(y_{-i})} \\
&= \pi(x_i|y_{-i})\pi(y_i, y_{-i}) = P_i(y, x)\pi(y)
\end{aligned}
$$

Poprawność próbnika Gibbsa II

Próbnik Gibbsa z losowym wyborem współrzędnych jest odwracalny.

$$P_{RS}(x, y) = \sum_{i=1}^{d} \frac{1}{n} P_i(x, y)\pi(x)$$

Każdy z P_i odwracalny to kombinacja wypukła jest odwracalna. Próbnik Gibbsa z deterministycznym wyborem współrzędnych jest nieodwracalny a zachowuje rozkład stacjonarny.

$$\begin{aligned} P_{SS}(x, y) = & \\ & P_1((x_1, \ldots, x_d), (y_1, x_2, \ldots, x_d)) \cdots \\ & P_i((y_1, \ldots, y_{i-1}, x_i, \ldots, x_d), (y_1, \ldots, y_i, x_{i+1}, \ldots, x_d)) \cdots \\ & P_d((y_d, \ldots, y_{d-1}, x_d), (y_1, \ldots, y_d)) \end{aligned}$$

Poprawność próbnika Gibbsa III

$$
\begin{aligned}
\pi \mathrm{P}_{\mathrm{SS}}(\mathrm{y}) &= \int \mathrm{P}_{\mathrm{SS}}(\mathrm{x}, \mathrm{y}) \pi(\mathrm{x}) \mathrm{dx} \\
&= \int \mathrm{P}_2 \cdots \mathrm{P}_\mathrm{d} \int \mathrm{P}_1((\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{d}), (\mathrm{y}_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{d})) \pi(\mathrm{x}) \mathrm{dx}_1 \mathrm{dx}_{-1} \\
&= \int \mathrm{P}_2((\mathrm{y}_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{d}), (\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{x}_\mathrm{d})) \cdots \mathrm{P}_\mathrm{d} \pi((\mathrm{y}_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{d})) \mathrm{dx}_{-1} \\
&= \cdots = \pi(\mathrm{y})
\end{aligned}
$$

Przykład 1

$$\pi(\mu, \tau | \mathrm{X}_1 \ldots, \mathrm{X}_\mathrm{n}) \propto \tau^{\mathrm{a}+\mathrm{n}/2-1}$$
$$\exp\left\{-\tau \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{X}_\mathrm{i} - \mu)^2}{2} - \frac{(\mu - \mathrm{m})^2}{2\mathrm{v}^2} - \tau \mathrm{b}\right\}$$

Pełne rozkłady warunkowe:

$$\pi(\mu | \tau, \mathrm{X}_1 \ldots, \mathrm{X}_\mathrm{n}) \sim \mathrm{N}\left(\frac{\mathrm{n}\bar{\mathrm{X}}\mathrm{v}^2 + \mathrm{m}\tau^{-1}}{\mathrm{n}\mathrm{v}^2 + \tau^{-1}}, \frac{\mathrm{v}^2\tau^{-1}}{\mathrm{n}\mathrm{v}^2 + \tau^{-1}}\right)$$
$$\pi(\tau | \mu, \mathrm{X}_1 \ldots, \mathrm{X}_\mathrm{n}) \sim \mathrm{Gamma}\left(\mathrm{a} + \mathrm{n}/2, \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{X}_\mathrm{i} - \mu)^2}{2} + \mathrm{b}\right)$$

Próbnik Gibbsa dla modelu hierarchicznego

$$\pi(\mu_1, \ldots, \mu_k, \tau_1, \ldots, \tau_k, \mu, \tau | X, a, b, \alpha, \lambda, m, v^2) \propto$$
$$\prod_{j=1}^{k} \left\{ \prod_{i=1}^{N_j} [p(X_{ij}|\mu_j, \tau_j)] \, \pi(\mu_j|\mu, \tau)\pi(\tau_j|a, b) \right\} \pi(\mu|m, v^2)\pi(\tau|\alpha, \lambda)$$

Pełne rozkłady warunkowe:

$$\pi(\mu_j|\tau_1, \ldots, \tau_k, \tau, \mu, , X_1 \ldots, X_n) \sim N\left(\frac{N_j \bar{X}_j \tau^{-1} + \mu \tau_j^{-1}}{N_j \tau^{-1} + \tau_j^{-1}}, \frac{\tau^{-1}\tau_j^{-1}}{N_j \tau^{-1} + \tau_j^{-1}} \right)$$

$$\pi(\tau_j|\mu_1, \ldots, \mu_k, \tau, \mu, X_1 \ldots, X_n) \sim \text{Gamma}\left(a + N_j/2, \frac{\sum_{i=1}^{N_j} (X_{ij} - \mu_j)^2}{2} + b \right)$$

$$\pi(\mu|\tau_1, \ldots, \tau_k, \mu_1, \ldots, \mu_k, \tau, X_1 \ldots, X_n) \sim N\left(\frac{k\bar{\mu}v^2 + m\tau^{-1}}{kv^2 + \tau^{-1}}, \frac{v^2\tau^{-1}}{kv^2 + \tau^{-1}} \right)$$

$$\pi(\tau|\tau_1, \ldots, \tau_k, \mu_1, \ldots, \mu_k, \mu, X_1 \ldots, X_n) \sim \text{Gamma}\left(\alpha + k/2, \frac{\sum_{i=1}^{k} (\mu_i - \mu)^2}{2} + \lambda \right)$$

Metropolis wewnątrz Gibbsa

Próbnik Gibbsa składa się z małych kroków postaci

$$\mathrm{P_i}(\mathrm{x}, \mathrm{dy}) = \pi(\mathrm{dy_i}|\mathrm{x_{-i}}).$$

Może się zdarzyć, że generowanie z jednego (wielu) rozkładów warunkowych $\pi(\cdot|\mathrm{x_{-i}})$ jest trudne. Wówczas krok $\mathrm{P_i}$ można zastąpić krokiem Metropolisa Hastingsa o rozkładzie docelowym $\pi(\cdot|\mathrm{x_{-i}})$.

