Proceedings Paper:
Genes, C., Esnaola, I., Perlaza, S.M. et al. (2 more authors) (2016) Recovering Missing Data via Matrix Completion in Electricity Distribution Systems. In: 2016 IEEE 17TH INTERNATIONAL WORKSHOP ON SIGNAL PROCESSING ADVANCES IN WIRELESS COMMUNICATIONS (SPAWC). 2016 IEEE 17TH INTERNATIONAL WORKSHOP ON SIGNAL PROCESSING ADVANCES IN WIRELESS COMMUNICATIONS (SPAWC), 03-06 Jul 2016, Edinburgh, UK. .

doi.org/10.1109/SPAWC.2016.7536744

Recovering Missing Data via Matrix Completion in Electricity Distribution Systems

Cristian Genes, Iñaki Esnaola, Samir M. Perlaza, Luis F. Ochoa, and Daniel Coca.

***Abstract*—The performance of matrix completion based recovery of missing data in electricity distribution systems is analyzed. Under the assumption that the state variables follow a multivariate Gaussian distribution the matrix completion recovery is compared to estimation and information theoretic limits. The assumption about the distribution of the state variables is validated by the data shared by Electricity North West Limited. That being the case, the achievable distortion using minimum mean square error (MMSE) estimation is assessed for both random sampling and optimal linear encoding acquisition schemes. Within this setting, the impact of imperfect second order source statistics is numerically evaluated. The fundamental limit of the recovery process is characterized using Rate-Distortion theory to obtain the optimal performance theoretically attainable. Interestingly, numerical results show that matrix completion based recovery outperforms MMSE estimator when the number of available observations is low and access to perfect source statistics is not available.**

I. Introduction

The electricity network is changing towards a locally controlled smart grid which incorporates an advanced sensing and management infrastructure. Energy sources such as solar or wind power are envisioned as integral elements of the network at the end-user level. As a result, the number of nonlinear loads is expected to increase, which results in larger perturbations in the electricity grid [1]. The complexity of the control strategies in the smart grid is expected to increase guided by the challenges posed by new and distributed energy sources. The implementation of advanced control strategies demands access to accurate and low latency data describing the state of the grid, which increases the performance requirements for the sensing infrastructure. The state estimation problem when data injection attacks are present is studied in [2], [3], [4], and [5]. Sensor failures, errors during data collection, unreliable transmission, and storage issues are just some of the causes of the operator having an incomplete set of observations of the state variables describing the grid. Given the size and complexity of the sensing infrastructure, tracking all these events is not feasible. It is therefore necessary to estimate the missing state variables using the available observations.

This research was supported in part by the European Commission under Marie Sklodowska-Curie Individual Fellowship No. 659316 (CYBERNETS).

Cristian Genes, Iñaki Esnaola, and Daniel Coca are with the Department of Automatic Control and Systems Engineering, University of Sheffield, S1 3JD, UK.

Samir M. Perlaza is with the CITI Lab of the Institut National de Recherche en Informatique et en Automatique (INRIA), Université de Lyon and Institut National des Sciences Appliquées (INSA) de Lyon. 6 Av. des Arts 69621 Villeurbanne, France.

Luis F. Ochoa is with the School of Electrical and Electronic Engineering, University of Manchester, M13 9PL, UK. (cgenes1@sheffield.ac.uk, esnaola@sheffield.ac.uk, samir.perlaza@inria.fr, luis_ochoa@ieee.org, and d.coca@sheffield.ac.uk).

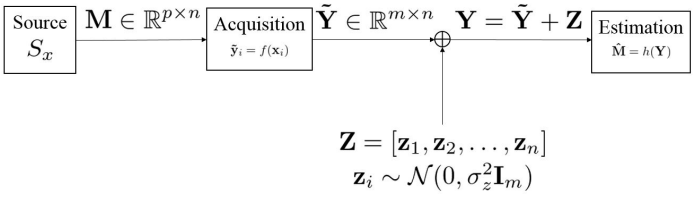

Fig. 1. Block diagram describing the system model.

Missing data recovery can be cast as a minimum mean square error (MMSE) estimation problem. However, this approach relies in access to prior information, specifically, second order statistics of the state variables. Therefore, in practical scenarios where perfect prior knowledge is not available to the operator, MMSE estimation based recovery is suboptimal [6]. In the smart grid context, the increased number of nonlinear loads affects the precision of the statistics postulated for the state variables model, and ultimately, the precision of MMSE based recovery.

Matrix completion offers an alternative approach to the problem of recovering missing observations by exploiting the statistical structure of the observations [7], [8]. In particular, the fact that correlated data vectors give rise to low rank data matrices is exploited in a convex optimization context. That being the case, it can be shown that the recovery of missing observations is feasible provided that a sufficient fraction of the observations is available [9], [10], [11], [12] and [13]. However, the results therein are based on the assumption that missing entries are not correlated, which is not always the case in practical scenarios. Within that setting, low rank minimization tools are proving useful in electricity grid settings [14], [15]. The case of correlated missing entries for phasor measurement units data is studied in [16].

