Low-Rank Kernel Learning for Electricity Market Inference

Vassilis Kekatos, Yu Zhang, and Georgios B. Giannakis
Dept. of ECE and DTC, University of Minnesota, Minneapolis, USA
Emails: {kekatos,zhan1220,georgios}@umn.edu

***Abstract*—Recognizing the importance of smart grid data analytics, modern statistical learning tools are applied here to wholesale electricity market inference. Market clearing congestion patterns are uniquely modeled as rank-one components in the matrix of spatiotemporally correlated prices. Upon postulating a low-rank matrix factorization, kernels across pricing nodes and hours are systematically selected via a novel methodology. To process the high-dimensional market data involved, a block-coordinate descent algorithm is developed by generalizing block-sparse vector recovery results to the matrix case. Preliminary numerical tests on real data corroborate the prediction merits of the developed approach.**

I. Introduction

In a typical whole-sale day-ahead electricity market, an independent system operator (ISO) collects bids submitted by generator owners and utilities. Compliant with network and reliability constraints, the grid is dispatched in the most economical way for every hour of the following day. Due to transmission grid limitations though, cheap electricity cannot be delivered everywhere across the grid, but out-of-merit energy sources have to be dispatched to balance the load. Hence, congestion together with heat losses lead to spatiotemporally-varying electricity prices, known as locational marginal prices (LMPs) [14].

Forecasting LMPs is undoubtedly an important decision making tool for traders [2]. In addition, conventional and particularly renewable asset owners plan their trading according to pricing predictions. Recently, ISOs broadcast their own market forecasts as proactive signals to relieve congestion [8]. At a larger geographical and time scale, electricity price analytics are pursued by government services to identify "transmission congestion corridors" [23].

To concretely formulate LMP forecasting, consider an electricity market over a network of pricing nodes and over the last T hours. In a day-ahead market, LMPs correspond to the cost of buying or selling electricity at each node and over one-hour periods for the following day. Setting LMPs as target variables, explanatory variables (features) can be LMPs from past days, load estimates, weather forecasts, scheduled outage capacity, and inter-area transfers.

Electricity market forecasting methods proposed so far aim at predicting *single-node* prices using either on time series models, neural networks and fuzzy logic systems, or combinatorial physical system modeling; see e.g., [6], [24], [25], and references therein. However, LMPs are not independent; but rather exhibit a transmission network-imposed dependence that is uniquely exploited next.

Distinct from existing approaches where predictors are trained on a per-node basis, an interconnection-wide inference framework is pursued in this work. Our first contribution is casting electricity market inference as a low-rank learning task [1], [3]. The matrix of spatio-temporal prices is modeled as a superposition of few congestion patterns and is subsequently learned using kernel-based trace-norm regularization. A systematic methodology for judiciously selecting kernels over space and time is the second contribution of this paper. Our novel analytic results extend kernel learning tools to low-rank multi-task models [18], [11]. The optimization problem derived is non-convex and involves high-dimensional price matrices. A block-coordinate descent algorithm converging to a stationary point of the postulated problem is our third contribution. Generalizing results from (block) compressed sensing [20], the resultant algorithm boils down to univariate minimizations and exploits the Kronecker product structure involved. The framework can be used for extrapolation in time (prediction) and space (node additions), or even imputation of missing entries. Prediction results using real data from the MISO market corroborate our findings.

Notation. Lower- (upper-) case boldface letters denote column vectors (matrices); calligraphic letters stand for sets. Symbols $\mathbf{A}^\top$ and $\mathrm{Tr}(\mathbf{A})$ denote transposition and matrix trace.

II. Problem Formulation

Consider a whole-sale electricity market over a set $\mathcal{N}$ of pricing nodes indexed by n. In a day-ahead market, locational marginal prices (LMPs) correspond to the cost of buying or selling electricity at each grid node and over one-hour periods for the following day [19]. Viewing price forecasting as an inference problem, LMPs are the targets to be learned. Explanatory variables (features) can be any data available at the time of forecasting relevant to the day-ahead market.

One could try designing per pricing node predictors; yet locational prices are not independent. They are collectively determined as the solution of the network-constrained economic dispatch and the unit commitment problems [9], [10]. Leveraging this network-imposed dependence, market forecasting is uniquely interpreted here as learning over a graph; see e.g.,

Work in this paper was supported by the Inst. of Renewable Energy and the Environment (IREE) under grant no. RL-0010-13, Univ. of Minnesota, and NSF Grant ECCS-1202135.

