White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/149034/

Version: Accepted Version

Proceedings Paper:
Genes, C., Esnaola, I., Perlaza, S. et al. (2019) Recovery of missing data in correlated smart grid datasets. In: 2019 IEEE Data Science Workshop (DSW). 2019 IEEE Data Science Workshop (DSW), 02-05 Jun 2019, Minneapolis, MN, USA. IEEE, pp. 263-269. ISBN: 978-1-7281-0708-0.

https://doi.org/10.1109/DSW.2019.8755592

© 2019 IEEE. This is an author-produced version of a paper subsequently published in 2019 IEEE Data Science Workshop (DSW). Uploaded in accordance with the publisher's self-archiving policy.

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

RECOVERY OF MISSING DATA IN CORRELATED SMART GRID DATASETS

Cristian Genes[1], *Iñaki Esnaola*[1,3], *Samir M. Perlaza*[2,3] *and Daniel Coca*[1].

[1]Department of Automatic Control and Systems Engineering, University of Sheffield, UK
[2]Institut National de Recherche en Informatique et en Automatique (INRIA), Lyon, France
[3]Department of Electrical Engineering, Princeton University, NJ, USA.

ABSTRACT

We study the recovery of missing data from multiple smart grid datasets within a matrix completion framework. The datasets contain the electrical magnitudes required for monitoring and control of the electricity distribution system. Each dataset is described by a low rank matrix. Different datasets are correlated as a result of containing measurements of different physical magnitudes generated by the same distribution system. To assess the validity of matrix completion techniques in the recovery of missing data, we characterize the fundamental limits when two correlated datasets are jointly recovered. We then proceed to evaluate the performance of Singular Value Thresholding (SVT) and Bayesian SVT (BSVT) in this setting. We show that BSVT outperforms SVT by simulating the recovery for different correlated datasets. The performance of BSVT displays the tradeoff behaviour described by the fundamental limit, which suggests that BSVT exploits the correlation between the datasets in an efficient manner.

Index Terms— smart grid, matrix completion, missing data recovery, correlated data

1. INTRODUCTION

The integration of residential low carbon energy sources such as solar or wind power generates bidirectional power flows that affect the stability of the smart grid [1]. The control strategies need to adapt to the new challenges posed by the additional distributed energy sources. In this context, the monitoring procedures are expected to manage the dynamic and unknown scenarios and to provide timely and accurate data describing the state of the grid. For example, the lack of data quality in power systems contributed towards several large-scale blackouts such as the 2003 U.S.-Canadian blackout [2] and the 2003 Italy blackout [3]. In addition, the integration of the Internet of things into the smart grid will significantly increase the number of datasets [4]. In practical scenarios, state estimation and monitoring systems face challenges like data injection attacks [5], [6], [7], [8], [9] or missing data [10], [11], [12]. Telemetry errors such as sensor failures or communication issues lead to incomplete sets of observations that do not fully describe the state of the grid. Therefore, it is vital to estimate the missing data based on the available observations. For instance, accurate measurements are necessary to implement centralized control schemes for voltage regulation in distribution systems [13].

Matrix completion (MC) is proposed in [14] as technique to recover missing data from partial observations. MC-based recovery exploits the fact that correlated state variable vectors give rise to approximately low rank data matrices. Specifically, in a convex optimization context, a low rank matrix is estimated given that a sufficient fraction of the entries is observed. See for instance [15] and [16]. However, when the number of observations is insufficient, the recovery of the data matrix is not possible. A potential way forward in this case is to attempt a joint recovery of multiple datasets by exploiting the fact that when datasets are correlated the rank of the resulting joint dataset grows in a sub-additive fashion. When the number of observations in one dataset is limited, this approach allows the estimation process to incorporate datasets produced by other sources in the system.

A framework for jointly recovering multiple datasets is provided in [17] where the MC setting is extended to the tensor case. Moreover, the singular value decomposition is extended to the tensor case in [18] which leads to the development of a tensor nuclear norm based algorithm in [19]. Alternatively, a collective MC framework is proposed in [20] to exploit the correlation between matrices with shared structure. However, the common structure constraint does not allow for sufficient generality in the definition of the correlation structure between datasets in a smart grid context.

This paper proposes an estimation setting in which data from multiple datasets is combined into a single data matrix that is recovered using MC-based algorithms. This allows the recovery process to exploit not only correlations within a dataset but also between datasets in the joint estimation paradigm. Specifically, the correlation between datasets is leveraged to facilitate the recovery when the number of observations in one dataset is limited. In addition, the fundamental limit of the joint recovery setting for two correlated datasets is characterized within an MC framework and, based on the geometry dictated by the fundamental limit, the joint recovery performance of two MC-based algorithms is benchmarked for different levels of correlation between the combined datasets. Numerical results show that the recently proposed Bayesian Singular Value Theresholding (BSVT) algorithm [12] is more effective in exploiting the correlation between datasets when compared to the Singular Value Theresholding (SVT) algorithm [21].

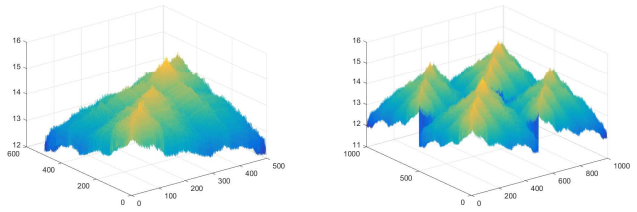
2. SYSTEM MODEL

Consider an electricity distribution system with N low voltage (LV) feeders. At the head of each feeder, a sensing unit measures various electrical magnitudes, e.g., voltage, intensity, active and reactive power at given time instants. These measures comprise the state variables that the operator uses for control, monitoring, and management purposes. The set of observations available to the operator is incomplete and corrupted by noise. The operator estimates the missing data based on the available observations. In the following, the analysis is carried out for a particular electrical magnitude, i.e., phase voltage.

