Some tensor decomposition methods for machine learning

Massimiliano Pontil

Istituto Italiano di Tecnologia and University College London

16 August 2016

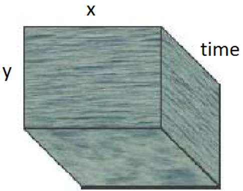
Outline

- Problem and motivation
- Tucker decomposition
- Two convex relaxations
- Applications
- Extensions

Problem

Learning a tensor from a set of linear measurements

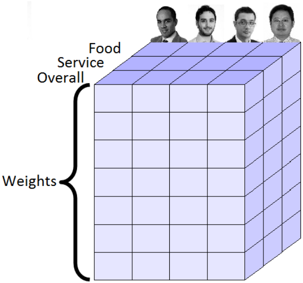
Main example: Tensor completion



- Video denoising/completion
- Context-aware recommendation
- Multisensor data analysis
- Entities-relationships learning
- ...

Multilinear multitask learning

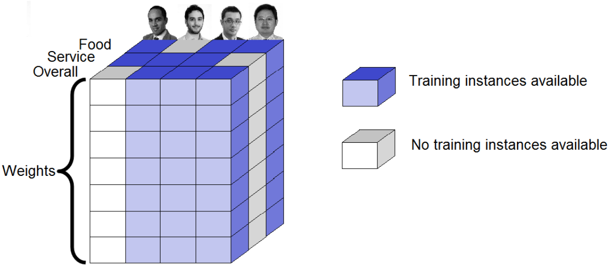
Tasks are associated with multiple indices, e.g. predict a rating given to different aspects of a restaurant by different critics



Tasks' regression vectors are "vertical" fibers of the tensor, e.g. (, 'food')

0-Shot Transfer Learning

Learning tasks for which no training instances are provided



Setting

We consider a supervised learning setting in which we wish to learn a tensor from a set of measurements

$$y = I(\mathcal{W}) + \epsilon$$

where $I : \mathbb{R}^{d_1 \times \cdots d_N} \rightarrow \mathbb{R}^m$ a linear (sampling) operator, e.g. a subset of the tensor entries

The number of parameters explodes with the order of the tensor, so regularization is key:

$$\underset{\mathcal{W}}{\text{minimize}}\ E(\mathcal{W}) + \lambda R(\mathcal{W})$$

for example $E(\mathcal{W}) = \|y - I(\mathcal{W})\|^2$

Matrix case

The matrix case has been thoroughly studied, particularly focusing on spectral regularizers which encourage low rank matrices

- low rank matrix completion
- multitask feature learning

Different notions of rank of a tensor! (some computationally intractable). Many are reviewed in [Kolda and Bader, 2009]

Which ones are simple and effective?

Objectives

We want some kind of guarantees

- statistical: bounds on the generalization / out-of-sample error
- optimization: convergence, rates
- function approximation

We discuss two approaches based on:

- B. Romera-Paredes, H. Aung, N. Bianchi-Berthouze, M. Pontil. **Multilinear multitask learning**. *30th International Conference on Machine Learning*, 2013
- B. Romera-Paredes & M. Pontil. **A new convex relaxation for tensor completion**. *Advances in Neural Information Processing Systems 26*, 2013

Function approximation

Learn function parametrized by a tensor from a sample

Example: $\mathbf{x} = (x^1, x^2, x^3) \in V_1 \times V_2 \times V_3$, three vector spaces

$$f(\mathbf{x}) = \langle \mathcal{W}, \phi_1(x^1) \otimes \phi_2(x^2) \otimes \phi_3(x^3) \rangle, \quad \mathcal{W} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$$

$\phi_i : V_i \to \mathbb{R}^{d_i}$ feature maps

$$\min_{\mathcal{W}} \sum_{i=1}^{m} Loss(y_i, f(\mathbf{x}_i)) + \gamma \Omega(\mathcal{W})$$

Dual problem using kernel methods

$$K(\mathbf{x}, \mathbf{t}) = K_1(x^1, t^1) K_2(x^2, t^2) K_3(x^3, t^3)$$

Matrix case: [Abernethy, Bach, Evgeniou, and Vert, JMLR 2009]

