# Load Forecasting via Low Rank Plus Sparse Matrix Factorization

Seung-Jun Kim and Geogios B. Giannakis
Department of Electrical and Computer Engineering
University of Minnesota
Minneapolis, MN 55455, USA
E-mail: {seungjun,georgios}@umn.edu

***Abstract*—Accurate imputation and prediction of load data are important prerequisites for many tasks of power systems, especially as renewables and plug-in electric vehicles penetrate the grid. A low-rank and sparse matrix factorization model is considered for load inference tasks to capture spatial as well as temporal structures in multi-site load data. The low-rank structure captures periodic patterns, and sparse matrix factors explain localized and clustered signatures. In order to predict load values for future time instants (and possibly for unforeseen sites), prior knowledge on correlations is necessarily incorporated in a nonparametric kernel-based learning framework. An efficient learning algorithm is also derived. Tests with real load data verify the efficacy of the proposed approach.**

## I. Introduction

Load forecasting is an essential grid informatics task for economic operation and planning of power systems. Forecasts of hourly, weekly and yearly time frames are needed for economic dispatch and unit commitment, as well as for long-term scheming of generation and transmission. In the future power systems, load forecasting will play an increasingly significant role, as the uncertainties associated with integration of volatile renewable energy resources will have to be addressed, and major impacts on the load pattern made by transportation electrification will have to be accounted for.

Typically, load forecasting aims at predicting future load levels based on the past values. Informative covariates such as weather data and various categorical features (including holidays and major events) can also be incorporated. A host of statistical inference methods have been developed for load forecasting in the past decades [1]. The approaches range from linear/nonlinear regression, stochastic time series methods based on ARMA, ARIMA, and ARIMAX models, Kalman filtering, principal component analysis (PCA) to neural networks [2], [3].

The main idea pursued in this work is to exploit both temporal and spatial structures in load data to perform inferences. As for the spatial dimension, $M$ sites are considered, at which load level measurements are made. At each site, load data are collected at regular intervals, say, every hour, for a given time duration $N$. However, due to various practical constraints, not all sites can make and communicate measurements at all times, giving rise to missing observations.

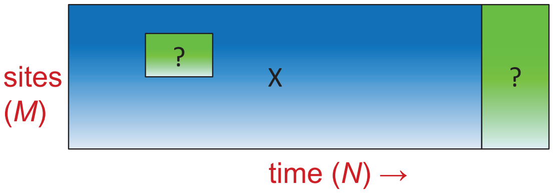


Fig. 1. Given load measurements and desired demands.

Two inference tasks are identified in this setup. First, the missing load levels can be interpolated based on the correlation structure learned from the available data. This is a load imputation task. Secondly, future demands of all sites may need to be predicted, which constitutes a load forecasting task. Fig. 1 depicts the available data as well as the missing part for which inference must be performed.

To accomplish the inference tasks, a novel matrix data model is put forth. Specifically, it is postulated that the matrix data is generated from a superposition of a low-rank matrix component plus a sparse matrix bifactor component. Each of these component models possesses well-documented merits in machine learning and signal processing applications [4], [5]. In the load data context, the low rank component can capture periodic and repetitive patterns in the load curves, while the sparse matrix factors admits a co-clustering interpretation and can account for localized and clustered load signatures.

The proposed model can be further extended to allow kernel-based nonparametric learning. Such an extension plays an instrumental role in load forecasting by accommodating prior knowledge on the spatio-temporal correlation structures of the data by means of kernels. An efficient and provably convergent learning algorithm is also derived.

The rest of this paper is organized as follows. The proposed model is discussed in Section II. The kernel extension of the model is delineated in Section III. A learning algorithm based on the model is developed in Section IV. Test results using real load data are reported in Section V, and conclusions are offered in Section VI.

## II. Low-Rank Plus Sparse Bifactor Model

To carry out the aforementioned inference tasks, it is necessary to postulate a certain low-dimensional structure that

This work was supported by NSF grant 1202135, and by the Institute of Renewable Energy and the Environment (IREE) at the University of Minnesota, under grant No. RL-0010-13.

load data adhere to. A promising approach is to arrange the load levels $x_{mn}$ corresponding to site $m \in \{1, 2, \ldots, M\}$ and time $n \in \{1, 2, \ldots, N\}$ in a matrix $\mathbf{X} \in \mathbb{R}^{M \times N}$ and hypothesize that $\mathbf{X}$ is of low rank. The low-rank structure is readily motivated by the fact that the load curves are influenced by a small number of latent factors, such as the ambient temperature and electricity prices. Moreover, load curves tend to have periodic patterns in daily, weekly, and yearly time scales, thus corroborating the low-rank assumption.