[15]. The market is further considered to be stationary only over the T most recent time periods comprising the set $\mathcal{T}$.

Adopting a kernel-based learning approach, the market could be regarded as a function $p : \mathcal{N} \times \mathcal{T} \to \mathbb{R}$ to be inferred. It is postulated that the price at node n and time t denoted by $p(n,t)$ belongs to the function space

$$\mathcal{P} := \left\{ p(n,t) = \sum_{\substack{n' \in \mathcal{N} \\ t' \in \mathcal{T}}} K_{\otimes}\left((n,t),(n',t')\right) a_{n't'} : a_{n't'} \in \mathbb{R} \right\}$$

defined by the kernel $K_{\otimes} : (\mathcal{N} \times \mathcal{T}) \times (\mathcal{N} \times \mathcal{T}) \to \mathbb{R}$. When $K_{\otimes}$ is a symmetric positive definite function, the function space $\mathcal{P}$ becomes a reproducing kernel Hilbert space (RKHS) equipped with a finite norm [4]

$$\|p\|^2_{\mathcal{K}_{\otimes}} := \sum_{n,n' \in \mathcal{N}} \sum_{t,t' \in \mathcal{T}} K_{\otimes}((n,t),(n',t')) a_{nt} a_{n't'}.$$

The sought p can be then found via the regularization [1], [12]

$$\min_{p \in \mathcal{P}} \ \|\mathbf{Z} - \mathbf{P}\|_F^2 + \mu \|p\|_{\mathcal{K}_{\otimes}} \tag{1}$$

where $\mathbf{P} \in \mathbb{R}^{N \times T}$ has entries $[\mathbf{P}]_{n,t} = p(n,t)$, and $\mathbf{Z}$ is the matrix of observed prices arranged accordingly. The regularizer $\|p\|_{\mathcal{K}_{\otimes}}$ constraints $p \in \mathcal{P}$ and facilitates generalization over unseen data. Balancing between the regularizer and the least-squares (LS) data fit is controlled by $\mu > 0$, a parameter typically tuned via cross-validation [12].

When $K_{\otimes}$ is additionally selected as the tensor product kernel $K_{\otimes}((n,t),(n',t')) := K(n,n')G(t,t')$, where $K : \mathcal{N} \times \mathcal{N} \to \mathbb{R}$ and $G : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ are kernels over nodes and hours, respectively; then every function in $\mathcal{P}$ admits the spatio-temporal decomposition [4], [1]

$$\mathcal{P} = \left\{ p(n,t) = \sum_{r=1}^{R} f_r(n) g_r(t), \ f_r \in \mathcal{H}_K, \ g_r \in \mathcal{H}_G \right\} \tag{2}$$

where $\mathcal{H}_{\mathcal{K}}$ and $\mathcal{H}_{\mathcal{G}}$ are the RKHSs defined accordingly by K and G, while R may be infinite [4].

Our modeling contribution is that the sought $p(n,t)$ admits a low-rank decomposition, i.e., $p(n,t)$ is decomposed in a small number of $p_r(n,t) = f_r(n)g_r(t)$ components. The market is determined by few congestion patterns $\{f_r(n)\}_{r=1}^R$ occurring when a transmission line has reached its capacity rating; congested lines contribute linearly to determining prices; while prices are modulated by temporal features. However, the $p(n,t)$ minimizing (1) will not be decomposable into a few p_r. To promote $p(n,t)$'s with a parsimonious spatiotemporal factorization, the *trace norm* $\|p\|_*$ is used in lieu of the$\|p\|_{K_{\otimes}}$ regularizer. Precisely, for every $p \in \mathcal{P}$, its $\|p\|_*$ can be expressed as [1]

$$\|p\|_* = \min_{\{f_r, g_r\}} \ \frac{1}{2}\left(\sum_{r=1}^{R} \|f_r\|^2_{\mathcal{K}} + \sum_{r=1}^{R} \|g_r\|^2_{\mathcal{G}} \right) \tag{3}$$
$$\text{s.to} \ \ p = \sum_{r=1}^{R} f_r g_r, \ f_r \in \mathcal{H}_K, \ g_r \in \mathcal{H}_G.$$

To build connections with low-rank matrix completion, consider $\mathcal{N}$ and $\mathcal{T}$ being Euclidean spaces, while the kernels $K(n,n')$ and $G(t,t')$ are selected as the Kronecker delta functions $\delta(n-n')$ and $\delta(t-t')$, respectively; see e.g., [5]. In this case, function $p(n,t)$ is fully described by matrix $\mathbf{P}$, $\|p\|_{\mathcal{K}_{\otimes}}$ is its Frobenius norm $\|\mathbf{P}\|_F$, and $\|p\|_*$ is its *nuclear norm* $\|\mathbf{P}\|_*$ (the sum of the matrix singular values), which has been widely used as a low-rank matrix regularizer [1].