Tensor case: [Signoretto, De Lathauwer, Suykens, Preprint 2013]

Basic operations

Tensor $\mathcal{W} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$, with entries $W_{i_1,\ldots,i_N}$

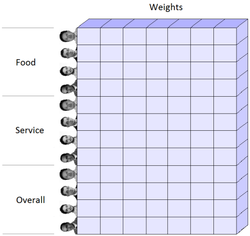
- **Mode-*n* fiber** is a vector in $\mathbb{R}^{d_n}$ formed by the elements of a tensor obtained by fixing all indices but the *n*-th one, e.g. in the above example $W_{i_1,i_2,:} \in \mathbb{R}^{d_3}$ is the (i_1, i_2)-th regression vector

Basic operations (cont.)

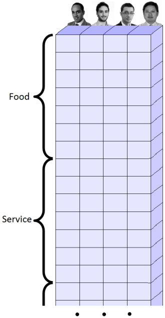
Matricization is the process of rearranging the tensor into a matrix

- **Mode-n matricization** $W_{(n)} \in \mathbb{R}^{d_n \times J_n}$, where $J_n = \prod_{k \neq n} d_k$, is the matrix, the column of which are the mode-n fibers of $\mathcal{W}$

$W_{(1)}$



$W_{(3)}$



Tucker rank

- $TR(\mathcal{W}) = (\mathrm{rank}(W_{(1)}), \dots, \mathrm{rank}(W_{(n)}))$
- A natural regularizer associated to this is the sum of the ranks of the matricizations:

$$R(\mathcal{W}) := \sum_{n} \mathrm{rank}(W_{(n)})$$

Tucker decomposition

$$\mathcal{W} = \mathcal{G} \times_1 A^{(1)} \times \cdots \times_N A^{(N)}$$

meaning that

$$\mathcal{W}_{i_1,\ldots,i_N} = \sum_{j_1=1}^{K_1} \cdots \sum_{j_N=1}^{K_N} \mathcal{G}_{j_1,\ldots,j_N} A^{(1)}_{i_1,j_1} \cdots A^{(N)}_{i_N,j_N}$$

$K_n \leq d_n$ (typically much smaller)

- $A^{(n)} \in \mathbb{R}^{d_n \times K_n}$, $n \in \{1, \ldots, N\}$ are the factor matrices
- $\mathcal{G} \in \mathbb{R}^{K_1 \times \cdots \times K_N}$ is called the core tensor and models the interaction between factors
- By construction $\mathrm{rank}(W_{(n)}) \leq K_n$

Nonconvex approach

The Tucker decomposition is invariant under multiplication and division of different factors by the same scalar. With the aim of avoiding this issue and reducing overfitting, we add Frobenius norm regularization terms to the components

$$\underset{\mathcal{G},\mathrm{A}^{(1)},\ldots,\mathrm{A}^{(\mathrm{N})}}{\text{minimize}} E(\mathcal{G} \times_1 A^{(1)} \cdots \times_N A^{(N)}) + \alpha \left[\|\mathcal{G}\|_{\mathrm{F}}^2 + \sum_{n=1}^{N} \|A^{(n)}\|_{\mathrm{F}}^2 \right]$$

where α is a regularization parameter

We solve this problem by alternate minimization

Each step is detailed in the paper (also code available)

Convex approach

An alternative approach is to relax the combinatorial problem

$$\underset{\mathcal{W}}{\operatorname{argmin}}\, E\left(\mathcal{W}\right) + \gamma \sum_{n=1}^{N} \operatorname{rank}\left(W_{(n)}\right)$$

The trace norm is a widely used convex surrogate for the rank. Therefore, we can consider the following convex relaxation:

$$\underset{\mathcal{W}}{\operatorname{argmin}}\, E\left(\mathcal{W}\right) + \gamma \sum_{n=1}^{N} \left\| W_{(n)} \right\|_{\mathrm{Tr}}$$

