Spatiotemporal Load Curve Data Cleansing and Imputation via Sparsity and Low Rank

Gonzalo Mateos and Georgios B. Giannakis
Dept. of ECE and Digital Technology Center
University of Minnesota
Minneapolis, MN 55455
Emails: {mate0058,georgios}@umn.edu

***Abstract*—The smart grid vision is to build an intelligent power network with an unprecedented level of situational awareness and controllability over its services and infrastructure. This paper advocates statistical inference methods to robustify power monitoring tasks against the outlier effects owing to faulty readings and malicious attacks, as well as against missing data due to privacy concerns and communication errors. In this context, a novel *load cleansing and imputation* scheme is developed leveraging the low intrinsic-dimensionality of spatiotemporal load profiles and the sparse nature of "bad data." A robust estimator based on principal components pursuit (PCP) is adopted, which effects a twofold sparsity-promoting regularization through an ℓ_1-norm of the outliers, and the nuclear norm of the nominal load profiles. After recasting the non-separable nuclear norm into a form amenable to distributed optimization, a *distributed* (D-) PCP algorithm is developed to carry out the imputation and cleansing tasks using a network of interconnected smart meters. Computer simulations and tests with real load curve data corroborate the convergence and effectiveness of the novel D-PCP algorithm.**

I. Introduction

The US power infrastructure has been recognized as the most important engineering achievement of the 20th century [18], yet it presently faces major challenges related to efficiency, reliability, security, environmental impact, sustainability, and market diversity issues [17]. The crystallizing vision of the smart grid (SG) aspires to build a cyber-physical network that can address such challenges by capitalizing on state-of-the-art information technologies in sensing, control, communication, optimization, and machine learning. Significant effort and investment have been committed to architect the necessary infrastructure by installing advanced metering systems, and establishing data communication networks throughout the grid. Accordingly, algorithms that optimally exploit such pervasive sensing and control capabilities are needed to make the necessary breakthroughs in the key problems in power grid monitoring and energy management. This is no easy endeavor though, in view of the challenges posed by increasingly distributed network operations under strict reliability requirements, also facing malicious cyber-attacks.

Statistical inference techniques are expected to play an increasingly instrumental role in power system monitoring, not only to meet the anticipated "big data" deluge as the installed base of phasor measurement units (PMUs) reaches out throughout the network, but also to robustify the monitoring tasks against "outlier" effects owing to faulty readings, malicious attacks, and communication errors, as well as against missing data due to privacy concerns and technical anomalies. In this context, a novel *load cleansing and imputation* scheme is developed in this paper, building on recent advances in sparsity-cognizant signal processing, low-rank matrix completion, and large-scale distributed optimization.

Load curve data refers to the electric energy consumption periodically recorded by meters at points of interest across the electric grid, e.g., end-user premises, buses, and substations. Accurate load profiles are critical assets aiding operational decisions in the envisioned SG system [5]. However, in the process of acquiring and transmitting such massive volumes of information, data are oftentimes corrupted or lost altogether. In a smart monitoring context for instance, incomplete load profiles emerge due to three reasons: (r1) PMU-instrumented buses are few; (r2) SCADA data become available at a considerably smaller time scale than PMU data; and (r3) regional operators are not willing to share all their variables [15].

In addition to dealing with missing data, a major requirement for grid monitoring is robustness to outliers, i.e., data not adhering to nominal models. Sources of so-termed "bad data" include meter failures, as well as strikes, unscheduled generator shutdowns, and extreme weather conditions [5]. Inconsistent data can also come from malicious (cyber-)attacks, or counterfeit meter readings [9].

In light of the aforementioned observations, the first contribution of this paper is on modeling spatiotemporal load profiles, accounting for the structure present to effectively impute missing data and devise robust load curve estimators stemming from convex optimization criteria (Section II). The aim is for *minimal-rank* cleansed load data, while also exploiting outlier *sparsity* across buses and time. An estimator tailored to these specifications is principal components pursuit (PCP, also known as robust principal component analysis) [3], [4], [20], which is outlined in Section III. While PCP has been widely adopted in computer vision, for voice separation in music, and unveiling network anomalies, its benefits to power systems engineering and monitoring remains largely unexplored. The second contribution pertains to developing a *distributed* PCP algorithm, to carry out the imputation and cleansing task using a network of interconnected smart meters (Section IV). This is possible by leveraging a general algorithmic framework for sparsity-regularized rank minimization put forth in [10].

978-1-4577-1703-1/12/$26.00 ©2012 IEEE

II. Modeling and Problem Statement

This section introduces the fundamental model for (possibly) incomplete and grossly corrupted load curve measurements, acquired by geographically-distributed meters monitoring the power grid. The communication network model needed to account for exchanges of information among smart meters is described as well. Lastly, the problem of load curve cleansing and imputation is formally stated.