3.36pt

2 Ukryte modele Markowa
- Dyskretne modele Markowa
- Obliczanie prawdopodobieństwa obserwacji
- Forward Backward
- Algorytm Viterbiego
- Algorytm FFBS
- Algorytm Bauma-Welcha
- Modele gaussowskie i filtr Kalmana

3 Algorytmy MCMC
- Podstawowe pojęcia statystyki bayesowskiej
- Metody Monte Carlo
- Algorytm Metropolisa-Hastingsa
- Próbnik Gibbsa
- Własności teoretyczne MCMC
- Adaptacyjne algorytmy MCMC
- Hamiltonian Monte Carlo

Potrzebujemy następujących własności łańcuchów Markowa

- Ergodyczność $\mathrm{P}(\mathrm{X_n} \in \mathrm{A}|\mathrm{X_1} = \mathrm{x}) \to \pi(\mathrm{A})$
- Prawo wielkich liczb $\hat{\mathrm{I}}_\mathrm{n} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{f}(\mathrm{X_i}) \to \pi\mathrm{f}$
- Centralne twierdzenie graniczne $\sqrt{\mathrm{n}}(\hat{\mathrm{I}}_\mathrm{n} - \pi\mathrm{f}) \to \mathrm{N}(0, \sigma^2)$

Ergodyczność dla skończonej przestrzeni stanów

TWIERDZENIE

Jeśli łańcuch Markowa o rozkładzie stacjonarnym π jest nieprzywiedlny i nieokresowy to jest ergodyczny.

π-nieprzywiedlność

Definicja

Łańcuch Markowa nazywamy π-nieprzywiedlnym, jeżeli dla każdego zbioru A, takim że $\pi(\mathrm{A}) > 0$, oraz dla każdego x istnieje n

$$\mathrm{P}^{\mathrm{n}}(\mathrm{x}, \mathrm{A}) > 0$$

Jeśli łańcuch Markowa jest π nieprzywiedlny to rozkład stacjonarny jest wyznaczony jednoznacznie.

Ergodyczność I

Definicja

Zbiór α nazywamy atomem jeżeli $\pi(\alpha) > 0$ oraz

$$\forall \mathrm{x} \in \alpha, \quad \mathrm{P}(\mathrm{x}, \mathrm{A}) = \nu(\mathrm{A}).$$

Definicja

Dla miar probabilistycznych μ, ν odległością pełnego wahania nazywamy

$$\|\mu - \nu\|_{\mathrm{tv}} = \sup_{\mathrm{A}} |\mu(\mathrm{A}) - \nu(\mathrm{A})|$$

Jeśli $\mathrm{X} \sim \mu$, $\mathrm{Y} \sim \nu$ to

$$\|\mu - \nu\|_{\mathrm{tv}} \leqslant \mathrm{P}(\mathrm{X} \neq \mathrm{Y})$$

Ergodyczność II

TWIERDZENIE

Jeśli łańcuch Markowa o rozkładzie stacjonarnym π jest π nieprzywiedlny i nieokresowy to

$$\|P^n(x,\cdot) - \pi(\cdot)\|_{tv} \to 0$$

Dowód:

Najpierw załóżmy, że łańcuch posidada atom α oraz zdefiniujmy parę łańcuhów Markowa $X_n, \tilde{X}_n$: $X_1 = x$, $\tilde{X}_1 \sim \pi$. Jeżeli $X_n \neq \tilde{X}_n$ to łańcuchy generowane są niezależnie zgodnie z $P(x, dy)$, natomiast jeśli $X_n = \tilde{X}_n$ to dalej trajektorie pozostają sklejone. Wówczas

$$\|P^n(x,\cdot) - \pi(\cdot)\|_{tv} \leqslant P(X_n \neq \tilde{X}_n) \leqslant 1 - P(\tau_{\alpha\times\alpha} \leqslant n),$$

gdzie

$$\tau_{\alpha\times\alpha} = \mathsf{min}\{k \geqslant 1 \colon X_k = \tilde{X}_k = \alpha\}.$$

Ergodyczność III

Z nieprzywiedlności i nieokresowości $P(\tau_{\alpha\times\alpha} \leqslant n) \to 1$
Jeżeli łańcuch nie posiada atomu to konstruujemy sztuczny atom:

Definicja

Zbiór C nazywamy małym zbiorem jeżeli istnieje n takie, że

$$P^n(x, A) \geqslant \beta\nu(A), \quad \forall x \in C$$

Konstrukcja rozszczepienia łańcucha dla $n = 1$. Definiujemy nowy łańcuch Markowa X_n, Γ_n na przestrzeni $X \times \{0, 1\}$

$$P(X_{n+1} \in A | X_n = x, \Gamma_n = 0) = \frac{P(x, A) - \beta\nu(A)1(x \in C)}{1 - \beta 1(x \in C)}$$

$$P(X_{n+1} \in A | X_n = x, \Gamma_n = 1) = \nu(A)$$

$$P(\Gamma_{n+1} = 1 | X_{n+1} = x) = \beta 1(x \in C)$$

Tak skonstruowany łańcuch ma atom $C \times \{1\}$ oraz rozkład brzegowy X_n jest taki sam jak pierwotnego łańcucha.