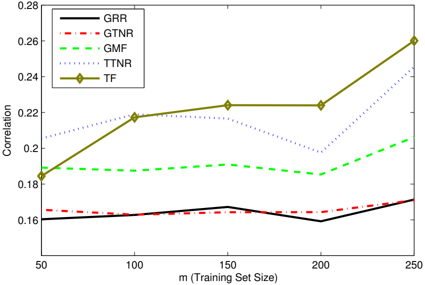
Tensor completion: [Liu et al, 2009, Gandy et al, 2011, Signoretto et al, 2012]

Discussion

In the Tucker decomposition the factors are explicit, hence we can add a new factor without relearning the full tensor

Space complexity: $\mathrm{O}\left(\sum_{n=1}^{N} d_n K_n + \prod_{n=1}^{N} K_n\right)$ which can be much smaller than that of the convex approach, particularly if $K_n \ll d_n$

Main drawback: no guarantee to find a local minimum



Alternating Direction Method of Multipliers (ADMM)

- The regularizer $\sum_{n=1}^{N} \left\| W_{(n)} \right\|_{\mathrm{Tr}}$ is composite
- ADMM decouples the problem
  - Introduce auxiliary tensors $\mathcal{B}^n$, $\forall n \in \{1, ..., N\}$ accounting for each term in the sum, adding the constraints $\mathcal{B}^n = \mathcal{W}$
  - Optimize the resultant Lagrangian w.r.t. each $\mathcal{B}^n$ only involves computing the proximal operator of the trace norm:

$$\mathrm{prox}_{\|\cdot\|_{\mathrm{Tr}}}(V) = \underset{X \in \mathbb{R}^{d \times d}}{\mathrm{argmin}} \frac{1}{2} \|X - V\|_{\mathrm{F}}^2 + \|X\|_{\mathrm{Tr}}$$

ADMM

Want to minimize

$$\frac{1}{\gamma}E(\mathcal{W})+\sum_{n=1}^{N}\Psi\left(W_{(n)}\right)$$

Decouple the regularization term

$$\min_{\mathcal{W},\mathcal{B}_1,\ldots,\mathcal{B}_N}\left\{\frac{1}{\gamma}E\left(\mathcal{W}\right)+\sum_{n=1}^{N}\Psi\left(B_{n(n)}\right)\ :\ \mathcal{B}_n=\mathcal{W},\ n=1,\ldots,N\right\}$$

Augmented Lagrangian:

$$\mathcal{L}\left(\mathcal{W},\mathcal{B},\mathcal{C}\right)=\frac{1}{\gamma}E\left(\mathcal{W}\right)+\sum_{n=1}^{N}\left[\Psi\left(B_{n(n)}\right)-\langle\mathcal{C}_n,\mathcal{W}-\mathcal{B}_n\rangle+\frac{\beta}{2}\left\|\mathcal{W}-\mathcal{B}_n\right\|_2^2\right]$$

ADMM (cont.)

$$\mathcal{L}\left(\mathcal{W},\mathcal{B},\mathcal{C}\right)=\frac{1}{\gamma}E\left(\mathcal{W}\right)+\sum_{n=1}^{N}\left[\Psi\Big(\big(B_{n(n)}\big)\Big)-\langle\mathcal{C}_n,\mathcal{W}-\mathcal{B}_n\rangle+\frac{\beta}{2}\left\|\mathcal{W}-\mathcal{B}_n\right\|_2^2\right]$$

Updating equations:

$$\begin{aligned}\mathcal{W}^{[i+1]} &\leftarrow \underset{\mathcal{W}}{\operatorname{argmin}}\ \mathcal{L}\left(\mathcal{W},\mathcal{B}^{[i]},\mathcal{C}^{[i]}\right)\\ \mathcal{B}_n^{[i+1]} &\leftarrow \underset{\mathcal{B}_n}{\operatorname{argmin}}\ \mathcal{L}\left(\mathcal{W}^{[i+1]},\mathcal{B},\mathcal{C}^{[i]}\right)\\ \mathcal{C}_n^{[i+1]} &\leftarrow \mathcal{C}_n^{[i]}-\left(\beta\mathcal{W}^{[i+1]}-\mathcal{B}_n^{[i+1]}\right)\end{aligned}$$

- 2nd step involves the computation of proximity operator of Ψ

Convergence properties of ADMM are detailed e.g. in [Eckstein and Bertsekas, Mathematical Programming 1992]