A. Spatiotemporal load curve data model

Let the $N \times 1$ vector $\mathbf{y}(t) := [y_{1,t}, \ldots, y_{N,t}]'$ ($'$ stands for transposition) collect the spatial load profiles measured by smart meters monitoring N network nodes (buses or residential premises), at a given discrete-time instant $t \in [1, T]$. Consider the $N \times T$ matrix of observations $\mathbf{Y} := [\mathbf{y}(1), \ldots, \mathbf{y}(T)]$. The n-th row of $\mathbf{Y}$ is the time series of energy consumption (load curve) measurements at node n, while the t-th column of $\mathbf{Y}$ represents a snapshot of the networkwide loads taken at time t. To model missing data, consider the set $\Omega \subseteq \{1, \ldots, N\} \times \{1, \ldots, T\}$ of index pairs (n, t) defining a sampling of the entries of $\mathbf{Y}$. Introducing the matrix sampling operator $\mathcal{P}_\Omega(\cdot)$, which sets the entries of its matrix argument not indexed by Ω to zero and leaves the rest unchanged, the (possibly) incomplete spatiotemporal load curve data in the presence of outliers can be modeled as

$$\mathcal{P}_\Omega(\mathbf{Y}) = \mathcal{P}_\Omega(\mathbf{X} + \mathbf{O} + \mathbf{E}) \tag{1}$$

where $\mathbf{X}$, $\mathbf{O}$, and $\mathbf{E}$ denote the nominal load profiles, the outliers, and small measurement errors, respectively. For nominal observations $y_{n,t} = x_{n,t} + e_{n,t}$, one has $o_{n,t} = 0$. Note that the model is inherently under-determined, since even for the (most favorable) case of full data, i.e., $\Omega \equiv \{1, \ldots, N\} \times \{1, \ldots, T\}$, there are twice as many unknowns in $\mathbf{X}$ and $\mathbf{O}$ as there is data in $\mathbf{Y}$. Estimating $\mathbf{X}$ and $\mathbf{O}$ becomes even more challenging when data are missing, since the number of unknowns remains the same, but the amount of data is reduced. In any case, estimation of $\{\mathbf{X}, \mathbf{O}\}$ from $\mathcal{P}_\Omega(\mathbf{Y})$ is an ill-posed problem unless one introduces extra structural assumptions on the model components to reduce the effective degrees of freedom.

Two cardinal properties of $\mathbf{X}$ and $\mathbf{O}$ are worth noting. First, common temporal patterns among the energy consumption of a few broad classes of loads (e.g., industrial, residential, seasonal) in addition to their (almost) periodic behaviors render most rows and columns of $\mathbf{X}$ linearly dependent, and thus $\mathbf{X}$ typically has *low-rank*. Second, outliers (or attacks) only occur sporadically in time and affect only a few buses, yielding a *sparse* matrix $\mathbf{O}$. Smoothness of the nominal load curves is a structural assumption somehow related to the low-rank property of $\mathbf{X}$, which was adopted in [5] to motivate a smoothing splines-based algorithm for cleansing. Approaches capitalizing on outlier- and "bad data-" sparsity can be found in e.g., [7] and [11].

B. Communication network model

Suppose that on top of the energy measurement functionality, the N networked smart meters are capable of performing simple local computations, as well as exchanging messages among directly connected neighbors. Single-hop communication models are appealing due to their simplicity, since one does not have to account for the routing. The network is naturally abstracted to an undirected graph $G(\mathcal{N}, \mathcal{L})$, where the vertex set $\mathcal{N} := \{1, \ldots, N\}$ corresponds to the network nodes, and the edges (links) in $\mathcal{L}$ represent pairs of nodes that are connected via a physical communication channel. Node $n \in \mathcal{N}$ communicates with its single-hop neighboring peers in $\mathcal{J}_n$, and the size of the neighborhood will be henceforth denoted by $|\mathcal{J}_n|$. The graph G is assumed connected, i.e., there exists a (possibly multihop) path that joins any pair of nodes in the network. This requirement ensures that the network is devoid of multiple isolated (connected) components, and allows for the data collected by e.g., smart meter n, namely the n-th row $(\mathbf{y}_n)'$ of $\mathbf{Y}$, to eventually reach every other node in the network. This way, even when only local interactions are allowed, the flow of information can propagate and eventually impact global network behaviors.

The importance of the network model will become apparent in Section IV.

C. Load curve cleansing and imputation

The load curve cleansing and imputation problem studied here entails identification and removal of outliers (or "bad data"), in addition to completion of the missing entries from the nominal load matrix, and denoising of the observed ones. To some extent, it is a joint estimation-interpolation (prediction)-detection problem. With reference to (1), given generally incomplete, noisy and outlier-contaminated spatiotemporal load data $\mathcal{P}_\Omega(\mathbf{Y})$, the cleansing and imputation tasks amount to estimating the nominal load profiles $\mathbf{X}$ and the outliers $\mathbf{O}$, by leveraging the low-rank property of $\mathbf{X}$ and the sparsity in $\mathbf{O}$.

Note that load cleansing and imputation are different from *load forecasting* [16], which amounts to predicting future load demand based on historical data of energy consumption and the weather conditions. Actually, cleansing and imputation are critical preprocessing tasks utilized to enhance the quality of load data, that would be subsequently used for load forecasting and optimum power flow [1].

