1 Partial Derivatives with respect to $Z_\alpha$

The optimization problem is minimization of the following objective

$$\text{minimize } D(X_{1:N_\alpha}||\hat{X}_{1:N_\alpha})$$

where the divergence function is separable for each observed tensor element as

$$D(X_{1:N_\alpha}||\hat{X}_{1:N_\alpha}) = \sum_\nu D_\nu(X_\nu||\hat{X}_\nu)$$

and each divergence is separable as the sum of scalar divergences $d(x||\mu)$ as

$$D_\nu(X_\nu||\hat{X}_\nu) = \sum_{u_\nu} d_\nu(X_\nu(u_\nu)||\hat{X}_\nu(u_\nu))$$

The optimization problem can be solved by various approaches. All of these requires us evaluating the gradient with respect to the individual tensor elements $Z_\alpha(u_\alpha)$ and model output tensors $\hat{X}_\nu$. In the general form the derivative of the $\beta$ divergence with respect to the second parameter is

$$\frac{\partial d_\beta(x||\hat{x})}{\partial \hat{x}} = -x\hat{x}^{-\beta} + \hat{x}^{1-\beta} = \frac{\hat{x} - x}{\hat{x}^\beta}$$

Typically in each iteration of an optimization procedure, we need to calculate the following derivative to update each model output tensor $\hat{X}_\nu$:

$$\frac{\partial d_{p_\nu}(X_\nu(u_\nu);\hat{X}_\nu(u_\nu))}{\partial \hat{X}_\nu(u_\nu)} = -\frac{X_\nu(u_\nu)}{X_\nu(u_\nu)^{p_\nu}} + \frac{\hat{X}_\nu(u_\nu)1^{-p_\nu}}{1 \hat{X}_\nu(u_\nu)^{p_\nu}} = \frac{\hat{X}_\nu(u_\nu) - X_\nu(u_\nu)}{X_\nu(u_\nu)^{p_\nu}}$$
Next each latent factor of the decomposition model is updated by the following derivative

$$\frac{\partial D(X_{1:N}; \hat{X}_{1:N})}{\partial Z_\alpha^\nu(v_\alpha)} = \sum_\nu \frac{1}{\phi_\nu} \sum_{u_\nu} \frac{\partial d_{\nu^\nu}(X_\nu(u_\nu); \hat{X}_\nu(u_\nu))}{\partial \hat{X}_\nu(u_\nu)} \frac{\partial \hat{X}_\nu(u_\nu)}{\partial Z_\alpha^\nu(v_\alpha)}$$

$$= \sum_\nu \left[ R(\nu, \alpha) \frac{1}{\phi_\nu} \sum_{v_\alpha} \left( \frac{\hat{X}_\nu(u_\nu) - X_\nu(u_\nu)}{X_\nu(u_\nu)p_\nu} \right) \prod_{\alpha' \neq \alpha} Z_\alpha'(v_{\alpha'})^{R(\nu, \alpha)} \right]$$

(1)

2 Algorithm Summary

**Algorithm 1:** Overview of the proposed method for coupled tensor factorization.

**Input:** Observed tensors: $X_1, X_2, \ldots, X_N$, Coupling matrix: $R$

**Output:** Latent factors: $Z_{1:N}$, Dispersions: $\phi_{1:N}$, Power parameters: $p_{1:N}$

Randomly initialize $Z_\alpha^{(0)}, \phi_\nu^{(0)}, p_\nu^{(0)}$ and set $i = 1$

// Estimate the latent factors

while not converged and $i \leq \text{MaxIter}$ do

// Run DIGD until convergence

while all blocks are not processed do

Pick a step size $\eta$

for each block $\gamma$ in $\sigma$ do in parallel

Form a stratum $\sigma$ (implicitly determined by the transfer schedule of the blocks)

Load the corresponding block of $X_\nu$

Run IGD on block $\gamma$ with step size $\eta$

// Update the local factors first, then update the shared factors

Update all $Z_\alpha^\nu(v_\alpha)$ with $v_\alpha \in B_1(I_\alpha, \gamma)$ by using $X_\nu(u_\nu)$ in $u_\nu \in B_1(I_{0,\nu}, \gamma)$