Let $\mathbf{\Omega} \in \{1, 0\}^{M \times N}$ be a matrix whose $(m, n)$-entry $\omega_{mn}$ is 0 if observation $x_{mn}$ is missing, and 1 otherwise. Then, interpolation of missing entries can be accomplished by fitting the data to the low-rank model. For instance, one can solve [4]

$$\hat{\mathbf{X}}_{\text{low-rank}} = \arg \min_{\mathbf{L} \in \mathbb{R}^{M \times N}} \frac{1}{2} \|\mathbf{\Omega} \odot (\mathbf{X} - \mathbf{L})\|_F^2 + \lambda \|\mathbf{L}\|_* \quad (1)$$

where $\odot$ is the entry-wise product, $\|\cdot\|_F$ denotes the Frobenius norm, and $\|\cdot\|_*$ the nuclear norm, which is the sum of singular values. The nuclear norm regularizer corresponds to a convex surrogate for the matrix rank, in the same token as the $\ell_1$-norm plays the role of a convex surrogate for the $\ell_0$-norm in various compressive sensing algorithms. Parameter $\lambda$ can be adjusted via cross-validation so that the rank of $\hat{\mathbf{X}}_{\text{low-rank}}$ matches the underlying rank of $\mathbf{X}$. Since (1) is a convex optimization problem, it can be solved effectively and optimally.

Another approach is to adopt a sparse matrix factorization model. Since the seminal paper by Lee and Seung [5], the advantageous features of nonnegative matrix factorization (NMF) models have been widely recognized. In particular, NMF models often provide readily interpretable part-based decomposition, which can be further promoted by encouraging sparsity in the factors [6]. Sparse matrix factorization models also admit a co-clustering interpretation, where the clusters on the rows and the clusters on the columns of a data matrix are simultaneously revealed [7], [8]. In the context of our load data matrix $\mathbf{X}$, row clusters can capture the fact that there are constituent power consumption patterns comprising the aggregate load [9]. Likewise, column clusters may disclose closely related groups of sites based on similar load patterns.

Sparse matrix factorization can be obtained, for instance, by solving the following optimization problem.

$$\min_{\mathbf{A} \in \mathbb{R}^{M \times \rho}, \mathbf{B} \in \mathbb{R}^{N \times \rho}} \frac{1}{2} \|\mathbf{\Omega} \odot (\mathbf{X} - \mathbf{A}\mathbf{B}^T)\|_F^2 + \mu_1 (\|\mathbf{A}\|_1 + \|\mathbf{B}\|_1) \quad (2)$$

where $\cdot^T$ denotes transposition, $\|\cdot\|_1$ is defined as the sum of the absolute values of the entries, and $\mu_1$ is a tuning parameter for controlling sparsity of the factors. If the entries of $\mathbf{X}$ are necessarily nonnegative, one can incorporate this by optimizing over $\mathbf{A} \in \mathbb{R}_+^{M \times \rho}$ and $\mathbf{B} \in \mathbb{R}_+^{N \times \rho}$. Although the global optimum for (2) is hard to come by due to nonconvexity of the bilinear model, locally optimal solutions can be readily obtained. Once factors $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ have been estimated, the missing entries can be recovered via $\hat{\mathbf{X}}_{\text{smf}} := \hat{\mathbf{A}}\hat{\mathbf{B}}^T$.

In this work, the goal is to combine the benefits of both low rank and sparse matrix factorization approaches by postulating that $\mathbf{X} \approx \mathbf{L} + \mathbf{A}\mathbf{B}^T$. A relevant optimization problem can then be cast as

$$\min_{\mathbf{L} \in \mathbb{R}^{M \times N}, \mathbf{A} \in \mathbb{R}^{M \times \rho}, \mathbf{B}^{N \times \rho}} \frac{1}{2} \|\mathbf{\Omega} \odot (\mathbf{X} - \mathbf{L} - \mathbf{A}\mathbf{B}^T)\|_F^2 + \lambda \|\mathbf{L}\|_* + \mu_1 (\|\mathbf{A}\|_1 + \|\mathbf{B}\|_1). \quad (3)$$

Given a (locally) optimal solution $\hat{\mathbf{L}}$, $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ to (3), one can reconstruct the entire data matrix including the missing entries by $\hat{\mathbf{X}} := \hat{\mathbf{L}} + \hat{\mathbf{A}}\hat{\mathbf{B}}^T$.

