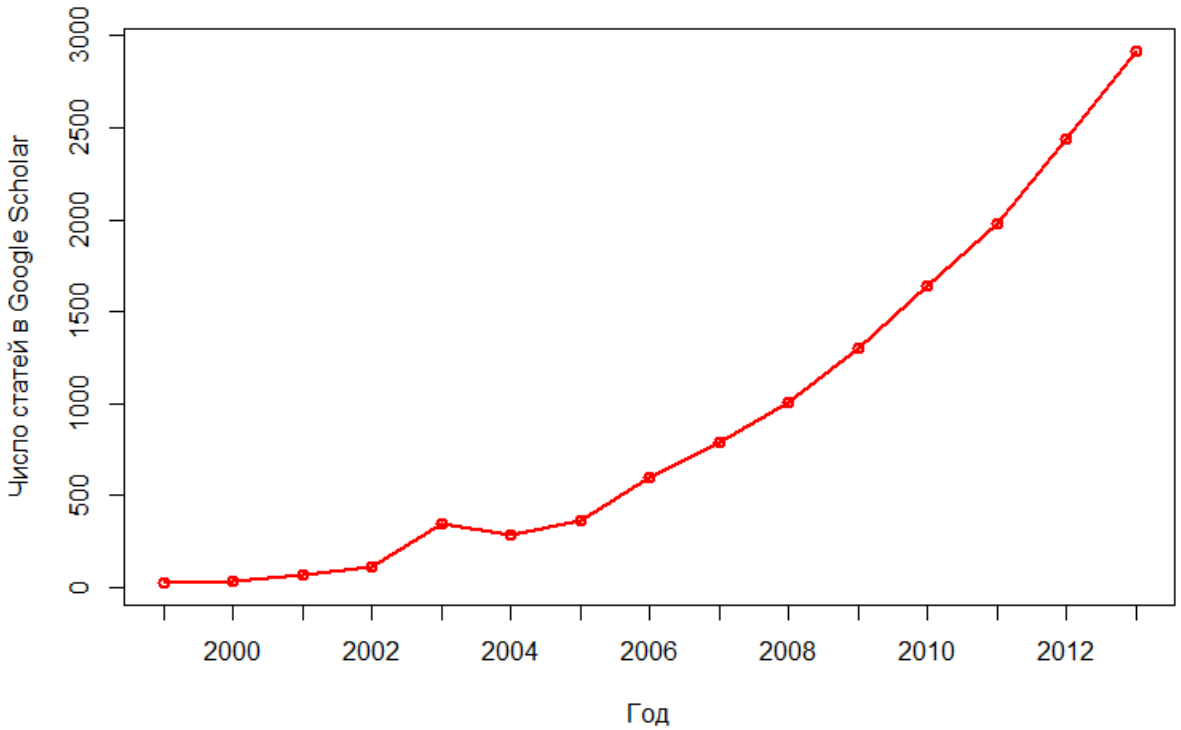
Неотрицательные матричные разложения

Рябенко Евгений
riabenko.e@gmail.com

Семинар «Стохастический анализ в задачах»
Независимый Московский университет
19 апреля 2014 г.

Неотрицательное матричное разложение (NMF)



$$P \approx AX \equiv Q,$$

$$P, Q \in \mathbb{R}_+^{m \times n}, \quad A \in \mathbb{R}_+^{m \times k}, \quad X \in \mathbb{R}_+^{k \times n}, \quad r < \min(m, n).$$

Оптимизационная задача:

$$(A^*, X^*) = \underset{A \geqslant 0,\, X \geqslant 0}{\operatorname{argmin}} \; D(P, AX). \tag{1}$$

Функции потерь (дивергенции)

$$D(P,Q) = \sum_{i=1}^{m}\sum_{j=1}^{n} d(p_{ij}, q_{ij}),$$

$$d(p,q) \geqslant 0,$$

$$d(p,q) = 0 \Leftrightarrow p = q.$$

Название	$d(p,q)$
норма l_1	$d_1(p,q) = \lvert p-q \rvert$
норма Фробениуса	$d_F(p,q) = (p-q)^2$
обобщённая дивергенция Кульбака-Лейблера	$d_{KL}(p,q) = p\ln\frac{p}{q} - p + q$
дивергенция Итакура-Саито	$d_{IS}(p,q) = \ln\frac{q}{p} + \frac{p}{q} - 1$
расстояние Хеллингера	$d_H(p,q) = \left(\sqrt{p} - \sqrt{q}\right)^2$
χ^2 Пирсона	$d_P(p,q) = \frac{(p-q)^2}{q}$
χ^2 Неймана	$d_N(p,q) = d_P(q,p) = \frac{(p-q)^2}{p}$

Больше примеров можно найти в [Cichocki et al., 2009].

Функции потерь (дивергенции)

Во многих случаях дивергенция — это замаскированное правдоподобие: существует такая функция плотности $p(P|Q)$, что

$$-\ln p(P|Q) = aD(P,Q) + b$$

для каких-то a и b.

Дивергенция	Порождающая модель	$p(P\|Q)$
Фробениуса	аддитивная гауссовская	$\prod_{ij} N\left(p_{ij}\|q_{ij},\sigma^2\right)$
Кульбака-Лейблера	пуассоновская	$\prod_{ij} P\left(p_{ij}\|q_{ij}\right)$
Итакура-Саито	мультипликативная гамма	$\prod_{ij} G\left(p_{ij}\|\alpha,\alpha/q_{ij}\right)$

Функции потерь (дивергенции)

α-дивергенция:

$$d_A^\alpha(p,q) = \begin{cases} \frac{1}{\alpha(\alpha-1)}\left(p^\alpha q^{1-\alpha} - \alpha p + (\alpha-1)q\right), & \alpha \neq 0,\ \alpha \neq 1, \\ p\ln\frac{p}{q} - p + q, & \alpha = 1, \\ q\ln\frac{q}{p} - q + p, & \alpha = 0. \end{cases}$$

Непрерывно соединяет хи-квадрат Пирсона, обобщённую дивергенцию Кульбака-Лейблера, расстояние Хеллингера и хи-квадрат Неймана.

β-дивергенция:

$$d_B^\beta(p,q) = \begin{cases} \frac{1}{\beta(\beta+1)}\left(p^{\beta+1} - q^{\beta+1} - (\beta+1)q^\beta(p-q)\right), & \beta \neq 0,\ \beta \neq -1, \\ p\ln\frac{p}{q} - p + q, & \beta = 0, \\ \ln\frac{q}{p} + \frac{p}{q} - 1, & \beta = -1. \end{cases}$$

Непрерывно соединяет обобщённую дивергенцию Кульбака-Лейблера, дивергенцию Итакура-Саито и норму Фробениуса.

Функции потерь (дивергенции)

АБ-дивергенция:

$$d_{AB}^{(\alpha,\beta)}(p,q) = \begin{cases} \frac{1}{\alpha\beta}\left(\frac{\alpha}{\alpha+\beta}p^{\alpha+\beta} + \frac{\beta}{\alpha+\beta}q^{\alpha+\beta} - p^{\alpha}q^{\beta}\right), & \alpha,\beta,\alpha+\beta \neq 0, \\ \frac{1}{\alpha^2}\left(p^{\alpha}\ln\frac{p^{\alpha}}{q^{\alpha}} - p^{\alpha} + q^{\alpha}\right), & \alpha \neq 0,\ \beta = 0, \\ \frac{1}{\alpha^2}\left(\ln\frac{q^{\alpha}}{p^{\alpha}} + \left(\frac{q^{\alpha}}{p^{\alpha}}\right)^{-1} - 1\right), & \alpha = -\beta \neq 0, \\ \frac{1}{\beta^2}\left(q^{\beta}\ln\frac{q^{\beta}}{p^{\beta}} - q^{\beta} + p^{\beta}\right), & \alpha = 0,\ \beta \neq 0, \\ \frac{1}{2}\left(\ln p - \ln q\right)^2, & \alpha = \beta = 0. \end{cases}$$

Объединяет α- и β-дивергенции, пересекающиеся в обобщённой дивергенции Кульбака-Лейблера.

Функции потерь (дивергенции)

Дивергенция Брегмана:

$$d_\phi(p, q) = \phi(p) - \phi(q) - \nabla\phi(q)(p - q),$$

$\phi(p)$ — произвольная строго выпуклая вещественнозначная функция, $\nabla\phi(q)$ — её производная по q.

