Automatic data cleaning via tensor factorization for large urban environmental sensor networks

Anonymous Author(s)
Affiliation
Address
email

Abstract

The US Environmental Protection Agency identifies that urban heat islands can negatively impact a community's environment and quality of life. Using low cost urban sensing networks, it is possible to measure the impacts of mitigation strategies in communities at a fine-grained scale, informing context-aware policies and infrastructure design. However, fine-grained city-scale data analysis is complicated by common, tedious data cleaning tasks such as removing outliers and imputing missing data. To address the challenge of data cleaning, this article introduces a robust low-rank tensor factorization method to automatically correct anomalies and impute missing entries for high-dimensional urban environmental datasets. We validate the method on a synthetically degraded National Oceanic and Atmospheric Administration temperature dataset, with a recovery error of 4%, and apply it to the Array of Things city-scale sensor network in Chicago, IL.

1 Introduction

Urban heat islands impact human health and cause socioeconomic disturbances [1, 2]. It is estimated that more than 8,000 premature deaths were attributed to elevated summer temperatures and prolonged heat waves from 1979 to 1999 in the US [1]. Other issues such as excessive energy consumption also arise [3]. As a consequence, a number of mitigation strategies have been proposed [1, 4, 5, 6]. To quantify the effects of the built environment on micro climate and other environmental impacts, many urban-scale environmental sensing initiatives are being developed [7, 8, 9, 10, 11]. These projects measure block-by-block micro-climate quantities to inform better green infrastructure investment, transportation planning and energy-saving designs. However, low-cost environmental sensors that facilitate dense instrumentation of urban communities are prone to errors, outliers, and missing data. Current approaches to clean the datasets prior to interpretation are often limited in functionality for which anomalies or missing data are independently addressed [12, 13, 14].

The main contribution of this work is to introduce a robust tensor factorization algorithm to automatically correct errors and impute missing data common to large distributed urban sensor networks (Section 2). We show that the proposed method is able to automatically correct outliers and impute missing data while preserving the normal variations of the dataset.

Two experiments (Section 3) demonstrate the approach. The first experiment begins with a complete (no missing data) *National Oceanic and Atmospheric Administration* (NOAA) temperature dataset [15], which is artificially degraded by injecting known outliers and also by removing some entries to simulate missing data. We demonstrate that the proposed tensor factorization approach correctly identifies the outliers and recovers accurate values for the missing data. The second experiment applies the method to the raw and incomplete temperature data from the *Array of Things* (AoT) urban sensing platform in Chicago, IL [16]. The recovered temperature data is validated by comparing to nearby NOAA readings when applicable, and the resulting clean data is available [17].

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

2 Tensor factorization

We briefly summarize our tensor factorization approach to remove outliers and impute missing data. We organize the raw data in a multi-dimensional array known as a tensor, which is a higher order generalization of a matrix. E.g., a third order tensor storing temperature data might arrange the timeseries data such that the first mode corresponds to each sensor, the second mode to each hour in a 24-hour period, and the third mode to each 24-hour period in the dataset.

Tensor factorization approaches [18, 19] exploit the fact that many large, noisy, and incomplete datasets actually have low intrinsic dimensionality. Assuming outliers appear sparsely in the raw data, we can reconstruct the underlying true complete data. Letting $\mathcal{B} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ denote the raw data tensor, our approach recovers a low dimensional tensor $\mathcal{X}$ (measured by the Tucker rank of the tensor [20]) containing the clean complete data, and a sparse (i.e., mostly zero entries) outlier tensor $\mathcal{E}$, such that $\mathcal{B} = \mathcal{X} + \mathcal{E}$ on the entries of $\mathcal{B}$ that are observed.

Recovering a low rank $\mathcal{X}$ and sparse outlier $\mathcal{E}$ from a corrupt $\mathcal{B}$ can be posed as a convex program:

$$\begin{aligned} \min_{\mathcal{X},\mathcal{E},\mathcal{O}} \quad & \sum_{i=1}^{N} \|\mathbf{X}_{(i)}\|_* + \lambda \|\mathbf{E}_{(2)}\|_{2,1} \\ \text{s.t.} \quad & \mathcal{B} = \mathcal{X} + \mathcal{E} + \mathcal{O}, \\ & \mathcal{O}_\Omega = 0. \end{aligned} \tag{1}$$