Although a formal analysis on the identifiability of the low-rank plus sparse matrix factorization model is an open problem, studies on related matrix models show that such models are often readily identifiable. For instance, model $\mathbf{X} = \mathbf{L} + \mathbf{S}$, where $\mathbf{L}$ is of low rank and $\mathbf{S}$ is sparse, has been studied in [10], [11], and model $\mathbf{X} = \mathbf{L} + \mathbf{R}\mathbf{S}$, where $\mathbf{R}$ is a given matrix has been investigated in [12]. Note that identifiability is particularly useful to establish uniqueness of factors when the factors themselves are of interest as for clustering applications. However, it may not be as critical for imputation and prediction tasks at hand.

## III. Kernelization

While formulations (1), (2) and (3) provide effective means to recover missing entries distributed sporadically in $\mathbf{X}$, they are not useful when entire columns (or rows) are missing. The rank is preserved when any linear combinations of the existing columns (or rows) are used for the missing columns (or rows). Likewise, the sparse factor model approach would end up with filling the missing entries with zeros. Since the prediction of an entire new column is precisely the load forecasting task described earlier, this hurdle must be circumvented.

Inspired by recent works on kernel-based learning for matrix data [13], [14], a nonparametric extension of (3) is pursued in this section, which facilitates the incorporation of correlation structures available as priori knowledge. To achieve this, the following characterization of the nuclear norm is first noted [14], [4]

$$\|\mathbf{L}\|_* = \min_{\mathbf{P} \in \mathbb{R}^{M \times r}, \mathbf{Q} \in \mathbb{R}^{N \times r}} \frac{1}{2} (\|\mathbf{P}\|_F^2 + \|\mathbf{Q}\|_F^2) \quad \text{subject to } \mathbf{L} = \mathbf{P}\mathbf{Q}^T \quad (4)$$

where $\operatorname{rank}(\mathbf{L}) \leq r$. Based on this, one can re-write (3) as

$$\min_{\substack{\mathbf{P} \in \mathbb{R}^{M \times r}, \mathbf{Q} \in \mathbb{R}^{N \times r}, \\ \mathbf{A} \in \mathbb{R}^{M \times \rho}, \mathbf{B} \in \mathbb{R}^{N \times \rho}}} \frac{1}{2} \|\mathbf{\Omega} \odot (\mathbf{X} - \mathbf{P}\mathbf{Q}^T - \mathbf{A}\mathbf{B}^T)\|_F^2 + \frac{\lambda}{2} (\|\mathbf{P}\|_F^2 + \|\mathbf{Q}\|_F^2) + \mu_1 (\|\mathbf{A}\|_1 + \|\mathbf{B}\|_1) \quad (5)$$

without sacrificing optimality.

Let $\mathcal{M} := \{1, 2, \ldots, M\}$ and $\mathcal{N} := \{1, 2, \ldots, N\}$. Generalization of (5) to a nonparametric setting can be attained by seeking functions $f : \mathcal{M} \times \mathcal{N} \to \mathbb{R}$ and $g : \mathcal{M} \times \mathcal{N} \to \mathbb{R}$, which belong to the following two families of functions, respectively:

$$\mathcal{F} := \left\{ f(m, n) = \sum_{i=1}^{r} p_i(m) q_i(n), p_i \in \mathcal{H}_p, q_i \in \mathcal{H}_q \right\} \quad (6)$$

$$\mathcal{G} := \left\{ g(m, n) = \sum_{i=1}^{\rho} a_i(m) b_i(n), a_i \in \mathcal{H}_a, b_i \in \mathcal{H}_b \right\} \quad (7)$$

where $\mathcal{H}_p$, $\mathcal{H}_q$, $\mathcal{H}_a$ and $\mathcal{H}_b$ are Hilbert spaces constructed from kernels $k_p(m, m')$, $k_q(n, n')$, $k_a(m, m')$ and $k_b(n, n')$, which are pre-specified over $m, m' \in \mathcal{M}$ and $n, n' \in \mathcal{N}$. Then, letting $f(m, n)$ and $g(m, n)$ represent the $(m, n)$-entries of $\mathbf{P}\mathbf{Q}^T$ and $\mathbf{A}\mathbf{B}^T$ in (5), respectively, the kernel-based learning of $f$ and $g$ amounts to solving