2.1. Source Model

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, denoted by $\mathbf{M}^{(s)} \in \mathbb{R}^{M \times N}$, contains the aggregated measurement vectors from all feeders $\mathbf{M}^{(s)} \triangleq [\mathbf{m}_1^{(s)}, \mathbf{m}_2^{(s)}, ..., \mathbf{m}_N^{(s)}]$, with the measurement vectors given by $\mathbf{m}_i^{(s)} \triangleq [m_{i,1}^{(s)}, m_{i,2}^{(s)}, ..., m_{i,M}^{(s)}]^\mathsf{T} \in \mathbb{R}^M$. The resulting data matrices $\mathbf{M}^{(\mathrm{A})}$, $\mathbf{M}^{(\mathrm{B})}$, $\mathbf{M}^{(\mathrm{C})}$ contain the voltage measurements on phase A, B and C respectively, at time instants $1, 2, ..., M$ for all N feeders.

Email addresses: c.genes@sheffield.ac.uk, esnaola@sheffield.ac.uk, samir.perlaza@inria.fr, and d.coca@sheffield.ac.uk



(a) Sample covariance matrix of the phase B voltage data matrix.

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

Fig. 1: Sample covariance matrices obtained using the real data provided by ENWL.

2.2. Real data model

Real data collected as part of the "Low Voltage Network Solutions" project run by Electricity North West Limited (ENWL) [22], is used in the following to model the statistical structure of the random process governing the phase voltage state variables. The dataset contains voltage measurements of phases A, B and C collected from 200 residential secondary substations across North West of England from June 2013 to January 2014. Each substation generates a daily file that contains the voltage measurements on all three phases.

An analysis of the distribution and sample covariance matrix of the phase A voltage measurements in the LV dataset under consideration is presented in [11]. Therein, it is shown that voltage measurements can be modelled as a multivariate Gaussian random process for $i \in \{1, 2, ..., N\}$. Specifically, we model the voltage measurements as $\mathbf{m}_i^{(s)} \sim \mathcal{N}(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$, and $\{\mathbf{m}_i\}$ for $i \in \{1, 2, ..., N\}$, is a sequence of independent and identically distributed random variables. Moreover, it is also shown in [11] that the sample covariance matrix for phase A exhibits a structure that is approximately Toeplitz. In addition, because the voltage data is correlated, the covariance matrix displays a high correlation across feeders and time instants. It is shown in [12] that the singular value decomposition of a 500×500 matrix with phase A voltage measurements has a large condition number [23].

Part of the LV data collected by ENWL is used to construct two complete data matrices $\mathbf{M}^{(\mathrm{B})}$ and $\mathbf{M}^{(\mathrm{C})}$ with $M = N = 500$ that contain phase B and phase C voltage measurements from the LV grid. The sample covariance matrix of the data matrix $\mathbf{M}^{(\mathrm{B})}$ is depicted in Fig. 1a. As expected, and in agreement with the observation in [11], the sample covariance matrix for the phase voltage data exhibits a structure that is approximately Toeplitz. In addition, when the phase B and phase C data matrices are combined into a single data matrix, i.e., $\mathbf{M}^{(\mathrm{BC})} = [\mathbf{M}^{(\mathrm{B})}, \mathbf{M}^{(\mathrm{C})}]^{\mathsf{T}}$, the resulting sample covariance matrix is depicted in Fig. 1b. Interestingly, the sample covariance for the combined matrix is a block matrix with four elements where each element exhibits a structure that is approximately Toeplitz. Based on this observation, the following section proposes a general model for correlated voltage datasets generated by different phases in smart grid systems.

2.3. Synthetic data model

A mathematical description of the model used to generate two correlated synthetic datasets follows. Let us denote the data matrix for the first dataset by $\mathbf{M}_1 \in \mathbb{R}^{M \times N}$ and the data matrix for the second dataset by $\mathbf{M}_2 \in \mathbb{R}^{M \times N}$. In this setting, the combined matrix is denoted by $\mathbf{M} \in \mathbb{R}^{2M \times N}$ given by

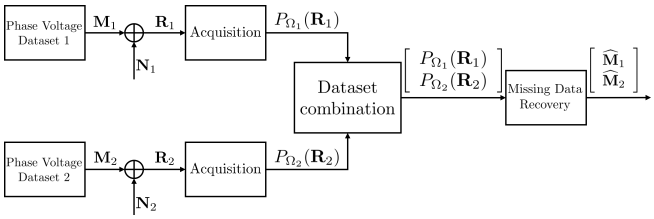


Fig. 2: Block diagram describing the system model for the joint recovery of two datasets.

$$\mathbf{M} \triangleq \begin{bmatrix} \mathbf{M}_1 \\ \mathbf{M}_2 \end{bmatrix}. \tag{1}$$

