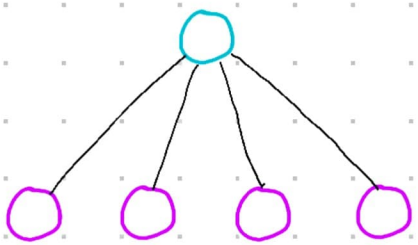
Stochastic Alternating Least Squares For Tensor Decomposition.

Luke Oeding Joint Work with Yanzhao Cao
Somak Das
Hans Werner Van Wyk

Auburn University



ArXiv: 2004.12530

A
B
C
D

A G T - - -
A C T - - -
A G T - - -
A G C

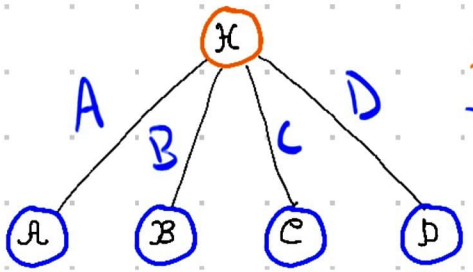
S

Data

Contingency table of occurrences of Quadruples of letters

T_{abcd} increased by 1 whenever see a,b,c,d

Motivating Example:



(so K=4)

Given Aligned Sequences of DNA from species that are assumed independent given a common ancestor

Determine the parameters of this model

$$ABT = \sum_s A_s \otimes B_s$$

The Model: $P_{abcd} = P(A=a, B=b, C=c, D=d)$

$$= \sum_{h \in \mathcal{H}} P(A=a, B=b, C=c, D=d \mid H=h) \cdot P(H=h)$$

$$= \sum_{h \in H} P(A=a \mid H=h) \cdot P(B=b \mid H=h) \cdots P(D=d \mid h=H) \cdot P(H=h)$$

$$P_{abcd} = \sum_{h \in H} A_{ah} B_{bh} C_{ch} D_{dh} \cdot \alpha_h$$

$$= \sum_{h \in H} \alpha_h A_h \otimes B_h \otimes C_h \otimes D_h$$

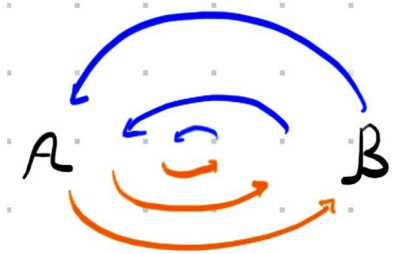
The Data: $T = (T_{abcd})$

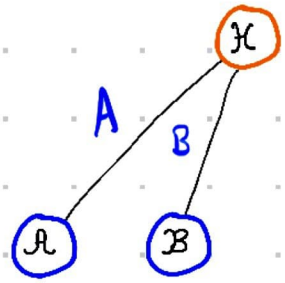
A:
B:
C:
D:

A	C	G	T	...
A	C	G	T	
C	T	A	T	
A	C	C	T	

2-factor Example

Alternating Least Squares





$\tilde{T} = (\tilde{T}_{ab}) = \frac{1}{n}(C_{ab})$ — n - Samples

$P = A B^T + E$

See [Li - Kinderman - Navasca 2013]

$$\min_{A,B} \| T - AB^T \|^2 + \frac{\lambda}{2}\left(\|A\|^2 + \|B\|^2\right)$$

Regularization

Not Convex, in general very hard.