It is worth mentioning that writing $\|\mathbf{P}\|_*$ equivalently as [21, Lemma 5.2]

$$\|\mathbf{P}\|_* = \min_{\mathbf{F},\mathbf{G}} \left\{ \frac{\|\mathbf{F}\|_F^2 + \|\mathbf{G}\|_F^2}{2} : \ \mathbf{P} = \mathbf{F}\mathbf{G}^\top \right\} \tag{4}$$

has been proved beneficial both algorithmically and for applying the Representer's Theorem in trace-norm regularization [21], [17], [5]. If $\mathbf{P} \in \mathbb{R}^{N \times T}$ and $R = \min\{N,T\}$, the variables $\mathbf{F}$ and $\mathbf{G}$ in (4) should have R columns. However, when (4) is used for matrix completion, the ensuing $\mathbf{P}$ is expected to be low-rank and smaller values for R are practically used.

To exploit the aforementioned low-rank factorization, market inference is posed as the regularization problem

$$\min_{p \in \mathcal{P}} \ \|\mathbf{Z} - \mathbf{P}\|_F^2 + \mu \sqrt{\|p\|_*}. \tag{5}$$

To derive efficient algorithms, problem (5) is transformed according to the following result (due to space limitations all proofs can be found in [13]):

Lemma 1. *The optimization in* (5) *is equivalent to*

$$\min_{p, \{f_r, g_r\}} \ \|\mathbf{Z} - \mathbf{P}\|_F^2 + \frac{\mu}{2}\left(\sum_{r=1}^{R} \|f_r\|^2_{\mathcal{K}} \right)^{\frac{1}{2}} + \frac{\mu}{2}\left(\sum_{r=1}^{R} \|g_r\|^2_{\mathcal{G}} \right)^{\frac{1}{2}}$$
$$\text{s.to } p = \sum_{r=1}^{R} f_r g_r, \ f_r \in \mathcal{H}_K, \ g_r \in \mathcal{H}_G. \tag{6}$$

The equivalence is based on (2) as well as the key result

$$\|p\|_*^{\frac{1}{2}} = \min_{p = \sum_r f_r g_r} \frac{1}{2}\left(\sum_{r=1}^{R} \|f_r\|^2_{\mathcal{K}} \right)^{\frac{1}{2}} + \frac{1}{2}\left(\sum_{r=1}^{R} \|g_r\|^2_{\mathcal{G}} \right)^{\frac{1}{2}} \tag{7}$$

which for Euclidean spaces and Kronecker delta kernels, yields interestingly

$$\|\mathbf{P}\|_*^{\frac{1}{2}} = \min_{\mathbf{F},\mathbf{G}} \left\{ \frac{\|\mathbf{F}\|_F^{\frac{1}{2}} + \|\mathbf{G}\|_F^{\frac{1}{2}}}{2} : \ \mathbf{P} = \mathbf{F}\mathbf{G}^\top \right\}. \tag{8}$$

III. Kernel Selection

Solving (6) presumes that kernels $\mathcal{K}$ and $\mathcal{G}$ are known. Practically though, the designer if often given candidate kernels and would like to determine which of them provide better inference results. The kernel selection approach of [18] is generalized here to the function regularization of (6). Two kernel function sets, $\{K_l\}_{l=1}^L$ and $\{G_m\}_{m=1}^M$, are provided for nodes and time. Consider the kernel spaces constructed as

$$\mathcal{K} := \text{Conv}\left(\{K_l\}_{l=1}^L\right), \ \mathcal{G} := \text{Conv}\left(\{G_m\}_{m=1}^M\right). \tag{9}$$

Optimizing the outcome of (6) over the convex combination weights in $\mathcal{K}$ and $\mathcal{G}$ provides a disciplined kernel design methodology. The following result shows how that can be accomplished without finding explicitly the weights.