Hence, the combined state variable matrix is defined as $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_N]$, where each state variable vector $\mathbf{m}_i \in \mathbb{R}^{2M}$ for $i \in \{1, 2, ..., N\}$ is generated by a multivariate Gaussian process with $\mathbf{0}$ mean and covariance matrix $\boldsymbol{\Sigma}$, i.e., $\mathbf{m}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. The covariance matrix $\boldsymbol{\Sigma}$ is a block matrix in which block $\boldsymbol{\Sigma}_{ll}$ is a Toeplitz matrix describing the covariance matrix of the dataset $l \in \{1, 2\}$. The resulting covariance matrix is given by

$$\boldsymbol{\Sigma} \triangleq \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \psi\boldsymbol{\Sigma}_{11} \\ \psi\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{22} \end{bmatrix}, \tag{2}$$

where $\boldsymbol{\Sigma}_{ll} \in \mathbb{R}^{M \times M}$ and $\psi \in [0, 1]$. In this framework, the elements of $\boldsymbol{\Sigma}_{ll}$ are defined as $(\boldsymbol{\Sigma}_{ll})_{i,j} \triangleq \rho^{\frac{1}{\zeta_{ll}}|i-j|}$, where $(\boldsymbol{\Sigma}_{ll})_{i,j}$ denotes the entry in row i and column j of the matrix $\boldsymbol{\Sigma}_{ll}$ with $i \in \{1, 2, \ldots, M\}$, $j \in \{1, 2, \ldots, M\}$, $\rho \in (0, 1)$ and ζ_{ll} a design parameter. Hence, the matrix $\boldsymbol{\Sigma}_{ll}$ is given by

$$\boldsymbol{\Sigma}_{ll} = \text{Toeplitz}(1, \ldots, \upsilon_{ll}), \tag{3}$$

where $\upsilon_{ll} \in [0, 1)$ obeys

$$\upsilon_{ll} = \rho^{\frac{1}{\zeta_{ll}}(M-1)}. \tag{4}$$

Fig. 2 describes the system model for the joint recovery of two datasets produced in an LV distribution system. In this setting, each phase voltage data matrix fully describes the state of the grid over M time instants and across N feeders. However, in the acquisition process, part of the measurements are lost and the ones that are available are corrupted by noise. The missing data recovery strategy needs to estimate the actual state of the grid for a noisy subset of observations.

2.4. Acquisition

The phase voltage measurements are assumed to be corrupted by additive white Gaussian noise (AWGN) such that for each dataset the resulting observations are given by $\mathbf{R}_l = \mathbf{M}_l + \mathbf{N}_l$, where $l \in \{1, 2\}$ denotes the number of datasets and

$$(\mathbf{N}_l)_{i,j} \sim \mathcal{N}(0, \sigma^2_{\mathbf{N}_l}), \tag{5}$$

where $i \in \{1, 2, ..., M\}$ and $j \in \{1, 2, ..., N\}$. Moreover, it is also assumed that only a fraction of the complete set of observations (entries in $\mathbf{R}_l$) are communicated to the operator. Denote by Ω_l the subset of observed entries of the dataset l, i.e., $\Omega_l \triangleq \{(i, j) : (\mathbf{R}_l)_{i,j}$ is observed$\}$. Formally, the 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

$$P_{\Omega_l}(\mathbf{R}_l) = \begin{cases} (\mathbf{R}_l)_{i,j}, & (i, j) \in \Omega_l, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The observations given by (6) describe all the data from dataset l that is available to the operator for estimation purposes. Therefore, the recovery of the missing data is performed from the observations $P_{\Omega_l}(\mathbf{R}_l)$. As depicted in Fig. 2, the acquisition step is performed independently for each dataset. After the acquisition step, the available observations from each dataset are combined into a single data matrix, i.e.,

$$P_{\Omega}(\mathbf{R}) = \begin{bmatrix} P_{\Omega_1}(\mathbf{R}_1) \\ P_{\Omega_2}(\mathbf{R}_2) \end{bmatrix} \in \mathbb{R}^{2M \times N}, \tag{7}$$

where Ω denotes the combined set of available observations from the two datasets. The resulting matrix $P_{\Omega}(\mathbf{R})$ is used for estimation purposes in the joint recovery paradigm.

2.5. Estimation

The estimation process for the combined matrix of measurements, based on the available observations from each dataset is modelled by the function $g : \mathbb{R}^{2M \times N} \to \mathbb{R}^{2M \times N}$, where Ω denotes the combined set of available observations from both datasets. The estimate $\widehat{\mathbf{M}} = g\big(P_{\Omega_1}(\mathbf{R}_1), P_{\Omega_2}(\mathbf{R}_2)\big)$ is obtained by solving an optimization problem based on a given optimality criterion. In the following, the optimality criterion is the normalized mean square error (NMSE) given by

$$\text{NMSE}\,(\mathbf{M}; g) = \frac{\mathbb{E}\left[\|\mathbf{M} - g\big(P_{\Omega_1}(\mathbf{R}_1), P_{\Omega_2}(\mathbf{R}_2)\big)\|_F^2\right]}{\|\mathbf{M}\|_F^2}, \tag{8}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

3. RECOVERING MISSING DATA USING MATRIX COMPLETION

Given a matrix $\mathbf{M}$ of size $2M \times N$, and observations $P_{\Omega}(\mathbf{M})$, the recovery of the missing entries is not feasible in the general case. However, when $\mathbf{M}$ is low rank or approximately low rank, it is shown in [14] that if the entries on Ω are sampled uniformly at random, the missing entries are recovered with high probability by solving the following optimization problem:

$$\begin{aligned} \underset{\mathbf{X}}{\text{minimize}} \quad & \|\mathbf{X}\|_* \\ \text{subject to} \quad & P_{\Omega}(\mathbf{X}) = P_{\Omega}(\mathbf{M}), \end{aligned} \tag{9}$$

where $\|\mathbf{X}\|_*$ denotes the nuclear norm of the matrix $\mathbf{X}$. To simplify the notation, let us assume that $2M \geq N$. We proceed to present the two MC-based algorithms used to assess the joint recovery performance. Namely, the SVT algorithm proposed in [21] and the BSVT approach presented in [12].