Prawo wielkich liczb I

TWIERDZENIE

Jeśli łańcuch Markowa z rozkładem stacjonarnym π jest π-nieprzywiedlny oraz posiada atom α to

$$\pi(\mathrm{A}) \propto \mathbb{E}_\alpha \sum_{\mathrm{k}=1}^{\tau_\alpha - 1} 1(\mathrm{X_i} \in \mathrm{A}),$$

gdzie

$$\tau_\alpha = \mathsf{min}\{\mathrm{k} > 1 \colon \mathrm{X_k} \in \alpha\}$$

Dowód dla α jednoelementowego

Prawo wielkich liczb II

Oznaczmy przez $\nu(\mathrm{A})$ prawą stronę. Najpierw załóżmy, że $\mathrm{A} \cap \alpha = \emptyset$

$$
\begin{aligned}
\nu \mathrm{P}(\mathrm{A}) &= \int_{\mathrm{X}} \mathbb{E}_\alpha \sum_{\mathrm{k}=1}^{\tau_\alpha - 1} 1(\mathrm{X_i} \in \mathrm{dx}) \mathrm{P}(\mathrm{x}, \mathrm{A}) \\
&= \int_{\mathrm{X}} \sum_{\mathrm{k}=1}^{\infty} \mathrm{P}_\alpha(\mathrm{X_k} \in \mathrm{dx}, \tau_\alpha > \mathrm{k}) \mathrm{P}(\mathrm{x}, \mathrm{A}) \\
&= \sum_{\mathrm{k}=1}^{\infty} \int_{\mathrm{X}} \mathrm{P}_\alpha(\mathrm{X_k} \in \mathrm{dx}, \tau_\alpha > \mathrm{k}) \mathrm{P}(\mathrm{x}, \mathrm{A}) \\
&= \sum_{\mathrm{k}=1}^{\infty} \mathrm{P}_\alpha(\mathrm{X_{k+1}} \in \mathrm{A}, \tau_\alpha > \mathrm{k}+1) \\
&= \nu(\mathrm{A})
\end{aligned}
$$

Prawo wielkich liczb III

$$
\begin{aligned}
\nu \mathrm{P}(\alpha) &= \int_{\mathrm{X}} \mathbb{E}_\alpha \sum_{\mathrm{k}=1}^{\tau_\alpha - 1} 1(\mathrm{X_i} \in \mathrm{dx}) \mathrm{P}(\mathrm{x}, \alpha) \\
&= \int_{\mathrm{X}} \sum_{\mathrm{k}=1}^{\infty} \mathrm{P}_\alpha(\mathrm{X_k} \in \mathrm{dx}, \tau_\alpha > \mathrm{k}) \mathrm{P}(\mathrm{x}, \alpha) \\
&= \sum_{\mathrm{k}=1}^{\infty} \int_{\mathrm{X}} \mathrm{P}_\alpha(\mathrm{X_k} \in \mathrm{dx}, \tau_\alpha > \mathrm{k}) \mathrm{P}(\mathrm{x}, \alpha) \\
&= \sum_{\mathrm{k}=1}^{\infty} \mathrm{P}_\alpha(\mathrm{X_{k+1}} = \alpha, \tau_\alpha = \mathrm{k}+1) \\
&= 1 = \nu(\alpha)
\end{aligned}
$$

Prawo wielkich liczb IV

TWIERDZENIE

Jeśli πf istnieje oraz oraz łańcuch jest π-nieprzywiedlny to

$$\frac{\sum_{k=1}^{n} f(X_i)}{n} \to \pi f, \quad \text{p.n.}$$

Niech T_i oznaczają kolejne czasy powrotu do α. Wówczas kawałki trajektori pomiędzy koljenymi czasami T_i

$$\Xi_k(f) = \sum_{k=T_{k-1}}^{T_k - 1} f(X_i)$$

są niezależne i z wyjatkiem pierwszej mają ten sam rozkład. Niech

$$T_{N(n)} \leqslant n \leqslant T_{N(n)+1}$$

Z PWL

$$T_k/k \to \nu(X)$$

Prawo wielkich liczb V

Z twierdzenia o trzech ciągach

$$\frac{n}{N(n)} \to \nu(X)$$

Dodatkowo jeśli $f \geqslant 0$ to

$$\sum_{k=1}^{N(n)} \Xi_k(f) \leqslant \sum_{i=1}^{n} f(X_i) \leqslant \sum_{k=1}^{N(n)+1} \Xi_k(f)$$

Ponownie z twierdzenia o trzech ciągach i PWL dla i.i.d

$$\frac{\sum_{k=1}^{n} f(X_i)}{n} \to \frac{\nu f}{\nu(X)} = \pi f$$

Centralne twierdzenie graniczne I

Centralne twierdzenie graniczne również mozna dowodzić techniką regeneracyjną. Jednak w tym przypadku, założenia oraz postać asymptotycznej wariancji łatwiej wyjaśnić korzystając z techniki martyngałowej.