$\phi(p)$	Дивергенция
$p^2/2$	Фробениуса
$p \ln p$	Кульбака-Лейблера
$-\ln p$	Итакура-Саито
$\begin{cases} \frac{1}{\beta(\beta+1)}\left(p^{\beta+1} - (\beta+1)p + \beta\right), & \beta > -1, \\ p\ln p - p + 1, & \beta = 0, \\ p - \ln p - 1, & \beta = -1 \end{cases}$	β-дивергенция

Проблемы NMF

- NMF NP-трудна [Vavasis, 2009].
- NMF некорректно поставлена: если $A_0,\ X_0$ — решение задачи (1), то пара $A = A_0 D,\ X = D^{-1} X_0$ тоже является решением (при условии, что матрица перехода D невырождена и сохраняет неотрицательность компонент разложения).
- Все $D\left(P, AX\right)$ не выпуклы по совокупности аргументов (A, X), поэтому чаще всего используют блочно-покоординатные методы минимизации:

  Вход: $A^0 \geqslant 0, X^0 \geqslant 0$

  Цикл *// t = 1, 2, ...*

  $X^t = f\left(P, A^{t-1}, X^{t-1}\right);$

  $\left(A^t\right)^T = f\left(P^T, \left(X^t\right)^T, \left(A^{t-1}\right)^T\right).$

- Лучшее, что можно гарантировать — сходимость к стационарной точке, определяемой условиями Каруша-Куна-Таккера:

$$\begin{aligned} &X^* \geqslant 0, \quad \nabla_X D\left(P, A^* X^*\right) \geqslant 0, \quad X^* \otimes \nabla_X D\left(P, A^* X^*\right) = 0, \\ &A^* \geqslant 0, \quad \nabla_A D\left(P, A^* X^*\right) \geqslant 0, \quad A^* \otimes \nabla_A D\left(P, A^* X^*\right) = 0. \end{aligned} \tag{2}$$

Условия единственности разложения

Бесполезные

- Достаточное: пусть матрица P имеет ранг r и у неё есть точное неотрицательное разложение, тогда оно единственно, если P содержит r ненулевых столбцов, в каждом из которых есть $r-1$ ноль, и профили разреженности всех $r-1$ строк, содержащих нули, различны [Gillis, 2012].
- Достаточное: определим

$$C \equiv \left\{ y \in \mathbb{R}^r \,\middle|\, y^T 1 \geqslant \sqrt{r-1}\, \|y\|_2 \right\},$$

$$bdC* \equiv \left\{ y \in \mathbb{R}^r \,\middle|\, y^T 1 = \|y\|_2 \right\},$$

$$\operatorname{cone}\left(X^T\right) \equiv \left\{ y = X^T \lambda \,|\, \lambda \geqslant 0 \right\},$$

$$\operatorname{cone}\left(X^T\right)^* \equiv \left\{ x \,\middle|\, y^T x \geqslant 0 \;\; \forall y \in \operatorname{cone}\left(X^T\right) \right\}.$$

Если $\operatorname{cone}\left(X^T\right) \supseteq C$ и

$$bdC* \cap \operatorname{cone}\left(X^T\right)^* = \{ \lambda e_k \,|\, \lambda \geqslant 0, k = 1, \ldots, r \},$$

и аналогичное верно для A^T, то разложение единственно [Huang et al., 2014]. Проверка этого условия — NP-трудная задача.

Условия единственности разложения

Относительно полезные

- Необходимое: носитель каждого столбца X (и строки A) не может являться подмножеством носителя любого другого столбца X (строки A) [Laurberg, Christensen, 2008].
- Достаточное: одна из двух матриц A и X содержит диагональную подматрицу размера r [Donoho, Stodden, 2004, Laurberg, Christensen, 2008].

NMF с нормой Фробениуса

Оптимизационная задача:

$$(A^*, X^*) = \underset{A \geqslant 0,\ X \geqslant 0}{\operatorname{argmin}} \|P - AX\|_F^2 . \tag{3}$$

Без ограничения неотрицательности решение можно было бы получить с помощью SVD.

Базовый метод — поочерёдный градиентный спуск:

$$x_{kj} \leftarrow x_{kj} - \nu_{kj} \frac{\partial D\left(P, AX\right)}{\partial x_{kj}},$$

$$a_{ik} \leftarrow a_{ik} - \eta_{ik} \frac{\partial D\left(P, AX\right)}{\partial a_{ik}}.$$

$k = 1, \ldots, r,\ \ i = 1, \ldots, m,\ \ j = 1, \ldots, n.$

Мультипликативные обновления

Один из первых методов был предложен в [Lee, Seung, 1999, 2001]. Идея: выбрать шаги градиентного спуска так, чтобы обновления стали мультипликативными, тогда будет сохраняться неотрицательность элементов матриц.

$$[\nabla_X]_{kj} = \frac{\partial D(P, AX)}{\partial x_{kj}} = \left[\nabla_X^+\right]_{kj} - \left[\nabla_X^-\right]_{kj},$$

$$\nu_{kj} = \frac{x_{kj}}{\left[\nabla_X^+\right]_{kj}},$$

$$x_{kj} \leftarrow x_{kj} - \frac{x_{kj}}{\left[\nabla_X^+\right]_{kj}} \left(\left[\nabla_X^+\right]_{kj} - \left[\nabla_X^-\right]_{kj}\right) =$$

$$= x_{kj} \frac{\left[\nabla_X^-\right]_{kj}}{\left[\nabla_X^+\right]_{kj}};$$

в матричном виде:

$$X \leftarrow X \otimes \nabla_X^- \oslash \nabla_X^+.$$

$\otimes$ — поэлементное умножение матриц, $\oslash$ — поэлементное деление.

Мультипликативные обновления

MU для нормы Фробениуса:

$$\nabla_X = A^T AX - A^T X,$$

$$X \leftarrow X \otimes \left(A^T P\right) \oslash \left(A^T AX\right),$$

$$x_{kj} \leftarrow x_{kj} \frac{\left[A^T P\right]_{kj}}{\left[A^T AX\right]_{kj}}.$$

Правая часть — глобальный минимум **дополнительной функции** (квадратичной функции, мажорирующей $D_F\left(P, AX\right)$ на текущей итерации) [Lee, Seung, 2001]. Следовательно, функция потерь монотонно невозрастает.

Является ли предельная точка мультипликативного алгоритма стационарной точкой задачи (3)?

Алгоритм может застревать в нестационарных точках вблизи нулей: если $x_{kj} = 0$, то он останется равным нулю, даже если $\left[\nabla_X\right]_{kj} < 0$.

Мультипликативные обновления

Варианты модификации мультипликативного алгоритма.

- Отделять элементы A и X небольшой положительной константой ε [Gillis, Glineur, 2012, Hibi, Takahashi, 2011]:

$$X \leftarrow \max\left(\varepsilon, X \otimes \left(A^T P\right) \oslash \left(A^T A X\right)\right).$$

Показано, что алгоритм с такими обновлениями сходится к стационарной точке модифицированной задачи:

$$(A_\varepsilon^*, X_\varepsilon^*) = \underset{A \geqslant \varepsilon,\, X \geqslant \varepsilon}{\operatorname{argmin}} \|P - AX\|_F^2, \tag{4}$$

Для $A = A_\varepsilon^* \otimes [A_\varepsilon^* > \varepsilon]$, $X = X_\varepsilon^* \otimes [X_\varepsilon^* > \varepsilon]$ условия стационарности исходной задачи выполняются с точностью до $\mathcal{O}(\varepsilon)$:

$$\forall k, j \quad \begin{bmatrix} x_{kj} = 0, & [\nabla_X]_{kj} \geqslant -\mathcal{O}(\varepsilon), \\ x_{kj} > 0, & \left|[\nabla_X]_{kj}\right| \leqslant \mathcal{O}(\varepsilon). \end{bmatrix}$$