$$
\begin{aligned}
\min_{f\in\mathcal{F}, g\in\mathcal{G}} \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N} \omega_{mn}\big(x_{mn} - f(m,n) - g(m,n)\big)^2 \\
+\mu_1 \sum_{i=1}^{\rho}\left[\sum_{m=1}^{M}|a_i(m)| + \sum_{n=1}^{N}|b_i(n)|\right] \\
+\frac{\lambda}{2}\sum_{i=1}^{r}(\|p\|_{\mathcal{H}_p}^2 + \|q\|_{\mathcal{H}_q}^2) + \frac{\mu_2}{2}\sum_{i=1}^{\rho}(\|a\|_{\mathcal{H}_a}^2 + \|b\|_{\mathcal{H}_b}^2) \quad (8)
\end{aligned}
$$

where $\|\cdot\|_{\mathcal{H}}$ denotes the norm associated with Hilbert space $\mathcal{H}$.

Recursive application of Representer's Theorem allows finite-dimensional parametrization of $p_i$, $q_i$, $a_i$ and $b_i$ without sacrificing optimality [14]. Specifically, they can be written as $p_i(m) = \sum_{m'=1}^{M} \pi_{m'i} k_p(m', m)$, $q_i(n) = \sum_{n'=1}^{N} \theta_{n'i} k_q(n', n)$, $a_i(m) = \sum_{m'=1}^{M} \alpha_{m'i} k_a(m', m)$ and $b_i(n) = \sum_{n'=1}^{N} \beta_{n'i} k_b(n', n)$. Substituting these into (8) and defining $\tilde{\mathbf{P}}$ from entries $\{\pi_{mi}\}$ (and likewise $\tilde{\mathbf{Q}}$, $\tilde{\mathbf{A}}$, and $\tilde{\mathbf{B}}$ from $\{\theta_{ni}\}$, $\{\alpha_{mi}\}$ and $\{\beta_{ni}\}$, respectively), as well as $\mathbf{K}_p$ from entries $\{k_p(m, m')\}$ (and similarly $\mathbf{K}_q$, $\mathbf{K}_a$ and $\mathbf{K}_b$) yields

$$
\begin{aligned}
\min_{\substack{\tilde{\mathbf{P}}\in\mathbb{R}^{M\times r}, \tilde{\mathbf{Q}}\in\mathbb{R}^{N\times r},\\ \tilde{\mathbf{A}}\in\mathbb{R}^{M\times\rho}, \tilde{\mathbf{B}}\in\mathbb{R}^{N\times\rho}}} \frac{1}{2}\|\mathbf{\Omega}\odot(\mathbf{X} - \mathbf{K}_p\tilde{\mathbf{P}}\tilde{\mathbf{Q}}^T\mathbf{K}_q - \mathbf{K}_a\tilde{\mathbf{A}}\tilde{\mathbf{B}}^T\mathbf{K}_b^T)\|_F^2 \\
+\frac{\lambda}{2}\left(\operatorname{tr}(\tilde{\mathbf{P}}^T\mathbf{K}_p\tilde{\mathbf{P}}) + \operatorname{tr}(\tilde{\mathbf{Q}}^T\mathbf{K}_q\tilde{\mathbf{Q}})\right) \\
+\mu_1(\|\mathbf{K}_a\tilde{\mathbf{A}}\|_1 + \|\mathbf{K}_b\tilde{\mathbf{B}}\|_1) \\
+\frac{\mu_2}{2}\left(\operatorname{tr}(\tilde{\mathbf{A}}^T\mathbf{K}_a\tilde{\mathbf{A}}) + \operatorname{tr}(\tilde{\mathbf{B}}^T\mathbf{K}_b\tilde{\mathbf{B}})\right). \quad (9)
\end{aligned}
$$