Równanie Poissona Funkcja $\hat{f}$ spełnia równanie poissona jeśli

$$\hat{f}(x) - P\hat{f}(x) = f(x) - \pi f$$

Jeśli $\hat{f}$ rozwiązanie równania poissona to

$$\begin{aligned}\sum_{k=1}^{n} f(X_i) - \pi f &= \sum_{k=1}^{n} \hat{f}(X_i) - P\hat{f}(X_i) \\ &= \sum_{k=2}^{n} \hat{f}(X_i) - P\hat{f}(X_{i-1}) + \hat{f}(X_1) + P\hat{f}(X_n)\end{aligned}$$

Ostatnie kawałki po podzieleniu przez $\sqrt{n}$ są pomijalne a pierwszy tworzy martyngał. I chcemy skorzystać z CTG dla

Centralne twierdzenie graniczne II

martyngałów. Dodatkwow można pokazać, że CTG nie zależy od rozkładu początkowego, można założyć $X_1 \sim \pi$

$$\mathbb{E}((\hat{f}(X_i) - P\hat{f}(X_{i-1})^2 | X_{i-1}) = P\hat{f}^2(X_{i-1}) - (P\hat{f}(X_{i-1}))^2$$

Stąd

$$\sigma^2 = \pi(\hat{f}^2 - (P\hat{f})^2)$$

Oraz

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} f(X_i) - \pi f \to N(0, \sigma^2)$$

Do precyzyjnego sformułowania będziemy potrzebowali geometrycznej ergodyczności. Najpierw jednka przyjżymy się asymptotycznej wariancji.

Asymptotyczna wariancja I

Jedynym kandydatem na rozwiązanie równania Poissona jest

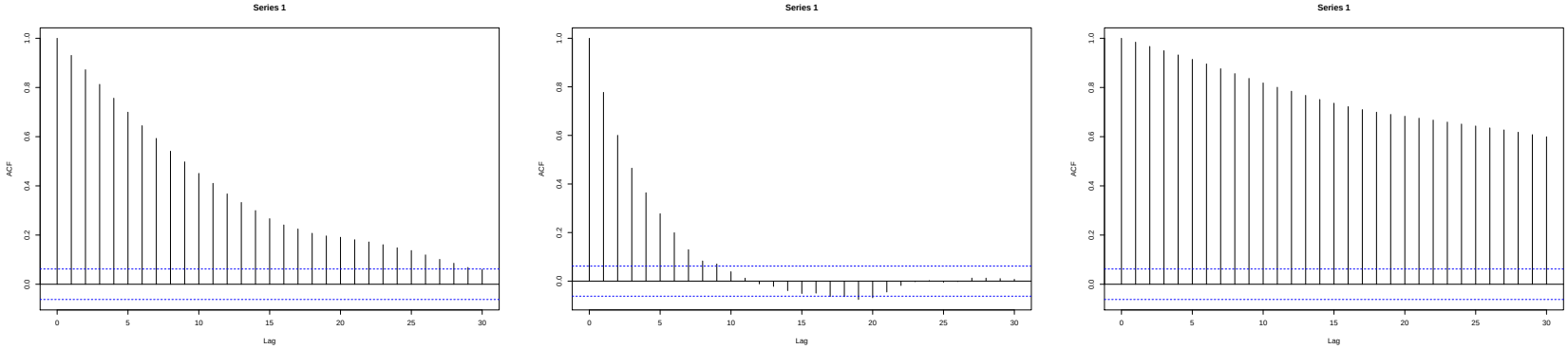
$$\hat{f} = \sum_{k=0}^{\infty} P^k f(x) - \pi f$$

Oznaczmy $\bar{f} = f - \pi f$ wówczas

$$\begin{aligned}
\sigma^2 &= \pi(\hat{f}^2 - (P\hat{f})^2) \\
&= \pi(\hat{f}^2 - (\hat{f} - \bar{f})^2) \\
&= \pi(2\bar{f}\hat{f} - \bar{f}^2) \\
&= 2\pi(\sum_{k=0}^{\infty} \bar{f}P^k\bar{f} - \bar{f}^2) \\
&= \pi(\bar{f}^2) + \sum_{k=1}^{\infty} \pi(\bar{f}P^k\bar{f}) \\
&= \mathrm{Var}_\pi(f(X)) + \sum_{k=2}^{\infty} \mathrm{cov}_\pi(X_1, X_k)
\end{aligned}$$

Asymptotyczna wariancja II

Stąd ważna w diagnostyce MCMC jest funkcja autokorelacji.