Мультипликативные обновления

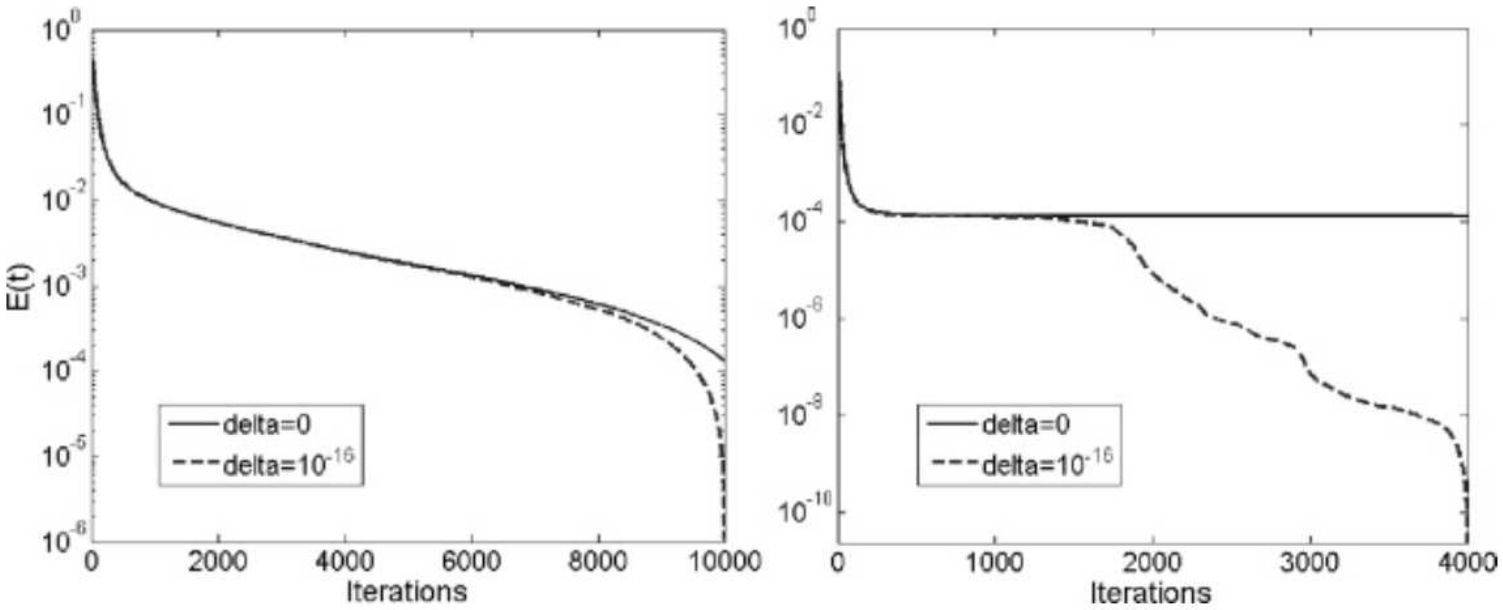
- Использовать оригинальные мультипликативные обновления, но после каждого шага реинициализировать небольшой положительной константой ε только те элементы, для которых соответствующая компонента градиента отрицательна [Chi, Kolda, 2012].

Мультипликативные обновления

$$e^t = \left\| P - A^t X^t \right\|_F,$$

$$E^t = \frac{e^t - e_{\min}}{e^0 - e_{\min}},$$

$e_{\min}$ — наименьшая итоговая ошибка при многократной инициализации.



[Gillis, Glineur, 2012]: E^t для обычного (delta=0) и модифицированного (delta=10^{-16}) мультипликативных алгоритмов на плотных (слева) и разреженных (справа) данных.

Мультипликативные обновления

Пусть Z — число ненулевых элементов матрицы P, тогда на обновление X требуется:

шаг	флоп
$M_1 = A^T P$	$2Zr$
$M_2 = A^T A$	$2mr^2$
$M_3 = M_2 X$	$2nr^2$
$X \leftarrow X \otimes M_1 \oslash M_3$	$2nr$

$r\,(2Z + 2mr + 2nr + 2n)$ флоп.

Больше всего операций $(2Zr + 2mr^2)$ требуется для вычисления $A^T P$ и $A^T A$; эффективнее вычислить их один раз и сделать несколько итераций по X.

Относительная стоимость первого обновления:

$$\rho_X = \frac{2r\,(Z + mr + nr + n)}{2nr\,(r+1)} = 1 + \frac{Z + mr}{n\,(r+1)};$$

при $Z \geqslant r\,(m+n) \quad \rho_X \geqslant 2\frac{r}{r+1}$.

Мультипликативные обновления

Число внутренних итераций l можно определять в соответствии с динамическим критерием:

$$\left\| X^{t,l} - X^{t,l-1} \right\|_F \leqslant \epsilon \left\| X^{t,1} - X^{t,0} \right\|_F ,$$

или брать $l = 1 + \alpha\rho_X$, подбирая $\alpha \geqslant 0$.

На практике хорошо работает комбинация этих двух методов [Gillis, Glineur, 2012].

Мультипликативные обновления

Методы с мультипликативными обновлениями очень популярны, потому что они:

- просты в реализации;
- хорошо масштабируются и легко приспосабливаются к работе с разреженными матрицами;
- были предложены в самой первой работе по NMF.

Известно, что скорость их сходимости невелика [Han et al., 2009]; однако её можно существенно увеличить, если обновлять A и X по несколько раз подряд [Gillis, Glineur, 2012].

Метод попеременных наименьших квадратов

Alternating Least Squares (ALS): на каждом шаге находится решение задачи наименьших квадратов по одной из компонент, а затем проецируется на неотрицательную область.

$$X \leftarrow \max\left(\underset{Y \in \mathbb{R}^{k \times n}}{\operatorname{argmin}} \|P - AY\|_F, 0\right) =$$
$$= \max\left(\left(A^T A\right)^{-1} A^T P, 0\right).$$

Проецирование портит решение; его можно улучшить, если на каждом шаге умножать обновляемую компоненту на

$$\alpha^* = \underset{\alpha \geqslant 0}{\operatorname{argmin}} \|P - \alpha AX\|_F = \frac{\langle PX^T, A \rangle}{\langle A^T A, XX^T \rangle}. \quad (5)$$

Метод очень грубый: итерационный процесс не сходится, функция потерь осциллирует [Gillis, 2014]. Можно использовать для инициализации других методов.

Метод попеременных неотрицательных наименьших квадратов

Alternating Nonnegative Least Squares (ANLS) [Berry et al., 2007]: на каждом шаге точно находится покомпонентный минимум в неотрицательной области.

$$X \leftarrow \underset{X \geqslant 0}{\operatorname{argmin}} \|P - AX\|_F .$$

Покомпонентные минимумы можно находить с помощью методов активных ограничений [Kim, Park, 2007, 2008, 2011], проекции градиента [Lin, 2007], квазиньютоновских методов [Zdunek, Cichocki, 2006], оптимального градиентного метода Нестерова [Guan et al., 2012].

Для сходимости метода блочно-покоординатного спуска с двумя блоками к стационарной точке достаточно, чтобы D была непрерывно дифференцируемой, а допустимое множество — декартовым произведением замкнутых выпуклых множеств [Grippo, Sciandrone, 2000].

Каждая итерация требует существенных вычислительных затрат: лучший — метод Нестерова — на $Kr^2\,(m+n)$ флоп больше MU. Можно использовать для уточнения решения, найденного более простыми методами.

Метод иерархических попеременных наименьших квадратов

Hierarchical alternating least squares (HALS) [Cichocki et al., 2007, Cichocki, Phan, 2009], а также [Ho, 2008, Gillis, Glineur, 2010, Li, Zhang, 2009]:
на каждом шаге находится точный минимум в неотрицательной области по столбцу A или строке X.

Пусть фиксированы все переменные, кроме x_{kj}.

$$x_{kj} \leftarrow \underset{x_{kj} \geqslant 0}{\operatorname{argmin}} \|P - AX\|_F = \max\left(0, \frac{A_{:k}^T P_{:j} - \sum_{l \neq k} A_{:l}^T A_{:l} x_{lj}}{A_{:k}^T A_{:k}}\right).$$

Это решение не зависит от других x в k-й строке, поэтому можно находить точный минимум сразу для всей строки:

$$X_{k:} \leftarrow \underset{X_{k:} \geqslant 0}{\operatorname{argmin}} \|P - AX\|_F = \max\left(0, \frac{A_{:k}^T P - \sum_{l \neq k} A_{:k}^T A_{:l} X_{l:}}{A_{:k}^T A_{:k}}\right).$$

Метод иерархических попеременных наименьших квадратов

Удобно записать обновления через матрицу частичных остатков:

$$P^{(k)} \equiv P - \sum_{l \neq k} A_{:l} X_{l:},$$

$$X_{k:} \leftarrow \max\left(0, \frac{A_{:k}^T P^{(k)}}{A_{:k}^T A_{:k}}\right).$$