Proximity Operator

Let $B = B_{n(n)}$ and where $A = (\mathcal{W} - \frac{1}{\beta}\mathcal{C}_n)_{(n)}$. Rewrite 2nd step as:

$$\hat{B} = \mathrm{prox}_{\frac{1}{\beta}\Psi}(A) := \operatorname*{argmin}_{B} \left\{ \frac{1}{2} \|B - A\|_2^2 + \frac{1}{\beta}\Psi(B) \right\}$$

Case of interest: $\Psi(B) = \psi(\sigma(B))$

By von Neumann's matrix inequality:

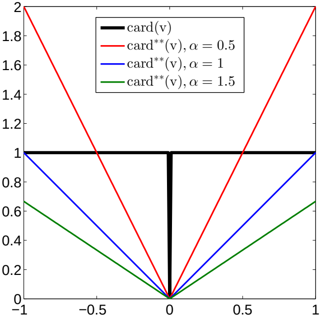
$$\mathrm{prox}_{\frac{1}{\beta}\Psi}(A) = U_A \mathrm{diag}\Big(\mathrm{prox}_{\frac{1}{\beta}\psi}(\sigma_{\mathrm{A}})\Big) V_A^\top$$

Rethinking the convex approach

Convex envelope of a function f on a set S is the largest convex function f^{**} majorized by f at every point in S

E.g: cardinality of a vector:

- $f(v) = card(v)$
- $S = \{v : ||v||_\infty \leq \alpha\}$
- $f^{**}(v) = ||v||_1/\alpha$



The smaller S, the tighter the convex envelope

In practice α is unknown and tuned by cross validation

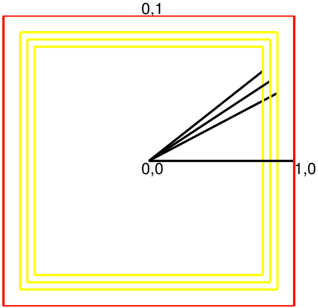
Rethinking the convex approach

The same reasoning applies to matrices

- $\|W\|_{\mathrm{Tr}}/\alpha$ is the convex envelope of $\mathrm{rank}(\mathrm{W})$ on the spectral unit ball of radius α [Fazel, Hindi, & Boyd, 2001]
- By using the regularizer $\sum_{n=1}^{N} \left\| W_{(n)} \right\|_{\mathrm{Tr}}$ we implicitly assume the same α for the different matricizations
- Difficulty with tensors: $\|W_{(n)}\|_{\infty}$ varies with n!

Rethinking the convex approach

- $\left\| W_{(n)} \right\|_{\infty}$ varies with n
- Let us consider $\mathcal{W} \in \mathbb{R}^{2\times2\times2\times2}$. Then:



■ Vectors of singular values of each matricization

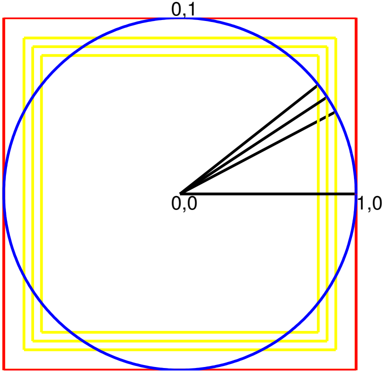
🟨 Smallest ℓ_{∞} ball containing each of the vectors

🟥 Smallest ℓ_{∞} ball containing all vectors

Rethinking the convex approach

- We are interested in convex functions on matrices which are invariant to the matricization operation
- The Frobenius norm is very appealing:
  - It is also a spectral function
- Therefore, we consider the set $S = \{W : \|W\|_{\mathrm{F}} \leq \alpha\}$
- Calculating the convex envelope of the rank on S can be reduced to calculating the convex envelope of $\mathrm{card}\,(\mathrm{v})$ on the set $\{v : \|v\|_2 \leq \alpha\}$, where v is the vector of singular values of W (follows by von Neumann's matrix inequality)