In this paper, the performance of different missing data recovery methods is studied. The viability of matrix completion as a recovery strategy when there are missing observations is compared to MMSE estimation based recovery. A mismatched covariance matrix scenario is proposed to study the trade-off between the amount of prior knowledge and the performance of different recovery techniques. In this framework, a comparison between matrix completion and MMSE estimation for different levels of mismatch is presented.

The main contributions in this work are summarized next. It is shown that the data set is approximately Gaussian distributed. In view of this, a Gaussian random process is proposed to model the state variables. The conditions for which matrix completion outperforms MMSE estimation are characterized. Interestingly, numerical results show that matrix

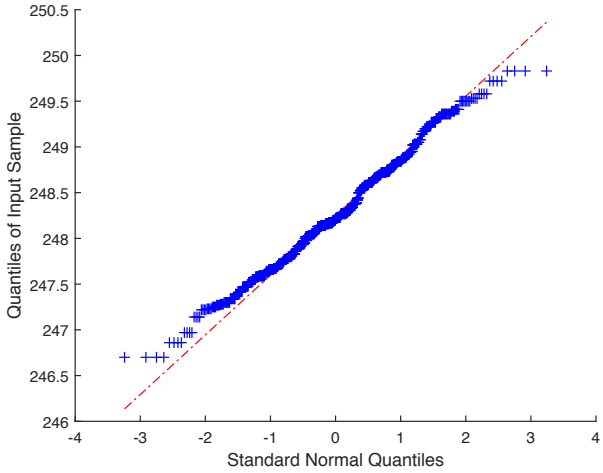

Fig. 2. QQ plot for the distribution of the voltage data provided by ENWL versus a Gaussian distribution.

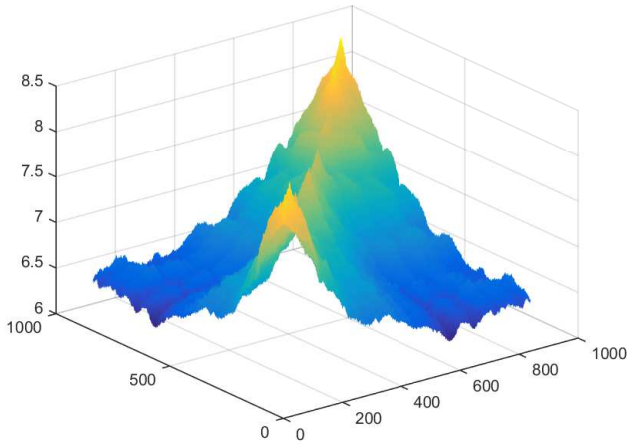

Fig. 3. Sample covariance matrix of the voltage data provided by ENWL.

completion performs better for moderate and high levels of mismatched statistics when more than half of the measurements are missing. MMSE estimation is also outperformed by matrix completion for moderate mismatch regimes when at least a quarter of the data is not available. Similarly, for high mismatch regimes matrix completion based recovery outperforms MMSE estimation for a wide range of missing data values.

II. System Model

Consider a electricity distribution secondary substation with n low voltage (LV) feeders. At the head of the feeder connected to the transformer a sensing unit measures various electrical magnitudes, e.g., voltage, intensity, active and reactive power in the feeder at a given time instant. These measures comprise the state variables that the operator uses for control, monitoring, and management purposes. The process of acquiring the state variables by the operator is referred to as the data acquisition process. Unfortunately, the presence of noise and missing data variables provides the operator with a set of incomplete and corrupted state variables. That being the case, the operator needs to estimate the missing entries based on the available observations with a given optimality criterion.

More specifically, the data acquisition process is modelled by the scheme depicted in Figure 1. In this setting, the realizations of the state variables produced during p time instants in the n feeders of the LV distribution system are arranged in the matrix $\mathbf{M} \in \mathbb{R}^{p\times n}$. The information source, S_x, expresses the statistical structure of the underlying stochastic processes governing the state variables. A subset of the state variables, $\widetilde{\mathbf{Y}} \in \mathbb{R}^{m\times n}$, $m < p$, is observed and corrupted by additive white Gaussian noise which results in the observations, $\mathbf{Y} \in \mathbb{R}^{m\times n}$, that are available to the operator for estimation purposes. The additive noise represents the thermal noise introduced by the sensors used in the LV feeders. The subset of state variables that are not observed accounts for the missing entries in the data acquisition process. The main challenge for the recovery procedure is to estimate the missing entries. A detailed description of the elements in the system follows.