III. Prinicipal Components Pursuit

An estimator matching nicely the specifications of the load curve cleansing and imputation problem stated in Section II-C, is the so-termed (stable) principal components pursuit (PCP) [3], [4], [20], that will be outlined here for completeness. PCP seeks the estimates $\{\hat{\mathbf{X}}, \hat{\mathbf{O}}\}$ as the minimizers of

$$\text{(P1)} \qquad \min_{\{\mathbf{X},\mathbf{O}\}} \frac{1}{2} \left\| \mathcal{P}_\Omega(\mathbf{Y} - \mathbf{X} - \mathbf{O}) \right\|_F^2 + \lambda_* \left\| \mathbf{X} \right\|_* + \lambda_1 \left\| \mathbf{O} \right\|_1$$

where the ℓ_1-norm $\|\mathbf{O}\|_1 := \sum_{n,t} |o_{n,t}|$ and the nuclear norm $\|\mathbf{X}\|_* := \sum_i \sigma_i(\mathbf{X})$ ($\sigma_i(\mathbf{X})$ denotes the i-th singular value of $\mathbf{X}$) are utilized to promote sparsity in the number of anomalies (nonzero entries) in $\mathbf{O}$, and the low rank of $\mathbf{X}$, respectively. The nuclear norm and ℓ_1-norm are the closest

convex surrogates to the rank and cardinality functions, which albeit the most natural criteria they are in general NP-hard to optimize [13], [6]. The tuning parameters $\lambda_1, \lambda_* \geq 0$ control the tradeoff between fitting error, rank, and sparsity level of the solution. When an estimate $\hat{\sigma}_v^2$ of the observation noise variance is available, guidelines for selecting λ_* and λ_1 have been proposed in [20]. The nonzero entries in $\hat{\mathbf{O}}$ reveal "bad data" across both buses and time.

Being convex (P1) is computationally appealing, and it has been demonstrated to attain good performance in theory and practice. For instance, in the absence of noise and when there is no missing data, remarkable exact recovery results were reported in [3] and [4]. Even when data are missing, it is possible to recover the low-rank component under some technical assumptions [3]. Theoretical performance guarantees in the presence of noise are also available [20].

Regarding algorithms, a PCP solver based on the accelerated proximal gradient method was put forth in [8], while the alternating-directions method of multipliers was employed in [19]. Implementing these *centralized* algorithms presumes that networked metering devices continuously communicate their local load measurements $y_{n,t}$ to a central monitoring and data analytics station, which uses their aggregation in $\mathcal{P}_\Omega(\mathbf{Y})$ to reject outliers and impute missing data. While for the most part this is the prevailing operational paradigm nowadays, it is fair to say there are limitations associated with this architecture. For instance, collecting all this information centrally may lead to excessive overhead in the communication network, especially when the rate of data acquisition is high at the meters. Moreover, minimizing the exchanges of raw measurements may be desirable to reduce unavoidable communication errors that translate to missing data. Performing the optimization in a centralized fashion raises robustness concerns as well, since the central data analytics station represents an isolated point of failure. These reasons motivate devising fully-distributed iterative algorithms for PCP, embedding the load cleansing and imputation functionality to the smart meters. This is the subject of the next section.

IV. In-network Cleansing and Imputation

A distributed (D-)PCP algorithm to solve (P1) using a network of smart meters (modeled as in Section II-B) should be understood as an iterative method, whereby each node carries out simple local (optimization) tasks per iteration $k = 1, 2, \ldots$, and exchanges messages only with its directly connected neighbors. The ultimate goal is for each node to form local estimates $\mathbf{x}_n[k]$ and $\mathbf{o}_n[k]$ that coincide with the n-th rows of $\hat{\mathbf{X}}$ and $\hat{\mathbf{O}}$ as $k \to \infty$, where $\{\hat{\mathbf{X}}, \hat{\mathbf{O}}\}$ is the solution of (P1) obtained when all data $\mathcal{P}_\Omega(\mathbf{Y})$ are centrally available. Attaining the centralized performance with distributed data is impossible if the network is disconnected.

To facilitate reduction of the computational complexity and memory storage requirements of the D-PCP algorithm sought, it is henceforth assumed that an upper bound $\text{rank}(\hat{\mathbf{X}}) \leq \rho$ is a priori available [recall $\hat{\mathbf{X}}$ is the estimated low-rank cleansed load profile obtained via (P1)]. As argued next, the smaller the value of ρ, the more efficient the algorithm becomes. Small values of ρ are well motivated due to the low intrinsic dimensionality of the spatiotemporal load profiles (cf. Section II-A). Because $\text{rank}(\hat{\mathbf{X}}) \leq \rho$, (P1)'s search space is effectively reduced and one can factorize the decision variable as $\mathbf{X} = \mathbf{P}\mathbf{Q}'$, where $\mathbf{P}$ and $\mathbf{Q}$ are $N \times \rho$ and $T \times \rho$ matrices, respectively. Adopting this reparametrization of $\mathbf{X}$ in (P1) and making explicit the distributed nature of the data (cf. Section II-B), one arrives at an equivalent optimization problem

$$\text{(P2)} \quad \min_{\{\mathbf{P},\mathbf{Q},\mathbf{O}\}} \sum_{n=1}^{N} \left[\frac{1}{2} \|\mathcal{P}_{\Omega_n}(\mathbf{y}_n - \mathbf{Q}\mathbf{p}_n - \mathbf{o}_n)\|_2^2 + \frac{\lambda_*}{N} \|\mathbf{P}\mathbf{Q}'\|_* + \lambda_1 \|\mathbf{o}_n\|_1 \right]$$

which is non-convex due to the bilinear term $\mathbf{P}\mathbf{Q}'$, and where $\mathbf{P} := [\mathbf{p}_1, \ldots, \mathbf{p}_N]'$. The number of variables is reduced from $2NT$ in (P1), to $\rho(N+T) + NT$ in (P2). The savings can be significant when ρ is small, and both N and T are large. Note that the dominant NT-term in the variable count of (P2) is due to $\mathbf{O}$, which is sparse and can be efficiently handled even when both N and T are large.