Upon defining $\mathbf{P} := \mathbf{K}_p\tilde{\mathbf{P}}$, $\mathbf{Q} := \mathbf{K}_q\tilde{\mathbf{Q}}$, $\mathbf{A} := \mathbf{K}_a\tilde{\mathbf{A}}$, and $\mathbf{B} := \mathbf{K}_b\tilde{\mathbf{B}}$, (9) is equivalent to

$$
\begin{aligned}
\min_{\mathbf{P},\mathbf{Q},\mathbf{A},\mathbf{B}} \frac{1}{2}\|\mathbf{\Omega}\odot(\mathbf{X} - \mathbf{P}\mathbf{Q}^T - \mathbf{A}\mathbf{B}^T)\|_F^2 \\
+\frac{\lambda}{2}\left(\operatorname{tr}(\mathbf{P}^T\mathbf{K}_p^{-1}\mathbf{P}) + \operatorname{tr}(\mathbf{Q}^T\mathbf{K}_q^{-1}\mathbf{Q})\right) \\
+\mu_1(\|\mathbf{A}\|_1 + \|\mathbf{B}\|_1) \\
+\frac{\mu_2}{2}\left(\operatorname{tr}(\mathbf{A}^T\mathbf{K}_a^{-1}\mathbf{A}) + \operatorname{tr}(\mathbf{B}^T\mathbf{K}_b^{-1}\mathbf{B})\right). \quad (10)
\end{aligned}
$$

## IV. Algorithms

Since (10) is not convex, it is in general hard to achieve a globally optimal solution. However, a locally optimal solution is easily obtained by employing a block coordinate descent (BCD) method, where the cost is minimized with respect to a block of variables, with the rest of the blocks fixed, and iterating block by block. Given the separable structure of (10), it can be shown that BCD update converges to a stationary point [15].

Here, it is proposed to perform the BCD updates for the following blocks of variables: the rows of $\mathbf{P}$, the rows of $\mathbf{Q}$, the entries of $\mathbf{A}$, and the entries of $\mathbf{B}$. The blocks have been selected so that a closed-form solution is available per update. For instance, solving for a row of $\mathbf{P}$ entails solving a quadratic problem of dimension $r$. Similarly, updating an entry of $\mathbf{A}$ can be performed through a simple soft thresholding operation.

Let $\mathbf{z}_i^T$ denote the $i$-th row of a generic matrix $\mathbf{Z}$. Then, with $\mathbf{Q}$, $\mathbf{A}$, and $\mathbf{B}$ as well as $\{\mathbf{p}_j\}_{j\neq i}$ fixed, the update for $\mathbf{p}_i$ for $i = 1, 2, \ldots, M$ can be done via solving

$$
\begin{aligned}
\min_{\mathbf{p}_i\in\mathbb{R}^r} \frac{1}{2}\|\mathbf{w}_i^T \odot (\mathbf{x}_i^T - \mathbf{p}_i^T\mathbf{Q}^T - \mathbf{a}_i^T\mathbf{B})\|_2^2 \\
+\frac{\lambda}{2}(k_p(i,i)\mathbf{p}_i^T\mathbf{p}_i + 2\sum_{j\neq i} k_p(i,j)\mathbf{p}_i^T\mathbf{p}_j) \quad (11)
\end{aligned}
$$

where $\mathbf{w}_i^T := [\omega_{i1}, \ldots, \omega_{iN}] \in \{1, 0\}^N$ indicates the availability of measurements $\{x_{in}\}_{n=1}^N$. Upon defining $\mathbf{D}_{w_i} := \operatorname{diag}(\mathbf{w}_i)$ as the diagonal matrix whose diagonal is equal to $\mathbf{w}_i$, the solution to (11) is given by

$$
\begin{aligned}
\mathbf{p}_i = \left[\mathbf{Q}^T\mathbf{D}_{w_i}\mathbf{Q} + \lambda k_p(i,i)\mathbf{I}_r\right]^{-1} \\
\left[\mathbf{Q}^T\mathbf{D}_{w_i}(\mathbf{x}_i - \mathbf{B}\mathbf{a}_i) - \lambda\sum_{j\neq i} k_p(i,j)\mathbf{p}_j\right] \quad (12)
\end{aligned}
$$

where $\mathbf{I}_r$ is the identity matrix of dimension $r \times r$. Likewise, upon defining $\mathbf{D}_{\omega_i} := \operatorname{diag}([\omega_{1i}, \ldots, \omega_{Mi}])$, and letting $\boldsymbol{\chi}_i$ denote the $i$-th column of $\mathbf{X}$, the update for $\mathbf{q}_i$, $i = 1, 2, \ldots, N$, is given by

$$
\begin{aligned}
\mathbf{q}_i = \left[\mathbf{P}^T\mathbf{D}_{\omega_i}\mathbf{P} + \lambda k_q(i,i)\mathbf{I}_r\right]^{-1} \\
\left[\mathbf{P}^T\mathbf{D}_{\omega_i}(\boldsymbol{\chi}_i - \mathbf{A}\mathbf{b}_i) - \lambda\sum_{j\neq i} k_q(i,j)\mathbf{q}_j\right]. \quad (13)
\end{aligned}
$$

