ROBUST STREAMING TENSOR FACTORISATION VIA ONLINE VARIATIONAL BAYESIAN INFERENCE

Avinandan Bose Spandan Senapati Ketan Rejawat

***Abstract*—Streaming Tensor Factorization is a powerful tool for processing high-volume and multi-way temporal data. It finds use in Internet Networks,recommender systems,image/video data analysis and Estimated Time of Arrival(ETA) predictions. This work puts forward the use of time-series and tensor completion algorithms. Our contribution is extending the previously existing Robust Streaming Tensor Factorization methods and incorporating first order auto-regressive time series in it.**

I. INTRODUCTION

Tensors are extensively used in various application domains including recommender systems, computer vision etc.Strategies such as CP factorizations, have been proposed to obtain low-rank representations of tensors. Streaming tensors appear sequentially in the time domain. Incorporating temporal relationships in tensor data analysis can give significant advantages in traffic prediction , recommender systems etc. Various Tensor Decomposition methods such as the Tucker and the CP decomposition have been proposed in the past few decades.They differ in the way they solve the numerical optimization problem. However these methods assume fixed rank and fail to capture outliers in the data. However streaming tensor factorisations follow bayesian inference that leads to an Automatic Rank Determination(ARD) and also enables one to capture the sparse outliers in the data.

We consider first low-rank robust subspace filtering approach for online tensor imputation and prediction. Different from the existing tensor completion formulations, we consider low-rank matrices whose underlying subspace evolves according to a first order Markovian Chain. As slices of tensor data matrix arrive sequentially over time, the low rank components as well as the state-space model are learned in an online fashion using the Variational Bayes formalism In this paper we discuss the traffic estimation problem and show that the Variational Bayesian approach incorporated with a first order auto regressive model can be used to learn the model in online setting which estimates the missing entries(i.e speed) when only a fraction of other entries is available. Robust version of the Variational Bayesian Subspace Filtering algorithm is proposed for outlier removal and data cleansing in a dynamic setting.

II. PROBABILISTIC MODEL

A. Details about Notations

Given data $\mathcal{Y}_{\Omega_T}$ in the form of a tensor of N dimensions, we wish to find its representation in it's CP Factorisation form. $\mathcal{Y}_{\Omega_T} =< A^{(1)}, A^{(2)}, ..., A^{(N)}; b_T > +S_{\Omega_T}$,where $A^{(1)}, .., A^{(N)}$ are the factor matrices which evolve over time ,b_t is temporal factor which follows the first order autoregressive model $p(b_t|J, b_{t-1}) = \mathcal{N}(b_t|Jb_{t-1}, I_r)$ and S_{Ω_T} accounts for the sparse outliers. The matrix B is of dimensions R $\times$ T, (R is the rank of the low-rank CP-decomposition and determined on its own,ie. not present) and the t^{th} row of B is b_t which is like a weighing factor in the CP decomposition for the factor matrices.We use Variational Bayesian Inference to compute the optimal values if the parameters.

B. Variational Inference

Variational Inference forms one of the means of Approximate Inference i.e. there are other forms of approximate inference like Markov Chain Monte Carlo Stimulation(MCMC) etc. Solving Optimisation Problems by Bayesian Inference can prove to be a handy alternative in many situations especially when convex optimisation methods can't be adopted for the same. In this we assume prior distributions over latent variables(variables that aren't observable but affect the model predictions inherently) and look for an iterative scheme to reach convergence of a function termed the Evidence Lower Bound(ELBO).

$$\log p(x) = \underbrace{\mathcal{L}_q}_{\text{ELBO}} + \underbrace{\mathrm{KL}(q(z)||p(z|x))}_{\text{KL Divergence serves as a measure of dissimilarity}} .$$

The optimal changes to the distribution q is obtained as,

$$q_j^* = \mathop{\mathbb{E}}_{i \neq j}\left[\ln p(X, z_1, z_2....z_n)\right]$$

where $z_1, z_2....z_n$ denote the latent variables in the model and q_z denotes the assumed prior distribution on the latent variable z. The conditional distributions are obtained from Bayes Theorem as,

$$p(z|x) = \frac{p(z,x)}{p(x)} = \frac{p(x|z)p(z)}{\int p(x|z)p(z)\mathrm{dz}}$$

C. Likelihood

$$\begin{aligned}\text{ELBO} &= p(Y_{\Omega_T}, \widetilde{D}_{t,\Omega_t})|A^{(1)}, A^{(2)}, A^{(3)}.....A^{(N)}, B, S_{\Omega_T}, \beta) \\ &= \mathcal{N}(y_{\omega_N}| < \hat{a_{i_1}}^{(1)}, \hat{a_{i_2}}^{(2)}..\hat{a_{i_n}}^{(n)}, b_T > + S_{\omega_N}, \beta^{-1})\end{aligned}$$

where $\widetilde{D}_t = Y_t - S_t$ Here $\widetilde{D}_t$ is the cleaned up data at time t after removing the sparse outliers and ω_N denotes the subscript $i_1, i_2, ...i_N$. We could hence use ω_N and the subscript Notation interchangably.