The objective function in problem (1) balances the tensor rank of $\mathcal{X}$, with the sparsity of the outliers $\mathcal{E}$ via λ, which is set according to [21, 22]. The term $\sum_i \|\mathbf{X}_{(i)}\|_*$ is a convex relaxation of the tensor rank of $\mathcal{X}$, where $\mathbf{X}_{(i)}$ is the mode-i matrix unfolding of $\mathcal{X}$, and $\|\cdot\|_*$ denotes the nuclear norm [19]. The $l_{2,1}$ norm $\|\cdot\|_{2,1}$ imposes a specific sparsity pattern on the outlier tensor $\mathcal{E}$, namely encouraging outliers to persist across one of the orders of the tensor [23]. For example, it can be used to model the observation that some sensors degrade and produce faulty data for extended periods of time. A compensation tensor $\mathcal{O}$, which is zero for entries in the observation set Ω, and free otherwise, is used to handle missing entries. Problem (1) is solved by singular value thresholding [24, 22] based on the *alternating direction method of multipliers* (ADMM) framework [17, 19].

3 Experiments

We briefly summarize two experiments in which we apply the proposed tensor factorization method to large temperature datasets. The first experiment is a complete NOAA [15] temperature dataset that we synthetically degrade, so that the recovery relative error can be computed. In the second experiment, we apply the method to Array of Things temperature data which contains missing data and outliers. We assess the quality of the recovery by comparing the correlation of AoT data with NOAA sensors when they are in close proximity.

Experiment 1. Synthetically degraded NOAA data. We apply tensor factorization on a complete NOAA dataset [15]. We use temperature data from April to September, 2018 recorded from stationary, high-end climate sensors located at 14 USCRN monitoring sites [25] in the US Midwest. The accessed 14 NOAA sensors record data hourly for 24 hours a day, for 183 days, which is arranged as $\mathcal{X} \in \mathbb{R}^{14 \times 24 \times 183}$. The raw NOAA data is used as the true temperature in this experiment, denoted $\mathcal{X}_{\text{true}}$, which is to be estimated from a degraded corrupted dataset $\mathcal{B}$.

To test the factorization method, we generate a synthetically corrupted dataset $\mathcal{B}$ from $\mathcal{X}_{\text{true}}$ that has missing data and erroneous values. The volume and structure of the outliers in $\mathcal{B}$ is inspired by the patterns observed in the AoT dataset used in Experiment 2 below. We degrade the data by randomly removing blocks of data ranging between 8-16 days, accounting for a total missing data rate of 15%. We randomly modify 2% of the entries to create outlier readings, also clustered in blocks of time.

Given $\mathcal{B}$, we solve problem (1) with $\lambda = 0.345$, where the decision variable $\mathcal{X}$ at optimality is the recovered dataset $\hat{\mathcal{X}}$. We compare the quality of the recovered dataset $\hat{\mathcal{X}}$ to the true dataset $\mathcal{X}_{\text{true}}$ by computing the *relative error*, RE $= \frac{\|\mathcal{X}_{\text{true}} - \hat{\mathcal{X}}\|_F}{\|\mathcal{X}_{\text{true}}\|_F}$, where $\|\cdot\|_F$ is the tensor Frobenius norm. The results are summarized in Table 1. The relative error of $\hat{\mathcal{X}}$ computed on all entries of the dataset is reduced from 15.84% in the corrupted dataset $\mathcal{B}$ to 3.85% in the recovered data. Restricted to

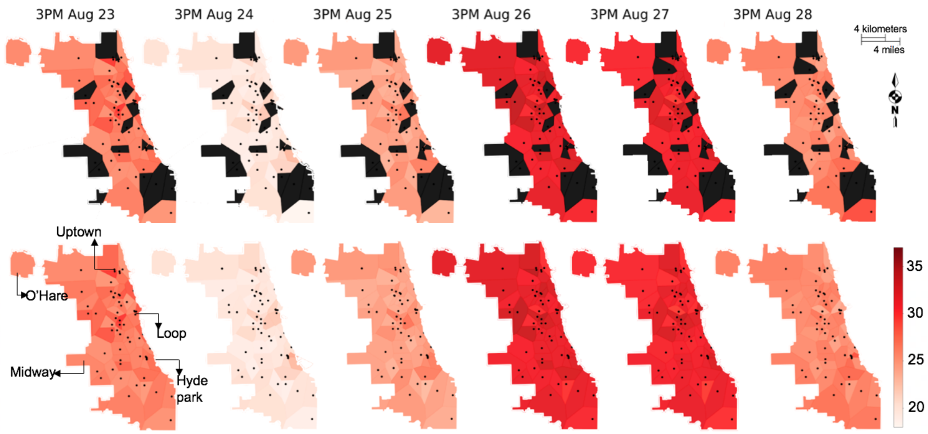

Figure 1: Voronoi heat maps (in °C) at 3PM from Aug. 23 to 28, 2018 produced by raw (top) and recovered (bottom) ambient air temperature data. Each dot marks an active AoT unit. Missing entries are highlighted in black (top).

