NMF cost functions

Factorization: $\mathbf{X} = \mathbf{WS}$ where $\mathbf{X} \in \mathbb{R}^{M \times N}$ $\mathbf{W} \in \mathbb{R}^{M \times K}$ $\mathbf{S} \in \mathbb{R}^{K \times N}$ $K < N$

NMF loss function: $$\mathbf{W}, \mathbf{S} = \min_{\mathbf{W},\mathbf{S} \geq 0} \mathbf{L}(\mathbf{X}, \mathbf{W}, \mathbf{S}) + \alpha\rho\left(\|\mathbf{W}\|_1 + \|\mathbf{S}\|_1\right) + \alpha(1-\rho)\left(\|\mathbf{W}\|_F^2 + \|\mathbf{S}\|_F^2\right)$$

where α = regularization constant

ρ = sparsity constant (a.k.a. L1-ratio)

$\mathbf{L}$ = generic loss function

Frobenius loss: $$\mathbf{L}_F = \|\mathbf{X} - \mathbf{WS}\|_F^2$$

KL divergence: $$\mathbf{L}_{KL} = \sum_{i,j} X_{i,j}\left(\log \frac{X_{i,j}}{(WS)_{i,j}} - 1\right) + (WS)_{i,j}$$

Probabilistic LSA (pLSA)

- Matrix factorization is a very practical approach to document retrieval, however it isn't principled in a statistical sense. At a high level, we start with an extremely sparse term-document matrix and ask, how can we measure similarity between documents? To do this we compute a projection onto a K dimensional space where the data is represented by a dense(r) matrix, and compute distances on it.
- pLSA uses a latent variable approach to model terms and documents as a joint distribution $p(w, d)$ by factoring it using a set of latent variables Z:

$$p(d, w) = \sum_{z \in Z} p(z)p(d \mid z)p(w \mid z)$$

- The standard MLE method to fit latent variable models is the EM algorithm (general form is shown here, pLSA actually uses a more involved EM-based approach):

E step: $$p(z \mid d, w) = \frac{p(z)p(d \mid z)p(w \mid z)}{\sum_{z' \in Z} p(z')p(d \mid z')p(w \mid z')}$$

M step:
$$p(w \mid z) \propto \sum_{d \in D} f_{d,w} p(z \mid d, w)$$
$$p(d \mid z) \propto \sum_{w \in W} f_{d,w} p(z \mid d, w)$$
$$p(z) \propto \sum_{d \in D} \sum_{w \in W} f_{d,w} p(z \mid d, w)$$

[1] Hoffman, 2000