A. Source Model for State Variables

Let $x_{j,i,l} \in \mathbb{R}$ be the value of the state variable j in feeder i at time l. The column vector $\mathbf{x}_{j,i} = [x_{j,i,1}, x_{j,i,2}, \ldots, x_{j,i,p}]^T$ contains the values of state variable j in feeder i at discrete time instants $l = 1, 2, \ldots, p$. The aggregated data describing state variable j in all feeders is given by the data matrix $\mathbf{M}_j = [\mathbf{x}_{j,1}\ \mathbf{x}_{j,2} \ldots \mathbf{x}_{j,n}] \in \mathbb{R}^{p\times n}$. In the remaining of the paper, the analysis is presented for a particular state variable, and therefore, the index j is dropped. That being the case, the data matrix $\mathbf{M}$ describes the state variable of interest in all feeders for time instants $l = 1, 2, \ldots, p$.

The vector of state variables, $\mathbf{x}_i$, is a realization of the discrete random process S_x. As part of the "Low Voltage Network Solutions" project run by Electricity North West Limited, measurements are collected every minute from 200 residential secondary substations across the North West of England from June 2013 to January 2014. Daily data files contain the following measurements: voltage, current, real and reactive power on all three phases. The analysis in this paper is particularized to the case in which the state variable under consideration is voltage, but can be easily extended to other state variables. Figure 2 shows the Q-Q plot of the voltage data comparing the LV data to a Gaussian distribution. It can be seen that the distribution is close to a Gaussian distribution up to a minor deviation around the tails. In view of this, the real data set used in this work suggests that S_x can be modelled as a multivariate Gaussian random process, i.e. $\mathbf{x}_i \sim \mathcal{N}(\mu, \mathbf{\Sigma})$, and $\{\mathbf{x}_i\}_{i=1}^n$ is an independent and identically distributed sequence. The sample covariance matrix obtained with the real data set is depicted in Figure 3. Interestingly, the covariance matrix exhibits a structure that is approximately Toeplitz, a feature that is usually observed in stationary autoregressive signals. The Toeplitz model resembles a physical temporal correlation where the correlation decreases as the temporal distance increases. This implies that the correlation

between two voltage observations in the same feeder depends on their separation in time.

B. Acquisition

The acquisition process is modelled by the function $f : \mathbb{R}^n \to \mathbb{R}^m$, where m is the number of observed entries for each vector of state variables $\mathbf{x}_i$. The observations from feeder i that are available to the operator are given by $\mathbf{y}_i \in \mathbb{R}^m$. Note that noise is modelled as additive Gaussian noise $\mathbf{z}_i \sim \mathcal{N}(0, \sigma_z^2 \mathbf{I}_m)$ and the observations of feeder i available to the operator are given by $\mathbf{y}_i = \tilde{\mathbf{y}}_i + \mathbf{z}_i$, where $\tilde{\mathbf{y}}_i = f(\mathbf{x}_i)$ are the noiseless observations of feeder i. The resulting set of noiseless observations are given by matrix $\widetilde{\mathbf{Y}} \in \mathbb{R}^{m \times n}$ which is formed as $\widetilde{\mathbf{Y}} = [\tilde{\mathbf{y}}_1, \tilde{\mathbf{y}}_2, \ldots, \tilde{\mathbf{y}}_n]$. Thus, the noisy set of observations available to the operator are given by $\mathbf{Y} = \widetilde{\mathbf{Y}} + \mathbf{Z}$, where $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n]$.

C. Estimation

The estimation process is modelled by the function $g : \mathbb{R}^m \to \mathbb{R}^n$ which produces the estimate $\hat{\mathbf{x}} = g(f(\mathbf{x}) + \mathbf{z})$. The optimality criterion for the reconstruction error is the mean square error (MSE) given by

$$\text{MSE}\left(\mathbf{x}; g(f(\mathbf{x}))\right) \triangleq \mathbb{E}\left[\|\mathbf{x} - g(f(\mathbf{x}) + \mathbf{z})\|_2^2\right]. \quad (1)$$

The optimal reconstruction strategy in the MSE sense, g^*, is the MMSE estimator given by the following conditional expectation:

$$\hat{\mathbf{x}}_{\mathsf{MMSE}} = g^*(\mathbf{y}) \triangleq \mathbb{E}[\mathbf{x}|\mathbf{y}, \mathbf{\Sigma}]. \quad (2)$$

For a given acquisition function, f, the MSE achievable via MMSE estimation is given by

$$\text{MMSE}(\mathbf{x}|f(\mathbf{x}) + \mathbf{z}) = \mathbb{E}\big[\|\mathbf{x} - \mathbb{E}[\mathbf{x}|f(\mathbf{x}) + \mathbf{z}]\|_2^2\big]. \quad (3)$$