3.1. Singular Value Theresholding

SVT is an MC-based algorithm [21] which produces a sequence of matrices $\mathbf{X}^{(k)}$ that converges to the unique solution of the following optimization problem:

$$\begin{aligned} \underset{\mathbf{X}}{\text{minimize}} \quad & \tau\|\mathbf{X}\|_* + \frac{1}{2}\|\mathbf{X}\|_F^2 \\ \text{subject to} \quad & P_{\Omega}(\mathbf{X}) = P_{\Omega}(\mathbf{M}). \end{aligned} \tag{10}$$

Note that when $\tau \to \infty$, the optimization problem in (10) converges to the nuclear norm minimization problem in (9). The iterations of the SVT algorithm are:

$$\begin{cases} \mathbf{X}^{(k)} = D_{\tau}(\mathbf{Y}^{(k-1)}), \\ \mathbf{Y}^{(k)} = \mathbf{Y}^{(k-1)} + \delta_s\big(P_{\Omega}(\mathbf{M}) - P_{\Omega}(\mathbf{X}^{(k)})\big), \end{cases} \tag{11}$$

where $\mathbf{Y}^{(0)} = \mathbf{0}$ is used for initialization, δ_s is the step size that obeys $0 < \delta_s < 2$, and the soft-thresholding operator, D_{τ} that shrinks the singular values of $\mathbf{Y}^{(k-1)}$ towards zero [21].

Interestingly, the choice of τ is important to guarantee a successful recovery, since large values guarantee a low-rank matrix estimate but for values larger than $\max_i(\sigma_i(\mathbf{Y}))$ all the singular values vanish. In [21], the proposed threshold is $\tau = 5N$. However, simulation results presented in [11] show that $\tau = 5N$ gives suboptimal performance when the number of missing entries is large. The main shortcoming of the SVT algorithm is the lack of guidelines for tuning the threshold τ. This problem is addressed in [12] where a new algorithm is proposed to adapt the recovery to the dataset by leveraging knowledge of the second order statistics.

3.2. Bayesian Singular Value Theresholding

BSVT is an MC-based algorithm [12] that is able to optimize the value of τ at each iteration using additional prior knowledge in the form of second order statistics. The optimization of the soft-theresholding step is performed using Stein's unbiased risk estimate (SURE) [24] for which a closed-form expression is presented in [25]. A detailed description of the BSVT algorithm is presented in [12] but we reproduce the algorithm below to aid with the presentation. Note that D_{LMMSE} represents the average noise per entry in Ω^c.

The main advantage of the BSVT algorithm is that the threshold is optimized at each iteration. This is achieved by incorporating the prior knowledge about the matrix in the form of second order statistics via the introduction of the SURE and linear minimum mean square error (LMMSE) steps. Admittedly, this approach requires additional knowledge that is not necessary when using the SVT algorithm. However, it is shown in [12] that the introduction of the prior knowledge enables a robust recovery of the missing entries.

Algorithm 1 Bayesian Singular Value Thresholding [12]

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

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

Set $\mathbf{Y}^0 = \mathbf{0}$
Set $\mathbf{Z}^0 = \mathbf{0}$
Set $\tau = 0$
Set $\Omega^c = \{1, 2, ..., 2M\} \times \{1, 2, ..., N\} \setminus \Omega$
for $k = 1$ to k_{max} **do**
  Compute $[\mathbf{U}, \mathbf{S}, \mathbf{V}] = \text{svd}(\mathbf{Z}^{(k-1)})$
  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 $\mathbf{Y}^{(k)} = \mathbf{Y}^{(k-1)} + \delta_b\big(P_{\Omega}(\mathbf{R}) - P_{\Omega}(\mathbf{X}^{(k)})\big)$
  Set $\mathbf{L}^{(k)} = \mathbf{\Sigma}_{\Omega^c\Omega}\mathbf{\Sigma}_{\Omega\Omega}^{-1}\mathbf{Y}^{(k)}$
  Set $\mathbf{Z}^{(k)} = \mathbf{Y}^{(k)} + \mathbf{L}^{(k)}$
  Set $\sigma^2_{\mathbf{Z}^{(k)}} = (\|\mathbf{Y}^{(k)} - P_{\Omega}(\mathbf{R})\|_F^2 + |\Omega^c| D_{\text{LMMSE}})/2MN$
  Set $\tau^{(k)} = \arg\min_{\tau} \text{SURE}(D_{\tau})(\mathbf{Z}^{(k)})$
end for
Set $\widehat{\mathbf{M}}_{\text{BSVT}} = \mathbf{X}^{(k)}$

4. JOINT RECOVERY OF MISSING DATA IN TWO DATASETS

In this section, an estimation framework for recovering missing data from different datasets is proposed. The estimation framework facilitates exploiting the correlation between datasets for a wide range

of correlation structures. We begin by noting that in the joint recovery case, there are two types of correlation between the entries of the combined matrix. First, the intra-correlation that refers to the correlation between the entries within each dataset. This is the type of correlation that is exploited in the independent recovery scenario, i.e., when the missing entries from each dataset are recovered using only available observations from that dataset. Second, the cross-correlation defined as the correlation between the data points from the two different datasets. In contrast to the independent recovery case, a joint recovery technique needs to account for both types of correlation. By considering the cross-correlation, the recovery process leverages on other types of data in order to recover the datasets with limited available observations.

