CS 598 EVS: Tensor Computations

Bayesian Methods for Tensor Decomposition

Edgar Solomonik

University of Illinois, Urbana-Champaign

Probabilistic Tensor Models

- Consider a tensor $\mathcal{T} \in \mathbb{Z}^{n\times n\times n}$ of count data,

$$\mathcal{T} = \sum_{l=1}^{M} \boldsymbol{e}_{i_l} \otimes \boldsymbol{e}_{j_l} \otimes \boldsymbol{e}_{k_l}$$

where (i_l, j_l, k_l) are random samples of a probability distribution $p(i, j, k)$[1]

  - *The empirical estimate is* $p(i, j, k) \approx t_{ijk}/M$*, but a large* M *needed for accuracy*
  - *Suppose there exists a latent random variable* $X \in \{1, \dots, R\}$ *such that* i*,* j*, and* k *are independent modulo the value of* X*, i.e.,* $p(i, j, k|X = l) = u_l(i)v_l(j)w_l(k)$
  - *This model is general, in particular* $R = O(n^2)$*, and has much fewer variables*

[1] Huang, Kejun, and Nicholas D. Sidiropoulos. "Kullback-Leibler principal component for tensors is not NP-hard." 2017 51st Asilomar Conference on Signals, Systems, and Computers. IEEE, 2017.

Optimization with KL divergence

- Tensor decompositions with Kullback-Liebler (KL) divergence are suitable for approximating $p(i,j,k)$
  - *Seek the best CP approximation* $[\![\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}]\!]$ *to* $\mathcal{T}$
  - *The KL divergence of distributions* $p(x)$ *and* $q(x)$ *is given by the expectation in* p *of* $\log p - \log q$

$$\sigma(p,q) = \int p(x)\log(p(x)/q(x))dx$$

  - *The maximum likelihood model of* $p(i,j,k) = t_{ijk}/M$ *in terms of KL divergence is then*

$$\min_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}\geqslant 0} -\sum_{i,j,k} p(i,j,k)\log(q(i,j,k)),$$

$$q(i,j,k) = q'(i,j,k|X)p(X), q'(i,j,k|X=l) = u_l(i)v_l(j)w_l(k)$$

  - *This corresponds to a nonnegative CP decomposition of* $\mathcal{T}$ *with* $a_{ir} = u_i(r)$*, etc.*

Optimal Rank-1 Approximation with KL divergence

- The rank-1 estimate with least KL divergence is easy to compute
  - *Consider the objective function in the rank-1 case*

$$
\begin{aligned}
\phi(u, v, w) &= -\sum_{i,j,k} t_{ijk} \log(u(i)v(j)w(k)) \\
&= -\sum_{i,j,k} t_{ijk} (\log u(i) + \log v(j) + \log w(k)) \\
&= -\begin{bmatrix} \boldsymbol{\alpha}^T & \boldsymbol{\beta}^T & \boldsymbol{\gamma}^T \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{u}} \\ \hat{\boldsymbol{v}} \\ \hat{\boldsymbol{w}} \end{bmatrix}, \\
\alpha_i &= \sum_{jk} t_{ijk}, \beta_j = \sum_{ik} t_{ijk}, \gamma_k = \sum_{ij} t_{ijk} \\
\hat{u}_i &= \log u(i), \hat{v}_i = \log v(i), \hat{w}_i = \log w(i)
\end{aligned}
$$

  - *This optimization problem is independent in $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$ and with normalization constraints the solution is $u(i) = \alpha_i / \|\boldsymbol{\alpha}\|_1$, $v(j) = \beta_j / \|\boldsymbol{\beta}\|_1$, and $w(k) = \gamma_k / \|\boldsymbol{\gamma}\|_1$, which are the marginal distributions of $p(i, j, k) = t_{ijk}/M$*

Probability Estimation from Samples

- Suppose we are interested in estimating a probability density $p(x, y, z)$ from samples $(x_k, y_k, z_k)_{k=1}^m$, using tensor products of orthogonal basis functions[2] $\psi_1, \ldots \psi_n$, so

$$p(x, y, z) = \sum_{i,j,k} c_{ijk}\psi_i(x)\psi_j(y)\psi_k(z)$$

  - *From orthogonality of the basis functions it follows that*

$$c_{ijk} = \int\int\int p(x, y, z)\psi_i(x)\psi_j(y)\psi_k(z)dxdydz$$

  - *Hence, we can get the expected value of $\mathcal{C}$ given the samples as*

$$\mathbb{E}[\mathcal{C}] = \frac{1}{m}\sum_{k=1}^{m} \boldsymbol{u}_k \otimes \boldsymbol{v}_k \otimes \boldsymbol{w}_k, u_k(i) = \psi_i(x_k), v_k(i) = \psi_i(y_k), w_k(i) = \psi_i(z_k)$$

  - *Tensor decompositions may then be optimized relative to this emperical distribution, and may be less sensitive to noise*

---

[2] Tang, Xun, and Lexing Ying. ”Solving high-dimensional Fokker-Planck equation with functional hierarchical tensor.” Journal of Computational Physics 511 (2024): 113110.