Problem (P2) is still not amenable for distributed implementation due to: (c1) the non-separable nuclear norm present in the cost function; and (c2) the global variable $\mathbf{Q}$ coupling the per-node summands. These two challenges are dealt with in the ensuing sub-sections.

A. A separable low-rank regularization

To address (c1), consider the following alternative characterization of the nuclear norm (see e.g. [14])

$$\|\mathbf{X}\|_* := \min_{\{\mathbf{P},\mathbf{Q}\}} \quad \frac{1}{2}\left(\|\mathbf{P}\|_F^2 + \|\mathbf{Q}\|_F^2\right), \quad \text{s. to} \quad \mathbf{X} = \mathbf{P}\mathbf{Q}'. \tag{2}$$

The optimization (2) is over all possible bilinear factorizations of $\mathbf{X}$, so that the number of columns ρ of $\mathbf{L}$ and $\mathbf{Q}$ is also a variable. Leveraging (2), the following reformulation of (P2) provides an important first step towards obtaining the D-PCP algorithm:

$$\text{(P3)} \quad \min_{\{\mathbf{P},\mathbf{Q},\mathbf{O}\}} \sum_{n=1}^{N} \left[\frac{1}{2} \|\mathcal{P}_{\Omega_n}(\mathbf{y}_n - \mathbf{Q}\mathbf{p}_n - \mathbf{o}_n)\|_2^2 + \lambda_1 \|\mathbf{o}_n\|_1 + \frac{\lambda_*}{2N}\left(N\|\mathbf{p}_n\|_2^2 + \|\mathbf{Q}\|_F^2\right) \right].$$

As asserted in [10, Lemma 1], adopting the separable Frobenius-norm regularization in (P3) comes with no loss of optimality relative to (P1), provided $\text{rank}(\hat{\mathbf{X}}) \leq \rho$. By finding the global minimum of (P3) [which could have considerably less variables than (P1)], one can recover the optimal solution of (P1). However, since (P3) is non-convex, it may have stationary points which need not be globally optimum. Interestingly, the next proposition shows that under relatively mild assumptions on $\text{rank}(\hat{\mathbf{X}})$ every stationary point of (P3) is globally optimum for (P1). For a proof (omitted here due to space limitations), see [10, App. A].

Proposition 1: *Let* $\{\bar{\mathbf{P}}, \bar{\mathbf{Q}}, \bar{\mathbf{O}}\}$ *be a stationary point of (P3). If* $\|\mathcal{P}_\Omega(\mathbf{Y} - \bar{\mathbf{P}}\bar{\mathbf{Q}}' - \bar{\mathbf{O}})\| < \lambda_*$*, then* $\{\hat{\mathbf{X}} := \bar{\mathbf{P}}\bar{\mathbf{Q}}', \hat{\mathbf{O}} := \bar{\mathbf{O}}\}$ *is the globally optimal solution of (P1).*

The qualification condition $\|\mathcal{P}_\Omega(\mathbf{Y} - \bar{\mathbf{P}}\bar{\mathbf{Q}}' - \bar{\mathbf{O}})\| < \lambda_*$ captures tacitly the role of ρ. In particular, for sufficiently small ρ the residual $\|\mathcal{P}_\Omega(\mathbf{Y} - \bar{\mathbf{P}}\bar{\mathbf{Q}}' - \bar{\mathbf{O}})\|$ becomes large and consequently the condition is violated [unless λ_* is large enough, in which case a sufficiently low-rank solution to (P1) is expected]. The condition on the residual also implicitly enforces $\text{rank}(\hat{\mathbf{X}}) \leq \rho$, which is necessary for the equivalence between (P1) and (P3).

B. Local variables and consensus constraints

To decompose the cost in (P3), in which summands inside the square brackets are coupled through the global variable $\mathbf{Q}$ [cf. (c2) at the beginning of Section IV], introduce auxiliary variables $\{\mathbf{Q}_n\}_{n=1}^N$ representing local estimates of $\mathbf{Q}$, one per smart meter n. To obtain a separable PCP formulation, use these estimates along with *consensus* constraints, that is

$$\text{(P4)} \quad \min_{\{\mathbf{P}_n, \mathbf{Q}_n, \mathbf{O}\}} \sum_{n=1}^N \Big[\frac{1}{2}\|\mathcal{P}_{\Omega_n}(\mathbf{y}_n - \mathbf{Q}_n\mathbf{p}_n - \mathbf{o}_n)\|_2^2 + \lambda_1\|\mathbf{o}_n\|_1 + \frac{\lambda_*}{2N}\left(N\|\mathbf{p}_n\|_2^2 + \|\mathbf{Q}_n\|_F^2\right)\Big] \quad \text{s. to} \quad \mathbf{Q}_n = \mathbf{Q}_m, \quad m \in \mathcal{J}_n,\ n \in \mathcal{N}.$$

Notice that (P3) and (P4) are equivalent optimization problems, since the network graph $G(\mathcal{N}, \mathcal{L})$ is connected by assumption. Even though consensus is a fortiori imposed within neighborhoods, it extends to the whole (connected) network and local estimates agree on the global solution of (P3). To arrive at the desired D-PCP algorithm, it is convenient to reparametrize the consensus constraints in (P4) as

$$\mathbf{Q}_n = \bar{\mathbf{F}}_n^m,\ \mathbf{Q}_m = \tilde{\mathbf{F}}_n^m,\ \text{and}\ \bar{\mathbf{F}}_n^m = \tilde{\mathbf{F}}_n^m, \quad m \in \mathcal{J}_n,\ n \in \mathcal{N} \tag{3}$$

where $\{\bar{\mathbf{F}}_n^m, \tilde{\mathbf{F}}_n^m\}_{n\in\mathcal{N}}$, are auxiliary optimization variables that will be eventually eliminated (cf. Remark 1).