Метод иерархических попеременных наименьших квадратов

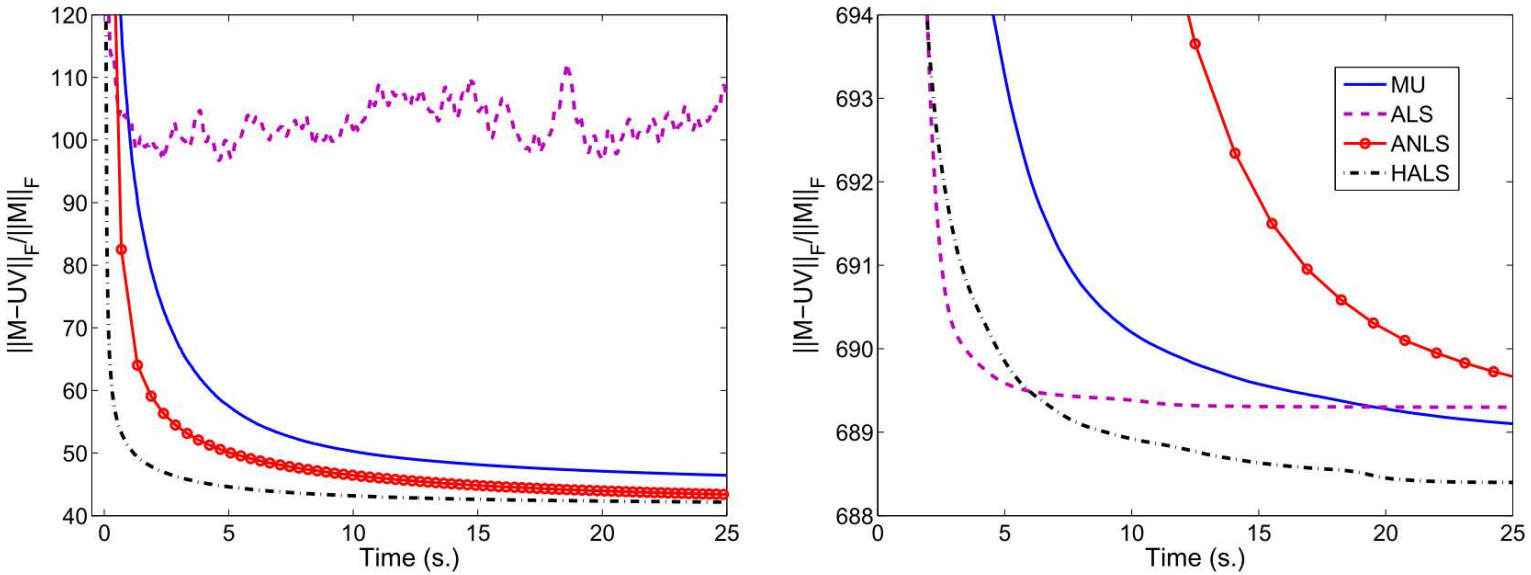
Особенности метода.

- Как и MU, HALS может застревать в нестационарных точках на границе. Можно использовать аналогичные модификации:

$$X_{k:} \leftarrow \max\left(\varepsilon, \frac{A_{:k}^T P - \sum_{l \neq k} A_{:k}^T A_{:l} X_{l:}}{A_{:k}^T A_{:k}}\right).$$

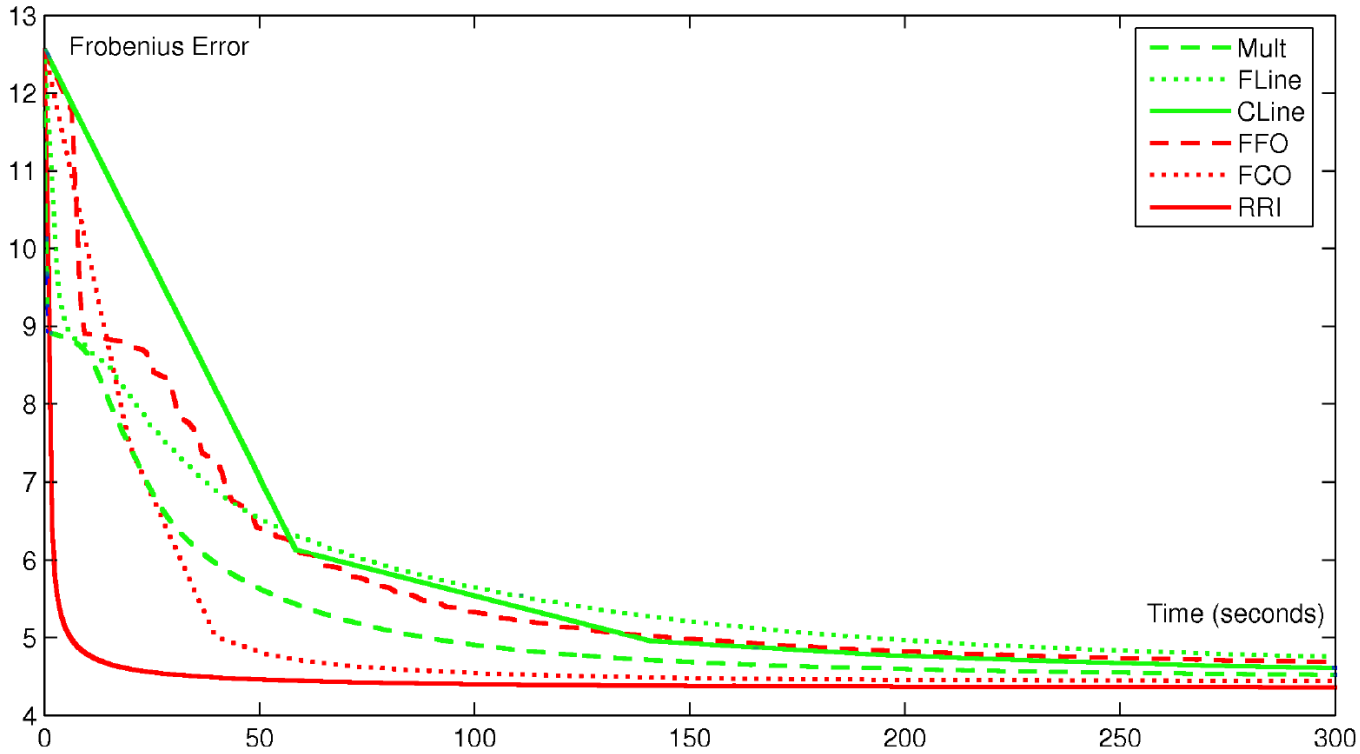
  Любая предельная точка такого алгоритма является стационарной точкой модифицированной задачи (4) [Gillis, Glineur, 2012].
- Метод чувствителен к начальному приближению, особенно к плохой нормировке: лучше использовать (5).
- Обновление X требует почти того же числа операций, что и MU ($r\left(2Z + 2mr + 2nr + n\right)$ флоп), но HALS сходится быстрее.
- Как и MU, HALS можно ускорить, обновляя X по несколько раз [Gillis, Glineur, 2012], но ещё выгоднее выбирать строки X по правилу типа Гаусса-Саутвелла (строки с минимальным градиентом) [Hsieh, Dhillon, 2011].

Сравнение алгоритмов



[Gillis, 2014]: зависимость относительной точности приближения $\frac{\|P-AX\|_F}{\|P\|_F}$ от времени работы рассмотренных алгоритмов на плотных (слева) и разреженных (справа) данных.

Сравнение алгоритмов



[Ho, 2008]: зависимость абсолютной точности приближения от времени работы рассмотренных алгоритмов на плотных данных. Mult — мультипликативные обновления, FLine, CLine, FFO, FCO — различные методы второго порядка, RRI — HALS.

Инициализация

Поскольку методы сходятся локально, начальное приближение играет большую роль. Следующие методы используются для построения начального приближения.

- Случайная инициализация.
- Кластеризация [Casalino et al., 2014]: центроиды, полученные при кластеризации столбцов P, инициализируют столбцы A, а X инициализируется с помощью матрицы принадлежности к кластерам ($x_{kj} \neq 0 \Leftrightarrow P_{:j}$ принадлежит кластеру k).
- SVD [Boutsidis, Gallopoulos, 2008]: для P находятся r наибольших сингулярных троек (u_k, s_k, v_k).

$A^0_{:1} = u_1\sqrt{s_1}, \quad X^0_{1:} = v_1\sqrt{s_1}$;

Цикл // $k = 2, \ldots, r$

Если $\left\|u_k^+\right\| \cdot \left\|v_k^+\right\| > \left\|u_k^-\right\| \cdot \left\|v_k^-\right\|$, **то**

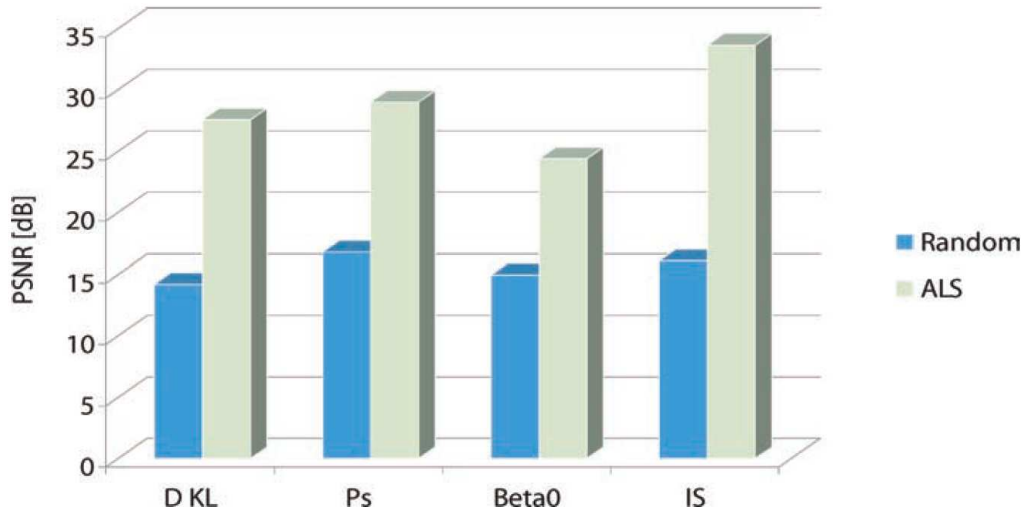
$$A^0_{:k} = u_k^+\sqrt{s_k\frac{\left\|v_k^+\right\|}{\left\|u_k^+\right\|}}, \quad X^0_{k:} = v_k^+\sqrt{s_k\frac{\left\|u_k^+\right\|}{\left\|v_k^+\right\|}},$$

иначе

$$A^0_{:k} = u_k^-\sqrt{s_k\frac{\left\|v_k^-\right\|}{\left\|u_k^-\right\|}}, \quad X^0_{k:} = v_k^-\sqrt{s_k\frac{\left\|u_k^-\right\|}{\left\|v_k^-\right\|}}.$$

Инициализация

- ALS [Cichocki et al., 2009]:



- Мультистарт [Cichocki et al., 2009]:
  1. с помощью ALS генерируются 10-20 пар матриц;
  2. делаются 10-20 итераций целевого метода на каждой паре;
  3. в качестве начального приближения выбирается пара с наименьшим значением функционала.