For a particular feeder i the operator produces the estimate $\hat{\mathbf{x}}_i$. Thus, it is easy to extend the previous estimation vector formulation to a matrix formulation where the estimate of the data matrix is given by $\widehat{\mathbf{M}} = h(\mathbf{Y})$ with the estimation function given by $h : \mathbb{R}^{m \times n} \to \mathbb{R}^{p \times n}$. Consequently, the MSE optimality criterion for the estimation of the data matrix is

$$\text{MSE}\left(\mathbf{M}; h(\mathbf{Y})\right) = \mathbb{E}\left[\|\mathbf{M} - h(\mathbf{Y})\|_F^2\right], \quad (4)$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Similarly, the MMSE estimate is obtained as

$$\widehat{\mathbf{M}}_{\mathsf{MMSE}} = h^*(\mathbf{Y}) = \mathbb{E}[\mathbf{M}|\mathbf{Y}, \mathbf{\Sigma}], \quad (5)$$

where $h^* : \mathbb{R}^{m \times n} \to \mathbb{R}^{p \times n}$ is the MMSE estimation function which yields a performance given by

$$\text{MMSE}(\mathbf{X}|\mathbf{Y}) = \mathbb{E}\big[\|\mathbf{X} - \mathbb{E}[\mathbf{M}|\mathbf{Y}, \mathbf{\Sigma}]\|_F^2\big]. \quad (6)$$

In practical settings the real covariance matrix $\mathbf{\Sigma}$ is not known during the recovery process due to the fact that source statistics need to be estimated by the operator. For that reason, practical systems operate with a postulated covariance matrix $\mathbf{\Sigma}^*$ which differs, in general, from the real covariance matrix. Note that for the case in which the estimator has access to perfect prior knowledge, it holds that $\mathbf{\Sigma} = \mathbf{\Sigma}^*$.

III. Matrix completion

Given a data matrix $\mathbf{M} \in \mathbb{R}^{p \times n}$, with $p \leq n$, let $\mathbf{M}_{i,j}$ with $(i,j) \in \Omega$ be the set of observations where Ω is the set of indices of the available entries. In general, the missing entries cannot be estimated without assuming additional structure about the data matrix $\mathbf{M}$. Remarkably, in [9] it is shown that most low-rank matrices can be recovered when the number of sampled entries obeys

$$m \geq C n^{1.25} r \log n, \quad (7)$$

where r is the rank of $\mathbf{M}$ and C is a positive constant with a probability of recovery of at least $1 - cn^{-3}\log n$ with c a positive constant. Let P_Ω be the orthogonal projector onto the span of matrices vanishing outside Ω so that the the entry (i,j) of $P_\Omega(\mathbf{X})$ is equal to $\mathbf{X}_{ij}$ if $(i,j) \in \Omega$ and zero otherwise. The missing entries are recovered by solving the optimisation problem

$$\begin{aligned} \underset{\mathbf{X}}{\text{minimize}} \quad & \text{rank}(\mathbf{X}) \\ \text{subject to} \quad & P_\Omega(\mathbf{X}) = P_\Omega(\mathbf{M}), \end{aligned} \quad (8)$$

where $\mathbf{X}$ is the decision variable. Unfortunately, solving this problem is computationally unfeasible. The optimisation problem is NP-hard and all known algorithms achieving the exact solution require time doubly exponential in the dimension of the matrix [17]. However, it can be shown that in some cases the optimization problem in (8) can be solved exactly via convex programming. Specifically, the following convex relaxation is proposed in [9] based on nuclear norm minimization:

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

where $\|\mathbf{X}\|_*$ refers to the nuclear norm of the matrix $\mathbf{X}$,

$$\|\mathbf{X}\|_* = \sum_{k=1}^{p} \sigma_k(\mathbf{X}), \quad (10)$$

and $\sigma_k(\mathbf{X})$ denotes the k-th largest singular value of $\mathbf{X}$.

There are several approaches to solve the nuclear norm minimization problem. A short classification based on the trade-offs between computational performance, theoretical guarantees, and numerical accuracy is provided in [7]. For small matrices, interior point methods can be used to provide accurate solutions. Methods like SeDuMi [18] or SDPT3 [19] use second-order information and are able to produce accurate solutions for matrix dimensions around 50. However, to reduce memory requirements the problem structure must be exploited. In [8] matrix sizes up to 350 can be recovered using interior point methods. Alternatively, singular value thresholding (SVT) is a simple, first-order algorithm proposed in [20]. For iteration k the algorithm produces the pair of matrices $(\mathbf{X}^k, \mathbf{Y}^k)$ by performing a soft-thresholding operation on the singular values of matrix $\mathbf{Y}^k$. The main advantage of this approach is that the algorithm makes use of minimal storage space by exploiting the sparsity of $\mathbf{Y}^k$ and has a low computational cost per iteration. It is shown in