Next, let $\tilde{\mathbf{a}}_i$ and $\tilde{\mathbf{b}}_i$ denote the $i$-th columns of $\mathbf{A}$ and $\mathbf{B}$, respectively. Also, let $\mathbf{y}_{ij} := \mathbf{x}_i - \mathbf{Q}\mathbf{p}_i - \sum_{k\neq j}\tilde{\mathbf{b}}_k a_{ik}$. Then, the update for $a_{ij}$ for $i = 1, \ldots, M$ and $j = 1, \ldots, N$, with all other variables fixed, amounts to solving

$$
\begin{aligned}
\min_{a_{ij}} \frac{1}{2}\left\|\mathbf{D}_{w_i}\left(\mathbf{y}_{ij} - \tilde{\mathbf{b}}_j a_{ij}\right)\right\|_2^2 + \mu_1|a_{ij}| \\
+\frac{\mu_2}{2}\left(k_a(i,i)a_{ij}^2 + 2\sum_{k\neq i} k_a(i,k)a_{ij}a_{kj}\right) \quad (14)
\end{aligned}
$$

which yields the solution in a closed form as

$$
a_{ij} = \frac{\text{soft\_th}_{\mu_1}\left(\tilde{\mathbf{b}}_j^T\mathbf{D}_{w_i}\mathbf{y}_{ij} - \mu_2\sum_{k\neq i} k_a(i,k)a_{kj}\right)}{\|\mathbf{D}_{w_i}\tilde{\mathbf{b}}_j\|_2^2 + \mu_2 k_a(i,i)} \quad (15)
$$

where the soft thresholding operator is defined as

$$
\text{soft\_th}_{\mu}(x) = \operatorname{sgn}(x)\max\{0, |x| - \mu\}. \quad (16)
$$

Similarly, upon defining $\tilde{\mathbf{y}}_{ij} := \boldsymbol{\chi}_i - \mathbf{P}\mathbf{q}_i - \sum_{k\neq j}\tilde{\mathbf{a}}_k b_{ik}$, the update for $b_{ij}$ is given by

$$
b_{ij} = \frac{\text{soft\_th}_{\mu_1}\left(\tilde{\mathbf{a}}_j^T\mathbf{D}_{\omega_i}\tilde{\mathbf{y}}_{ij} - \mu_2\sum_{k\neq i} k_b(i,k)b_{kj}\right)}{\|\mathbf{D}_{\omega_i}\tilde{\mathbf{a}}_j\|_2^2 + \mu_2 k_b(i,i)}. \quad (17)
$$

The overall algorithm is tabulated in Table I.

TABLE I. OVERALL BCD ALGORITHM.

```
Initialize $\mathbf{P}$, $\mathbf{Q}$, $\mathbf{A}$ and $\mathbf{B}$ randomly
Repeat
    For $i = 1, 2, \ldots, M$
        Update $\mathbf{p}_i$ via (12)
    Next $i$
    For $i = 1, 2, \ldots, N$
        Update $\mathbf{q}_i$ via (13)
    Next $i$
    For $i = 1, \ldots, M$ and $j = 1, \ldots, N$
        Update $a_{ij}$ via (15)
    Next $i$ and $j$
    For $i = 1, \ldots, M$ and $j = 1, \ldots, N$
        Update $b_{ij}$ via (17)
    Next $i$ and $j$
Until convergence
```

## V. TESTS WITH REAL DATA

To test the performance of the proposed algorithm, an hourly load dataset for 17 sites, spanning 48 weeks was employed. Matrix $\mathbf{X}$ of size (17 sites $\times$ 672 hours [4 weeks]) was used throughout. The dataset was divided in time into three parts, 16 weeks long each. The first part was utilized for estimating covariance matrices, which were used as kernels. The second part was used for tuning the model parameters such as $\lambda$, $\mu_1$ and $\mu_2$. The test error was calculated on the remaining part of the dataset. To minimize any seasonal influence, each part was subdivided into four 4-week chunks, and each chunk was taken from different periods of the dataset. To be precise, the first part was taken from weeks 1–4, 13–16, 21–24, and 37–40; the second part from weeks 5–8, 17–20, 29–32, and 41–44; and the third part consisted of the rest.