Для методов, требующих отделения A и X от нуля, инициализацию можно модифицировать.

NMF с обобщённой дивергенцией Кульбака-Лейблера

Оптимизационная задача:

$$(A^*, X^*) = \underset{A \geqslant 0, X \geqslant 0}{\operatorname{argmin}} D_{KL}(P, AX) = \\ = \underset{A \geqslant 0, X \geqslant 0}{\operatorname{argmin}} \sum_{i=1}^{m} \sum_{j=1}^{n} \left(p_{ij} \ln \frac{p_{ij}}{[AX]_{ij}} - p_{ij} + [AX]_{ij} \right). \quad (6)$$

Мультипликативные обновления

Алгоритм с мультипликативными обновлениями записывается по аналогии [Lee, Seung, 2001]:

$$\nabla_X = A^T J_{m\times n} - A^T \left(P \oslash (AX)\right),$$

$$X \leftarrow X \otimes \left(A^T \left(P \oslash (AX)\right)\right) \oslash \left(A^T J_{m\times n}\right),$$

$$[\nabla_X]_{kj} = \sum_{i=1}^{m} a_{ik} - \sum_{i=1}^{m} a_{ik} \frac{p_{ij}}{q_{ij}},$$

$$x_{kj} \leftarrow x_{kj} \frac{\sum\limits_{i=1}^{m} a_{ik} \frac{p_{ij}}{q_{ij}}}{\sum\limits_{i=1}^{m} a_{ik}}.$$

Обновления минимизируют дополнительную функцию, функция потерь монотонно невозрастает. Однако предельная точка алгоритма может не быть стационарной.

Квазиньютоновский метод

Единственный (?) предложенный метод второго порядка [Zdunek, Cichocki, 2006]:

$$X \leftarrow \max\left(0, X - H_X^{-1}\nabla_X\right),$$
$$H_X = \operatorname{diag}\left\{h_{X,j}, j = 1, \ldots, m\right\} \in \mathbb{R}^{kr \times kr},$$
$$h_{X,j} = A^T \operatorname{diag}\left\{\left[P \oslash (Q \otimes Q)\right]_{:,j}\right\} A \in \mathbb{R}^{r \times r}.$$

Чтобы обращать гессиан, делается регуляризация по методу Левенберга-Марквардта.

Алгоритм записан для α-дивергенции (при $\alpha = 1$ она превращается в дивергенцию Кульбака-Лейблера).

Связь с PLSA

Мультипликативный алгоритм для NMF с обобщённой дивергенцией Кульбака-Лейблера тесно связан с EM-PLSA.

E-шаг PLSA:

$$n_{dwt} = n_{dw} \frac{\phi_{wt}\theta_{td}}{\sum\limits_{s\in T} \phi_{ws}\theta_{sd}} \quad \rightsquigarrow \quad a_{ik}x_{kj}\frac{p_{ij}}{q_{ij}};$$

M-шаг:

$$\theta_{td} \leftarrow \frac{\sum\limits_{w\in W} n_{dwt}}{\sum\limits_{w\in W}\sum\limits_{s\in T} n_{dwt}} \quad \rightsquigarrow \quad x_{kj} \leftarrow x_{kj} \frac{\sum\limits_{i=1}^{m} a_{ik}\frac{p_{ij}}{q_{ij}}}{\sum\limits_{l=1}^{r} x_{lj} \sum\limits_{i=1}^{m} a_{il}\frac{p_{ij}}{q_{ij}}}.$$

KL-NMF

$$\frac{x_{kj}\sum\limits_{i=1}^{m} a_{ik}\frac{p_{ij}}{q_{ij}}}{\sum\limits_{i=1}^{m} a_{ik}}$$

EM-PLSA

$$\frac{x_{kj}\sum\limits_{i=1}^{m} a_{ik}\frac{p_{ij}}{q_{ij}}}{\sum\limits_{l=1}^{r} x_{lj} \sum\limits_{i=1}^{m} a_{il}\frac{p_{ij}}{q_{ij}}}$$

Связь с PLSA

[Ding et al., 2006, 2008]: PLSA и KL-NMF минимизируют один и тот же функционал, но по-разному, и поэтому могут сходиться в разные локальные минимумы.

Как показывают эксперименты, PLSA может выбираться из локальных минимумов, в которых застревает KL-NMF, и наоборот $\Rightarrow$ можно использовать гибридный алгоритм.

Применение к стохастическим матрицам

NMF может применяться к матрицам вероятностей, нормированных, например, по столбцам:

$$\sum_{i=1}^{n} p_{ij} = 1 \ \ \forall j.$$

Показано, что в стационарных точках задачи (6) суммы по строкам и столбцам матриц P и $Q^* = A^* X^*$ совпадают [Morup, Hansen, 2010], то есть, модельная матрица Q^* тоже является стохастической. Однако алгоритмов, гарантирующих сходимость к стационарной точке, нет.

Применение к стохастическим матрицам

Ещё больше проблем возникает, когда матрицы A и X тоже должны быть стохастическими:

$$\sum_{i=1}^{n} a_{ik} = 1, \quad \sum_{k=1}^{r} x_{kj} \quad \forall j, k.$$

Как это явно учесть при оптимизации?

- Нормировка [Lee, Seung, 1999]: после каждого шага обновлявшаяся матрица нормируется:

$$x_{kj} \leftarrow \frac{x_{kj}}{\sum_{l=1}^{r} x_{lj}}.$$

  Может нарушаться монотонность убывания функции потерь.
- Репараметризация [Zhu et al., 2013]: перейдём от матрицы X к Y:

$$x_{kj} = \frac{y_{kj}}{\sum_{l=1}^{r} y_{lj}},$$

$$\frac{\partial x_{kj}}{\partial y_{lj}} = \frac{\delta_{kl}}{\sum_{l=1}^{r} y_{lj}} - \frac{y_{kj}}{\left(\sum_{l=1}^{r} y_{lj}\right)^2},$$

Применение к стохастическим матрицам

$$[\nabla_Y]_{kj} = \underbrace{\frac{\left[\nabla_X^+\right]_{kj}}{\sum\limits_{l=1}^{r} y_{lj}} + \frac{\left[\nabla_X^- Y^T\right]_{kk}}{\left(\sum\limits_{l=1}^{r} y_{lj}\right)^2}}_{\left[\nabla_Y^+\right]_{kj}} - \underbrace{\left(\frac{\left[\nabla_X^+\right]_{kj}}{\sum\limits_{l=1}^{r} y_{lj}} + \frac{\left[\nabla_X^- Y^T\right]_{kk}}{\left(\sum\limits_{l=1}^{r} y_{lj}\right)^2} \right)}_{\left[\nabla_Y^-\right]_{kj}},$$

$$y_{kj} \leftarrow y_{kj} \frac{\left[\nabla_X^-\right]_{kj} + \sum\limits_{l=1}^{r} \left[\nabla_X^+\right]_{lj} x_{lj}}{\left[\nabla_X^+\right]_{kj} + \sum\limits_{l=1}^{r} \left[\nabla_X^-\right]_{lj} x_{lj}},$$

$$x_{kj} \leftarrow \frac{y_{kj}}{\sum\limits_{l=1}^{r} y_{lj}}.$$

Для дивергенции Кульбака-Лейблера:

$$y_{kj} \leftarrow y_{kj} \frac{\sum\limits_{i=1}^{m} a_{ik} \frac{p_{ij}}{q_{ij}} + \sum\limits_{l=1}^{r} \sum\limits_{i=1}^{m} a_{il} x_{lj}}{\sum\limits_{i=1}^{m} a_{ik} + \sum\limits_{l=1}^{r} \sum\limits_{i=1}^{m} a_{il} x_{lj} \frac{p_{ij}}{q_{ij}}} = \frac{\sum\limits_{i=1}^{m} \left(a_{ik} \frac{p_{ij}}{q_{ij}} + q_{ij} \right)}{\sum\limits_{i=1}^{m} \left(a_{ik} + p_{ij} \right)}.$$