Convex envelope of the cardinality of a vector in the ℓ_2 ball



■ Vectors of singular values of each matricization

🟨 Smallest ℓ_∞ ball containing each of the vectors

🟥 Smallest ℓ_∞ ball containing all vectors

🟦 Smallest ℓ_2 ball containing all vectors

Convex envelope of the cardinality of a vector in the ℓ_2 ball

- **Aim**: derive convex envelope of $f_\alpha(x) = \mathrm{card}(x)$ on the set $\{x \in \mathbb{R}^d \mid \|x\|_2 \leq \alpha\}$
- The conjugate of f_α, $\forall s \in \mathbb{R}^d$, is

$$f_\alpha^*(s) = \sup_{x \in \mathcal{B}_2} x^\top s - \mathrm{card}(x) = \max_{r \in \{0,\ldots,d\}} \{\alpha \|s_{1:r}\|_2 - r\}$$

- Biconjugate of f_α:

$$f_\alpha^{**}(v) = \sup_{s \in \mathbb{R}^d} s^\top v - f_\alpha^*(s), \quad \|v\|_2 \leq \alpha$$

We do now know how to compute f^{**}. We can compute it and its proximal operator by projected sub-gradient descent

Quality of Relaxation (cont.)

Lemma. If $\|x\|_2 = \alpha$ then $\omega_\alpha(x) = \mathrm{card}(x)$.

Let

$$\Omega_\alpha(\mathcal{W}) = \sum_{n=1}^{N} \omega_\alpha(\sigma(W_{(n)})), \qquad \|\mathcal{W}\|_{\mathrm{tr}} = \sum_{n=1}^{N} \|\sigma(W_{(n)})\|_1$$

Implication:

Theorem. If $\mathcal{W}$ satisfies (a,b,c) below then $\Omega_{p_{\min}}(\mathcal{W}) > \|\mathcal{W}\|_{\mathrm{tr}}$

a) $\|W_{(n)}\|_\infty \leq 1 \ \forall n$

b) $\|\mathcal{W}\|_2 = \sqrt{p_{\min}}$

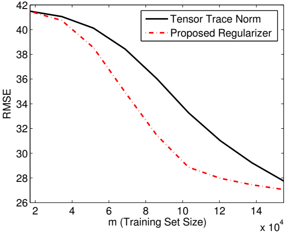
c) $\min_n \mathrm{rank}(W_{(n)}) < \max_n \mathrm{rank}(W_{(n)})$

On the other hand, ω_1 is the convex envelope of card on ℓ_2 unit ball, so:

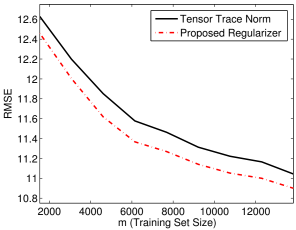
$$\Omega_1(\mathcal{W}) \geq \|\mathcal{W}\|_{\mathrm{tr}}, \quad \forall\ \mathcal{W} : \|\mathcal{W}\|_2 \leq 1$$

Experiments on tensor completion

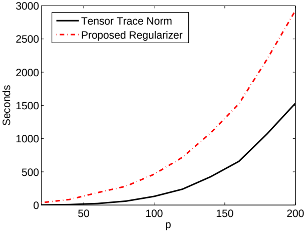
Video compression
($160 \times 112 \times 32 \times 3$ tensor)



Exam score prediction
($139 \times 11 \times 3 \times 3 \times 2$ tensor)



Time comparison:



Extensions and further work

- Latent tensor norm
- Tensor nuclear norm and an open problem
- Controlling the rank of all matricizations (quantum physics)
- Kernel methods
- Statistical analysis

Latent tensor nuclear norm

Defined by the variational problem [Tomioka and Suzuki 2013, Wimalawarne et al 2014]

$$\|\mathcal{W}\|_{\mathrm{LNN}} = \inf\left\{\sum_{j=1}^{m}\|\sigma(V_j)\|_1 \;\middle|\; \sum_{j=1}^{m} M_j^* V_j = \mathcal{W}\right\}.$$

where M_j^* is the adjoint of the matricization operator

The associated unit ball is

$$\mathrm{conv}\{\mathcal{W} \mid \mathrm{rank}(W^{(n)}) \leq 1, \|W^{(n)}\|_\infty \leq 1,\ \forall n\}$$