Lemma 2. *Minimizing* (6) *over the kernel spaces* $\mathcal{K}$ *and* $\mathcal{G}$ *defined in* (9) *is equivalent to solving*

$$\min_{p\in\mathcal{P}'} \|\mathbf{Z}-\mathbf{P}\|_F^2+\mu\sum_{l=1}^{L}\sqrt{\sum_{r=1}^{R}\|f_{lr}\|_{\mathcal{K}_l}^2}+\mu\sum_{m=1}^{M}\sqrt{\sum_{r=1}^{R}\|g_{mr}\|_{\mathcal{G}_m}^2} \tag{10}$$

over $\mathcal{P}' := \Big\{p(n,t)=\sum_{r=1}^{R} f_r(n)g_r(t) : f_r=\sum_{l=1}^{L} f_{lr},\ f_{lr}\in\mathcal{H}_{\mathcal{K}_l},\ g_r=\sum_{m=1}^{M} g_{mr},\ g_{mr}\in\mathcal{H}_{\mathcal{G}_m}\Big\}$, *where* $\{\mathcal{H}_{\mathcal{K}_l}\}$ *and* $\{\mathcal{H}_{\mathcal{G}_m}\}$ *are the function spaces defined* K_l *and* G_m.

The result asserts that kernel learning over $\mathcal{P}$ boils down to the functional minimization in (10) where now $p\in\mathcal{P}'$. Practically solving (10) requires transforming the functional to a vector minimization which can be accomplished as follows.

Lemma 3. *The functional minimization in* (10) *is equivalent to solving the vector minimization problem*

$$\min_{\mathbf{P},\{\mathbf{B}_l\},\{\mathbf{\Gamma}_m\}} \|\mathbf{Z}-\mathbf{P}\|_F^2+\mu\sum_{l=1}^{L}\|\mathbf{B}_l\|_{\mathbf{K}_l}+\mu\sum_{m=1}^{M}\|\mathbf{\Gamma}_m\|_{\mathbf{G}_m}$$
$$\text{s.to}\quad \mathbf{P}=\sum_{l=1}^{L}\sum_{m=1}^{M}\mathbf{K}_l\mathbf{B}_l\mathbf{\Gamma}_m^\top\mathbf{G}_m \tag{11}$$

where $\mathbf{K}_l\in\mathbb{S}_{++}^N$ ($\mathbf{G}_m\in\mathbb{S}_{++}^T$) *is the spatial (temporal) kernel matrix, and* $\|\mathbf{X}\|_{\mathbf{B}}^2:=\mathrm{Tr}(\mathbf{X}^\top\mathbf{B}\mathbf{X})$ *for positive definite* $\mathbf{B}$.

The key for showing Lemma 3 is that minimizing (10) over a specific f_{lr} is actually a functional minimization regularized by an increasing function of $\|f_{lr}\|_{\mathcal{K}_l}$. Hence, according to the Representer's Theorem (see e.g., [3], [12]), the minimizing f_{lr} can be expressed as $f_{lr}(n)=\sum_{n'=1}^{N}K_l(n,n')\beta_{lr,n'}$ for some $\beta_{lr,n'}\in\mathbb{R}$. In other words, f_{lr} is a linear combination of the kernel K_l evaluated only at the observed $n'\in\mathcal{N}$. The result holds for all f_{lr}'s. Every g_{mr} can be similarly expressed as $g_{mr}(t)=\sum_{t'=1}^{T}G_m(t,t')\gamma_{mr,t'}$ for some $\gamma_{mr,t'}\in\mathbb{R}$. Hence, the functional minimization is converted to the vector minimization (11) over the coefficients $\beta_{lr,n'}$ and $\gamma_{mr,t'}$, collected in matrices $\{\mathbf{B}_l\in\mathbb{R}^{N\times R}\}_{l=1}^L$ and $\{\mathbf{\Gamma}_m\in\mathbb{R}^{T\times R}\}_{m=1}^M$.

Since (11) admits low-rank minimizers anyway, the column dimension of $\{\mathbf{B}_l\}$ and $\{\mathbf{\Gamma}_m\}$ could be possibly restricted to a small R_0. If the $\mathbf{P}$ minimizing (11) over this restricted feasible set turns out to be of rank smaller than R_0, the restriction comes at no loss of optimality; see also [5], [1], [18], [16]. The dimension R will be henceforth set to 20.

IV. Block-Coordinate Descent Algorithm

Problem (11) not only is nonconvex, but it involves multiple high-dimensional matrices. The block-coordinate descent (BCD) algorithm developed next scales well with the problem dimensions and converges to a stationary point of (11). According to the BCD methodology, the initial optimization variable is partitioned into blocks. Per block minimizations retaining the rest of the variables fixed are then iterated cyclically over blocks.