The kernel matrices were constructed in the following way. For $\mathbf{K}_p$ and $\mathbf{K}_a$, each row of $\mathbf{X}$ was first normalized to have unit variance, and then the covariance of the columns of $\mathbf{X}$ was calculated. Both $\mathbf{K}_p$ and $\mathbf{K}_a$ were set equal to this covariance matrix. For $\mathbf{K}_q$ and $\mathbf{K}_b$, a product kernel was employed. For an hour index $n_i$, $i = 1, 2$, let $h_i \in \{1, 2, \ldots, 24\}$ denote the hour of a day, $d_i \in \{1, 2, \ldots, 7\}$ the day of a week, and $w_i \in \{1, 2, \ldots, 4\}$ the corresponding week number. Then, $k_q(n_1, n_2)$ was constructed as [16]

$$k_q(n_1, n_2) = k_h(h_1, h_2) k_d(d_1, d_2) 0.8^{|w_1 - w_2|}. \qquad (18)$$

where $k_h(h_1, h_2)$ and $k_d(d_1, d_2)$ are the covariances estimated from the data. Kernel $k_b$ was set equal to $k_q$.

Fig. 2 depict the load curves of all sites for week 12, where the load levels for the last 24 hours were forecast using the low-rank plus sparse matrix factorization model with $r = \rho = 10$. Here, we actually used sparse NMF by constraining $\mathbf{A}$ and $\mathbf{B}$ to have nonnegative entries. The solid curves are the true load values, and the dashed ones correspond to reconstructed or forecast values obtained from $\mathbf{PQ}^T + \mathbf{AB}^T$. It can be seen that the forecast values (inside the red box) are quite close to the true ones. The total normalized root-mean-square error including both reconstruction and forecast was 0.11.

The low rank component $\mathbf{PQ}^T$ and the NMF component $\mathbf{AB}^T$ are depicted in Fig. 3. It can be seen that the low rank component in Fig. 3(a) captures repetitive patterns, while the NMF component in Fig. 3(b) more localized variations.

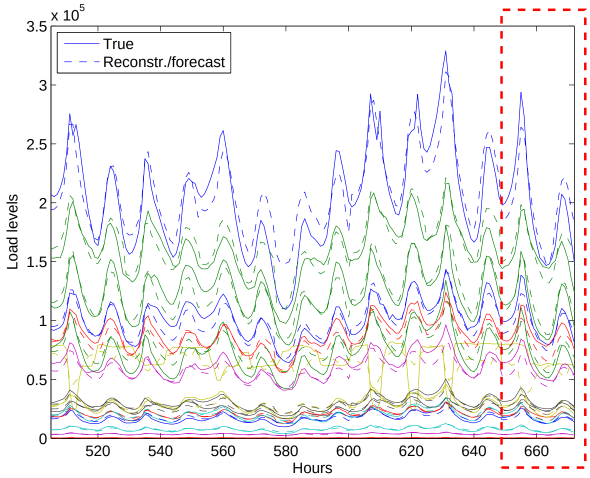
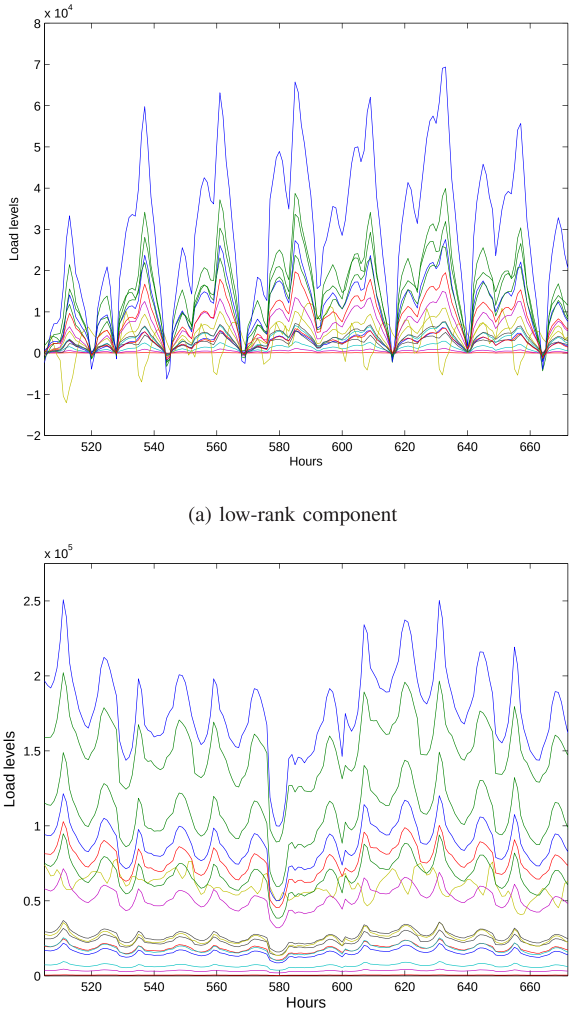