If have A approximate B: Regularized ordinary least Squares

$$\hat{b}_u^t = (A^T A + \lambda I)^{-1} A^T T_u$$

If have B approximate A: ROLS

$$\hat{a}_v^t = (B^T B + \lambda I)^{-1} B^T T_v^z$$

Pseudo code

Guess A

While $\|T - AB^T\|^2 + \frac{\lambda}{2}(\|A\|^2 + \|B\|^2) > \varepsilon$

- Approximate $\hat{B}$ holding A fixed
  - $B \leftarrow \hat{B}$
- Approximate $\hat{A}$ holding B fixed
  - $A \leftarrow \hat{A}$

Note ROLS steps don't increase error

Options For ROLS

- use a closed form solution
- use SVD
- use an iterative method (Gauss-Seidel)
- use a stochastic method
- Reformulate to take advantage of Sparsity

Scale Ambiguity and Regularization

$$T \approx A B^T$$

$$= (AU)(U^T B^T) \qquad UU^T = I$$

$$T = (A\ \mathrm{Diag}(d_1 .. d_n))(\mathrm{Diag}(d_1^{-1} .. d_n^{-1})\ B)$$

Option 1: Restrict cols of A, B to be all on a unit sphere.

Option 2: Regularize

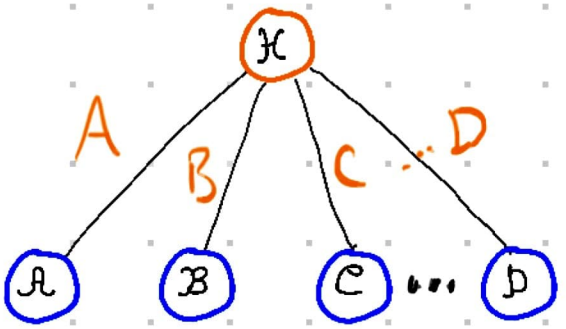
Introduce penalty term

$$+ \frac{\lambda}{2}\left(\|A\|^2 + \|B\|^2\right)$$

$$\|AD\|^2 \le \|A\|^2 \|D\|^2$$

$$\|D^{-1}B\|^2 \le \|D^{-1}\|^2 \|B\|^2$$

d-factors



[Carrol-Chang 1970]

$$(T_{abcd}) \approx (P_{abcd}) = A \odot B \odot C \odot \cdots \odot D$$

"highly" non-linear

$$= \sum_{h \in H} \lambda_h \, A_h \otimes B_h \cdots \otimes D_h$$

$$\underset{A,B,C,\ldots D}{\text{Minimize}} \; \| T - A \odot B \odot C \odot \cdots \odot D \|^2 + \tfrac{1}{2}\left(\|A\|^2 + \|B\|^2 + \|C\|^2 + \cdots + \|D\|^2 \right)$$

Guess A, B, C

While objective > tol

$D \leftarrow$ solve ROLS for D with A, B, C fixed

$A \leftarrow$ solve ROLS for A with B, C, D fixed

$B \leftarrow$ solve ROLS for B with A, C, D fixed

$C \leftarrow$ solve ROLS for C with A, B, D fixed

Stochastic Methods

Warmup Real Time Averaging

Given $\{x_1, x_2, \ldots \quad x_n, \ldots \} = x$

estimate the average (Expected Value)

$$E(x) \approx \frac{1}{n} \sum x_i = y_n$$

Sample and update $\quad y_{n+1} = \frac{n}{n+1} y_n + \frac{1}{n+1} x_{n+1}$

Deterioration of Sparsity

$\vec{x_i}$ only with few non-zero terms

Large sample → Dense Avergl.

Basic Newton Method for Minimization

Assume $f(x)$ is C^2 (use 2nd order Taylor approximates)

guess x_1

Iterate $x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$ ← t

until $\|f(x_k)\|^2$ stops improving

Learning Rate γ

$$x_{k+1} = x_k + \gamma t$$

Higher Dimensions:

$$f'(x) = \nabla f(x) = g(x)$$
$$f''(x) = \nabla^2 f(x) = H(x)$$

guess x_1

update $x_{k+1} = x_k - \gamma [H(x)]^{-1} g$

Stochastic Newton

$H_i = \nabla^2 f_i$

$g_i = \nabla f_i$

If $f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$ — n large

update

$$x_{k+1} = x_k - \frac{1}{n}\left(\sum_{i=1}^{n} H_i(x_k)\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} g_i(x_k)\right)$$