[20] that the sequence $\mathbf{X}^k$ converges to the unique solution of the following optimisation problem

$$\begin{aligned} \underset{\mathbf{X}}{\text{minimize}} \quad & \tau\|\mathbf{X}\|_* + \frac{1}{2}\|\mathbf{X}\|_F^2 \\ \text{subject to} \quad & P_\Omega(\mathbf{X}) = P_\Omega(\mathbf{M}), \end{aligned} \tag{11}$$

which converges to the problem described in (9) for $\tau \to \infty$. The iterations steps of the algorithm are described below:

$$\begin{cases} \mathbf{X}^k = D_\tau(\mathbf{Y}^{k-1}), \\ \mathbf{Y}^k = \mathbf{Y}^{k-1} + \delta_k P_\Omega(\mathbf{M} - \mathbf{X}^k), \end{cases} \tag{12}$$

where the initialization point is chosen as $\mathbf{Y}^0 = \mathbf{0}$, δ_k is a sequence of positive step sizes, and the soft-thresholding operator, D_τ, is defined as follows. For a matrix $\mathbf{X} \in \mathbb{R}^{p\times n}$ of rank r with singular value decomposition given by

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \quad \mathbf{\Sigma} = \text{diag}(\{\sigma_i\}_{1\le i\le r}), \tag{13}$$

where $\mathbf{U}$ and $\mathbf{V}$ are matrices with orthogonal columns of size $p \times r$ and $n \times r$, respectively, and σ_i are the singular values of the matrix $\mathbf{X}$, the soft-thresholding operator is defined as

$$D_\tau(\mathbf{X}) := \mathbf{U}D_\tau(\mathbf{\Sigma})\mathbf{V}^T, \quad D_\tau(\mathbf{\Sigma}) = \text{diag}(\{(\sigma_i - \tau)_+\}), \tag{14}$$

where $t_+ = \max(0, t)$. That is, the operator applies a soft-thresholding rule to the singular values of $\mathbf{X}$, shrinking these towards zero. Large values of τ guarantee that the result is a low-rank matrix. However, for values that are larger than $\max(\sigma_i)$ the soft-thresholding operator vanishes all the singular values. Clearly, the choice of τ is important to guarantee a successful recovery. In [20], it is proposed to set the value of the threshold $\tau = 5n$ to let the term $\tau\|\mathbf{M}\|_*$ dominate the term $\frac{1}{2}\|\mathbf{M}\|_F^2$. Using standard random matrix theory, it can be shown that the Frobenius norm of $\mathbf{M}$ concentrates around $n\sqrt{r}$ and the nuclear norm concentrates around nr [20]. Therefore, setting $\tau = 5n$ guarantees that on the average, the value of $\tau\|\mathbf{M}\|_*$ is 10 times that of $\frac{1}{2}\|\mathbf{M}\|_F^2$ as long as the rank is bounded away from the dimension n.

IV. Performance limits

The performance of matrix completion based recovery using SVT is compared to three different performance limits. First, the distortion of the MMSE estimator with access to perfect second order statistics is studied. The performance of an optimal linear encoder (OLE) that operates with the same number of measurements is also assessed. Last, the information theoretic limit given by the optimal performance theoretically attainable (OPTA) is characterized.

A. Minimum Mean Squared Error

The MMSE estimation performance is given by

$$\text{MMSE}(\mathbf{M}|P_\Omega(\mathbf{M})) = \mathbb{E}\big[\|\mathbf{M} - \mathbb{E}[\mathbf{M}|P_\Omega(\mathbf{M}) + \mathbf{Z}, \mathbf{\Sigma}]\|_F^2\big], \tag{15}$$

where P_Ω is the sampling operator defined in Section III and $\mathbf{\Sigma}$ is the covariance matrix available to the operator. The performance of the MMSE estimator depends on the quality of $\mathbf{\Sigma}$. For a multivariate Gaussian source the MMSE distortion is given by:

$$\mathsf{D}_{\text{MMSE}} = \frac{1}{n}\text{Tr}(\mathbf{\Sigma}_{\Omega^c\Omega^c} - \mathbf{\Sigma}_{\Omega^c\Omega}\mathbf{\Sigma}_{\Omega\Omega}^{-1}\mathbf{\Sigma}_{\Omega\Omega^c}), \tag{16}$$

where Ω is the set of observed entries, Ω^c is the set of missing entries, $\mathbf{\Sigma}_{\Omega^c\Omega}$ is the cross-covariance matrix between the entries in Ω^c and the entries in Ω and $\mathbf{\Sigma}_{\Omega^c\Omega^c}$ is the auto-covariance matrix of the entries in Ω^c . Similarly, $\mathbf{\Sigma}_{\Omega\Omega^c}$ is the cross-covariance matrix between the entries in Ω and the entries in Ω^c and $\mathbf{\Sigma}_{\Omega\Omega}$ is the auto-covariance matrix of the entries in Ω.