D. Prior Distributions

$A^{(1)}, A^{(2)}, ..., A^{(N)}$ are the factor matrices in the CP decomposition of input data which evolve over time.

$$\prod_{i=1}^{N} p(A^{(n)}|\lambda) = \prod_{k=1}^{N}\prod_{i_k}^{I_k} \mathcal{N}(\hat{a_{i_k}}^{(k)}|0, \Lambda^{-1}) \tag{1}$$

$$\text{where } \Lambda = \text{Diag}(\lambda)$$

$$p(\lambda) = \prod_{i=1}^{r} \Gamma(\lambda|c_0, d_0) \tag{2}$$

B is a matrix whose rows are the temporal components.Its rows follow a first order auto-regressive model.The following 3 equations describe this.

$$p(B|J) = N(b_i; \mu_1, \zeta_1) \prod_{t=2}^{T} \mathcal{N}(b_t; Jb_{t-1}, I_r) \tag{3}$$

$$p(J|\nu) = \prod_{i=1}^{r}\prod_{j=1}^{r} \mathcal{N}(J_{ij}|0, \nu_i^{-1} I) \tag{4}$$

$$p(\nu) = \prod_{i=1}^{r} \frac{1}{\nu_i} \tag{5}$$

S_{Ω_T} is a tensor which represents the sparse outliers in the input data.

$$p(S_{\Omega_T}|\gamma) = \prod_{i \in \Omega_T} \mathcal{N}(S_{i_1 i_2 ... i_N}|0, \gamma_{i_1 i_2 .... i_N}) \tag{6}$$

$$p(\gamma) = \prod_{i \in \Omega_T} \Gamma(\gamma_{i_1, i_2 ... i_N}|a_0^{\gamma}, b_0^{\gamma}) \tag{7}$$

β is the precision factor in our likelihood function.

$$p(\beta) = \Gamma(\beta|a_\beta, b_\beta) \tag{8}$$

III. Variational Bayesian Updates

A. Factor Matrices Updates

$$q(A^{(n)}) = \prod_{i_n=1}^{I_n} \mathcal{N}(\hat{a}_{i_n}^{(n)}|\bar{a}_{i_n}^{(n)}, V_{i_n}^{(n)}) \tag{9}$$

$$V_{i_n}^{(n)} = (\mathop{\mathbb{E}}_{q}[\beta]\Sigma_{t=i}^{T} \mathop{\mathbb{E}}_{q}[A_{i_n}^{(\backslash n)T} A_{i_n}^{(\backslash n)}]_{\Omega t} + \mathop{\mathbb{E}}_{q}[\Lambda])^{-1} \tag{10}$$

$$\begin{aligned}\bar{a}_{i_n}^{(n)} &= V_{i_n}^{(n)}(\mathop{\mathbb{E}}_{q}[A_{i_n}^{(\backslash n)T}]_{\Omega_T} \text{vec}(y_{\Omega_T} - \mathop{\mathbb{E}}_{q}[S_{\Omega_T}])) + \\ &= \Sigma_{t=i}^{T-1} \mathbb{E}_q[A_{i_n}^{(\backslash n)T}]_{\Omega_T} \text{vec}(\widetilde{D}_{t,\Omega_t})\end{aligned}$$

$$\mathop{\mathbb{E}}_{q}[A_{i_n}^{(\backslash n)}]_{\Omega_T} = (\mathop{\mathbb{E}}_{q}[\bigodot_{j \neq n} A^{(j)}])_{I_{i_n}} \tag{11}$$

The matrix $A_{i_n}^{(\backslash n)} is \prod_{j \neq n} I_j R$ and the Indicator function I_{i_n} samples the row $(i_1, ..., i_{n-1}, i_{n+1}, ..., i_{N+1})$ if the entry $(i_1, ..., i_{n-1}, i_{n+1}, .., i_{N+1})$ is in Ω_t. The Matrix $A^{(N+1)}$ is the temporal component and is equal to B,and we will look at its update seperately in Section 3.4 which is different from the updates of the non-temporal components.