Tempo zbieżności

Jednostajna ergodyczość:

$$\|\mathrm{P}^{\mathrm{n}}(\mathrm{x},\cdot)-\pi(\cdot)\|_{\mathrm{tv}} \leqslant \mathrm{M}\rho^{\mathrm{n}}, \quad \rho<1$$

Geometryczna ergodyczność:

$$\|\mathrm{P}^{\mathrm{n}}(\mathrm{x},\cdot)-\pi(\cdot)\|_{\mathrm{tv}} \leqslant \mathrm{M}(\mathrm{x})\rho^{\mathrm{n}}, \quad \rho<1$$

Jednostajna ergodyczność

TWIERDZENIE

Jednostajana ergodycznośc jest równoważna warunkowi, że cała przestrzeń jest małym zbiorem czyli

$$\mathrm{P^n(x, A)} \geqslant \beta\mu(\mathrm{A})$$

$$\mathrm{P(X_{kn} \neq Y_{kn})} \leqslant \mathrm{P}(\tau > \mathrm{kn}) = (1-\beta^2)^{\mathrm{k}} = \rho^{\mathrm{kn}},$$

gdzie $\rho = (1-\beta^2)^{1/\mathrm{n}}$

Geometryczna ergodyczność

TWIERDZENIE

Łańcuch Markowa jest geometrycznie ergodyczny wtedy i tylko wtedy gdy istnieje mały zbiór C

$$\mathrm{P}^{\mathrm{n}}(\mathrm{x}, \mathrm{A}) \geqslant \beta\mu(\mathrm{A})1(\mathrm{x} \in \mathrm{C})$$

oraz istnieje funkcja $\mathrm{V} \geqslant 1$ taka, że

$$\mathrm{PV}(\mathrm{x}) \leqslant \lambda \mathrm{V}(\mathrm{x}) + \mathrm{K}1(\mathrm{x} \in \mathrm{C})$$

dla $\lambda < 1$

Powrót do CTG

CTG jest spełnione jeżeli łańcuch jest geometrycznie ergodyczny oraz jeden z warunkow jest spełniony:

- łańcuch jest odwracalny i $\pi(f^2) < \infty$
- $\pi(|f|^{2+\delta}) < \infty$
- $\sup_x \frac{f^2(x)}{V(x)} < \infty$

3.36pt

2 Ukryte modele Markowa
- Dyskretne modele Markowa
- Obliczanie prawdopodobieństwa obserwacji
- Forward Backward
- Algorytm Viterbiego
- Algorytm FFBS
- Algorytm Bauma-Welcha
- Modele gaussowskie i filtr Kalmana

3 Algorytmy MCMC
- Podstawowe pojęcia statystyki bayesowskiej
- Metody Monte Carlo
- Algorytm Metropolisa-Hastingsa
- Próbnik Gibbsa
- Własności teoretyczne MCMC
- Adaptacyjne algorytmy MCMC
- Hamiltonian Monte Carlo

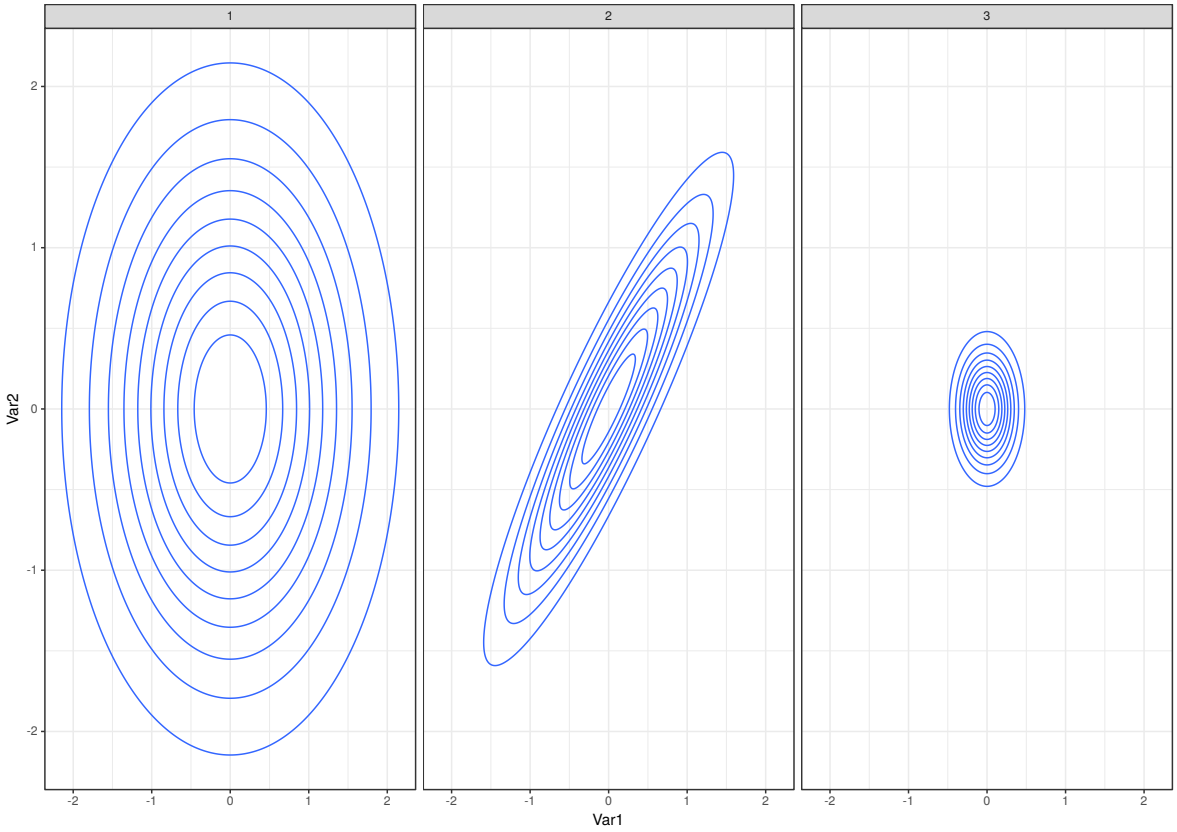
Problemy z algorytmem Metropolisa

RWM:

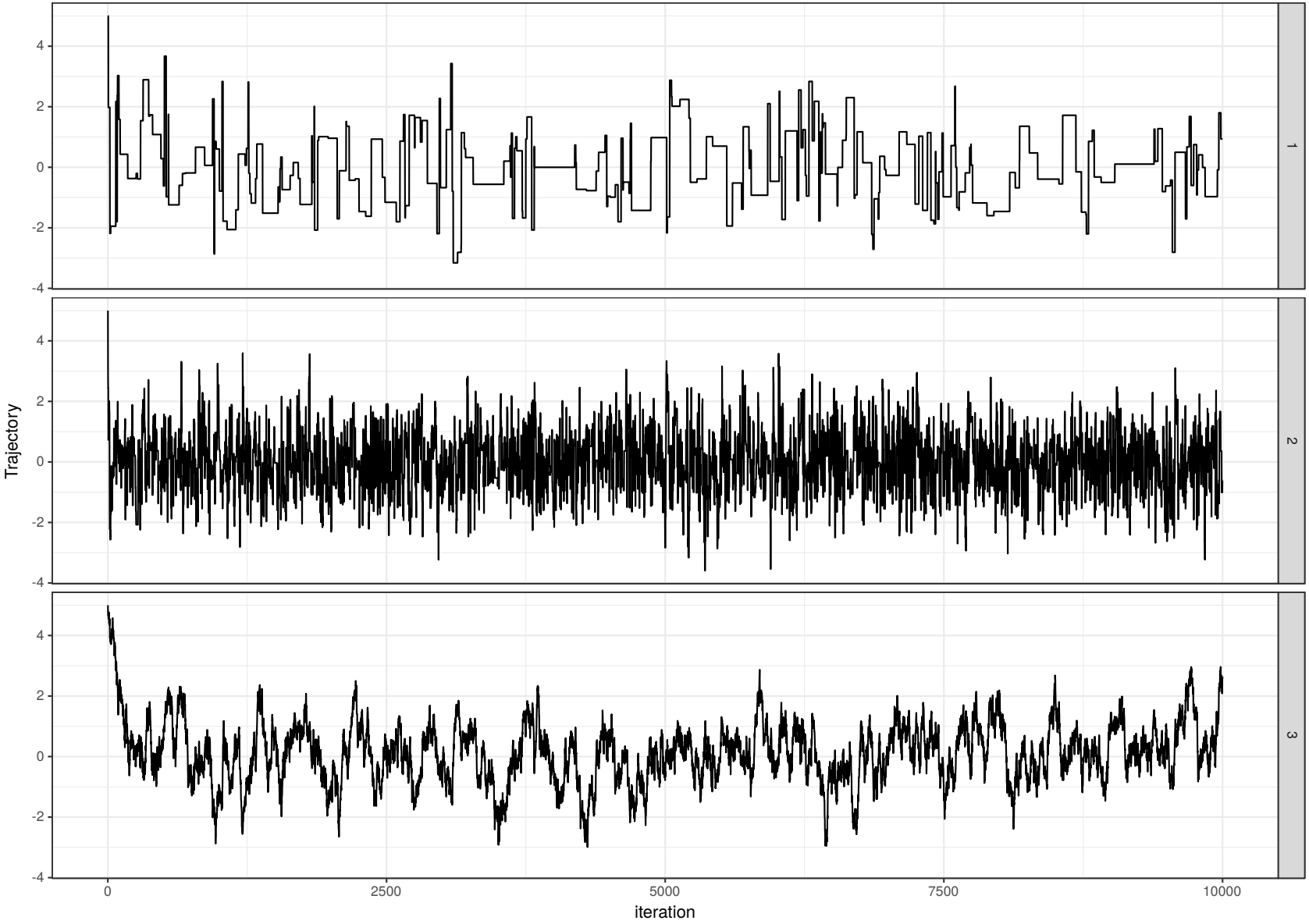
Propozycja $Y \sim N(X_n, \Sigma)$

Jak wybrać Σ?