B. Optimal Linear Encoder

In the case in which f is a linear transformation $\mathbf{P}$ and $\widetilde{\mathbf{Y}} = \mathbf{PM}$, the MMSE estimation performance is

$$\text{MMSE}(\mathbf{M}|\mathbf{PM}+\mathbf{Z}) = \mathbb{E}\big[\|\mathbf{M}-\mathbb{E}[\mathbf{M}|\mathbf{PM}+\mathbf{Z}, \mathbf{\Sigma}]\|_F^2\big]. \tag{17}$$

For Gaussian sources, the average distortion per sample is given by for any given linear projection matrix $\mathbf{P}$

$$\mathsf{D}_{\text{OLE}} = \frac{1}{n}\text{Tr}\big(\mathbf{\Sigma} - \mathbf{\Sigma}\mathbf{P}^T(\mathbf{P}\mathbf{\Sigma}\mathbf{P}^T + \sigma_z^2\mathbf{I}_m)^{-1}\mathbf{P}\mathbf{\Sigma}\big), \tag{18}$$

where $\mathbf{P} \in \mathbb{R}^{m\times p}$. The design of the optimal matrix $\mathbf{P}$ (in the MMSE sense) is described in [21].

C. Optimal Performance Theoretically Attainable

The optimal performance theoretically attainable is governed by the Rate-Distortion function of the distribution describing the state variables. The Rate-Distortion function determines the achievable distortion for a given number of observations. The trade-off between the number of available observations and the achievable distortion is determined by the rate-distortion function.

The Rate-Distortion function of a multivariate Gaussian source is given by the following parametric equations [22]

$$\begin{cases} R(\theta) &= \frac{1}{n}\sum_{i=0}^{n-1}\max(0, \frac{1}{2}\log\frac{\lambda_i}{\theta}) \\ \mathsf{D}(\theta) &= \frac{1}{n}\sum_{i=0}^{n-1}\min(\theta, \lambda_i), \end{cases} \tag{19}$$

where R is the source rate in nats/symbol, D is the mean squared error distortion per entry, λ_i is the ith largest eigenvalue of $\mathbf{\Sigma}$, and θ is a parameter.

Since the acquisition process introduces additive white Gaussian noise (AWGN) in the observations, the optimal performance theoretically attainable is given by

$$R(\mathsf{D}) < C, \tag{20}$$

where C is the capacity of the AWGN channel. Thus, the OPTA is given by

$$R(\mathsf{D}) \le \frac{m}{2pn}\log_{10}(1+\gamma), \tag{21}$$

where the signal to noise ratio, γ, is defined as:

$$\gamma = \frac{\frac{1}{n}\text{Tr}(\mathbf{\Sigma})}{\sigma_z^2}. \tag{22}$$

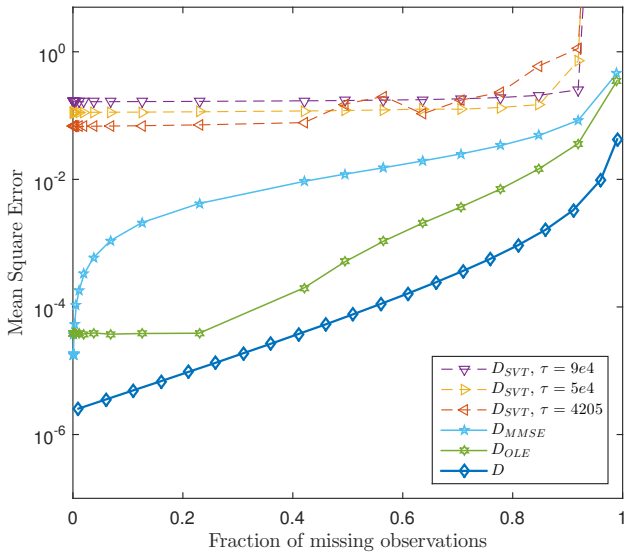

Fig. 4. Real data recovery performance using singular value thresholding, $\mathsf{D}_{\mathsf{SVT}}$, MMSE estimation, $\mathsf{D}_{\mathsf{MMSE}}$, MMSE estimation with the optimal linear encoder, $\mathsf{D}_{\mathsf{OLE}}$, and the OPTA, D, when SNR $= 20$ dB.

