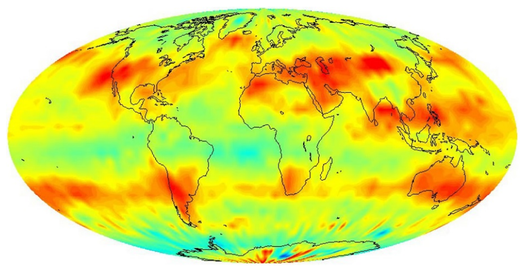
Fast Cokriging via Low Rank Tensor Learning

Mohammad Taha Bahadori†, Rose Yu†, and Yan Liu
University of Southern California
{mohammab,qiyu,yanliu.cs}@usc.edu

1. Cokriging

Definition Cokriging is the task of interpolating the data of certain variables for unknown locations by taking advantage of the observations of variables from known locations [2, 3].

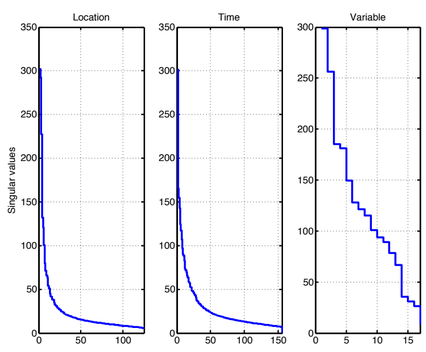


Settings Formally, denote the complete data for P locations over T time stamps with M variables as $\mathcal{X} \in \mathbb{R}^{P\times T\times M}$.
We only observe the measurements for a subset of locations $\Omega \subset \{1,\dots,P\}$ and the side information such as longitude and latitude.

Goal Given the measurements $\mathcal{X}_\Omega$ and the side information, the goal is to estimate a tensor $\mathcal{W} \in \mathbb{R}^{P\times T\times M}$ that satisfies $\mathcal{W}_\Omega = \mathcal{X}_\Omega$ for known locations and provides a good estimate for the other locations.

4. Tensor Rank

Evidence for low-rankness



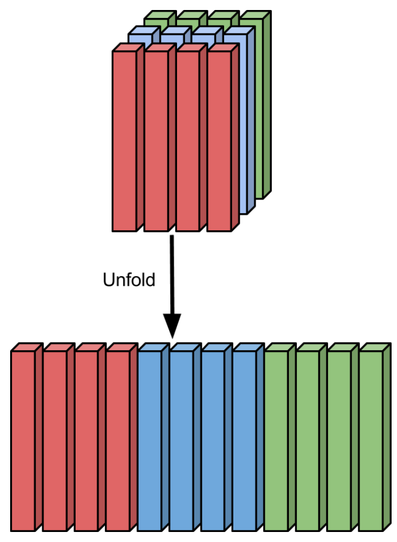
There are several definitions for the tensor rank [4, 5]:

CP Rank The straightforward generalization of the matrix rank: $\mathcal{W} = \sum_{i=1}^{r} \sigma_i \otimes_{n=1}^{N} \mathbf{u}_{i,n}$.

Tucker Rank A rank based on decomposition of a tensor to a core tensor and a matrix for each mode.

Mode-n Rank Sum of the rank of its mode-n unfolding $\mathcal{W}_{(n)}$.

Tensor unfolding



We choose mode-n rank because of its computational advantages.
Mode-n rank for a tensor $\mathcal{W}$ with N mode is defined as:

$$\text{mode-n rank}(\mathcal{W}) = \sum_{n=1}^{N} \text{rank}(\mathcal{W}_{(n)}).$$

2. Challenges

- Efficient and accurate predictions for large scale climate data.
- Incorporating complex correlations among variables and locations.
- We need to avoid complex parametric covariance models which are both slow to fit and restrictive.

Our Approach:

- Low-rank tensor formulation to capture correlations.
- A fast greedy low-rank tensor learning algorithm with theoretical guarantees.

5. Greedy Low-rank Learning

The main idea of the algorithm is to (1) unfold the tensor into a matrix, (2) seek for its rank-1 approximation, and (3) then fold back into a tensor with same dimensionality.

The key observation is that rank-1 matrix learning step can be performed efficiently as follows:

Lemma 1 *Consider the following rank-1 estimation problem*

$$\widehat{A}_1 = \operatorname*{argmin}_{A,\text{rank}(A)=1} \{\|Y - AX\|_F^2\}$$

where $Y \in \mathbb{R}^{q\times n}$ and $X \in \mathbb{R}^{p\times n}$. The optimal solution $\widehat{A}_1$ can be written as $\widehat{A}_1 = \widehat{\mathbf{u}}\widehat{\mathbf{v}}^\top$, $\|\widehat{\mathbf{v}}\|_2 = 1$ where $\widehat{\mathbf{v}}$ is the dominant eigenvector of the following generalized eigenvalue problem:

$$(XY^\top YX^\top)\mathbf{v} = \lambda(XX^\top)\mathbf{v}$$

and $\widehat{\mathbf{u}}$ can be computed as

$$\widehat{\mathbf{u}} = \frac{1}{\widehat{\mathbf{v}}^\top XX^\top \widehat{\mathbf{v}}} YX^\top \widehat{\mathbf{v}}.$$

