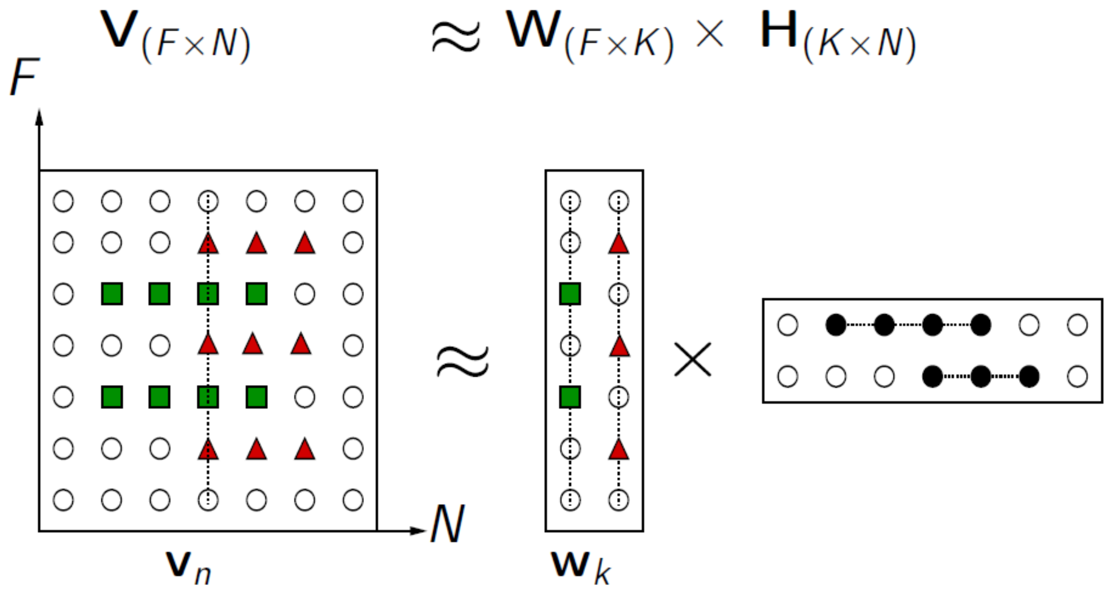
论文标题：

algorithms-for-non-negative-matrix-factorization

要点：

用下图直观描述矩阵分解：



白圆圈表示 0，黑圆圈表示 1。

矩阵 V 被称为：数据矩阵（data matrix）

矩阵 W 被称为：解释变量，基，字典，模式，主题（explanatory variables，basis，dictionary，pattern，topics）

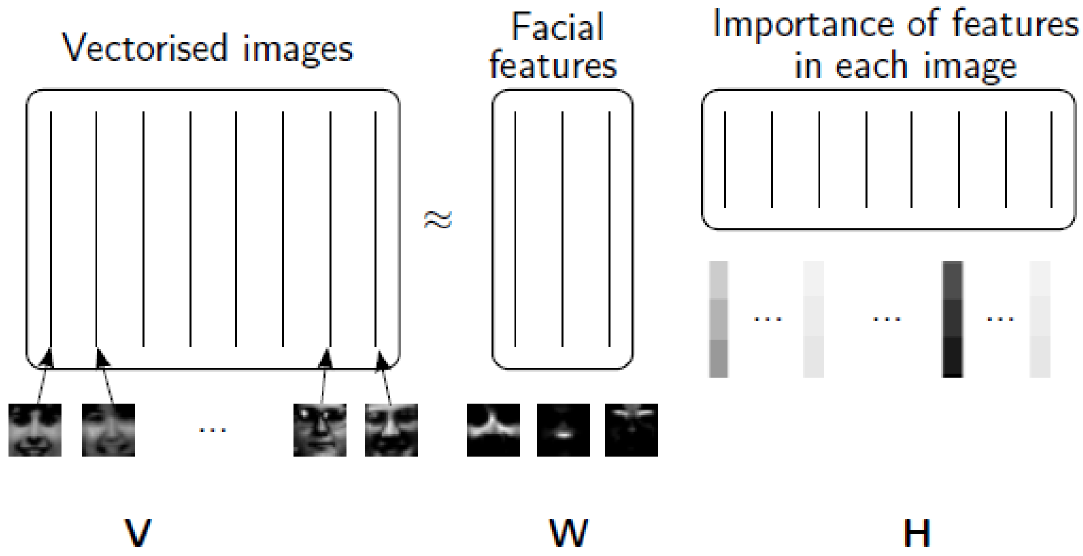
矩阵 H 被称为：回归因子，激活系数，扩展系数（regressors，activation coefficients，expansion coefficients）

非负矩阵分解中，有一个很重要的约束是非负，如下图：

$$\mathbf{V} \approx \mathbf{WH};$$

- $\mathbf{W} = [w_{fk}]$ s.t. $w_{fk} \geq 0$
  and
- $\mathbf{H} = [h_{kn}]$ s.t. $h_{kn} \geq 0.$