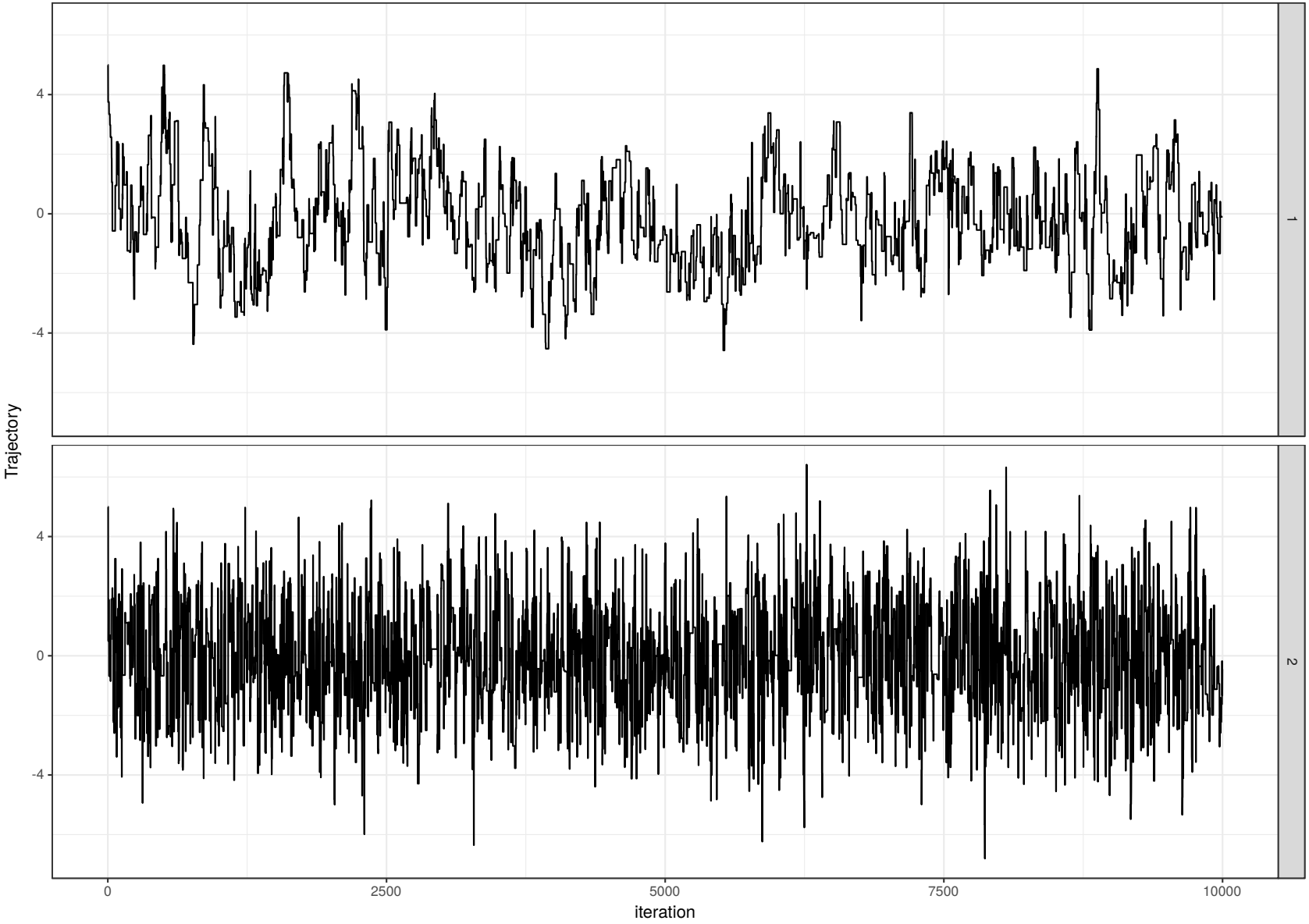
- Wielkość.
- Kształt.



Wielkość



Kształt



“Optymalny” wybór Σ

- Optymalana $\Sigma = \ell^2 \mathrm{Cov}_\pi(\mathrm{X})$, gdzie ℓ^2 można zdefiniować na dwa sposoby.
- $\ell^2 \approx \frac{2.38^2}{\mathrm{d}}$, gdzie d wymiar.
- ℓ^2 takie, że średnie prawdopodobieństwo akceptacji wynosi 0.234

Uzasadnienie teoretyczne I

Jeśli rozpatrzy się

$$\pi(x) = \prod_{i=1}^{d} f(x_i)$$

To można pokazać że dla $\Sigma = \frac{\ell^2}{d} I$ łańcuch Markowa generowany przez algorytm Metropolisa poprzeskalowaniu zbiega wraz z wymiarem do dyfuzji. Dokładniej

$$X_{\lfloor dt \rfloor} \to Y_t$$

gdzie

$$dY_t = \sqrt{h(l)} dW_t - h(l) \frac{1}{2} V'(Y_t)\, dt,$$

$$V(x) = -\log(f(x)),\ h(l) = 2l^2 \Phi\left(-\frac{l\sqrt{I}}{2}\right), I = \int (V')^2 \exp(-V).$$

Uzasadnienie teoretyczne II

Dodatkowo prawdopodobieństwo akceptacji zbiega do

$$\mathrm{a}(\mathrm{l}) = 2\Phi\left(-\frac{\ell\sqrt{\mathrm{I}}}{2}\right)$$

Funkcja h(l) odpowiada za szybkość dyfuzji i chcemy ją z maksymalizować.

$$\ell^* = \frac{2.38}{\sqrt{\mathrm{I}}} \text{ oraz } \mathrm{a}(\ell^*) = 0.234$$

Adaptacyjny RWM

Chcemy skonstrować algorytm, który automatycznie znajduje optymalną macierz Σ. W tym celu w każdym kroku algorytmu będziemy estymować macierz kowariancji oraz odpowiednio ją skalować. niech

$$\mu_n = \frac{n-1}{n}\mu_{n-1} + \frac{1}{n}X_n$$

$$\Gamma_n = \frac{n-1}{n}\Gamma_{n-1} + \frac{1}{n}(X_n - \mu_n)(X_n - \mu_n)^T$$