The variable blocks are selected here in the order $\{\mathbf{B}_1,\ldots,\mathbf{B}_L,\mathbf{\Gamma}_1,\ldots,\mathbf{\Gamma}_M\}$. The per block minimizations involved are detailed next. Consider minimizing (11) over a specific $\mathbf{B}_l$, while all other variables are fixed to their most recent values $\{\hat{\mathbf{B}}_{l'}\}_{l'\neq l}$ and $\{\hat{\mathbf{\Gamma}}_m\}_{m=1}^M$. Upon rearranging terms in (11), block $\mathbf{B}_l$ can be updated as

$$\hat{\mathbf{B}}_l=\arg\min_{\mathbf{B}_l}\ \|\mathbf{Z}_l^B-\mathbf{K}_l\mathbf{B}_l\mathbf{H}^\top\|_F^2+\mu\|\mathbf{B}_l\|_{\mathbf{K}_l} \tag{12}$$

where $\mathbf{H}:=\sum_{m=1}^{M}\mathbf{G}_m\hat{\mathbf{\Gamma}}_m$ and $\mathbf{Z}_l^B:=\mathbf{Z}-\sum_{l'\neq l}\mathbf{K}_{l'}\hat{\mathbf{B}}_{l'}\mathbf{H}^\top$. Similarly, a particular $\mathbf{\Gamma}_m$ can be updated as

$$\hat{\mathbf{\Gamma}}_m=\arg\min_{\mathbf{\Gamma}_m}\ \|\mathbf{Z}_m^\Gamma-\mathbf{F}\mathbf{\Gamma}_m^\top\mathbf{G}_m\|_F^2+\mu\|\mathbf{\Gamma}_m\|_{\mathbf{G}_m} \tag{13}$$

where $\mathbf{F}:=\sum_{l=1}^{L}\mathbf{K}_l\hat{\mathbf{B}}_l$ and $\mathbf{Z}_m^\Gamma:=\mathbf{Z}-\sum_{m'\neq m}\mathbf{F}\mathbf{\Gamma}_{m'}^\top\mathbf{G}_{m'}$.

Problems (12) and (13) exhibit the same canonical form that can be efficiently solved according to the following lemma generalizing the results of [20] to the matrix case.

Lemma 4. *Let* $\mathbf{A}\in\mathbb{R}^{d_1\times d_3}$, $\mathbf{B}\in\mathbb{S}_{++}^{d_1}$, $\mathbf{C}\in\mathbb{R}^{d_3\times d_2}$, *and* $\mu>0$. *The convex optimization problem*

$$\min_{\mathbf{X}}\ \|\mathbf{A}-\mathbf{B}\mathbf{X}\mathbf{C}^\top\|_F^2+\mu\|\mathbf{X}\|_{\mathbf{B}} \tag{14}$$

has a unique minimizer $\hat{\mathbf{X}}$ *provided by the solution of*

$$\mathbf{B}\hat{\mathbf{X}}\mathbf{C}^\top\mathbf{C}+\frac{\mu^2}{4\hat{w}}\hat{\mathbf{X}}=\mathbf{A}\mathbf{C} \tag{15}$$

if $\|\mathbf{B}^{1/2}\mathbf{A}\mathbf{C}\|_F>\mu/2$; *or,* $\hat{\mathbf{X}}=\mathbf{0}$, *otherwise. The scalar* $\hat{w}>0$ *in* (15) *is the minimizer of the convex problem*

$$\hat{w}:=\arg\min_{w\geq 0}\ w-\sum_{i=1}^{d_1}\sum_{j=1}^{d_2}\frac{[\mathbf{W}]_{ij}^2\lambda_i\mu_j w}{\lambda_i\mu_j w+\mu^2/4} \tag{16}$$

where $\mathbf{W}:=\mathbf{U}_B^\top\mathbf{A}\mathbf{U}_C$; $(\mathbf{U}_B,\{\lambda_i\}_{i=1}^{d_1})$ *are the eigenpairs of* $\mathbf{B}$; *and* $(\mathbf{U}_C,\{\mu_j\}_{j=1}^{d_2})$ *the non-zero eigenpairs of* $\mathbf{C}\mathbf{C}^\top$.