C. The D-PCP algorithm

To tackle (P4), associate Lagrange multipliers $\bar{\mathbf{M}}_n^m$ and $\tilde{\mathbf{M}}_n^m$ with the first pair of consensus constraints in (3). Introduce the quadratically *augmented* Lagrangian function [2]

$$\begin{aligned}\mathcal{L}_c\left(\mathcal{V}_1, \mathcal{V}_2, \mathcal{V}_3, \mathcal{M}\right) =& \sum_{n=1}^N \Big[\frac{1}{2}\|\mathcal{P}_{\Omega_n}(\mathbf{y}_n - \mathbf{Q}_n\mathbf{p}_n - \mathbf{o}_n)\|_2^2 \\ &+\lambda_1\|\mathbf{o}_n\|_1 + \frac{\lambda_*}{2N}\left(N\|\mathbf{p}_n\|_2^2 + \|\mathbf{Q}_n\|_F^2\right)\Big] \\ &+ \sum_{n=1}^N \sum_{m\in\mathcal{J}_n} \Big(\langle \bar{\mathbf{M}}_n^m, \mathbf{Q}_n - \bar{\mathbf{F}}_n^m\rangle + \langle \tilde{\mathbf{M}}_n^m, \mathbf{Q}_m - \tilde{\mathbf{F}}_n^m\rangle\Big) \\ &+ \frac{c}{2}\sum_{n=1}^N \sum_{m\in\mathcal{J}_n} \Big(\|\mathbf{Q}_n - \bar{\mathbf{F}}_n^m\|_F^2 + \|\mathbf{Q}_m - \tilde{\mathbf{F}}_n^m\|_F^2\|_F^2\Big)\end{aligned} \tag{4}$$

where $c > 0$ is a penalty parameter, and the primal variables are split into groups $\mathcal{V}_1 := \{\mathbf{Q}_n\}_{n=1}^N$, $\mathcal{V}_2 := \{\mathbf{p}_n\}_{n=1}^N$ and $\mathcal{V}_3 := \{\mathbf{o}_n, \bar{\mathbf{F}}_n^m\ , \tilde{\mathbf{F}}_n^m\}_{n\in\mathcal{N}}^{m\in\mathcal{J}_n}$. For notational brevity, collect all Lagrange multipliers in $\mathcal{M} := \{\bar{\mathbf{M}}_n^m, \tilde{\mathbf{M}}_n^m\}_{n\in\mathcal{N}}^{m\in\mathcal{J}_n}$. Note that the remaining constraints in (3), namely $C_V := \{\bar{\mathbf{F}}_n^m = \tilde{\mathbf{F}}_n^m,\ m \in \mathcal{J}_n,\ n \in \mathcal{N}\}$, have not been dualized.

To minimize (P4) in a distributed fashion, the alternating-direction method of multipliers (AD-MoM) will be adopted here. The AD-MoM is an iterative augmented Lagrangian method especially well suited for parallel processing [2], which has been proven successful to handle optimization tasks stemming from general distributed estimators of deterministic and (non-)stationary random signals; see e.g., [12]. The proposed solver entails an iterative procedure comprising four steps per iteration $k = 1, 2, \ldots$, which amount to a block-coordinate descent method cycling over $\mathcal{V}_1 \to \mathcal{V}_2 \to \mathcal{V}_3$ to minimize $\mathcal{L}_c$, and dual variable updates [10]. At each step while minimizing the augmented Lagrangian, the variable groups not being updated are treated as fixed, and are substituted with their most up to date values.

Reformulating the estimator (P1) to its equivalent form (P4) renders the augmented Lagrangian in (4) highly decomposable. The separability comes in two flavors, both with respect to the variable groups $\mathcal{V}_1$-$\mathcal{V}_3$, as well as across the network nodes $n \in \mathcal{N}$. This leads to highly parallelized, simplified recursions to be run by the networked smart meters. Specifically, it is shown in [10, App. B] that application of AD-MoM yields the D-PCP algorithm tabulated as Algorithm 1. Per iteration, each smart meter updates: [S1] a local matrix of dual prices $\mathbf{S}_n[k]$; [S2]-[S3] local cleansed load estimates $\mathbf{Q}_n[k+1]$ and $\mathbf{p}_n[k+1]$ obtained as solutions to respective unconstrained quadratic problems; and [S4] its local outlier vector, through a sparsity-promoting soft-thresholding operation. The $(k+1)$-st iteration is concluded after smart meter n transmits $\mathbf{Q}_n[k+1]$ to its single-hop neighbors in $\mathcal{J}_n$. Regarding communication cost, $\mathbf{Q}_n[k+1]$ is a $T \times \rho$ matrix and its transmission does not incur significant overhead for small values of ρ. Observe that the dual variables need not be exchanged, and the communication cost does not depend on the size of the network N.