In an MC setting, the minimum number of observations required depends on the size and the rank of the matrix [14]. The combination of the datasets into a single matrix increases the size of the matrix, and therefore, the fundamental limit for the joint recovery case depends on the tradeoff between the size and the rank of the combined matrix, and the number of observations available for each dataset. Note that the rank of the combined matrix depends on both the intra- and the cross-correlation. The following lemma provides lower and upper bounds for the rank of the combined matrix based on the individual rank of the matrices. To that end, let us denote rank of the matrices by $\text{rank}(\mathbf{M}_1) = r_1$, $\text{rank}(\mathbf{M}_2) = r_2$, and $\text{rank}(\mathbf{M}) = r$.

Lemma 1. *Let $\mathbf{M}_1 \in \mathbb{R}^{M\times N}$ and $\mathbf{M}_2 \in \mathbb{R}^{M\times N}$. Define the combined matrix $\mathbf{M} = \begin{bmatrix} \mathbf{M}_1 \\ \mathbf{M}_2 \end{bmatrix} \in \mathbb{R}^{2M\times N}$. Then, the following holds:*

$$\max(r_1, r_2) \leq r \leq r_1 + r_2. \tag{12}$$

Proof. Please refer to [26]. □

Based on the insight provided by Lemma 1, the intra-correlation determines the rank of the matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ which define the lower and upper bounds on r. A smaller value of intra-correlation in one of the datasets results in a larger lower bound for r. On the other hand, the cross-correlation governs the value of r within the limits defined by Lemma 1. Indeed, a larger value of cross-correlation results in a value of r that is closer to the lower bound while a smaller value of cross-correlation generates a combined matrix with a rank that is closer to the upper bound. In other words, the intra-correlation defines the limit values of r for which recovery is feasible and the cross-correlation governs the value of r within the limit.

In [27] it is shown that the low rank matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ can be successfully recovered independently when the number of observations for the first matrix, denoted by k_1, satisfies

$$k_1 > (M + N - r_1)r_1, \tag{13}$$

and the number of available observations for the second matrix, denoted by k_2, obeys

$$k_2 > (M + N - r_2)r_2. \tag{14}$$

This result is based on the assumption that for the random matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ there exist the σ-measures μ_1 and μ_2, respectively, and that both measures admit a Lebesgue decomposition. For the combined matrix $\mathbf{M}$, the σ-measure is obtained as the product of the measures of $\mathbf{M}_1$ and $\mathbf{M}_2$ [28], which yields $\mu = \mu_1 \times \mu_2$. Moreover, since μ is a σ-measure it also admits a Lebesgue decomposition [29] and [30]. Hence, the result in [27] applies for the combined matrix $\mathbf{M}$ without any additional assumptions, i.e.,

$$k_1 + k_2 > (2M + N - r)r. \tag{15}$$

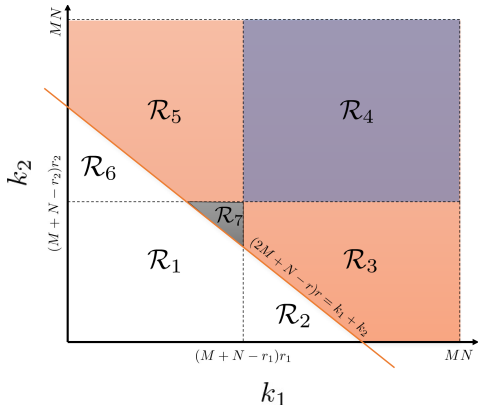


Fig. 3: Example of recovery regions imposed by the fundamental limit for the recovery of the matrices $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}$.

Fig. 3 depicts the inequalities in (13), (14) and (15) that describe the lower bound on the number of observations required to recover the matrices $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}$, respectively. The bounds divide the (k_1, k_2) plane into seven regions that correspond to different recovery scenarios for the independent and joint estimation settings. The seventh region, i.e., $\mathcal{R}_7$, corresponds to the case in which the independent recovery of each dataset is not possible but the joint recovery is feasible. In other words, the existence of region $\mathcal{R}_7$ is equivalent to the case in which it is beneficial to jointly recover the two datasets. Note that when the rank of the combined matrix increases, the line described by $(2M + N - r)r = k_1 + k_2$ is shifted towards larger values which for values of r larger than a given threshold induces an empty region $\mathcal{R}_7$. The value of the threshold is given by the sufficient condition in Theorem 1. It is also worth noting that based on the regions depicted in Fig. 3, the number of cases in which the joint recovery is feasible is larger than the number of cases in which the independent recovery of the datasets is possible. The following theorem provides the necessary and sufficient conditions to guarantee that the joint recovery is beneficial.

Theorem 1. *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*

$$1 - \frac{\max(r_1, r_2)}{\min(r_1, r_2)} > \frac{\min(r_1, r_2) - N}{M}, \tag{16}$$

and the rank of the combined matrix satisfies

$$\begin{aligned} r <& M + \frac{1}{2}N - \frac{1}{2}(M + N - 2r_1 - 2r_2) \\ &\times \left(1 + \frac{3M^2 + 2MN - 8r_1r_2}{(M + N - 2r_1 - 2r_2)^2}\right)^{1/2}. \end{aligned} \tag{17}$$

Proof. Please refer to [26]. □

Note that the necessary condition from Theorem 1, that is the inequality in (16) depends only on the matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ and does not depend on the combined matrix $\mathbf{M}$. In contrast, the sufficient condition in (17) provides an upper bound for the rank of the combined matrix such that the total number of observations required for the joint recovery is fewer when compared to the independent recovery case. Theorem 1 provides a necessary and sufficient condition for the joint recovery of two data matrices to be beneficial.