Lemma 4 asserts that if $\|\mathbf{B}^{1/2}\mathbf{A}\mathbf{C}\|_F\leq\mu/2$, the sought $\hat{\mathbf{X}}$ is set to zero. This property reveals that some of the $\{\hat{\mathbf{B}}_l\}$ and $\{\hat{\mathbf{\Gamma}}_m\}$ minimizing (11) will be zero, thus effecting kernel selection. When $\|\mathbf{B}^{1/2}\mathbf{A}\mathbf{C}\|_F>\mu/2$, the non-zero $\hat{\mathbf{X}}$ can be computed in $\mathcal{O}(d_1^3+d_2^3)$ numerical operations after rewriting (15) as a Sylvester equation. The scalar $\hat{w}$ involved in (15) can be found by any algorithm solving the univariate minimization in (16).

The BCD algorithm for solving (11) is tabulated as Alg. 1. The subroutine SolveCanonical$(\mathbf{A},\mathbf{B},\mathbf{C},\mu)$ returns the minimizer of (14). Due to the separability of the non-differentiable cost over the chosen variable blocks and the uniqueness of the per block minimizers, the BCD iterates are guaranteed to converge to a stationary point of (11) [22].

Algorithm 1 BCD algorithm for solving (11)

Input: $\mathbf{Z}$, $\{\mathbf{K}_l\}_{l=1}^L$, $\{\mathbf{G}_m\}_{m=1}^M$, R, μ

Randomly initialize $\{\hat{\mathbf{B}}_l\}_{l=1}^L$ and $\{\hat{\mathbf{\Gamma}}_m\}_{m=1}^M$
Compute $\mathbf{F}=\sum_{l=1}^L \mathbf{K}_l\hat{\mathbf{B}}_l$ and $\mathbf{H}=\sum_{m=1}^M \mathbf{G}_m\hat{\mathbf{\Gamma}}_m$
repeat
  for $l=1\rightarrow L$ **do**
    Update $\mathbf{F}=\mathbf{F}-\mathbf{K}_l\hat{\mathbf{B}}_l$
    Define $\mathbf{Z}_l^B=\mathbf{Z}-\mathbf{F}\mathbf{H}^\top$
    $\hat{\mathbf{B}}_l=$ SOLVECANONICAL($\mathbf{Z}_l^B$,$\mathbf{K}_l$,$\mathbf{H}$,μ)
    Update $\mathbf{F}=\mathbf{F}+\mathbf{K}_l\hat{\mathbf{B}}_l$
  end for
  for $m=1\rightarrow M$ **do**
    Update $\mathbf{H}=\mathbf{H}-\mathbf{G}_m\hat{\mathbf{\Gamma}}_m$
    Define $\mathbf{Z}_m^\Gamma=\mathbf{Z}-\mathbf{F}\mathbf{H}^\top$
    $\hat{\mathbf{\Gamma}}_m=$ SOLVECANONICAL($(\mathbf{Z}_m^\Gamma)^\top$,$\mathbf{G}_m$,$\mathbf{F}$,$\mu$)
    Update $\mathbf{H}=\mathbf{H}+\mathbf{G}_m\hat{\mathbf{\Gamma}}_m$
  end for
until convergence.

Output: $\{\hat{\mathbf{B}}_l\}_{l=1}^L$, $\{\hat{\mathbf{\Gamma}}_m\}_{m=1}^M$

V. Numerical Tests

The proposed multi-kernel learning approach with the BCD solver was tested using real data from the Midwest ISO (MISO) market. Day-ahead hourly LMPs were collected across $N=1,732$ nodes for three consecutive months from June 1 to August 31, 2012, a total of 2,208 hours.

Two pools of $K=5$ nodal and $L=5$ temporal kernels were constructed as briefly outlined next. Kernels $\mathbf{K}_1$ and $\mathbf{K}_2$ were selected as the regularized and the diffusion Laplacian kernel of a surrogate of the nodal connectivity graph; $\mathbf{K}_3$ is a Gaussian kernel of the categorical features using the information of nodal names and types, i.e., generator, load, interface, and hub. Kernel $\mathbf{K}_4$ was chosen to be the identity matrix capturing potential independence, while $\mathbf{K}_5$ the sample covariance of historical prices. Regarding temporal kernels, several features were utilized including yesterday's same-hour LMPs; load, outage, and weather forecasts; as well as categorical features such as hour of the day, day of the week, and a holiday indicator. Kernels $\{\mathbf{G}_m\}_{m=1}^5$ were designed by plugging these features into the linear and the Gaussian kernel for different bandwidth values and feature subsets.

