Recovery of Missing Data in Correlated Smart Grid Datasets

Cristian Genes, Iñaki Esnaola, Samir Perlaza, and Daniel Coca

c.genes@sheffield.ac.uk

June 3, 2019

Department of Automatic Control and Systems Engineering

The University Of Sheffield.

Overview

Introduction

- The integration of low carbon energy sources increases the performance requirements for the monitoring and control procedures
- Control strategies require timely and accurate data describing the state of the grid
- Challenges for the data acquisition system:
  - data injection attacks
  - **missing data** due to telemetry errors such as
    - sensor failures
    - unreliable communication
    - data storage issues
- It is vital to develop estimation procedures for the missing data using the **available observations**

Introduction

- Observations from different datasets:
  - different electrical magnitudes from the same network
  - data from other interdependent infrastructure systems
  - data from interdependent processes, e.g. weather forecast
- Joint recovery of multiple datasets is possible using
  - tensor extension of MC-based algorithms [Wang, Aggarwal, and Aeron, Oct. 2017]
  - collective MC framework [Gunasekar, Yamada, Yin, and Chang, Feb. 2015]
- *Can we use classical MC algorithms to jointly recover missing data from multiple datasets?*
- *When is it beneficial to include data from other sources in the missing data recovery process?*

System model

- Electricity distribution network with N low voltage feeders
- Each feeder includes a sensing unit that measures the electrical magnitudes of operational interest at predetermined time instants
- These measurements include power, intensity, **voltage** on phases A, B, and C, and support the operator in controlling, monitoring, and managing the network
- For a given phase voltage state variable, let $m_{i,j}^{(s)}$ be the corresponding value on phase $s \in \{A, B, C\}$, at feeder $i \in \{1, 2, ..., N\}$ and time $j \in \{1, 2, ..., M\}$
- The matrix with the measurements for phase s is denoted by $\mathbf{M}^{(s)} \in \mathbb{R}^{M \times N}$

System model

- Consider two datasets that are contained in the matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ respectively, with $\mathbf{M}_1, \mathbf{M}_2 \in \mathbb{R}^{M \times N}$
- Denote $\text{rank}(\mathbf{M}_1) = r_1$ and $\text{rank}(\mathbf{M}_2) = r_2$
- The combined matrix $\mathbf{M}$ is given by

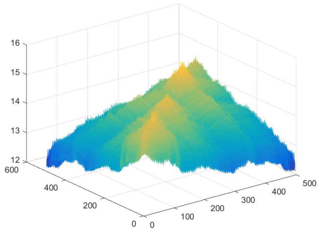
$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_1 \\ \mathbf{M}_2 \end{bmatrix}$$

- where $\mathbf{M} \in \mathbb{R}^{2M \times N}$ and $\text{rank}(\mathbf{M}) = r$

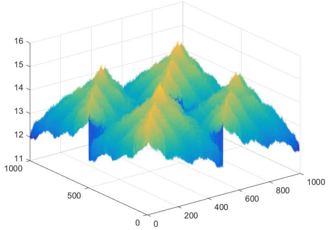
Real data model

- Real data collected from 200 residential secondary substations across North West of England from June 2013 to January 2014 as part of the "Low Voltage Network Solutions" project run by Electricity North West Limited (ENWL)
- Using two complete data matrices $\mathbf{M}^{(B)}$ and $\mathbf{M}^{(C)}$ with $M = N = 500$ that contain phase B and phase C voltage measurements from the grid
- Based on the properties of $\mathbf{M}^{(B)}$ and $\mathbf{M}^{(C)}$:
  - The voltage data is modelled as a multivariate Gaussian random process where the sample covariance matrix exhibits a structure that is approximately Toeplitz
  - The data matrix is approximately low rank

Real data model



Sample covariance matrix of the phase B voltage data matrix



Sample covariance matrix of the combined phase B and C voltage data matrices.