Before moving on, a clarification on the notation used in Algorithm 1 is due. To define matrix $\mathbf{\Omega}_n$ in [S2]-[S4], observe first that the local sampling operator can be expressed as $\mathcal{P}_{\Omega_n}(\mathbf{z}) = \boldsymbol{\omega}_n \odot \mathbf{z}$, where $\odot$ denotes Hadamard product, and the binary masking vector $\boldsymbol{\omega}_n \in \{0, 1\}^T$ has entries equal to 1 if the corresponding entry of $\mathbf{z}$ is observed, and 0 otherwise. It is then apparent that the Hadamard product can be replaced with the usual matrix-vector product as $\mathcal{P}_{\Omega_n}(\mathbf{z}) = \mathbf{\Omega}_n\mathbf{z}$, where $\mathbf{\Omega}_n := \text{diag}(\boldsymbol{\omega}_n)$. Operators $\otimes$ and $\text{vec}[\cdot]$ denote Kronecker product and matrix vectorization, respectively. Finally, the soft-thresholding operator is $\mathcal{S}_{\lambda_1}(\cdot) := \text{sign}(\cdot)\max(|\cdot| - \lambda_1, 0)$.

Remark 1 (Elimination of redundant variables): Careful inspection of Algorithm 1 reveals that the redundant auxiliary variables $\{\bar{\mathbf{F}}_n^m, \tilde{\mathbf{F}}_n^m, \tilde{\mathbf{M}}_n^m\}_{n\in\mathcal{N}}^{m\in\mathcal{J}_n}$ have been eliminated. Each smart meter, say the n-th, does not need to *separately* keep track of all its non-redundant multipliers $\{\bar{\mathbf{M}}_n^m\}_{m\in\mathcal{J}_n}$, but only update their respective (scaled) sums $\mathbf{S}_n[k] := 2\sum_{m\in\mathcal{J}_n}$

Algorithm 1 : D-PCP at smart meter $n \in \mathcal{N}$

input $\mathbf{y}_n, \mathbf{\Omega}_n, \lambda_*, \lambda_1$, and c.
initialize $\mathbf{S}[0] = \mathbf{0}_{T\times\rho}$, and $\mathbf{Q}_n[1]$, $\mathbf{p}_n[1]$ at random.
for $k = 1, 2, \ldots$ **do**
Receive $\{\mathbf{Q}_m[k]\}$ from neighbors $m \in \mathcal{J}_n$.
[S1] Update local dual variables:
$\mathbf{S}_n[k] = \mathbf{S}_n[k-1] + c\sum_{m\in\mathcal{J}_n}(\mathbf{Q}_n[k] - \mathbf{Q}_m[k])$.
[S2] Update first group of local primal variables:
$$\mathbf{Q}_n[k+1] = \text{unvec}\Big(\{(\mathbf{p}_n[k]\mathbf{p}_n'[k]) \otimes \mathbf{\Omega}_n + (\lambda_*/N + 2c|\mathcal{J}_n|)\mathbf{I}_{\rho T}\}^{-1} \times \Big\{(\mathbf{p}_n[k] \otimes \mathbf{\Omega}_n)(\mathbf{y}_n - \mathbf{o}_n[k]) - \text{vec}(\mathbf{S}_n[k]) + \text{vec}(c\textstyle\sum_{m\in\mathcal{J}_n}(\mathbf{Q}_n[k] + \mathbf{Q}_m[k]))\Big\}\Big).$$
[S3] Update second group of local primal variables:
$\mathbf{p}_n[k+1] = \{\mathbf{Q}_n'[k+1]\mathbf{\Omega}_n\mathbf{Q}_n[k+1] + \mathbf{I}_\rho\}^{-1}\mathbf{Q}_n'[k+1](\mathbf{y}_n - \mathbf{o}_n[k])$.
[S4] Update third group of local primal variables:
$\mathbf{o}_n[k+1] = \mathcal{S}_{\lambda_1}(\mathbf{\Omega}_n(\mathbf{y}_n - \mathbf{Q}_n[k+1]\mathbf{p}_n[k+1]))$.
Transmit $\mathbf{Q}_n[k+1]$ to neighbors $m \in \mathcal{J}_n$.
end for
return $\mathbf{Q}_n[\infty], \mathbf{p}_n[\infty], \mathbf{o}_n[\infty]$.

$\bar{\mathbf{M}}_n^m[k]$.

When employed to solve non-convex problems such as (P4), AD-MoM offers no convergence guarantees. However, there is ample experimental evidence in the literature which supports convergence of AD-MoM, especially when the non-convex problem at hand exhibits "favorable" structure. For instance, (P4) is bi-convex and gives rise to the strictly convex optimization subproblems each time $\mathcal{L}_c$ is minimized with respect to one of the group variables, which admit unique closed-form solutions per iteration (cf. [S2]-[S4] in Algorithm 1). This observation and the linearity of the constraints suggest good convergence properties for the D-PCP algorithm – extensive numerical tests including those presented in Section V demonstrate that this is indeed the case. While a formal convergence proof is subject of ongoing investigation, the following proposition proved in [10] asserts that upon convergence, the D-PCP algorithm attains consensus and global optimality.