B. Posterior Distribution of λ

$$q(\lambda) = \prod_{r=1}^{R} \Gamma(\lambda_r|c_M^r, d_M^r) \tag{12}$$

$$c_M^r = c_0 + 1 + \frac{\sum_{i=1}^{N} I_i}{2} \tag{13}$$

$$d_M^r = d_0 + 1 + \frac{1}{2}\sum_{i=1}^{N+1} \mathop{\mathbb{E}}_{q}[(a_r^n)^T a_r^n] \tag{14}$$

$$\mathop{\mathbb{E}}_{q}[\Lambda] = \text{Diag}(\frac{c_M^1}{d_M^1}, \frac{c_M^2}{d_M^2}, ... \frac{c_M^R}{d_M^R}) \tag{15}$$

where R is the learned Rank and M denotes the number of iterations.

C. Updates for J

$$q_{j_{i\cdot}} = \mathcal{N}(j_i.|\mu_i^J, \Xi_i^J) \tag{16}$$

$$\Xi_i^J = (\text{Diag}(\hat{\nu}) + \sum_{t=1}^{\tau} \Sigma_{t,t-1}^B)^{-1} \tag{17}$$

$$\mu_i^J = [\Xi_i^J \Sigma_{t,t-1}^B]_{.i} \tag{18}$$

where Ξ_i^J corresponds to the Covariance Matrix corresponding to the i th row of J and τ denotes current time.

$$\Sigma_{t,t-1}^B = \mu_t^B (\mu_{t-1}^B)^T + \Xi_{t,t-1}^B \tag{19}$$

$\Xi^B_{t,t-1}$ denotes the Covariance Matrix corresponding to the $t-1$ th Column of B which has been obtained already at time $t-1$

D. Updates for B

Define

$$\hat{\beta} = \frac{|\Omega_t|}{\sum_{t=1}^{\tau}\left(\sum_{i_1,i_2...i_N\in\Omega_t}\left(\Psi_t^1+\Psi_t^2\right)\right)} \quad (20)$$

$$\Psi_t^1 = y^2_{i_1,i_2.....i_N} - 2(y_{i_1,i_2....i_N} - [\mu^S_{i_1....i_N}]_{t=t})(\mu_i^A)^T\mu_\tau^B \\ -2y_{i_1,i_2.....i_N}[\mu^S_{i_1,i_2.....i_N}]_{t=t}$$

$$\Psi_t^2 = ([\mu_{i_1,i_2.....i_N}]^S_{t=t})^2 + \Xi^S_{i_1,i_2.....} + \text{Trace}(\Sigma_i^A\Sigma^B_{t,t})$$

$y_{i_1,i_2....i_N} - [\mu_{i_1....i_N}]^S_{t=t} = [D_{i_1,i_2....i_N}]_{t=t}$ for $t \le \tau-1$ and for $t=\tau$ we expresss it otherwise. Also $\mu^S_{i_1,i_2.....i_N}$ corresponds to the mean of the Univariate Normal corresponding to the $i_1, i_2....., i_n$ entry of S.

Now we describe how to define A a matrix of dimensions $\prod_{i=1}^N I_i \times R$, and subsequently defining the mean and covariance for each of its rows and this serves useful for the update of B.

$A = A^{(1)} \bigodot A^{(2)} \bigodot .....A^{(N)}$ and $\bigodot$ denotes the Khatri Rao product.

$$\mu^A = \left[< A^{(1)} > \bigodot < A^{(2)} > \bigodot < A^{(3)} > \bigodot ... < A^N > \right]$$

and μ_i^A denotes mean corresponding to i th row of A that is and the angular brackets denote the expectations of each Factor Matrix.
Define

$$\Xi_i^A = \prod_{k=1}^N{}_{\text{elementwise}}(V_{i_k}^{(k)} + \bar{a}^k_{i_k}(\bar{a}^k_{i_k})^T) - \prod_{k=1}^N{}_{\text{elementwise}}\bar{a}^k_{i_k}(\bar{a}^k_{i_k})^T$$

Ξ_i^A denotes to Covariance Matrix of ith row of A and $i_1, i_2, .., i_N$ are in the same order as the rows are selected in the Khatri Rao Product that gives A. The next few lines define some terms which are used to reduce the clumsiness of the expressions.