Применение к стохастическим матрицам

- Релаксация [Zhu et al., 2013]: применим метод множителей Лагранжа:

$$\tilde{D}\left(X, \{\lambda_j\}_{j=1}^n\right) = D(X) - \sum_{j=1}^{n} \lambda_j \left(\sum_{k=1}^{r} x_{kj} - 1\right),$$

$$\frac{\partial \tilde{D}}{\partial x_{kj}} = \left[\nabla_X^+\right]_{kj} - \left[\nabla_X^-\right]_{kj} - \lambda_j,$$

$$x'_{kj} \leftarrow x_{kj} \frac{\left[\nabla_X^-\right]_{kj} + \lambda_j}{\left[\nabla_X^+\right]_{kj}},$$

$$\sum_{l=1}^{r} x'_{lj} = 1 \Leftrightarrow \lambda_j = \frac{1 - \sum\limits_{l=1}^{r} x_{lj} \left[\nabla_X^-\right]_{lj} / \left[\nabla_X^+\right]_{lj}}{\sum\limits_{l=1}^{r} x_{lj} / \left[\nabla_X^+\right]_{lj}} \equiv \frac{1 - z_{kj}}{y_{kj}},$$

$$x'_{kj} \leftarrow x_{kj} \frac{\left[\nabla_X^-\right]_{kj} y_{kj} + 1 - z_{kj}}{\left[\nabla_X^+\right]_{kj} y_{kj}},$$

$$x_{kj} \leftarrow x_{kj} \frac{\left[\nabla_X^-\right]_{kj} y_{kj} + 1}{\left[\nabla_X^+\right]_{kj} y_{kj} + z_{kj}}.$$

Применение к стохастическим матрицам

Для дивергенции Кульбака-Лейблера:

$$y_{kj} = \sum_{l=1}^{r} \frac{x_{lj}}{\sum\limits_{i=1}^{m} a_{il}},$$

$$z_{kj} = \sum_{l=1}^{r} x_{lj} \frac{\sum\limits_{i=1}^{m} a_{ik} \frac{p_{ij}}{q_{ij}}}{\sum\limits_{i=1}^{m} a_{il}},$$

$$x_{kj} \leftarrow x_{kj} \frac{1 + y_{kj} \sum\limits_{i=1}^{m} a_{ik} \frac{p_{ij}}{q_{ij}}}{z_{kj} + y_{kj} \sum\limits_{i=1}^{m} a_{ik}}.$$

NMF с АБ-дивергенцией

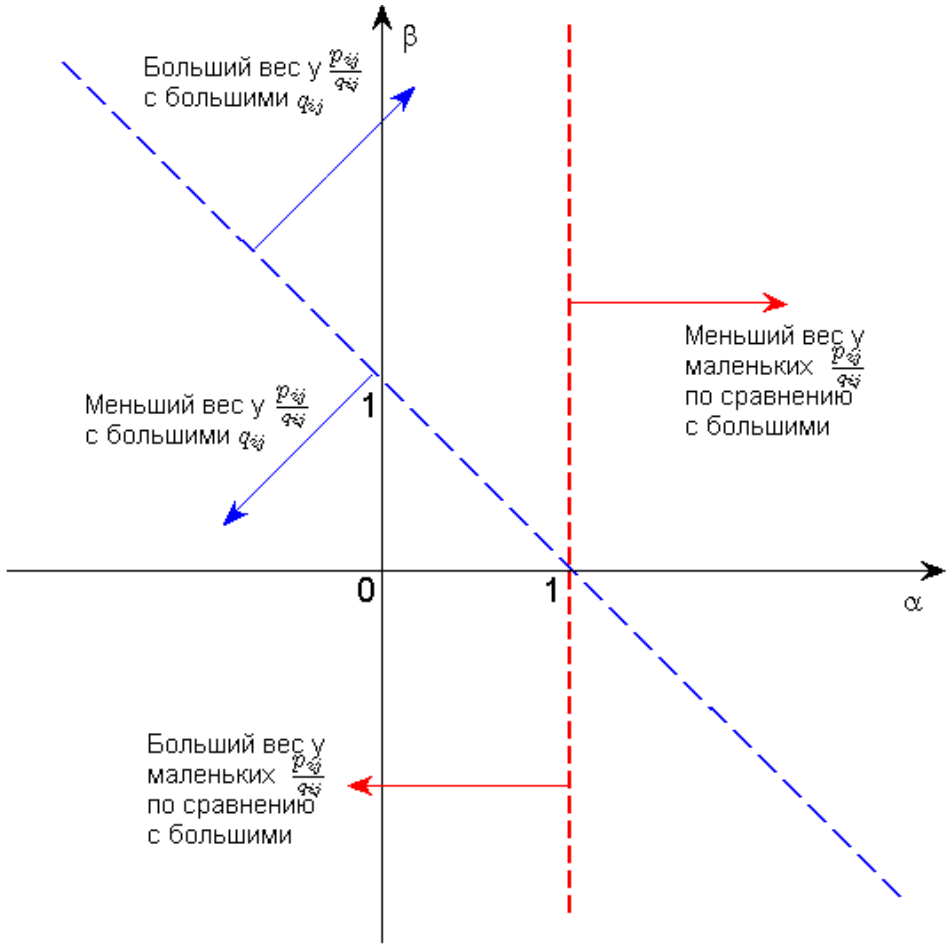
Оптимизационная задача:

$$(A^*, X^*) = \underset{A \geqslant 0, X \geqslant 0}{\operatorname{argmin}} D_{AB}^{(\alpha,\beta)}(P, AX).$$

Обобщаемые дивергенции:

(α, β)	Дивергенция
$\alpha + \beta = 1$	α
$(1, -)$	β
$(1, -1)$	Итакура-Саито
$(1, 1)$	норма Фробениуса
$(0.5, 0.5)$	расстояние Хеллингера
$(2, -1)$	χ^2 Пирсона
$(-1, 2)$	χ^2 Неймана
$(0, 0)$	лог-евклидово расстояние

NMF с АБ-дивергенцией



Влияние параметров α и β на вклад отношений $\frac{p_{ij}}{q_{ij}}$ в получаемые оценки.

Мультипликативные обновления

При $\alpha \neq 0$ работает мультипликативный алгоритм [Cichocki et al., 2011]:

$$X \leftarrow X \otimes \left(\left(A^T Z\right) \oslash \left(A^T Q^{[\alpha+\beta-1]}\right)\right)^{[\omega(\alpha,\beta)]},$$
$$Z = P^{[\alpha]} \otimes Q^{[\beta-1]},$$
$$\omega(\alpha,\beta) = \begin{cases} \frac{1}{1-\beta}, & \frac{\beta}{\alpha} < \frac{1}{\alpha} - 1, \\ \frac{1}{\alpha}, & \frac{\beta}{\alpha} \in \left[\frac{1}{\alpha} - 1, \frac{1}{\alpha}\right], \\ \frac{1}{\alpha+\beta-1}, & \frac{\beta}{\alpha} > \frac{1}{\alpha}. \end{cases}$$

$[\cdot]$ — поэлементное возведение в степень.

При $\alpha = \beta = 1$ обновления совпадают с мультипликативными обновлениями для нормы Фробениуса, при $\alpha = 1$, $\beta = 0$ — для обобщённой дивергенции Кульбака-Лейблера.

Как и все мультипликативные алгоритмы, может застревать в нестационарных точках вблизи нулей. Можно использовать отделение от нуля константой ε: в предельной точке модифицированного алгоритма условия стационарности выполняются с точностью до $\mathcal{O}(\varepsilon)$ [Рябенко, 2014].

NMF с дивергенцией Брегмана

Оптимизационная задача:

$$(A^*, X^*) = \underset{A \geqslant 0, X \geqslant 0}{\operatorname{argmin}} D_\phi(P, AX) =$$

$$= \underset{A \geqslant 0, X \geqslant 0}{\operatorname{argmin}} \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\phi(p_{ij}) - \phi(q_{ij}) - \nabla\phi(q_{ij})(p_{ij} - q_{ij})\right).$$

Мультипликативные обновления

[Dhillon, Sra, 2005]:

$$x_{kj} \leftarrow x_{kj} \frac{\sum\limits_{i=1}^{m} \nabla^2\phi\left(q_{ij}\right) p_{ij} a_{ik}}{\sum\limits_{i=1}^{m} \nabla^2\phi\left(q_{ij}\right) q_{ij} a_{ik}}.$$

Обновления минимизируют дополнительную функцию, функция потерь монотонно невозрастает.