Table 1: Performance summary. $\mathcal{B}$ denotes the raw corrupted data, and $\hat{\mathcal{X}}$ is the recovered data. Relative error (RE) reported for the NOAA experiment where $\mathcal{X}_{\text{true}}$ is known; the Pearson correlation coefficient r between AoT and NOAA when a nearby NOAA sensor is present.

	NOAA experiment				Array of Things experiment	
	$\text{RE}_{\text{uncorrupted}}$	$\text{RE}_{\text{outliers}}$	$\text{RE}_{\text{missing}}$	RE_{mean}	r_{present}	r_{missing}
$\mathcal{B}$	0	114%	-	15.84%	0.883	-
$\hat{\mathcal{X}}$	3.26%	6.87%	5.82%	3.85%	0.888	0.930

only the missing data entries, the relative error of $\hat{\mathcal{X}}$ is 5.82%, demonstrating the method is able to accurately impute missing data even when sensors report no data for long periods of time. Similarly, the relative error on the entries that are identified as outliers by the method (i.e., the nonzero entries of $\mathcal{E}$) has a low error of 6.87%, down from 114% on the same entries in the corrupt data tensor $\mathcal{B}$. We note that the zero relative error of $\mathcal{B}$ on the uncorrupted entries is an artifact of the fact that $\mathcal{B}$ was created directly from $\mathcal{X}_{\text{true}}$. The precision and recall of the outlier entries are both 1.

Experiment 2. Array of Things data. The method is next applied to Array of Things, a dense urban sensor network in Chicago [7] that collects real-time data on urban environment, infrastructure, and activity for research and public use. We construct an AoT temperature tensor as $\mathcal{X} \in \mathbb{R}^{345\times 24\times 183}$, representing 345 temperature sensors aggregated hourly, for 24 hours a day and for 183 days, matching the period of the NOAA data. Problem (1) is solved with $\lambda = 0.345$. Approximately 15% of the AoT data is missing in this period, and the outlier rate identified by our algorithm is about 1%.

Due to the lack of a ground truth dataset in this experiment, each AoT sensor is quantitatively compared to its closest NOAA sensor. Because the temperature field is spatially varying, we use the Pearson correlation coefficient to quantify the agreement between recovered AoT temperature readings and the nearest NOAA sensor. Figure 1 shows temperature variation in Chicago near a hot period on Aug 26-27 in the raw and recovered dataset. The quantitative results (Table 1), show that there is generally a high correlation on the AoT data that is present (r_{present}) before and after recovery. The correlation coefficient of the imputed missing data is $r_{\text{missing}} = 0.930$, indicating the method successfully imputes the missing temperature data.

Discussion & Conclusion. We proposed a method to automatically clean environmental data on two temperature datasets using tensor factorization. Our next steps are to create improved validation datasets for AoT to more rigorously quantify the quality of the recovery. We are also interested to extend the approach to accommodate other environmental sensors co-located on the AoT platform. Ultimately the cleaned data will assist its use by city planners and urban scientists interested in neighborhood-specific heat mitigation strategies to reduce adverse impacts [26].

References

[1] U.S. Environmental Protection Agency. Reducing urban heat islands: Compendium of strategies. `https://www.epa.gov/heat-islands/heat-island-compendium`.

[2] A. Crimmins, J. Balbus, J. L. Gamble, C. B. Beard, J. E. Bell, D. Dodgen, R. J. Eisen, N. Fann, M. D. Hawkins, S. C. Herring, L. Jantarasami, D. M. Mills, S. Saha, M. C. Sarofim, J. Trtanj, L. Ziska, and USGCRP. *The Impacts of Climate Change on Human Health in the United States: A Scientific Assessment*. U.S. Global Change Research Program, Washington, DC, 2016.

[3] M. Santamouris. On the energy impact of urban heat island and global warming on buildings. *Energy and Buildings*, 82:100 – 113, 2014.