Wersja I:

$$\Sigma_n = \frac{2.38^2}{d}\Gamma_n + \epsilon I$$

Wersja II:

$$T_n = T_{n-1} + \frac{1}{n}(\alpha(X_n, Y_n) - 0.234)$$

$$\Sigma_n = \exp(T_n)\Gamma_n + \epsilon I$$

Uwagi o implementacji adaptacyjnego RWM I

Generowanie wielowymiarowgo rozkładu gaussowskiego

Jeśli $X = (X_1, \ldots, X_d)$ i.i.d N(0,1) oraz $\Sigma = LL^T$ to

$$LX \sim N(0, \Sigma)$$

Typowo używa się rozkładu Choleskiego. Dla którego istnieje one-rank-update Jeśli

$$M_{n+1} = M_n + xx^T$$

To odpowiadające im rozkłady Choleskiego

$$L_{n+1} = ORU(L_n, x).$$

W adaptacyjnym RWM wystarczy pamiętać L_n rozkład Choleskiego Γ_n i

$$L_{n+1} = ORU(\frac{\sqrt{n}}{\sqrt{n+1}} L_n, \frac{1}{\sqrt{n+1}}(X_{n+1} - \mu_{n+1}))$$

Uwagi o implementacji adaptacyjnego RWM II

Generowanie propozycji:

$$\mathrm{Y_n} = \mathrm{X_n} + \mathsf{exp}(\mathrm{T_n}/2)\mathrm{L_n}\mathrm{Z_n} + \sqrt{\epsilon}\mathrm{Z'_n}$$

gdzie $\mathrm{Z_n}$ i $\mathrm{Z'_n}$ niezależne $\mathrm{N}(0, \mathrm{I})$.

3.36pt

2 Ukryte modele Markowa
- Dyskretne modele Markowa
- Obliczanie prawdopodobieństwa obserwacji
- Forward Backward
- Algorytm Viterbiego
- Algorytm FFBS
- Algorytm Bauma-Welcha
- Modele gaussowskie i filtr Kalmana

3 Algorytmy MCMC
- Podstawowe pojęcia statystyki bayesowskiej
- Metody Monte Carlo
- Algorytm Metropolisa-Hastingsa
- Próbnik Gibbsa
- Własności teoretyczne MCMC
- Adaptacyjne algorytmy MCMC
- Hamiltonian Monte Carlo

Układ Hamiltonowski

$$H(p, x) = K(p) + U(x)$$

K - energia kinetyczna, U- energia potencjalna Układ Hamiltonowski

$$\frac{dx_i}{dt} = \frac{\partial H}{\partial p_i}$$
$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial x_i}$$

Rozwiązanie zachowuje energię.

Związek z rozkładami prawdopodobieństwa

Rozkład Gibbsa

$$\pi(\mathrm{p}, \mathrm{x}) \propto \mathsf{exp}(-\mathrm{H}(\mathrm{p}, \mathrm{x}))$$

Rozkład brzegowy dla dowolnego K

$$\pi(\mathrm{x}) \propto \mathsf{exp}(-\mathrm{U}(\mathrm{x}))$$

Rozkład warunkowych

$$\pi(\mathrm{p}|\mathrm{x}) \propto \mathsf{exp}(-\mathrm{K}(\mathrm{p})) = \pi(\mathrm{p})$$

Idealne Hamiltonowskie Monte Carlo

$X_n \to X_{n+1}$

- Wyolsuj $P \sim \pi(p)$.
- Zgodnie z Hamiltonowską dynamiką

$$(P, X_n) = (x_0, p_0) \to (x_T, p_T) = (X_{n+1}, P')$$

Potrzeba K zadające rozkład łatwy do generowania, oraz potrzeba umieć obliczyć rozwiązanie układu hamiltonowskiego w czasie T.

Hamiltonian MCMC I

$$K(p) = \frac{1}{2} p^T p$$

Czyli $\pi(p) = N(0, I)$. Dodatkowo potrzeba numerycznie rozwiązywać układ hamiltonowski.

Leapfrog integrator:

Dla $1 \leqslant n \leqslant T/\epsilon$

$$\begin{aligned} p_{n+1/2} &= p_n - \frac{\epsilon}{2} \frac{\partial U}{\partial x}(x_n) \\ x_{n+1} &= x_n + \epsilon p_{n+1/2} \\ p_{n+1} &= p_{n+1/2} - \frac{\epsilon}{2} \frac{\partial U}{\partial x}(x_{n+1}) \end{aligned}$$

Hamiltonian MCMC II

Jeżeli zamienimy znak $p_T \to -p_T$ to leapfrog integrator jest odwracalny. I żeby usunąć błąd numerycznego rozwiązania wystarczy dodać akcpetacje Metropolisa

$$\alpha((x_0, p_0), (x_T, -p_T)) = \mathsf{exp}(H(x_T, p_T) - H(x_0, p_0)) \wedge 1$$

Hamiltonian MCMC III

Algorytm

1. Generujemy $\mathrm{p} \sim \mathrm{N}(0, \mathrm{I})$
2. Przybliżamy rozwiązanie układu hamiltonowskiego (leapfrog integrator) $\mathrm{X_n} = \mathrm{x_0}, \mathrm{p} = \mathrm{p_0} \to \mathrm{x_T}, \mathrm{p_T}$
3. $$\mathrm{X_{n+1}} = \begin{cases} \mathrm{x_T} & \text{z prawdopodobieństwem } \alpha((\mathrm{x_0}, \mathrm{p_0}), (\mathrm{x_T}, -\mathrm{p_T})) \\ \mathrm{X_n} & \text{z prawdopodobieństwem } 1 - \alpha((\mathrm{x_0}, \mathrm{p_0}), (\mathrm{x_T}, -\mathrm{p_T})) \end{cases}$$