При $\phi(p) = p^2/2$ и $\phi(p) = p \ln p$ обновления совпадают с обновлениями MU для нормы Фробениуса и обобщённой дивергенции Кульбака-Лейблера соответственно.

Быстрый алгоритм с использованием разложения Тейлора

Scalar Block Coordinate Descent (sBCD) [Li et al., 2012]: показано, что для дивергенции Брегмана справедливо следующее представление:

$$E_t(P,Q) \equiv \sum_{i=1}^{m}\sum_{j=1}^{n} |p_{ij} - q_{ij}|^t, \quad t = 1, 2, \ldots,$$

$$P^{(k)} \equiv P - \sum_{l \neq k} A_{:l} X_{l:},$$

$$D_\phi(P,Q) = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{t=2}^{\infty} \frac{\nabla^t \phi(q_{ij})}{t!} \left(-\operatorname{sgn}(q_{ij} - p_{ij})\right)^t E_t\left(p_{ij}^{(k)}, q_{ij}\right).$$

Приравнивая к нулю производные $D_\phi(P,Q)$ в таком представлении, получим следующие обновления:

$$x_{kj} \leftarrow \frac{\sum_{i=1}^{m} \nabla^2 \phi(q_{ij})\, p_{ij}^{(k)} a_{ik}}{\sum_{i=1}^{m} \nabla^2 \phi(q_{ij})\, a_{ik} a_{ik}}.$$

Быстрый алгоритм с использованием разложения Тейлора

Как и MU, sBCD — градиентный метод, но его шаги равномерно больше:

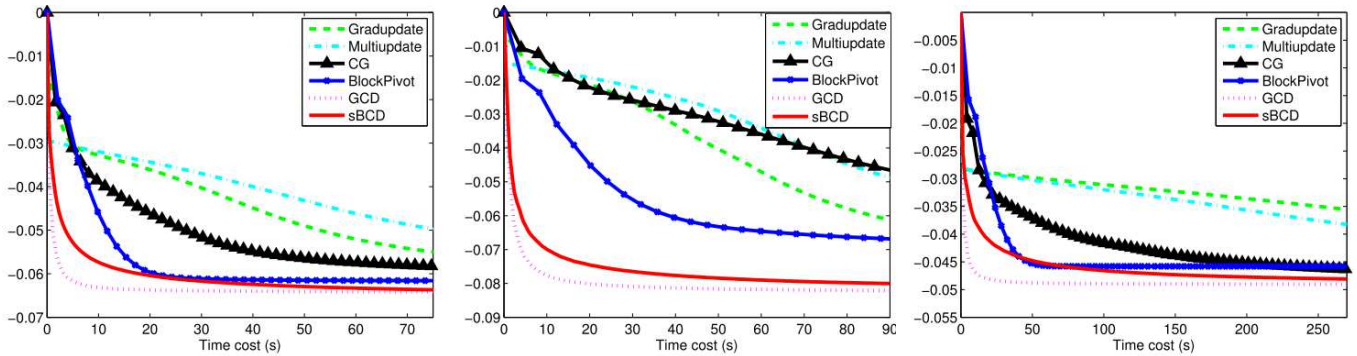
$$\text{sBCD:}\quad x_{kj} = x_{kj} - \frac{1}{\left[(A \otimes A)^T \nabla^2\phi(AX)\right]_{kj}} \frac{\partial D_\phi(P, AX)}{\partial x_{kj}},$$

$$\text{MU:}\quad x_{kj} = x_{kj} - \frac{1}{\left[A^T(\nabla^2\phi(AX) \otimes (AX))\right]_{kj}} \frac{\partial D_\phi(P, AX)}{\partial x_{kj}}.$$

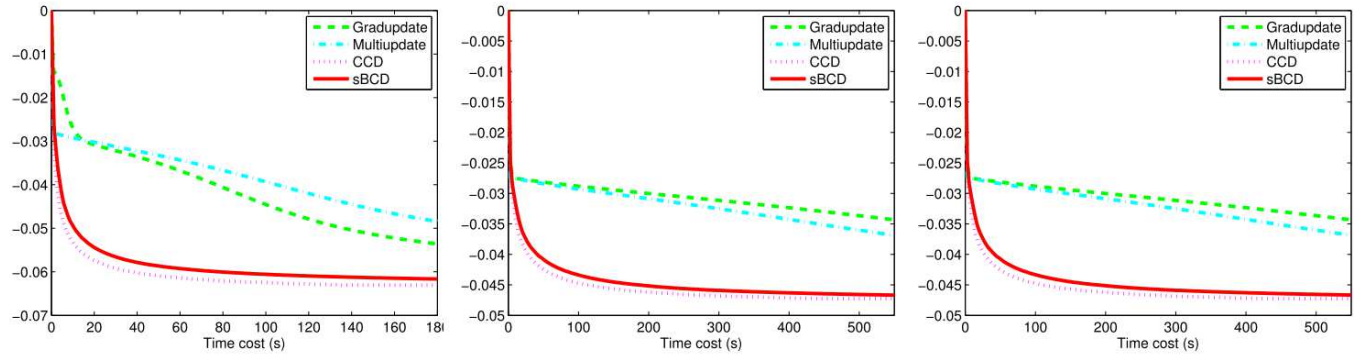
При $\phi(p) = p^2/2$ обновления sBCD совпадают с обновлениями HALS; при $\phi(p) = p \ln p$ имеем:

KL-NMF	PLSA	sBCD
$\dfrac{x_{kj}\sum_{i=1}^{m} a_{ik}\frac{p_{ij}}{q_{ij}}}{\sum_{i=1}^{m} a_{ik}}$	$\dfrac{x_{kj}\sum_{i=1}^{m} a_{ik}\frac{p_{ij}}{q_{ij}}}{\sum_{l=1}^{r} x_{lj}\sum_{i=1}^{m} a_{il}\frac{p_{ij}}{q_{ij}}}$	$\dfrac{\sum_{i=1}^{m} a_{ik}\frac{p_{ij}^{(k)}}{q_{ij}}}{\sum_{i=1}^{m} a_{ik}\frac{a_{ik}}{q_{ij}}}$

Сравнение алгоритмов



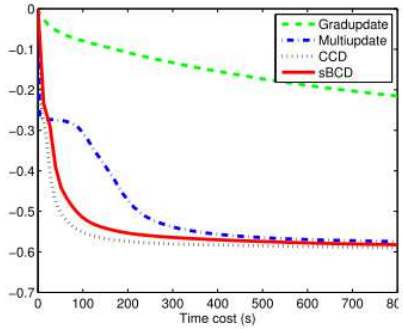
(a) Frobenius, RD1 (b) Frobenius, RD2 (c) Frobenius, RD3



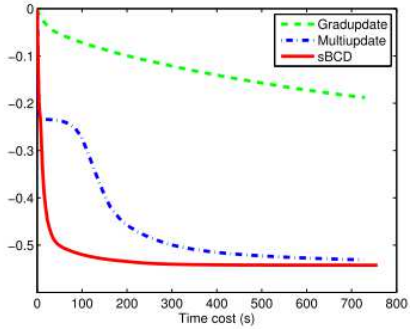
(d) KL, RD1 (e) KL, RD2 (f) KL, RD3

[Li et al., 2012]: зависимость относительной точности приближения $\log_{10} \frac{D_\phi(P, AX)}{D_\phi(P, A_0 X_0)}$ от времени работы алгоритмов на плотных данных. RD1: $m = 2000$, $n = 1000$, $r = 30$; RD2: $m = 2000$, $n = 1000$, $r = 60$;, RD2: $m = 3000$, $n = 2000$, $r = 30$. Gradupdate — градиентный метод с фиксированным шагом, Multupdate — MU, CG — ANLS с методом сопряжённых градиентов, BlockPivot — ANLS с методом активных ограничений, GCD/CCD — жадный градиентный метод с/без отбора важных переменных из [Hofmann, 1999].