5. NUMERICAL RESULTS

This section presents a numerical evaluation of the joint recovery performance for two datasets. The matrices used for the simulations are generated using the model described in Section 2.3 and the size of the matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ is fixed such that $M = 50$ and $N = 100$, respectively. Hence, the joint matrix $\mathbf{M}$ is a square matrix of size 100. The range of rank values selected aims to characterize the joint recovery in two scenarios: when the two combined matrices have the same rank, i.e., $r_1 = 6$ and $r_2 = 6$, and when the ratio between the two rank values is small, i.e., $r_1 = 6$ and $r_2 = 9$. To facilitate this, the synthetic data model presented in Section 2.3 is used to generate correlated data matrices with the rank values of interest.

5.1. Simulation framework

Using the mathematical model defined in Section 2.3, the covariance matrix for the combined matrix $\mathbf{M}$ is given by (2) where the intra-correlation between the state variables in $\mathbf{M}_1$ is modelled by the υ_{11} parameter in (4), the intra-correlation between the state variables in $\mathbf{M}_2$ is modelled by the υ_{22} parameter in (4) and the cross-correlation between $\mathbf{M}_1$ and $\mathbf{M}_2$ is modelled by ψ in (2). The numerical analysis shows that a larger value of υ_{ll} results in a more correlated matrix $\mathbf{M}_l$ and consequently a smaller value for r_l, where $l \in \{1, 2\}$. Moreover, the cross-correlation between $\mathbf{M}_1$ and $\mathbf{M}_2$ increases with the value of ψ which leads to a smaller value for r within the bounds defined by Lemma 1.

The matrix $\mathbf{M}$ generated using the model in (2) is not exactly low rank. Instead, it can be well approximated by a low rank matrix. Let us denote by $\widetilde{\mathbf{M}}(r)$ the low rank approximation of rank r obtained by vanishing the smallest $N - r$ singular values of the matrix $\mathbf{M}$. In the following, r is defined as the minimum value for which the NMSE between the matrix $\mathbf{M}$ and the low rank approximation of rank r, i.e., $\widetilde{\mathbf{M}}(r)$, is below 10^{-3}. Consequently, the model in (2) is used to generate data matrices $\mathbf{M}$ such that the low rank approximations $\widetilde{\mathbf{M}}_1(r_1)$, $\widetilde{\mathbf{M}}_2(r_2)$ and $\widetilde{\mathbf{M}}(r)$ have the intended ranks. Moreover, the low rank approximation of the combined matrix, i.e., $\widetilde{\mathbf{M}}(r)$, is used to evaluate the numerical performance of both BSVT and SVT in exploiting the correlation between the two datasets. The recovery performance of both algorithms is averaged over ten realizations of Ω, where the locations of the available entries are sampled uniformly at random in each dataset.

In the following, a numerical analysis for the joint recovery performance of SVT and BSVT is presented for the cases in which the rank values for the combined matrices are: $r_1 = 6, r_2 = 6, r = 9$ and $r_1 = 6, r_2 = 9, r = 10$. The choice of rank for the combined matrix resembles a high cross-correlation case in which the value of r satisfies the condition imposed by Theorem 1. We consider a low noise regime for which SNR=50 dB in both datasets to emphasize the impact of the intra- and cross-correlation in the recovery process, where the SNR in dataset $l \in \{1, 2\}$ is defined as

$$\text{SNR}_l \triangleq 10\log_{10} \frac{\frac{1}{M}\text{Tr}(\mathbf{\Sigma}_{ll})}{\sigma^2_{\mathbf{N}_l}}, \tag{18}$$

where $\mathbf{\Sigma}_{ll}$ is defined in (3) and $\sigma^2_{\mathbf{N}_l}$ is described in (5). A wider range of rank values and noise regimes is presented in [31]. The efficiency in exploiting the cross-correlation between the combined datasets is evaluated by comparing the recovery performance across different sampling regimes in which the number of observations in Ω is constant but the ratio between the number of available entries in each dataset varies. The cross-correlation is successfully exploited when the recovery error is similar across different sampling regimes.

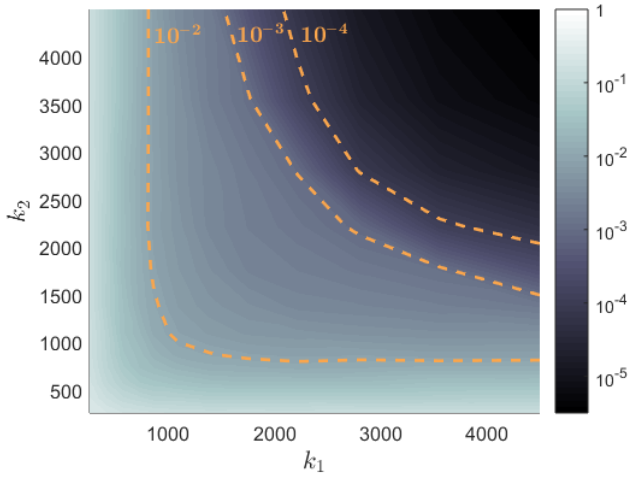


Fig. 4: Joint recovery error using BSVT, measured by NMSE, when $r_1 = 6$, $r_2 = 6$, $r = 9$ and SNR=50 dB.

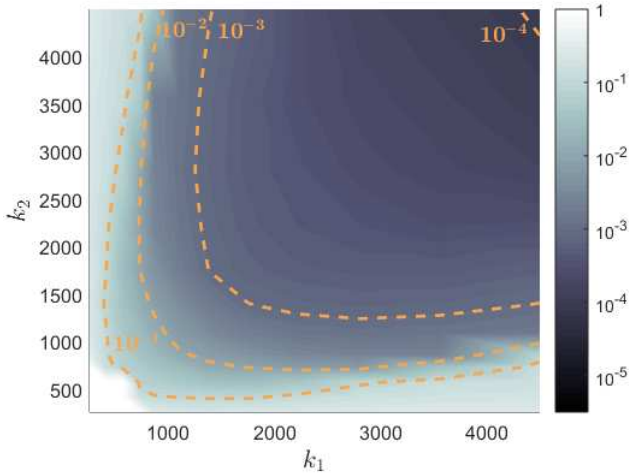


Fig. 5: Joint recovery error using SVT, measured by NMSE, when $r_1 = 6$, $r_2 = 6$, $r = 9$ and SNR=50 dB.