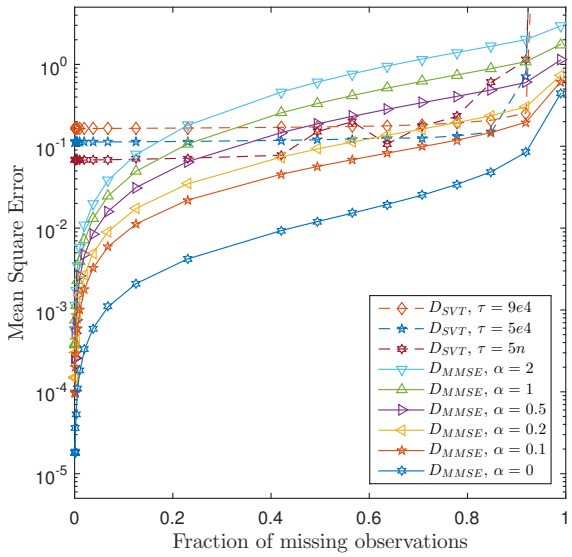

Fig. 5. Real data recovery performance using singular value thresholding, $\mathsf{D}_{\mathsf{SVT}}$, and MMSE estimation, $\mathsf{D}_{\mathsf{MMSE}}$, for different values of mismatch when SNR $= 20$ dB.

V. Numerical results

Different recovery techniques are numerically assessed using real data. To assess the recovery error only the complete files containing voltage state variables are used. The test matrix, $\mathbf{M}$, is 841×841 (841 measurements describing the state of the grid over a period of 3.5 hours from each of the 841 files). Each column contains voltage measurements describing the state of the grid in different days and for different feeders. The recovery of missing data based on SVT is evaluated for different values of τ. The value $\tau = 5n$ is proposed in [20] following the reasoning described in Section III, while other values of τ are obtained by numerical optimization. The performance of the SVT based recovery is defined in terms of the distortion of the error given by

$$\mathsf{D}_{\mathrm{SVT}} = \frac{1}{n^2}\|\mathbf{M} - \widehat{\mathbf{M}}\|_F^2. \tag{23}$$

Numerical results in this section are obtained for a logarithmic signal to noise ratio value of $10\log_{10}\gamma = 20$ dB.

A. Perfect prior knowledge

The case in which perfect knowledge of the second order statistics is available to the operator is studied. Figure 4 depicts the achievable distortion when SVT, MMSE estimation with perfect prior knowledge, OLE, and the OPTA are considered. Interestingly, SVT distortion is close to the optimal distortion achievable by MMSE estimation when the fraction of missing entries is greater than 0.9. Note that the SVT based recovery performs the closest to the fundamental limit right before the phase transition of the SVT approach takes place. This implies that operating in a regime in which the matrix completion approach is efficient imposes low robustness guarantees, i.e., the operating point is close to the phase transition.

B. Mismatched covariance matrix

In practical scenarios, postulated statistics available to the operator do not match the real statistics. To study this case, D^* is numerically assessed in the presence of different levels of mismatch.

In this case, the distortion of the estimator is given by $\mathsf{D}^* = \mathsf{D}_0 + \mathsf{D}_{\mathsf{A}}$, where D_0 is the distortion when perfect prior knowledge is available and D_{A} is the excess distortion incurred by the system in the mismatched case.

A Gaussian Wishart perturbation model is introduced to assess the performance of mismatched estimators [23], [24]. The postulated covariance matrix is given by:

$$\mathbf{\Sigma}^* = \mathbf{\Sigma} + \alpha\mathbf{A}, \tag{24}$$

where $\mathbf{\Sigma}^*$ is the postulated mismatched covariance matrix, $\mathbf{A} = \mathbf{H}\mathbf{H}^T$ with $\mathbf{H} \in \mathbb{R}^{n\times n}$, and the entries of $\mathbf{H}$ are distributed as $\mathcal{N}(0, n^{-1})$ so that $\frac{1}{n}\mathbb{E}[\mathrm{Tr}(\mathbf{A})] = 1$. The strength of the mismatch is determined by α.

Figure 5 shows the performance of the MMSE estimator for different levels of mismatch and the SVT based recovery for different threshold values as a function of the number of missing observations. It can be seen that SVT based recovery outperforms MMSE estimation in the moderate mismatch regime, i.e., $\alpha \geq 0.5$, for a wide range of the fraction of missing observations. Remarkably, the setting in which the SVT recovery outperforms MMSE estimation extends to the moderate missing data regime. Consequently, the SVT approach is the best performing strategy even when it operates away from the phase-transition point, which provides additional robustness guarantees, i.e., the number of missing observations can change without inducing catastrophic errors in the recovery process.

VI. Conclusion

This paper presents matrix completion using SVT as an alternative to MMSE estimation when the statistics of the data are not known perfectly. Using real data of a electricity distribution grid, the distortion introduced by MMSE estimation and SVT recovery is numerically assessed. The availability of second order statistics in a practical setting is modelled by considering access to a mismatched covariance matrix. It is numerically shown, that under source uncertainty, matrix completion recovery outperforms classical Bayesian estimation. However, an analysis of the information theoretic limits shows that better alternatives need to be devised when the number of missing observations is low. Still, the SVT recovery operates with minimum prior knowledge, i.e., the data matrix admits a low rank approximation. In contrast to that, MMSE estimation requires access to accurate second order statistics which is an unrealistic assumption in a real system. Therefore, matrix completion based recovery is a viable alternative for recovering missing samples in distribution grids.