Synthetic data model

- The combined data matrix is given by

$$\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_N] \quad \text{where} \quad \mathbf{m}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$$

and the covariance matrix $\mathbf{\Sigma}$ is given by

$$\mathbf{\Sigma} \triangleq \begin{bmatrix} \mathbf{\Sigma}_{11} & \psi \mathbf{\Sigma}_{11} \\ \psi \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{22} \end{bmatrix}$$

where $\psi \in [0, 1]$ and $\mathbf{\Sigma}_{ll} = \text{Toeplitz}(1, \ldots, \upsilon_{ll})$ with $\upsilon_{ll} = \rho^{\frac{1}{\zeta_{ll}}(M-1)}$

- ζ_{ll} determines the **intra-correlation** of the matrix $\mathbf{M}_l$
- ψ determines the **cross-correlation** between $\mathbf{M}_1$ and $\mathbf{M}_2$

Acquisition

- Measurements are corrupted by additive white Gaussian noise (AWGN) such that
$$\mathbf{R}_l = \mathbf{M}_l + \mathbf{N}_l$$
where $l \in \{1,2\}$ denotes the number of datasets and
$$(\mathbf{N}_l)_{i,j} \sim \mathcal{N}(0, \sigma^2_{\mathbf{N}_l})$$
where $i \in \{1,2,...,M\}$ and $j \in \{1,2,...,N\}$
- The data acquisition process is modelled by the functions $P_{\Omega_l} : \mathbb{R}^{M\times N} \to \mathbb{R}^{M\times N}$ with $l \in \{1,2\}$ and
$$P_{\Omega_l}(\mathbf{R}_l) = \begin{cases} (\mathbf{R}_l)_{i,j}, & (i,j) \in \Omega_l, \\ 0, & \text{otherwise} \end{cases}$$

Estimation

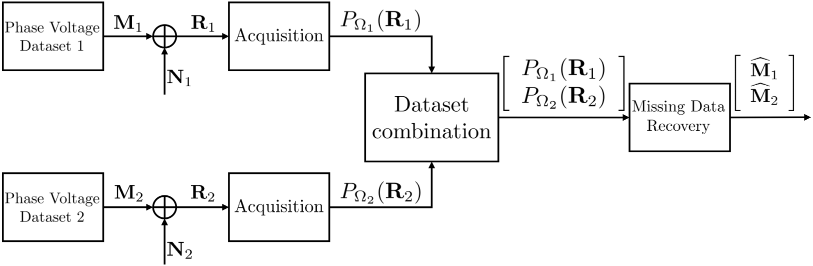
- The estimation process is modelled by the function $g : \mathbb{R}^{2M \times N} \to \mathbb{R}^{2M \times N}$
- The estimate is given by

$$\widehat{\mathbf{M}} = g\left(P_{\Omega_1}(\mathbf{R}_1), P_{\Omega_2}(\mathbf{R}_2)\right)$$

- The optimality criterion is the normalized mean square error (NMSE)

$$\mathrm{NMSE}\left(\mathbf{M}; g\right) = \frac{\mathbb{E}\left[\left\|\mathbf{M} - g\left(P_{\Omega_1}(\mathbf{R}_1), P_{\Omega_2}(\mathbf{R}_2)\right)\right\|_F^2\right]}{\|\mathbf{M}\|_F^2}$$

System model diagram



Block diagram describing the system model for the joint recovery of two datasets

Recovering missing data using matrix completion

When **M** is low rank or approximately low rank, the missing entries are recovered with high probability by solving the following optimization problem:

$$\begin{aligned} \underset{\mathbf{X}}{\text{minimize}} \quad & \|\mathbf{X}\|_* \\ \text{subject to} \quad & P_\Omega(\mathbf{X}) = P_\Omega(\mathbf{M}) \end{aligned}$$

- **Singular Value Theresholding** [Cai, Candès, and Shen, Mar. 2010]
  - \+ low computational cost
  - \- requires parameter tuning
- **Bayesian Singular Value Theresholding** [Genes, Esnaola, Perlaza, Ochoa, and Coca, May 2018]
  - \+ optimizes parameter at each iteration
  - \- requires prior knowledge (second order statistics)

Bayesian Singular Value Thresholding

Input: set of observations Ω, observed entries $P_\Omega(\mathbf{R})$, mean $\mathbf{0}$, covariance matrix $\boldsymbol{\Sigma}$, step size δ_b, tolerance ϵ, and maximum iteration count $k_{\max}$