Fig. 4 depicts the performance of the BSVT algorithm when $r_1 = 6$, $r_2 = 6$, $r = 9$ and SNR=50 dB. Interestingly, the contour lines for the 10^{-4} and 10^{-3} recovery error exhibit a similar shape to the line depicted by $(2M + N - r)r = k_1 + k_2$ in Fig. 3. This suggests that the BSVT algorithm successfully exploits the cross-correlation in that region and obtains a similar recovery performance tradeoff when the ratio between k_1 and k_2 varies for a fixed value of $k_1 + k_2$. In contrast, the contour lines for 10^{-3}, 10^{-2} and 10^{-1} SVT recovery error depicted in Fig. 5 exhibit a similar shape to the region $\mathcal{R}_4$ in Fig. 3 which corresponds to the independent recovery area in which the cross-correlation is not exploited. Based on this observation, it is reasonable to assume that SVT is not effective in exploiting the cross-correlation as the recovery error changes with the ratio between k_1 and k_2 for a fixed total number of observations.

Fig. 6 depicts the performance of the BSVT algorithm when $r_1 = 6$, $r_2 = 9$, $r = 10$ and SNR=50 dB. In line with the case discussed in Fig. 4, the contour lines for the 10^{-4} and 10^{-3} recov-

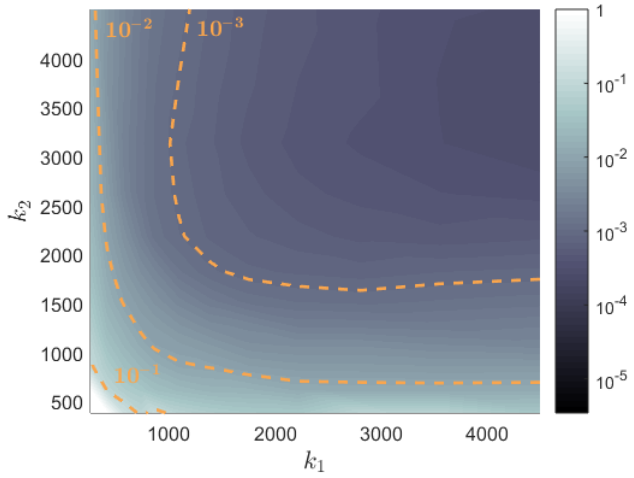


Fig. 7: Joint recovery error using SVT, measured by NMSE, when $r_1 = 6, r_2 = 9, r = 10$ and SNR=50 dB.

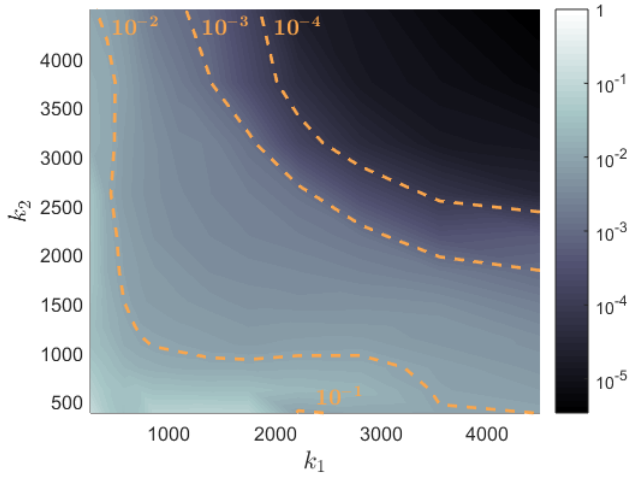


Fig. 6: Joint recovery error using BSVT, measured by NMSE, when $r_1 = 6, r_2 = 9, r = 10$ and SNR=50 dB.

ery error exhibit a similar shape to the fundamental limit in Fig. 3. This suggests that the BSVT approach is able to exploit the cross-correlation between the combined datasets in the almost noiseless regime for both rank cases considered. In Fig. 7 the performance of the SVT algorithm is depicted for the case in which $r_1 = 6, r_2 = 9$, $r = 10$ and SNR=50 dB. In this case, the shape of the contour lines for 10^{-3} and 10^{-2} recovery error is similar to the shape of the region $\mathcal{R}_4$ in Fig. 3 which suggests that the SVT algorithm is not efficient in exploiting cross-correlation. Consequently, the BSVT algorithm is able to exploit the cross-correlation between the combined datasets more effectively when compared to the SVT approach in the almost noiseless regime. The gain in recovery performance is facilitated by the prior knowledge incorporated in the structure of the BSVT algorithm.

6. CONCLUSION

The fundamental limits for the joint recovery of two datasets have been characterized in terms of the rank of the single and combined data matrices. Theoretical conditions are derived for the case in which the joint recovery of two datasets requires less observations compared to the independent recovery case. Based on the insight provided by the fundamental limit, the number of cases in which the joint recovery is feasible is significantly larger when compared to the independent recovery setting.

A model for correlated datasets is proposed. Numerical results show that the correlation between different types of data is exploited by leveraging the information provided by the dataset with fewer missing entries to enable the recovery of the other dataset. Moreover, in contrast to the SVT algorithm, the performance of the BSVT approach matches the geometry imposed by the fundamental limit which suggests that BSVT is indeed better suited to exploit the correlation between datasets.