References

[1] H. Maaß, H. K. Cakmak, F. Bach, R. Mikut, A. Harrabi, W. Süß, W. Jakob, K.-U. Stucky, U. G. Kühnapfel, and V. Hagenmeyer, "Data processing of high-rate low-voltage distribution grid recordings for smart grid monitoring and analysis," *EURASIP Journal on Advances in Signal Processing*, Dec. 2015.

[2] M. Ozay, I. Esnaola, F. T. Y. Vural, S. R. Kulkarni, and H. V. Poor, "Sparse attack construction and state estimation in the smart grid: Centralized and distributed models," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 7, Jul. 2013.

[3] A. Tajer, S. Kar, H. V. Poor, and S. Cui, "Distributed joint cyber attack detection and state recovery in smart grids," in *Proc. of the IEEE International Conference on Smart Grid Communications*, Oct. 2011.

[4] T. T. Kim and H. V. Poor, "Strategic protection against data injection attacks on power grids," *IEEE Trans. Smart Grid*, Jun. 2011.

[5] O. Kosut, L. Jia, R. J. Thomas, and L. Tong, "Malicious data attacks on smart grid state estimation: Attack strategies and countermeasures," in *Proc. of the First IEEE International Conference on Smart Grid Communications*, Oct. 2010, pp. 220–225.

[6] I. Esnaola, A. Tulino, and J. Garcia-Frias, "Linear analog coding of correlated multivariate Gaussian sources," *IEEE Trans. Commun.*, vol. 61, no. 8, pp. 3438–3447, Aug. 2013.

[7] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 3, pp. 471–501, Aug. 2010.

[8] Z. Liu and L. Vandenberghe, "Interior-point method for nuclear norm approximation with application to system identification," *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 3, pp. 1235–1256, Nov. 2009.

[9] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, Apr. 2009.

[10] B. Recht, "A simpler approach to matrix completion," *J. Mach. Learn. Res.*, pp. 3413–3430, Dec. 2011.

[11] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inf. Theory*, pp. 2053–2080, May 2010.

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

[13] R. H. Keshavan, S. Oh, and A. Montanari, "Matrix completion from a few entries," in *Proc of the IEEE International Symposium on Information Theory*, Jun. 2009, pp. 324–328.

[14] M. Wang, J. H. Chow, P. Gao, X. T. Jiang, Y. Xia, S. G. Ghiocel, B. Fardanesh, G. Stefopolous, Y. Kokai, N. Saito, and M. Razanousky, "A low-rank matrix approach for the analysis of large amounts of power system synchrophasor data," in *Proc. of the Hawaii International Conference on System Sciences*, Hawaii, USA, Jan. 2015, pp. 2637–2644.

[15] V. Kekatos, Y. Zhang, and G. B. Giannakis, "Electricity market forecasting via low-rank multi-kernel learning," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 6, pp. 1182–1193, Dec. 2014.

[16] 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.*, pp. 1006–1013, Mar. 2016.

[17] A. Chistov and D. Grigoriev, "Complexity of quantifier elimination in the theory of algebraically closed fields," *in Proc. of the Symposium on Mathematical Foundation of Computer Science*, Sep. 1984.

[18] J.F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optimization Methods & Software*, vol. 11-2, no. 1-4, SI, pp. 625–653, 1999.

[19] K.C. Toh, M.J. Todd, and R.H. Tutuncu, "SDPT3 - A MATLAB software package for semidefinite programming, version 1.3," *Optimization Methods & Software*, vol. 11-2, no. 1-4, SI, pp. 545–581, 1999.

[20] 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.

[21] K.-H. Lee and D. Petersen, "Optimal linear coding for vector channels," *IEEE Trans. Commun.*, vol. 24, no. 12, pp. 1283–1290, Dec. 1976.

[22] A. Kolmogorov, "On the Shannon theory of information transmission in the case of continuous signals," *IRE Transactions on Information Theory*, vol. 2, no. 4, pp. 102–108, Dec. 1956.

[23] I. Esnaola, A. M. Tulino, and H. V. Poor, "Mismatched mmse estimation of multivariate Gaussian sources," in *Proc. of the IEEE International Symposium on Information Theory*, Jul. 2012, pp. 716–720.

[24] S. Verdu, "Mismatched estimation and relative entropy," *IEEE Trans. Inf. Theory*, pp. 3712–3720, Aug. 2010.