The Greedy Algorithm

Input: Data tensor X_Ω, Laplacian L, and stopping error η
Output: N mode tensor $\mathcal{W}$
Initialize $\mathcal{W} \leftarrow 0$
repeat
  for $n = 1$ **to** N **do**
    $B_n \leftarrow \text{argmin}_{B:\text{rank}(B)=1} \mathcal{L}(\mathcal{W}_{(n)} + B; X_\Omega, L)$
    $\Delta_n \leftarrow \mathcal{L}(\mathcal{W}_{(n)}; \mathcal{Y}, \mathcal{Z}) - \mathcal{L}(\mathcal{W}_{(n)} + B_n; X_\Omega, L)$
  end for
  $n^* \leftarrow \text{argmax}_n\{\Delta_n\}$
  if $\Delta_{n^*} > \eta$ **then**
    $\mathcal{W} \leftarrow \mathcal{W} + \text{fold}(B_{n^*}, n^*)$
  end if
  $\mathcal{W} \leftarrow \operatorname*{argmin}_{\substack{\text{row}(\mathcal{A}_{(1)})\subseteq\text{row}(\mathcal{W}_{(1)}) \\ \text{col}(\mathcal{A}_{(1)})\subseteq\text{col}(\mathcal{W}_{(1)})}} \mathcal{L}(\mathcal{A}; X_\Omega, L)$
until $\Delta_{n^*} < \eta$

We also provide a bound on the difference between the loss function at the greedy algorithm solution and the globally optimal solution.

Theorem 1 *The solution of the algorithm at its kth iteration step satisfies the following inequality:*

$$\mathcal{L}(\mathcal{W}_k; \mathcal{X}_\Omega, L) - \mathcal{L}(\mathcal{W}^*; \mathcal{X}_\Omega, L) \le \frac{C^2}{k+1},$$

where C is a constant that depends on $\mathcal{X}_\Omega$, L, and $\mathcal{W}^*$, the global minimizer of the problem in Eq. (1).

3. Our Formulation

Existing work have identified two key consistency principles [2]:

Global consistency the data in the common structure are likely to be similar.

Local consistency the data in close neighborhood locations are likely to be similar.

We propose to perform cokriging by finding the value of tensor $\mathcal{W}$ in the following optimization problem:

$$\min_{\mathcal{W}} \left\{ \|\mathcal{W}_\Omega - \mathcal{X}_\Omega\|_F^2 + \mu \sum_{m=1}^{M} \text{tr}(\mathcal{W}_{:,:,m}^\top L \mathcal{W}_{:,:,m}) \right\} \quad (1)$$
$$\text{subject to} \quad \text{rank}(\mathcal{W}) \le \rho,$$

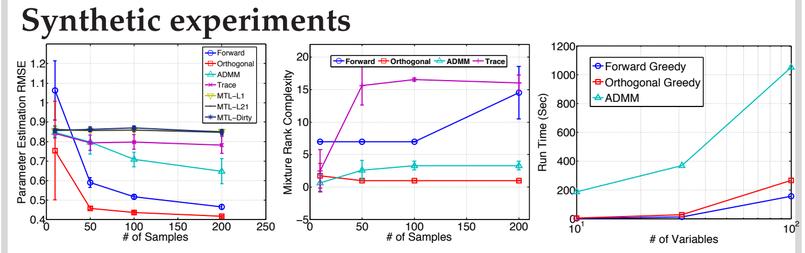
where L is the Laplacian matrix constructed from the location information and $\mu, \rho > 0$ are the local and global consistency tradeoff parameters.
— Two main elements in our formulation:

Low-rank constraint enforces the global consistency principle. It finds the principal components of the tensor and reduces the complexity of the model.

Laplacian regularizer enforces the local consistency principle. It has a smoothing effect on the data using the relational information among the locations. The Laplacian matrix is defined as $L = D - A$, where A is a kernel matrix constructed by pairwise similarity and diagonal matrix $D_{i,i} = \sum_j (A_{i,j})$.

6. Experiments

Synthetic experiments



RMSE | Complexity | Run-time

Climate Data Experiments We use two datasets:

USHCN The U.S. Historical Climatology Network Monthly dataset consists of monthly climatological data of three variables in 108 stations spanning from year 1915 to 2000.

CCDS The Comprehensive Climate Dataset (CCDS) is a collection of climate records of North America. It contains monthly observations of 17 variables. The observations were interpolated on a 2.5×2.5 degree grid with 125 observation locations.

Cokriging accuracy We present the cokriging RMSE of 6 methods averaged over 10 runs. We randomly pick 10% of locations and eliminate the measurements of all variables over the whole time span for each data set. Then, we produce the estimates for all variables of each time stamp. We repeat the procedure ten times and report the average prediction RMSE for all time stamps over 10 random runs.

Dataset	ADMM	Greedy	Simple	Ordinary	MTGP
USHCN	0.8051	**0.7210**	0.8760	0.7803	1.0007
CCDS	0.8292	**0.4532**	0.7634	0.7312	1.0296

Run time comparison with baselines (in seconds)

Dataset	USHCN	CCDS
ORTHO	**75.47**	**21.38**
ADMM	235.73	45.62

References

1 M. T. Bahadori†, R. Yu†, and Y. Liu, "Fast Multivariate Spatio-temporal Analysis via Low Rank Tensor Learning," Accepted as spotlight in NIPS 2014. († Equal contributions)
2 Cressie, Noel, and Christopher K. Wikle. Statistics for spatio-temporal data. John Wiley & Sons, 2011.
3 N. Cressie and G. Johannesson. Fixed rank kriging for very large spatial data sets. JRSS B (Statistical Methodology), 70(1):209âĂŞ226, 2008.
4 T. Kolda and B. Bader, “Tensor decompositions and applications,” SIAM review, 2009.
5 S. Gandy, B. Recht, and I. Yamada, “Tensor completion and low-n-rank tensor recovery via convex optimization,” Inverse Problems, 2011.