Proposition 2: *Suppose iterates* $\{\mathbf{Q}_n[k], \mathbf{p}_n[k], \mathbf{o}_n[k]\}_{n\in\mathcal{N}}$ *generated by Algorithm 1 converge to* $\{\bar{\mathbf{Q}}_n, \bar{\mathbf{p}}_n, \bar{\mathbf{o}}_n\}_{n\in\mathcal{N}}$. *If* $\{\hat{\mathbf{X}}, \hat{\mathbf{O}}\}$ *is the optimal solution of (P1), then* $\bar{\mathbf{Q}}_1 = \bar{\mathbf{Q}}_2 = \ldots = \bar{\mathbf{Q}}_N$. *Also, if* $\|\mathcal{P}_\Omega(\mathbf{Y} - \bar{\mathbf{P}}\bar{\mathbf{Q}}_1' - \bar{\mathbf{O}})\| < \lambda_*$, *then* $\{\hat{\mathbf{X}} = \bar{\mathbf{P}}\bar{\mathbf{Q}}_1', \hat{\mathbf{O}} = \bar{\mathbf{O}}\}$.

V. Numerical Tests

This section corroborates convergence and gauges performance of the D-PCP algorithm, when tested using synthetic and real load curve data.

A. Synthetic data test

A network of $N = 25$ smart meters is generated as a realization of the random geometric graph model, meaning nodes are randomly placed on the unit square and two nodes communicate with each other if their Euclidean distance is less than a prescribed communication range of $d_c = 0.4$. The time horizon is $T = 600$. Entries of $\mathbf{V}$ are independent and identically distributed (i.i.d.), zero-mean, Gaussian with variance $\sigma^2 = 10^{-3}$; i.e., $v_{l,t} \sim \mathcal{N}(0, \sigma^2)$. Low-rank spatiotemporal load profiles with rank $r = 3$ are generated from the bilinear factorization model $\mathbf{X} = \mathbf{W}\mathbf{Z}'$, where $\mathbf{W}$ and $\mathbf{Z}$ are $N \times r$ and $T \times r$ matrices with i.i.d. entries drawn from Gaussian distributions $\mathcal{N}(0, 100/N)$ and $\mathcal{N}(0, 100/T)$, respectively. Every entry of $\mathbf{O}$ is randomly drawn from the set $\{-1, 0, 1\}$ with $\Pr(o_{n,t} = -1) = \Pr(o_{n,t} = 1) = 5 \times 10^{-2}$. To simulate missing data, a sampling matrix $\mathbf{\Omega} \in \{0, 1\}^{N\times T}$ is generated with i.i.d. Bernoulli distributed entries $o_{n,t} \sim \text{Ber}(0.75)$ (25% missing data on average). Finally, measurements are generated as $\mathcal{P}_\Omega(\mathbf{Y}) = \mathbf{\Omega} \odot (\mathbf{X} + \mathbf{O} + \mathbf{V})$ [cf. (1)], and smart meter n has available the n-th row of $\mathcal{P}_\Omega(\mathbf{Y})$.

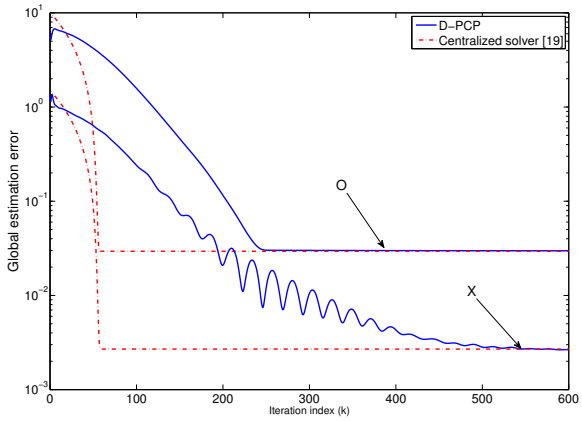

Fig. 1. Convergence of the D-PCP algorithm.

To experimentally corroborate the convergence and optimality (as per Proposition 2) of the D-PCP algorithm, Algorithm 1 is run with $c = 1$ and compared with the centralized benchmark (P1), obtained using the solver in [19]. Parameters $\lambda_1 = 0.0141$ and $\lambda_* = 0.346$ are chosen as suggested in [20]. For both schemes, Fig. 1 shows the evolution of the global estimation errors $e_X[k] := \|\mathbf{X}[k] - \mathbf{X}\|_F/\|\mathbf{X}\|_F$ and $e_O[k] := \|\mathbf{O}[k] - \mathbf{O}\|_F/\|\mathbf{O}\|_F$. It is apparent that the D-PCP algorithm converges to the centralized estimator, and as expected the convergence is slower since there is a delay associated with the information flow throughout the network.

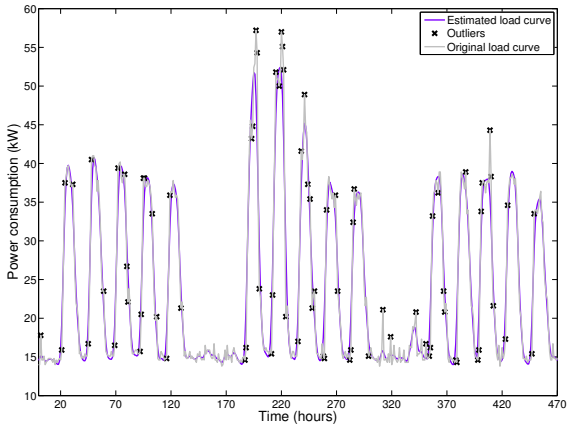

Fig. 2. Goverment building load curve data cleansing.