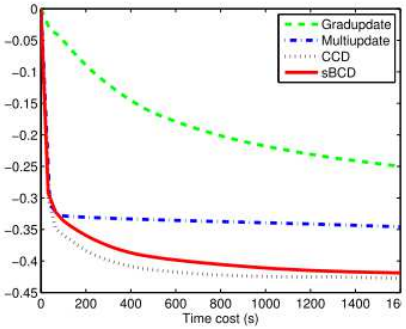
Сравнение алгоритмов



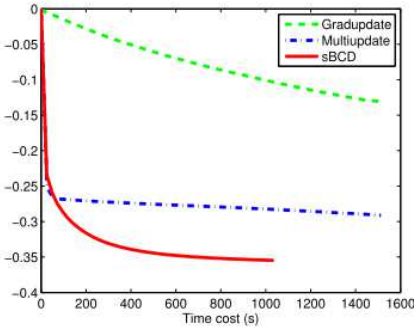
(a) KL, Face Image



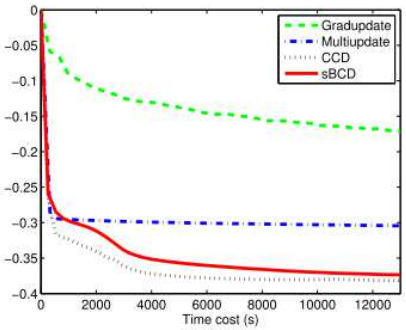
(b) IS, Face Image



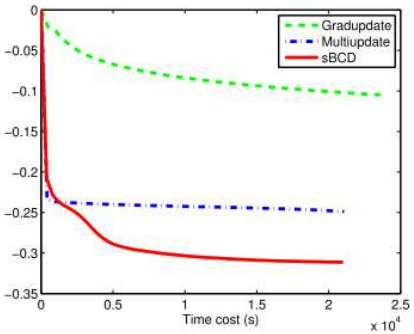
(c) KL, Movielens



(d) IS, Movielens



(e) KL, Netflix



(f) IS, Netflix

[Li et al., 2012]: то же. Face Image: $m = 10304$, $n = 400$, $r = 20$; Movielens: $m = 71567$, $n = 65133$, $r = 20$; Netflix: $m = 480189$, $n = 17770$, $r = 20$; Movielens и Netflix очень разреженные. Gradupdate — градиентный метод с фиксированным шагом, Multupdate — MU, CCD — жадный градиентный метод из [Hofmann, 1999].

Библиография I

Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P., Plemmons, R. J. (2007). *Algorithms and applications for approximate nonnegative matrix factorization*. Computational Statistics & Data Analysis, 52(1), 155–173.

Boutsidis, C., Gallopoulos, E. (2008). *SVD based initialization: A head start for nonnegative matrix factorization*. Pattern Recognition, 41, 1350–1362.

Casalino, G., Del Buono, N., Mencar, C. (2014). *Subtractive clustering for seeding non-negative matrix factorizations*. Information Sciences, 257, 369–387.

Chi, E., Kolda, T. (2012). *On tensors, sparsity, and nonnegative factorizations*. SIAM Journal on Matrix Analysis and Applications, 1–28.

Cichocki, A., Zdunek, R., Amari, S. (2007). *Hierarchical ALS algorithms for nonnegative matrix and 3D tensor factorization*. In M. E. Davies, C. J. James, S. A. Abdallah, M. D. Plumbley (Eds.), Independent Component Analysis and Signal Separation (pp. 169–176). Springer Berlin Heidelberg.

Cichocki, A., Phan, A.-H. (2009). *Fast local algorithms for large scale nonnegative matrix and tensor factorizations*. IEICE Transactions on Fundamentals of Electronics Communications and Computer Sciences, E92-A(3), 708–721.

Cichocki, A., Zdunek, R., Phan, A. (2009). *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. Tokyo: John Wiley & Sons.

Cichocki, A., Cruces, S., Amari, S. (2011). Generalized Alpha-Beta Divergences and Their Application to Robust Nonnegative Matrix Factorization. Entropy, 13(1), 134–170.

Библиография II

Ding, C., Li, T., Peng, W. (2006).*Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence chi-square statistic, and a hybrid method*. Proceedings of the National Conference on Artificial Intelligence, 342–347.

Ding, C., Li, T., Peng, W. (2008). *On the equivalence between Non-negative Matrix Factorization and Probabilistic Latent Semantic Indexing*. Computational Statistics & Data Analysis, 52(8), 3913–3927.

Dhillon, I. S., Sra, S. (2005). *Generalized nonnegative matrix approximations with Bregman divergences*. In Y. Weiss, B. Scholkopf, J. C. Platt (Eds.), Neural Information Processing Systems (pp. 283–290). Vancouver, Canada.

Donoho, D., Stodden, V. (2004). *When does non-negative matrix factorization give a correct decomposition into parts?* In S. Thrun, L. K. Saul, B. Sch?lkopf (Eds.), Advances in Neural Information Processing Systems (pp. 1141–1148).

Gillis, N., Glineur, F. (2010). *Using underapproximations for sparse nonnegative matrix factorization*. Pattern Recognition, 43(4), 1676–1687.

Gillis, N., Glineur, F. (2012). *Accelerated multiplicative updates and hierarchical ALS algorithms for nonnegative matrix factorization*. Neural Computation, 24(4), 1085–105.

Gillis, N. (2012). *Sparse and unique nonnegative matrix factorization through data preprocessing*. The Journal of Machine Learning Research, 13(1), 3349?3386.

Gillis, N. (2014). *The Why and How of Nonnegative Matrix Factorization*. arXiv Preprint.

Grippo, L., Sciandrone, M. (2000). *On the convergence of the block nonlinear Gauss–Seidel method under convex constraints*. Operations Research Letters, 26(3), 127–136.

Библиография III

Guan, N., Tao, D., Luo, Z., Yuan, B. (2012). *NeNMF: An Optimal Gradient Method for Nonnegative Matrix Factorization*. IEEE Transactions on Signal Processing, 60(6), 2882–2898.

Han, J., Han, L., Neumann, M., Prasad, U. (2009). *On the rate of convergence of the image space reconstruction algorithm*. Operators and Matrices, 25(1), 1986–1987.

Hibi, R., Takahashi, N. (2011). *A modified multiplicative update algorithm for euclidean distance-based nonnegative matrix factorization and its global convergence*. Neural Information Processing, 7063, 655–662.

Ho, N.-D. (2008). *Nonnegative matrix factorization algorithms and applications*. PhD thesis, University catholique de Louvain.

Hofmann, T. (1999). *Probabilistic latent semantic indexing*. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99 (pp. 50–57). New York, NY, USA: ACM Press.

Hsieh, C.-J., Dhillon, I. S. (2011). *Fast coordinate descent methods with variable selection for non-negative matrix factorization*. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11 (p. 1064). New York, NY, USA: ACM Press.

Huang, K., Sidiropoulos, N. D., Swami, A. (2014). *Non-Negative Matrix Factorization Revisited: Uniqueness and Algorithm for Symmetric Decomposition*. IEEE Transactions on Signal Processing, 62(1), 211–224.

Kim, H., Park, H. (2007). *Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis*. Bioinformatics (Oxford, England), 23(12), 1495–502.

Библиография IV

Kim, H., Park, H. (2008). *Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method*. SIAM Journal on Matrix Analysis and Applications, 30(2), 713–730.

Kim, J., Park, H. (2011). *Fast nonnegative matrix factorization: An active-set-like method and comparisons*. SIAM Journal on Scientific Computing, 33(6), 3261–3281.

Laurberg, H., Christensen, M. G., Plumbley, M. D., Hansen, L. K., Jensen, S. H. (2008). *Theorems on positive data: on the uniqueness of NMF*. Computational Intelligence and Neuroscience, (2), 764206.

Lee, D. D., Seung, S. H. (1999). *Learning the parts of objects by non-negative matrix factorization*. Nature, 401(6755), 788–791.

Lee, D. D., Seung, S. H. (2001). *Algorithms for non-negative matrix factorization*. In Advances in Neural Information Processing Systems (pp. 556–562).

Li, L., Lebanon, G., Park, H. (2012). *Fast bregman divergence NMF using taylor expansion and coordinate descent*. In ACM SIGKDD international conference on Knowledge discovery and data mining (p. 307). New York, NY, USA: ACM Press.

Li, L., Zhang, Y.-J. (2009). *FastNMF: highly efficient monotonic fixed-point nonnegative matrix factorization algorithm with good applicability*. Journal of Electronic Imaging, 18(3), 33004–33012.

Lin, C.-J. (2007). *Projected gradient methods for nonnegative matrix factorization*. Neural Computation, 19(10), 2756–79.

Morup, M., Hansen, L. K. (2010). *Archetypal analysis for machine learning*. In IEEE International Workshop on Machine Learning for Signal Processing (pp. 172–177).

Библиография V

Vavasis, S. A. (2009). *On the complexity of nonnegative matrix factorization*. SIAM Journal on Optimization, 20, 1364–1377.

Zdunek, R., Cichocki, A. (2006). *Non-negative matrix factorization with quasi-Newton optimization*. In Artificial Intelligence and Soft Computing (Vol. 4029, pp. 870–879).

Zhu, Z., Yang, Z., Oja, E. (2013). *Multiplicative Updates for Learning with Stochastic Matrices*. In Scandinavian Conference on Image Analysis, SCIA (pp. 143–152). Espoo, Finland.

Рябенко, Е. А. (2014). *Мультипликативный метод неотрицательного матричного разложения с АБ-дивергенцией и его сходимость*. Машинное обучение и анализ данных, 1(7), 800–816.