Output: $\widehat{\mathbf{M}}_{\mathrm{BSVT}}$

Set $\mathbf{Y}^0 = \mathbf{0}$
Set $\mathbf{Z}^0 = \mathbf{0}$
Set $\tau = 0$
Set $\Omega^c = \{1, 2, \ldots, 2M\} \times \{1, 2, \ldots, N\} \setminus \Omega$
for $k = 1$ to $k_{\max}$ **do**
  Compute $[\mathbf{U}, \mathbf{S}, \mathbf{V}] = \mathrm{svd}(\mathbf{Z}^{(k-1)})$
  Set $\mathbf{X}^{(k)} = \sum_{j=1}^{N} \max(0, \sigma_j(\mathbf{Z}^{(k-1)}) - \tau^{(k-1)})\mathbf{u}_j\mathbf{v}_j$
  if $\|P_\Omega(\mathbf{X}^{(k)} - \mathbf{R})\|_F / \|P_\Omega(\mathbf{R})\|_F \leq \epsilon$ **then break**
  end if
  Set $\mathbf{Y}^{(k)} = \mathbf{Y}^{(k-1)} + \delta_b(P_\Omega(\mathbf{R}) - P_\Omega(\mathbf{X}^{(k)}))$
  Set $\mathbf{L}^{(k)} = \boldsymbol{\Sigma}_{\Omega^c\Omega}\boldsymbol{\Sigma}_{\Omega\Omega}^{-1}\mathbf{Y}^{(k)}$
  Set $\mathbf{Z}^{(k)} = \mathbf{Y}^{(k)} + \mathbf{L}^{(k)}$
  Set $\sigma^2_{\mathbf{Z}^{(k)}} = (\|\mathbf{Y}^{(k)} - P_\Omega(\mathbf{R})\|_F^2 + |\Omega^c| D_{\mathrm{LMMSE}})/2MN$
  Set $\tau^{(k)} = \arg\min_{\tau} \mathrm{SURE}(D_\tau)(\mathbf{Z}^{(k)})$
end for
Set $\widehat{\mathbf{M}}_{\mathrm{BSVT}} = \mathbf{X}^{(k)}$

Joint recovery of missing data in two datasets

The joint recovery exploits two types of correlation

- **intra-correlation**: correlation between the entries within each dataset
- **cross-correlation**: correlation between the data points from different datasets

In an MC setting, the minimum number of observations required depends on the size and the rank of the matrix

Lemma

The rank of the combined matrix is bounded by

$$\max(r_1, r_2) \leq r \leq r_1 + r_2$$

Joint recovery of missing data in two datasets

Riegler, E., Stotz, D., and Bölcskei, H. (Jun. 2015). Information-theoretic limits of matrix completion. In Proc. *of the 2015 IEEE International Symposium on Information Theory (ISIT), pages 1836–1840.*

The minimum number of entries required to recover $\mathbf{M}_1$ and $\mathbf{M}_2$

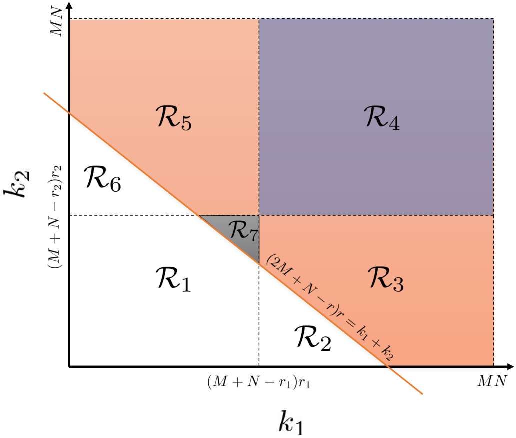
$$k_1 > (M + N - r_1)r_1$$

$$k_2 > (M + N - r_2)r_2$$

Applying the same result on the combined matrix $\mathbf{M}$ gives

$$k_1 + k_2 > (2M + N - r)r$$

Graphical interpretation of the fundamental limit



Joint recovery of missing data in two datasets

Theorem

Let $\mathbf{M}_1, \mathbf{M}_2 \in \mathbb{R}^{M \times N}$, with rank r_1 and r_2. Then, the joint recovery of the two matrices requires fewer observations than the independent recovery if

$$1 - \frac{\max(r_1, r_2)}{\min(r_1, r_2)} > \frac{\min(r_1, r_2) - N}{M},$$

and the rank of the combined matrix satisfies

$$r < M + \frac{1}{2}N - \frac{1}{2}(M + N - 2r_1 - 2r_2)\sqrt{1 + \frac{3M^2 + 2MN - 8r_1 r_2}{(M + N - 2r_1 - 2r_2)^2}}$$

Numerical results

- Joint recovery performance for $M = 50$, $N = 100$
- Simulations assume a signal to noise ratio value of SNR$= 50$ dB for both datasets, where the SNR for dataset l is given by

$$\mathrm{SNR}_l \triangleq 10\log_{10} \frac{\frac{1}{M}\mathrm{Tr}(\mathbf{\Sigma}_{ll})}{\sigma^2_{\mathbf{N}_l}}$$