Fig. 2. True and reconstructed/forecast load curves.



(a) low-rank component

(b) NMF component

Fig. 3. Reconstructed/forecast low rank and sparse NMF components.

## VI. Conclusion

A novel low-rank plus sparse bifactor matrix model was proposed for load forecasting to exploit spatio-temporal structure inherent in multi-site load data. The low-rank component was motivated by periodic patterns in load curves and a small number of latent factors influencing load variations. The sparse (nonnegative) matrix factors can capture localized signatures. In order to perform the load forecasting task, prior information on correlation structures was incorporated in a kernel-based learning framework. An efficient and provably convergent BCD update algorithm was derived to solve the learning problem. Preliminary tests with real load data showed promising forecasting performance. Rigorous comparison with existing methods, as well as incorporation of exogenous data such as weather data are left for future work.

## References

[1] 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.

[2] I. Moghram and S. Rahman, "Analysis and evaluation of five short-term load forecasting techniques," *IEEE Trans. Power Syst.*, vol. 4, no. 4, pp. 1484–1491, Oct. 1989.

[3] J. W. Taylor and P. E. McSharry, "Short-term load forecasting methods: An evaluation based on European data," *IEEE Trans. Power Syst.*, vol. 22, no. 4, pp. 2213–2219, Nov. 2007.

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

[5] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[6] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Machine Learning Res.*, vol. 5, pp. 1457–1469, Nov. 2004.

[7] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proc. of the ACM SIGIR Conf.*, Toronto, Canada, Jul.-Aug. 2003, pp. 267–273.

[8] E. E. Papalexakis, N. D. Sidiropoulos, and R. Bro, "From $k$-means to higher-way co-clustering: multilinear decomposition with sparse latent factors," *IEEE Trans. Sig. Proc.*, vol. 61, no. 2, pp. 493–506, Jan. 2013.

[9] H. Gonçalves, A. Ocneanu, and M. Bergés, "Unsupervised disaggregation of appliances using aggregated consumption data," in *Proc. of the 1st KDD Workshop on Data Maining Applications and Sustainability (SustKDD)*, San Diego, CA, Aug. 2011.

[10] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. of the ACM*, vol. 58, no. 3, article no. 11, May 2011.

[11] J. Wright, A. Ganesh, K. Min, and Y. Ma, "Compressive principal component pursuit," *Information and Inference*, vol. 2, no. 1, pp. 32–68, Jun. 2013.

[12] M. Mardani, G. Mateos, and G. B. Giannakis, "Recovery of low-rank plus compressed sparse matrices with application to unveiling traffic anomalies," *IEEE Trans. Info. Theory*, vol. 59, no. 8, pp. 5186–5205, Aug. 2013.

[13] J. Abernethy, F. Bach, T. Evgeniou, and J.-P. Vert, "A new approach to collaborative filtering: Operator estimation with spectral regularization," *J. Machine Learning Res.*, vol. 10, pp. 803–826, Mar. 2009.

[14] J. A. Bazerque and G. B. Giannakis, "Nonparametric basis pursuit via sparse kernel-based learning: A unifying view with advances in blind methods," *IEEE Sig. Proc. Mag.*, vol. 30, no. 4, pp. 112–125, Jul. 2013.

[15] P. Tseng, "Convergence of block coordinate descent method for non-differentiable minimization," *J. Optimiz. Theory Applicat.*, vol. 109, pp. 475–494, Jun. 2001.

[16] V. Kekatos, S. Veeramachaneni, M. Light, and G. B. Giannakis, "Day-ahead electricity market forecasting," in *Proc. of the IEEE PES Innovative Smart Grid Technol. Conf.*, Washington, DC, Feb. 2013.