$$\Sigma_i^A = \mu_i^A(\mu_i^A)^T + \Xi_i^A$$
$$\Sigma^B_{\tau,\tau} = \mu_\tau^B(\mu_\tau^B)^T + \Xi^B_{\tau,\tau}$$

where $\Sigma^B_{\tau,\tau}$is an rt $\times$ rt matrix

$$\Xi_B = \hat{\beta}\cdot\text{Diag}(\Sigma^A_{(1)},\Sigma^A_{(2)}..\Sigma^A_{(\tau)}) + \underbrace{\begin{bmatrix}\zeta_1^{-1} & -\hat{J} & .. & .. & 0\\ -\hat{J} & I_r+\Sigma_J & -\hat{J} & .. & ..\\ .. & .. & .. & -\hat{J} & I_r\end{bmatrix}}_{\mathcal{X}} \quad (21)$$

where $\mathcal{X}$ is a triagonal matrix and $\Sigma^A_{(t)} = \sum_{i\in\Omega_t}\Sigma_i^A$

$$\mu_B = \Xi_B\begin{bmatrix}\hat{\beta}\sum_{i\in\Omega_1}[y_{i_1,i_2....i_n}]_{t=1}\cdot\mu_i^A + \zeta_1^{-1}\mu_1\\ \hat{\beta}\sum_{i\in\Omega_2}[y_{i_1,i_2.....i_n}]_{t=2}\cdot\mu_i^A\\ \cdots\\ \cdots\\ \cdots\\ \hat{\beta}\sum_{i\in\Omega_\tau}[y_{i_1,i_2.....i_n}]_{t=\tau}\cdot\mu_i^A\end{bmatrix} \quad (22)$$

where $\hat{J} := \mathbb{E}_q[J|y_\Omega]$

E. Posterior Distribution of Sparse S

$$q(S_{\Omega_T}) = \prod_{i_1i_2..i_N\in\Omega_T}\mathcal{N}(S_{i_1i_2..i_N}|\mu^S_{i_1i_2..i_N},\sigma^2_{i_1i_2..i_N}) \quad (23)$$

$$\mu^S_{i_1i_2..i_N} = \sigma^2_{i_1i_2..i_N}\,\mathbb{E}_q[\beta](\mathcal{Y}_{i_1i_2..i_N} - \mathbb{E}_q[< \hat{a}^{(1)}_{i_1},...,\hat{a}^{(N)}_{i_N}; b_T >]) \quad (24)$$

$$\sigma^2_{i_1i_2..i_N} = (\mathop{\mathbb{E}}_q[\gamma_{i_1i_2..i_N}] + \mathop{\mathbb{E}}_q[\beta])^{-1} \quad (25)$$

F. Posterior Distribution of γ

$$q(\gamma_{i_1,i_2,i_3.....i_N}) = \Gamma\left(\gamma_{i_1,i_2,i_3...i_N}|a_0^\gamma + \frac{1}{2}, b^{\gamma_{i_1},...,i_N}\right) \quad (26)$$

$$< \gamma_{i_1,i_2...i_N} >= \frac{a_0^\gamma + \frac{1}{2}}{b^{\gamma_{i_1,i_2...i_N}}} \quad (27)$$

where

$$b^{\gamma_{i_1},...,i_N} = b_0^\gamma + \tfrac{1}{2}\cdot\left((\mu_{i_1,i_2.....i_N})^2 + \sigma^2_{i_1,i_2.....i_N}\right)$$

G. Posterior Distribution of β

$$q^\beta = \Gamma(\beta|a_0,b_0) \quad (28)$$

$$a_0 = a_\beta + \frac{|\Omega_\tau|}{2} \quad (29)$$

$$b_0 = b_\beta + \frac{\mathbb{E}_q[||\mathcal{Y} - < \hat{a}^{(1)}_{i_1},\hat{a}^{(2)}_{i_2},...,\hat{a}^{(N)}_{i_N}; b_T > -S_T||^2_F]}{2} \quad (30)$$

H. Update for ν

$$\hat{\nu}_i = \frac{T}{\sum_{k \in \text{rows}} [\mu_k^J]_i^2 + [\Xi_k^J]_{ii}} \tag{31}$$

$$\nu = \Gamma(\nu | \Theta, \kappa) \tag{32}$$

such that

$$\frac{\Theta}{\kappa} = \hat{\nu}_i \tag{33}$$

where $[\mu_k^J]_i$ denotes the i th term of μ_k^J and $[\Xi_k^J]_{ii}$ denotes the ii th term of Ξ_k^J. $\hat{\nu}_i$ denotes the updated Mean of the Gamma Distribuion of ν.

REFERENCES

[1] Cole Hawkins, Zheng Zhang, Variational Bayesian Inference for Robust Streaming Tensor Factorization and Completion.

[2] Charul, Uttkarsha Bhatt, Pravesh Biyani, Ketan Rajawat, Traffic estimation and prediction via Online Variational Bayesian Subspace Filtering