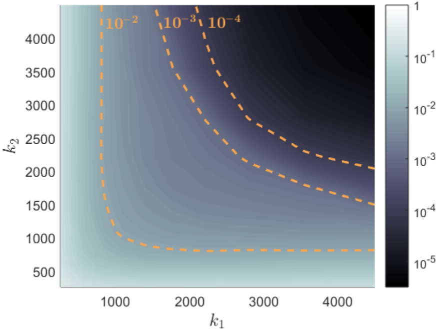
- Locations of the available entries are sampled uniformly at random in each dataset
- The recovery is performed using the SVT and BSVT algorithms

Simulation framework

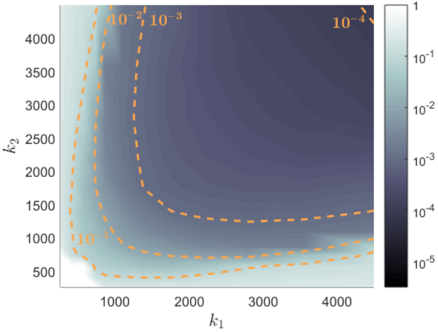
- The synthetic data model is used to generate correlated data matrices $\mathbf{M}$ such that the NMSE between the data matrix and the low rank approximation is below 10^{-3}
- The rank values of interest are:
  - $r_1 = 6,\ r_2 = 6,\ r = 9$
  - $r_1 = 6,\ r_2 = 9,\ r = 10$
- More rank and SNR scenarios in
  C. Genes, Novel Matrix Completion Methods for Missing Data Recovery in Urban Systems, Ph.D. thesis, University of Sheffield, 2018.

Numerical results for $r_1 = 6$, $r_2 = 6$, $r = 9$

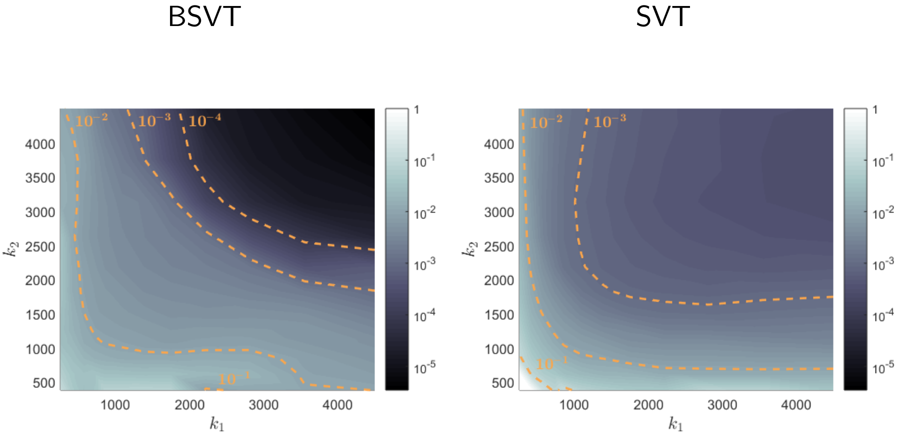
BSVT



SVT



Numerical results for $r_1 = 6$, $r_2 = 9$, $r = 10$



Numerical results

- The BSVT recovery performance follows the geometry dictated by the fundamental limit for the **joint** recovery of two correlated datasets
- This suggest that the BSVT algorithm is able to exploit the cross-correlation
- The SVT recovery performance follows the geometry dictated by the fundamental limit for the **independent** recovery of two correlated datasets
- This suggests that SVT is not efficient in exploiting the cross-correlation

Conclusions

- We study the fundamental limits for the joint recovery of two datasets in terms of the rank of the single and combined data matrices
- The joint recovery is feasible in more cases when compared to the independent recovery
- A model for generating correlated synthetic datasets has been proposed
- In comparison with SVT, the BSVT algorithm is better suited to exploit the correlation between different types of data in a missing data recovery setting

Thanks!

References

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, Mar. 2010.

C. Genes, I. Esnaola, S. M. Perlaza, L. F. Ochoa, and D. Coca. Robust recovery of missing data in electricity distribution systems. *IEEE Trans. Smart Grid*, May 2018.

S. Gunasekar, M. Yamada, D. Yin, and Y. Chang. Consistent collective matrix completion under joint low rank structure. In *Proc. International Conference on Artificial Intelligence and Statistics*, pages 306–314, Feb. 2015.

W. Wang, V. Aggarwal, and S. Aeron. Efficient low rank tensor ring completion. In *Proc. of the IEEE International Conference on Computer Vision*, pages 5697–5705, Oct. 2017.