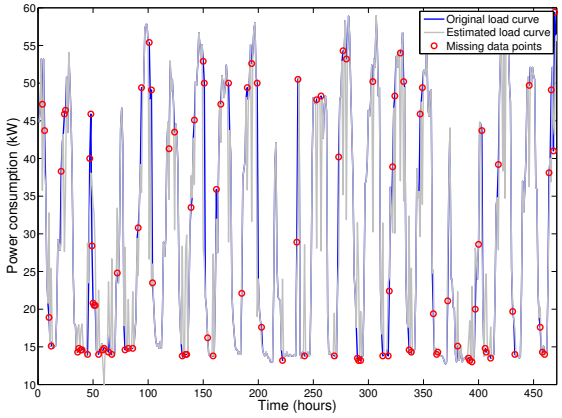

Fig. 3. Grocery store load curve imputation.

B. Real load curve data test

Next, the D-PCP algorithm is tested on real load curve data. The dataset consists of power consumption measurements (in kW) for a government building, a grocery store, and three schools ($N = 5$) collected every fifteen minutes during a period of more than five years, ranging from July 2005 to October 2010. Data is downsampled by a factor of four, to yield one measurement per hour. For the present experiment, only a subset of the whole data is utilized for concreteness, where $T = 470$ was chosen corresponding to a 470 hour period. For the government building case, a snapshot of the available load curve data spanning the studied three-week period is shown in gray in Fig. 2. Weekday activity patterns can be clearly discerned from those corresponding to weekends, as expected for most government buildings; but different, e.g., for the load profile of the grocery store in Fig. 3.

To run the D-PCP algorithm, an underlying communication graph was generated as in Section V-A. A randomly chosen subset of 25% of the measurements was removed to model missing data. For the government building data, Fig. 2 depicts the cleansed load curves that closely follow the measurements, but are smooth enough to avoid overfitting the abnormal energy peaks on the so-termed "building operational shoulders." Indeed, these peaks are in most cases identified as outliers. The effectiveness in terms of imputation of missing data is illustrated in Fig. 3 (identified outliers are not shown here); note how the cleansed (blue) load curve goes through the (red) missing data points. The relative error in predicting missing data is around 7%.

ACKNOWLEDGMENT

The authors would like to thank NorthWrite Energy Group and Prof. Vladimir Cherkassky (Dept. of ECE, University of Minnesota) for providing the data analysed in Section V.

REFERENCES

[1] A. R. Bergen and V. Vittal, *Power System Analysis*, Upper Saddle River, NJ: Prentice Hall, 2000.

[2] D. Bertsekas and J. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Athena-Scientific, 1999.

[3] E. J. Candes, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM*, vol. 58, no. 1, pp. 1–37, 2011.

[4] V. Chandrasekaran, S. Sanghavi, P. R. Parrilo, and A. S. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM J. Optim.*, vol. 21, no. 2, pp. 572–596, 2011.

[5] J. Chen, W. Li, A. Lau, J. Cao, and K. Eang, "Automated load curve data cleansing in power systems," *IEEE Trans. Smart Grid*, vol. 1, pp. 213–221, Sep. 2010.

[6] A. Chistov and D. Grigorev, "Complexity of quantifier elimination in the theory of algebraically closed fields," in *Math. Found. of Computer Science*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 1984, vol. 176, pp. 17–31.

[7] O. Kosut, L. Jia, 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] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma, "Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix," UIUC Technical Report UILU-ENG-09-2214, July 2009.

[9] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," in *Proc. of Conf. on Computer and Communications Security*, Chicago, IL, Nov. 2009, pp. 9–13.

[10] M. Mardani, G. Mateos, and G. B. Giannakis, "In-network sparsity-regularized rank minimization: Applications and algorithms," *IEEE Trans. Signal Process.*, 2012, see also arXiv:1203.1507v1 [cs.MA].

[11] G. Mateos and G. B. Giannakis, "Robust nonparametric regression via sparsity control with application to load curve data cleansing," *IEEE Trans. Signal Process.* vol. 60, no. 4, pp. 1571 - 1584, April 2012.

[12] G. Mateos, J. A. Bazerque, and G. B. Giannakis, "Distributed sparse linear regression," *IEEE Trans. Signal Process.*, vol. 58, pp. 5262–5276, Oct. 2010.

[13] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, pp. 227–234, 1995.

[14] B. Recht and C. Re, "Parallel stochastic gradient algorithms for large-scale matrix completion," *Optimization Online*, 2011.

[15] K. M. Rogers, R. D. Spadoni, and T. J. Overbye, "Identification of power system topology from synchrophasor data," in *Proc. of Power Systems Conference and Exposition*, Phoenix, AZ, Mar. 2011.

[16] M. Shahidehpour, H. Yamin, and Z. Li, *Market Operations in Electric Power Systems: Forecasting, Scheduling, and Risk Management*. New York, NY: Wiley-IEEE Press, 2002.

[17] U.S. Department of Energy, "The smart grid: An introduction," 2008, [Online.] Available: http://www.oe.energy.gov/SmartGridIntroduction.htm.

[18] W. Wulf, "Great achievements and grand challenges," The Brattle Group, Freeman, Sullivan and Co., and Global Energy Partners, LLC, Tech. Rep. 3/4, Fall 2010, [Online.] Available: http://www.greatachievements.org/.

[19] X. M. Yuan, J. Yang, "Sparse and low-rank matrix decomposition via alternating direction methods," *Pacific Journal of Optimization,* 2012 (to appear).

[20] Z. Zhou, X. Li, J. Wright, E. Candes, and Y. Ma, "Stable principal component pursuit," in *Proc. of Intl. Symp. on Information Theory*, Austin, TX, Jun. 2010, pp. 1518–1522.