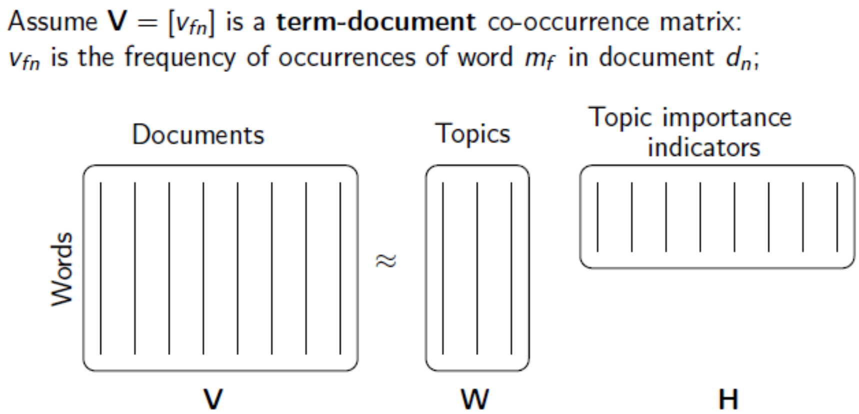
下图是一个例子，用 NMF 来解释人脸特征：



类似的，如果用 NMF 解释音频的 spectrogram 的分解：

V 就是表示 spectrogram 本身；W 表示 spectra features；H 表示 Importance of features in each spectra of the spectrogram

还有自然语言理解中的主题模型：



下面解释各矩阵的含义：

V 是一个 F *N 的数据矩阵，N 表示观测点/样本/列/特征向量的数目，F 表示特征/行的数目

$\mathbf{v}_n = (v_{1n}, \cdots, v_{Fn})^T$ 代表第 n 个观测点的特征向量，属于 $\mathbb{R}_+^F$

W 是一个 F*K 的字典矩阵，w_{fk}称为系数，w_k则称为 k 个元素中的基向量或者字典向量。

H 是一个 K*N 的激活/扩展矩阵，列向量h_n为观测点v_n的激活系数，行向量h_k与

基向量w_k相关，因为h_k中的每个激活系数均是在 w 中的第 k 个特征上的反应。

下面是对非负矩阵分解的优化目标：

NMF approximation $\mathbf{V} \approx \mathbf{WH}$ is usually obtained through:

$$\min_{\mathbf{W},\mathbf{H} \geq 0} D(\mathbf{V}|\mathbf{WH}),$$

where $D(\mathbf{V}|\widehat{\mathbf{V}})$ is a *separable matrix divergence*:

$$D(\mathbf{V}|\widehat{\mathbf{V}}) = \sum_{f=1}^{F} \sum_{n=1}^{N} d(v_{fn}|\hat{v}_{fn}),$$

在 x, y>=0 时，d(x |y)用来刻画标量的散度：

1、 d(x |y)在 x 和 y 上连续；

2、 d(x |y)>=0 对所有的 x, y>=0 成立

3、 d(x |y) = 0 当且仅当 x=y

下面是几种流行的标量散度公式及，凸特性及规模不变性(scale invariant)：

Euclidean (EUC) distance (Lee and Seung, 1999)

$$d_{EUC}(x|y) = (x-y)^2$$

Kullback-Leibler (KL) divergence (Lee and Seung, 1999)

$$d_{KL}(x|y) = x \log \frac{x}{y} - x + y$$

Itakura-Saito (IS) divergence (Févotte et al., 2009)

$$d_{IS}(x|y) = \frac{x}{y} - \log \frac{x}{y} - 1$$

Divergence $d(x\|y)$	EUC	KL	IS
Convex on x	yes	yes	yes
Convex on y	yes	yes	no

（凸特性仅分别 w.r.t .x 或者 y，但不会同时对 x 和 y）

$$\begin{aligned} d_{EUC}(\lambda x|\lambda y) &= \lambda^2 d_{EUC}(x|y) \\ d_{KL}(\lambda x|\lambda y) &= \lambda d_{KL}(x|y) \\ d_{IS}(\lambda x|\lambda y) &= d_{IS}(x|y) \end{aligned}$$

IS 散度的 scale invariant 特性在 audio spectra 上可以提供更高的准确性

下面给出目标的优化算法：

$$\min_{\mathbf{W},\mathbf{H}\geq 0} D(\mathbf{V}|\mathbf{WH}) \Leftrightarrow \min_{\theta} C(\theta)\,; \;\; C(\theta) \stackrel{\text{def}}{=} D(\mathbf{V}|\mathbf{WH})$$

where $\theta \stackrel{\text{def}}{=} \{\mathbf{W}, \mathbf{H}\}$

Alternating optimization a.k.a block-coordinate descent (one iteration):

- update **W**, given **H** fixed,
- update **H**, given **W** fixed.

比较著名的优化方法是：Multiplicative update rules（把梯度拆成正减去负）

Multiplicative update (MU) rule for **H** (similarly for **W**) is defined as:

$$h_{kn} \leftarrow h_{kn} \, [\nabla_{h_{kn}} C(\theta)]_- \, / \, [\nabla_{h_{kn}} C(\theta)]_+ \,,$$

where

$$\nabla_{h_{kn}} C(\theta) = [\nabla_{h_{kn}} C(\theta)]_+ - [\nabla_{h_{kn}} C(\theta)]_- \,,$$

and the summands are both nonnegative.

NOTE: The nonnegativity of **W** and **H** is guaranteed by construction.

上面的例子比较复杂，可以用一个一元函数 C(h)作为直观的例子，把 C 对 h 的梯度拆成两部分：

$$\nabla_h C(h) = \nabla_+ - \nabla_-$$

如果$\nabla_h C(h) > 0$，也就是导数为正，则$\nabla_+ > \nabla_-$（$\frac{\nabla_+}{\nabla_-} > 1$）；反之，$\nabla_+ < \nabla_-$（（$\frac{\nabla_+}{\nabla_-} < 1$））。前者，$h\frac{\nabla_+}{\nabla_-} > h$，迫使更新时 h 向右移动，后者则迫使 h 向左移动。如下图两个紫色的移动箭头都使得函数逐渐逼向**最小值**。注意在此过程中 h 保持非负。