Pass the shared factor block to another node (will determine the next $\sigma$)

end

end

// Estimate the dispersion and power parameters

while all blocks are not processed do

for each block $\gamma$ in $\sigma$ do in parallel

Form a stratum $\sigma$ (implicitly determined by the transfer schedule of the blocks)

Load the corresponding block of $X_\nu$

Compute the related parts of the output tensors at each node:

$$\hat{X}_\nu^{(i)}(u_\nu) = \sum_{u_\nu} \prod_{\alpha} Z_\alpha^\nu(v_\alpha)^{R(\nu, \alpha)} \quad \text{for all } \nu \in [N_x], u_\nu \in B_1(I_{0,\nu}, \gamma)$$

Compute the related part of the dispersion updates and likelihoods for the grid search

Pass the shared factor block to another node (will determine the next $\sigma$)

end

Send the intermediate dispersion updates and the likelihoods to the responsible node of the site

The responsible nodes aggregate the results and compute the new $\phi_\nu$ and $p_\nu$:

$$\phi_\nu^{(i)} = \arg \max_{\phi} \log P(X_{\nu}|\phi, X_{\nu}^{(i)}, p^{(i-1)} \phi(\phi)) \quad \text{for all } \nu \in [N_x]$$

$$p_\nu^{(i)} = \arg \max_{p} \log P(X_{\nu}|\hat{X}_{\nu}^{(i)}, \phi^{(i)}, p) \quad \text{for all } \nu \in [N_x]$$

The responsible nodes broadcast the new $\phi_\nu$ and $p_\nu$ to the relevant nodes

$i \leftarrow i + 1$
3 Exponential Dispersion Models and the Tweedie Family

An exponential dispersion model (EDM) can be defined by a two parameter density as follows [1]:

\[ P(x; \theta, \phi) = h(x, \phi) \exp \left\{ \frac{1}{\phi} (\theta x - \kappa(\theta)) \right\} \]  

(2)

where \( \theta \) is the canonical parameter, \( \phi \) is the dispersion parameter and \( \kappa \) is the cumulant (log-partition) function ensuring normalization. Here, \( h(x, \phi) \) is the base measure and is independent of the canonical parameter.

EDMs are a studied in particular as the response distribution of the generalized linear models [2]. For an EDM, we can verify that the mean \( \hat{x} \) and the variance \( \text{Var}\{x\} \) are obtained directly by differentiating \( \kappa(\cdot) \):

\[ \kappa'(\theta) = \langle x \rangle_{p(x; \theta, \phi)} = \hat{x}, \quad \kappa''(\theta) = \frac{1}{\phi} \text{Var}\{x\} = v(\hat{x}). \]

Here \( v(\hat{x}) \) is also known as the variance function [3][1].

In this paper, we focus on a particular EDM, namely The Tweedie family \( TW_p(x; \hat{x}, \phi) \). Tweedie distributions specify the variance function as \( v(\hat{x}) = \hat{x}^p \) [1]. The variance function is related to the \( p \)th power of the mean, therefore it is called a power variance function. Note that this choice directly dictates the form of \( \hat{x} \) and \( \kappa(\theta) \) that can be solved as

\[ \hat{x}(\theta) = \left\{ \begin{array}{ll}
\frac{1}{2-p} \left( (1-p)\theta \right)^{\frac{1-p}{p}} \exp(\theta) & p \neq 1 \\
\phi & p = 1
\end{array} \right. \]  

(3)

\[ \kappa(\theta) = \left\{ \begin{array}{ll}
\frac{1}{2-p} \left( (1-p)\theta \right)^{\frac{3-p}{p}} \log(-\theta) & p \neq 1, 2 \\
\phi & p = 1
\end{array} \right. \]  

(4)

Here, different choices for \( p \) yield well-known important distributions such as the Gaussian (\( p = 0 \)), Poisson (\( p = 1 \)), compound Poisson (\( 1 < p < 2 \)), Gamma (\( p = 2 \)) and inverse Gaussian (\( p = 3 \)) distributions. Excluding the interval \( 0 < p < 1 \) for which no EDM exists, for all other values of \( p \) not mentioned above, one obtains Tweedie stable distributions [1].