7. REFERENCES

[1] R. A. Walling, R. Saint, R. C. Dugan, J. Burke, and L. A. Kojovic, "Summary of distributed resources impact on power delivery systems," *IEEE Trans. Power Del.*, vol. 23, no. 3, pp. 1636–1644, Jul. 2008.

[2] US-Canada Power System Outage Task Force, *Final report on the August 14, 2003 Blackout in the United States and Canada: causes and recommendations*, Apr. 2004.

[3] Union for the Co-ordination of Transmission of Electricity Investigation Committee, "Final report of the Investigation Committee on the 28 September 2003 Blackout in Italy," *Union for the Co-ordination of Transmission of Electricity*, Apr. 2004.

[4] N. Bui, A. P Castellani, P. Casari, and M. Zorzi, "The internet of energy: a web-enabled smart grid system," *IEEE Network*, vol. 26, no. 4, Jul. 2012.

[5] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Transactions on Information and System Security*, vol. 14, no. 1, pp. 13, May 2011.

[6] T. T. Kim and H. V. Poor, "Strategic protection against data injection attacks on power grids," *IEEE Trans. Smart Grid*, vol. 2, no. 2, pp. 326–333, Jun. 2011.

[7] O. Kosut, L. Jia, R. J. Thomas, and L. Tong, "Malicious data attacks on the smart grid," *IEEE Trans. Smart Grid*, vol. 2, no. 4, pp. 645–658, Dec. 2011.

[8] M. Ozay, I. Esnaola, F. T. Y. Vural, S. R. Kulkarni, and H. V. Poor, "Machine learning methods for attack detection in the smart grid," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 8, pp. 1773–1786, Aug. 2016.

[9] K. Sun, I. Esnaola, S. M. Perlaza, and H. V. Poor, "Information-theoretic attacks in the smart grid," in *Proc. IEEE International Conference on Smart Grid Communications*, Oct. 2017, pp. 455–460.

[10] P. Gao, M. Wang, S. G. Ghiocel, J. H. Chow, B. Fardanesh, and G. Stefopoulos, "Missing data recovery by exploiting low-dimensionality in power system synchrophasor measurements," *IEEE Trans. Power Syst.*, vol. 31, no. 2, pp. 1006–1013, Mar. 2016.

[11] C. Genes, I. Esnaola, S. M. Perlaza, L. F. Ochoa, and D. Coca, "Recovering missing data via matrix completion in electricity distribution systems," in *Proc. IEEE International Workshop on Signal Processing Advances in Wireless Communications*, Jul. 2016, pp. 1–6.

[12] 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.

[13] Y. Isozaki, S. Yoshizawa, Y. Fujimoto, H. I. Ishii, I. Ono, T. Onoda, and Y. Hayashi, "Detection of cyber-attacks against voltage control in distribution power grids with PVs," *IEEE Trans. Smart Grid*, vol. 7, no. 4, pp. 1824–1835, Jul. 2016.

[14] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, pp. 717, Apr. 2009.

[15] E. J. Candès and Y. Plan, "Matrix completion with noise," *Proc. IEEE*, vol. 98, no. 6, pp. 925–936, Jun. 2010.

[16] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2053–2080, May 2010.

[17] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 208–220, Jan. 2013.

[18] M. E. Kilmer, K. Braman, N. Hao, and R. C. Hoover, "Third-order tensors as operators on matrices: A theoretical and computational framework with applications in imaging," *SIAM Journal on Matrix Analysis and Applications*, vol. 34, no. 1, pp. 148–172, Feb. 2013.

[19] Z. Zhang, G. Ely, S. Aeron, N. Hao, and M. Kilmer, "Novel methods for multilinear data completion and de-noising based on tensor-SVD," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 3842–3849.

[20] 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*, Feb. 2015, pp. 306–314.

[21] J. F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, Mar. 2010.

[22] Electricity North West Limited, "Low voltage network solutions," [Online]. Available: https://www.enwl.co.uk/innovation/smaller-projects/low-carbon-networks-fund/low-voltage-network-solutions/, Accessed: 13 Feb. 2019.

[23] G. A. F. Seber, *A matrix handbook for statisticians*, Wiley, Hoboken, NJ, 2007.

[24] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *The Annals of Statistics*, vol. 9, no. 6, pp. 1135–1151, Nov. 1981.

[25] E. J. Candès, C. A. Sing-Long, and J. D. Trzasko, "Unbiased risk estimates for singular value thresholding and spectral estimators," *IEEE Trans. Signal Process.*, vol. 61, no. 19, pp. 4643–4657, Oct. 2013.

[26] C. Genes, I. Esnaola, S. M. Perlaza, and D. Coca, "Recovery of missing data in correlated smart grid datasets," *arXiv preprint*, 2019.

[27] E. Riegler, D. Stotz, and H. Bölcskei, "Information-theoretic limits of matrix completion," in *Proc. IEEE International Symposium on Information Theory*, Jun. 2015, pp. 1836–1840.

[28] P. R. Halmos, *Measure theory*, Springer Science & Business Media, New York, NY, 2013.

[29] E. Hewitt and K. Stromberg, *Real and abstract analysis: a modern treatment of the theory of functions of a real variable*, Springer-Verlag, Berlin, Germany, 2013.

[30] H. L. Royden and P. Fitzpatrick, *Real analysis*, Pearson, New York, NY, 2018.

[31] C. Genes, *Novel Matrix Completion Methods for Missing Data Recovery in Urban Systems*, Ph.D. thesis, University of Sheffield, 2018.