Market prices in $\mathbf{Z}$ were centered upon subtracting the per-hour sample mean to cope with cyclostationarity. Hence, instead of forecasting the absolute nodal prices, the developed predictor will forecast the mean-compensated ones, which are of interest in practice since bilateral transactions depend on exactly such nodal differentials [7]. Consider the possible non-stationarity, hourly prices of a day were predicted using the market data of the previous week. The regularization parameter μ was tuned via cross-validation over the first two weeks. To enhance convergence speed and numerical accuracy of the solution to problem (16), the solver `fmincon` with the interior-point method in Matlab was used for solving the univariate optimization problem.

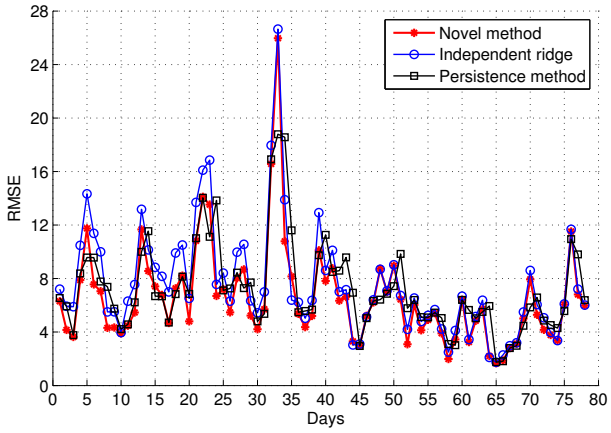

Fig. 1. RMSE comparison of forecasting methods. The RMSEs averaged across 78 evaluation days are 6.53 (red), 7.55 (blue), and 7.20 (black).

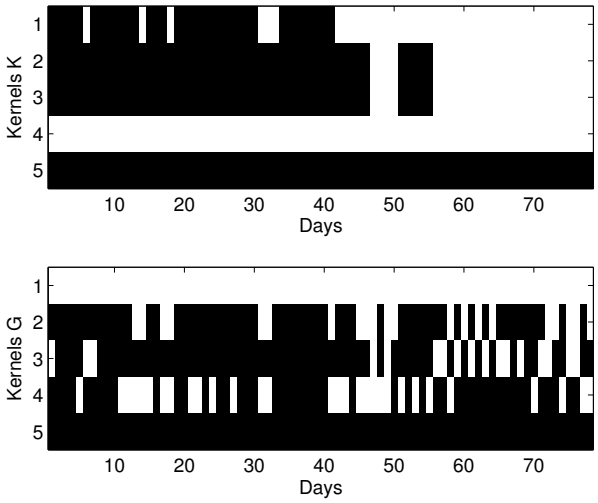

Fig. 2. Kernel selection: a black (white) square indicates that the respective kernel has been selected (eliminated) while forecasting that specific day.

The forecasting performance is provided in Fig. 1. Specifically, three methods were tested: (i) the novel low-rank multi-kernel learning method; (ii) the ridge regression forecast where each node predictor is independently obtained by solving $\min_{\mathbf{a}} \|\mathbf{z}-\mathbf{G}_1\mathbf{a}\|_2^2+\mu\mathbf{a}^T\mathbf{G}_1\mathbf{a}$; and (iii) the persistence method which simply repeats yesterday's prices. Clearly, the derived low-rank and sparsity-leveraging multi-kernel forecast attains almost consistently the lowest root mean-square error (RMSE).

Figure 2 shows the kernel selection capability of the novel multi-kernel learning approach. Checking whether the obtained $\{\|\mathbf{B}_l\|_{\mathbf{K}_l}\}_{l=1}^L$ and $\{\|\mathbf{\Gamma}_m\|_{\mathbf{G}_m}\}_{m=1}^M$ are zero or not, indicates whether the corresponding kernels, $\{\mathbf{K}_l\}$ and $\{\mathbf{G}_m\}$ have been eliminated. Interestingly, out of the 10 kernels, the identity kernel $\mathbf{K}_4$ and the linear kernel $\mathbf{G}_1$ with a prescribed bandwidth were consistently not selected.

Figure 3 depicts the singular values of 78 successive matrices $\mathbf{Z}$ in decreasing order. The fast-decaying distribution implies that most of the market information could be possibly captured by the top 20 singular values. Such an observation

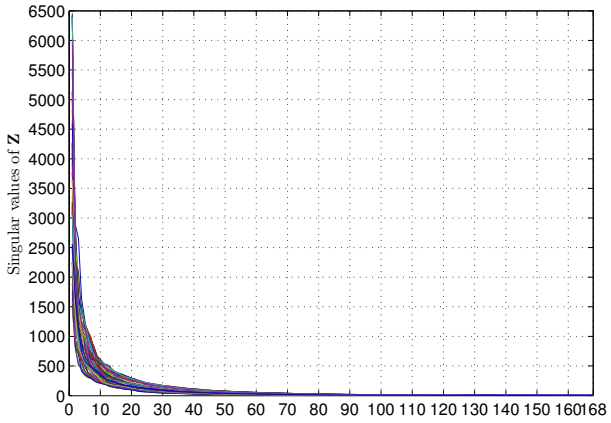

Fig. 3. Sorted singular values for 78 matrices $\mathbf{Z}$ appearing in (11).