Latent tensor nuclear norm

If we change the above to

$$\mathrm{conv}\{\mathcal{W} \mid \mathrm{rank}(W^{(n)}) \leq k, \|W^{(n)}\|_p \leq 1\}$$

the induced norm becomes [Combettes et al 2016]

$$\|\mathcal{W}\|_{\mathrm{LNN}} = \inf \left\{ \sum_{j=1}^{m} \|\sigma(V_j)\|_{p,k} \;\Big|\; \sum_{j=1}^{m} M_j^* V_j = \mathcal{W} \right\}$$

where $\|\cdot\|_{p,k}$ is the (p,k)-support norm [McDonald, Pontil, Stamos 2016]. Its unit ball is

$$\mathrm{conv}\{x \in \mathbb{R}^d \mid \mathrm{card}(x) \leq k,\ \|x\|_p \leq 1\}$$

Ongoing experiments...

Tensor nuclear norm

Defined by

$$\|\mathcal{W}\|_{TNN,p} = \inf\left\{\|\lambda\|_p \mid \sum_{k=1}^{K} \lambda_k u_k^{(1)} \otimes \cdots \otimes u_k^{(N)} = \mathcal{W}\right\}$$

where the infimum is over $K \in \mathbb{N}$, $\lambda = (\lambda_1, \ldots, \lambda_K) \in \mathbb{R}^K$ and vectors $u_k^{(n)} \in \mathbb{R}^{d_n}$ such that $\|u_r^{(n)}\| = 1$, $\forall n, k$

Originally proposed and claimed to be a norm in [Lim & Comon, 2010], however [Friedland & Lim, 2016] shows this is well defined and a norm only when $p = 1$, but it is always zero if $p > 0$ disproving the claim in the former paper

Tensor nuclear norm

[Friedland & Lim, 2016] also shows that computing $\|\mathcal{W}\|_{TNN}$ is NP-hard.

Assume for simplicity $d_1 = \cdots = d_N = d$

Claim 1. We may restrict w.l.o.g. $K \leq d^{N-1}$

Claim 2. We can rewrite the norm as

$$\|\mathcal{W}\|_{TNN} = \inf \left\{ \sum_{k=1}^{K} \prod_{n=1}^{N} \|v_k^{(n)}\| : \sum_{k=1}^{K} \lambda_k v_k^{(1)} \otimes \cdots \otimes v_k^{(N)} = \mathcal{W} \right\}$$

where now the $v_k^{(n)} \in \mathbb{R}^d$ are not constrained to have unit norm

Tensor nuclear norm

Claim 3. Let

$$\varphi(\mathcal{W}) = E(\mathcal{W}) + \gamma\|\mathcal{W}\|_{TNN}$$

and

$$h(V^{(1)}, \ldots, V^{(N)}) = \varphi(V^{(1)} \otimes \cdots \otimes V^{(N)})$$

If $(V^{(1)}, \ldots, V^{(N)})$ is a local minima of h then

$$\mathcal{W} = V^{(1)} \otimes \cdots \otimes V^{(N)}$$

is a local minima of φ
True for $N = 2$ (easy to show)

Controlling the rank of all matricizations

Let $\mathbf{c} \subset \{1, \ldots, \}$ and $W^{(\mathbf{i},\bar{\mathbf{i}})}$ be the matricization obtained by taking the modes in $\mathbf{c}$ as rows and those in $\bar{\mathbf{c}}$ as columns

Fact: if

$$\max_{\mathbf{c} \subset \{1,\ldots,N\}} \operatorname{rank}(W^{(\mathbf{c},\bar{\mathbf{c}})}) \leq k$$

then $\mathcal{W}$ has a compact decomposition called "matrix-product-state" in quantum physics [Vidal 2003]"

$$W_{i_1,\ldots,i_n} = \operatorname{trace}(A^1_{i_1} \cdots A^N_{i_n})$$

where A^n_i are $k \times k$ matrices, so we can describe the tensor by $O(Nk^2)$ parameters. However, with this method we do not have control of $\max_n \operatorname{rank}(W(n))$

If $N \leq 6$ we may still obtain a latent norm convex relaxation

THANK YOU!