↑ hard to compute

Instead only update a sample of the directions

Keep a sequence $x_k^1 \cdots x_k^n$ ← subseq. ← iterate

Sample $S \subseteq \{1, \ldots, n\}$

Only need to recompute $H_i(x_k^i)$ and $g_i(x_k^i)$

for $i \in S$

CP Decomposition from Samples

$\#\mathcal{C} = k$

k fixed

Suppose $\mathcal{X}$ is a random tensor, we observe samples $\mathcal{X}^1, \dots, \mathcal{X}^m$

And want to find a CP decomp (factor matrices)

$$\hat{\mathcal{X}} = \sum_{i=1}^{k} a_1^i \otimes \cdots \otimes a_d^i =: [\![A_1, \dots, A_d]\!] = [\![x]\!]$$

Solving $\quad \underset{\hat{\mathcal{X}} \in \mathbb{R}^{dnk}}{\operatorname{argmin}} \; \mathbb{E}\left[\|\mathcal{X} - \hat{\mathcal{X}}\|^2 \right]$ ave decomp

equivalently, $\quad \underset{A_1, \dots, A_d \in \mathbb{R}^{kn}}{\operatorname{argmin}} \left\| \mathbb{E}[\mathcal{X}] - [\![A_1 \cdots A_d]\!] \right\|^2$ decomp ave tensor

We don't have access to $\mathcal{X}$, only samples of $\mathcal{X}$

Regularized Version: $\quad x = (A_1 \cdots A_d)$

$$\underset{x \in \mathbb{R}^{knd}}{\operatorname{argmin}} \; \mathbb{E}\left(\| \mathcal{X} - [\![x]\!] \|^2 + \lambda \|x\|^2 \right) \qquad \lambda > 0$$

SALS

Block stochastic Newton minimization

[Maehara, Hayashi, Kawarabayashi 2016]

(other instances of Randomized tensor decomp

[Kolda - Hong, Battaglino - Ballard - Kolda 2018]

(Dense w/ sparse samples) (Randomized methods for OLS))

Fix Rank Say r

Given Samples $\underline{\mathcal{X}}^1, \ldots, \underline{\mathcal{X}}^m$ of $\underline{\mathcal{X}}$

Find a minimizer for (regularized) CP decomp of $\mathbb{E}[\underline{\mathcal{X}}]$

Variables $x = [A_1, \ldots, A_d]$ ← A_i is $n \times r$

① Guess x^1

For $k = 1, 2, \ldots$ do

Generate Random Sample $\underline{\mathcal{X}}^k = [\underline{\mathcal{X}}^{k,1}, \ldots, \underline{\mathcal{X}}^{k,m_k}]$

For $i = 1 \ldots p$ do

Compute sample gradient $\tilde{g}^{k,i}$ and Hessian $H^{k,i}$

Compute step size $\alpha^{k,i}$

Update block i :

$$x_i^{k+1} = x_i^k - \alpha^{k,i} \left(H^{k,i}\right)^{-1} \tilde{g}^{k,i}$$

So

$$x^{k,i} = \left(X_1^{k+1}, \ldots, X_i^{k+1}, X_{i+1}^{k}, \ldots, X_p^{k} \right)$$

End for

$x^{k+1} = x^{k,p}$

End for

Theorem (Cao-Das-Oeding-Van Wyk)

If the observed data is bounded,

Then the SALS algorithm for the regularized CP decomposition of the average of a random tensor converges

↑ To a minimizer

Bounded data Assumption

$$\|\mathcal{X}\| \leq M \quad \text{a.s. on } \Omega$$

Boundedness and multilinearity

→ Lipschitz Continuity of gradients and Hessians

→ Bounded invertibility of Hessians

→ Bounds on expectations of iterates

$$\mathbb{E}\,\|x_i^{k+1}\|$$

Regularization ensures:

1. Minimizers exist
2. Hessians stay positive Definite (and invertable)
3. Iterates are bounded in terms of bounds on Σ

Numerical Experiments study effects of

- Noise
- Learning Rate
- Batch Size
- Complexity