For \( p \in \{0, 1, 2, 3\} \) the densities are given as follows:

\[ TW_0(x; \hat{x}, \phi) = (2\pi\phi)^{-\frac{1}{2}} \exp\left( -\frac{1}{\phi} \frac{(x - \hat{x})^2}{2} \right) \]  

(5)

\[ TW_1(x; \hat{x}, \phi) = \frac{\hat{x}}{e^{x} \Gamma(\phi) \phi} \exp\left( -\frac{1}{\phi} \frac{x \log \hat{x} x - \hat{x} + \hat{x}}{2} \right) \]  

(6)

\[ TW_2(x; \hat{x}, \phi) = \frac{1}{\Gamma(\phi) \phi} \exp\left( -\frac{1}{\phi} \frac{x (x \log \hat{x} x - 1)}{2} \right) \]  

(7)

\[ TW_3(x; \hat{x}, \phi) = (2\pi x^3 \phi)^{-\frac{1}{2}} \exp\left( -\frac{1}{\phi} \frac{(x - \hat{x})^2}{2x \hat{x}^2} \right) \]  

(8)

Note that, the Poisson distribution in its well-known form, is an exponential dispersion model with unitary dispersion (\( \phi = 1 \)). This distribution is called over-dispersed (\( \phi > 1 \)) or under-dispersed (\( \phi < 1 \)) when the nominal variance is not sufficient to determine the variance of the observations [2]. When we introduce a dispersion parameter to the Poisson distribution, the domain of the probability distribution is re-defined on the integer multiples of \( \phi \): \( \hat{x} \in \{0, \phi, 2\phi, 3\phi, \ldots\} \). This can be interpreted as the data are scaled by \( \phi \) at each iteration.

For the remaining cases of \( p \), the probability density functions cannot be written in closed-form analytical forms. However, they can be expressed as infinite series that is defined as follows: [1]

\[ TW_p(x; \hat{x}, \phi) = \frac{1}{x^{\xi_p}} \left( \sum_{k=1}^{\infty} V_k \right) \exp\left\{ \frac{1}{\phi} \left( \hat{x}^{1-p} x - \frac{\hat{x}^{2-p}}{1-p} \right) \right\} \]  

(9)
and $\xi_p = 1$ for $p \in (1, 2)$ and $\xi_p = \pi$ otherwise.

The Tweedie density with $p \in (1, 2)$ coincides with the compound Poisson distribution \(^{[1]}\). The compound Poisson distribution has a support for continuous positive data and a discrete probability mass at zero. For $x = 0$, the density function is defined as $TW_p(x; \cdot) = \exp(\bar{x}^{2-p}/(\phi(p-2)))$ and for $x > 0$, it follows the form of Equation \(^{[2]}\) where the terms $V_k$ for this distribution is defined as follows:

$$V_k = \frac{x^{-k\alpha}(p-1)^{k\alpha}\phi^{k(\alpha-1)}}{(2-p)^k\Gamma(k+1)\Gamma(-k\alpha)} \tag{10}$$

where $\alpha = (2-p)/(1-p)$.

The cases $p < 0$ and $p > 2$ of the Tweedie class correspond to Tweedie stable distributions. For the Tweedie models with $p < 0$ and $p > 2$, the terms $V_k$ are defined as follows:

$$V_k = \frac{\Gamma(1 + \frac{k}{\alpha})\phi^{\frac{k}{\alpha}}(-1)^k \sin(k\pi \alpha)}{\Gamma(k+1)(1-p)^k(2-p)^{-\frac{k}{\alpha}x^{-k}}}, \quad V_k = \frac{\Gamma(1 + \alpha k)\phi^{(1-\alpha)}(\alpha)(-1)^k \sin(-k\pi \alpha)}{\Gamma(k+1)(p-1)^{-\alpha k}(p-2)^{\alpha k}} \tag{11}$$

References