[4] C. Rosenzweig, W. Solecki, L. Parshall, S. Gaffin, B. Lynn, R. Goldberg, J. Cox, and S. Hodges. Mitigating new york city's heat island with urban forestry, living roofs, and light surfaces. *86th AMS Annual Meeting*, 01 2006.

[5] D. Li, E. Bou-Zeid, and M. Oppenheimer. The effectiveness of cool and green roofs as urban heat island mitigation strategies. *Environmental Research Letters*, 9(5):055002, may 2014.

[6] W. D. Solecki, C. Rosenzweig, L. Parshall, G. Pope, M. Clark, J. Cox, and M. Wiencke. Mitigation of the heat island effect in urban New Jersey. *Global Environmental Change Part B: Environmental Hazards*, 6(1):39–49, 2005.

[7] C. E. Catlett, P. H. Beckman, R. Sankaran, and K. Galvin. Array of things: A scientific research instrument in the public way: Platform design and early lessons learned. In *Proceedings of the 2Nd International Workshop on Science of Smart City Operations and Platforms Engineering*, SCOPE '17, pages 26–33, New York, NY, USA, 2017. ACM.

[8] R. N. Murty, G. Mainland, I. Rose, A. R. Chowdhury, A. Gosain, J. Bers, and M. Welsh. Citysense: An urban-scale wireless sensor network and testbed. In *2008 IEEE Conference on Technologies for Homeland Security*, pages 583–588, May 2008.

[9] S. Eisenman, N. D. Lane, E. Miluzzo, R. A. Peterson, G. Ahn, and A. T. Campbell. Metrosense project: People-centric sensing at scale. *Proc. of Workshop on World-Sensor-Web (WSW)*, 01 2009.

[10] L. Sanchez, L. Muñoz, J. Antonio Galache, P. Sotres, J. R. Santana, V. Gutierrez, R. Ramdhany, A. Gluhak, S. Krco, E. Theodoridis, and D. Pfisterer. Smartsantander: IoT experimentation over a smart city testbed. *Computer Networks*, 61:217 – 238, 2014. Special issue on Future Internet Testbeds – Part I.

[11] K. Kloeckl, O. Senn, and C. Ratti. Enabling the real-time city: Live singapore! *Journal of Urban Technology*, 19(2):89–112, 2012.

[12] M. Daszykowski, K. Kaczmarek, Y. Vander Heyden, and B. Walczak. Robust statistics in data analysis — a review: Basic concepts. *Chemometrics and Intelligent Laboratory Systems*, 85(2):203 – 219, 2007.

[13] D. J. Hill and B. S. Minsker. Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. *Environmental Modelling & Software*, 25(9):1014 – 1022, 2010. Thematic Issue on Sensors and the Environment – Modelling & ICT challenges.

[14] M. Yu. Smirnov and G. D. Egbert. Robust principal component analysis of electromagnetic arrays with missing data. *Geophysical Journal International*, 190(3):1423–1438, 09 2012.

[15] National Centers for Environmental Information. Global summary of the year (GSOY), version 1. `https://www.ncei.noaa.gov/access/search/data-search/global-summary-of-the-year`.

[16] University of Chicago. Array of Things file browser. `https://afb.plenar.io/data-sets/chicago-complete`, 2019.

[17] A. Author(s). Data and source code for the article, "Automatic data cleaning via tensor factorization for large urban environmental sensor networks". Github URL removed during double blind review, 2019.

[18] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

[19] D. Goldfarb and Z. Qin. Robust low-rank tensor recovery: Models and algorithms. *SIAM Journal on Matrix Analysis and Applications*, 35(1):225–253, 2014.

[20] L. R Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[21] P. Zhou and J. Feng. Outlier-robust tensor PCA. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2263–2271, 2017.

[22] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[23] Y. Hu and D. B. Work. Robust tensor recovery with fiber outliers for traffic events. *arXiv preprint arXiv:1908.10198*, 2019.

[24] J. F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[25] What's a USCRN station? `https://www.ncei.noaa.gov/news/what-is-a-uscrn-station`. Accessed: 2019-08-29.

[26] M. P. Silva, A. Sharma, M. Budhathoki, R. Jain, and C. E. Catlett. Neighborhood scale heat mitigation strategies using array of things (AoT) data in Chicago. In *AGU Fall Meeting Abstracts*, 2018.