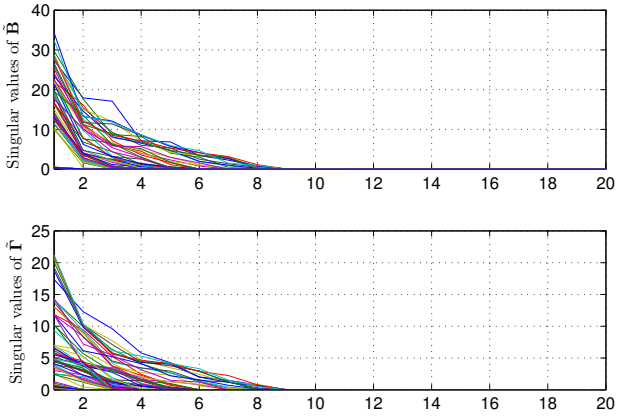

Fig. 4. Sorted singular values for 78 matrices $\tilde{\mathbf{B}} := [\mathbf{B}_1^\top, \ldots, \mathbf{B}_L^\top]^\top$ and $\tilde{\mathbf{\Gamma}} := [\mathbf{\Gamma}_1^\top, \ldots, \mathbf{\Gamma}_M^\top]^\top$, where the predictors $\{\mathbf{B}_l\}$ and $\{\mathbf{\Gamma}_m\}$ are the optimal solutions to problem (11).

not only justifies the trace norm regularization in (5), but also hints at fixing R to 20 for a good complexity-performance tradeoff. Finally, Fig. 4 shows the sorted singular values of matrices $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{\Gamma}}$ as obtained by solving (11). Clearly, the rank of $\tilde{\mathbf{B}}$'s and $\tilde{\mathbf{\Gamma}}$'s is no more than 10 in all 78 predictions, which justifies the prescribed choice of $R = 20$.

VI. Conclusions

The contribution of this paper is three-fold. On the application side, upon recognizing that electricity prices exhibit low-rank properties, market inference was postulated as a trace-norm regularized kernel-based learning task. The optimization framework derived enables market data extrapolation, imputation, and extrapolation. On the learning side, this paper develops a systematic kernel selection methodology under a collaborative filtering or matrix completion setup. Algorithmically, a BCD-based solver handling efficiently high-dimensional market data and converging to a stationary point was developed. Our findings were corroborated using real market data.

References

[1] 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, 2009.

[2] N. Amjady and M. Hemmati, "Energy price forecasting - problems and proposals for such predictions," *IEEE Power Energy Mag.*, vol. 4, no. 2, pp. 20–29, Mar./Apr. 2006.

[3] A. Argyriou, C. A. Michelli, and M. Pontil, "When is there a representer theorem? Vector versus matrix regularizers," *J. Machine Learning Res.*, vol. 10, pp. 2507–2529, 2009.

[4] N. Aronszajn, "Theory of reproducing kernels," *Trans. of the American Mathematical Society*, vol. 68, no. 3, pp. 337–404, May 1950.

[5] J. A. Bazerque and G. B. Giannakis, "Nonparametric basis pursuit via sparse kernel-based learning," *IEEE Signal Process. Mag.*, vol. 12, pp. 112–125, Jul. 2013.

[6] A. J. Conejo, M. A. Plazas, R. Espinola, and A. B. Molina, "Day-ahead electricity price forecasting using the wavelet transform and ARIMA models," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 1035–1042, May 2005.

[7] S. J. Deng and S. S. Oren, "Electricity derivatives and risk management," *Energy*, vol. 31, no. 6, pp. 940–953, 2006.

[8] Electric Reliability Council of Texas (ERCOT), "Ercot launches wholesale pricing forecast tool," July 11, 2012. [Online]. Available: http://www.ercot.com/news/press_releases/show/26244

[9] G. B. Giannakis, V. Kekatos, N. Gatsis, S.-J. Kim, H. Zhu, and B. Wollenberg, "Monitoring and optimization for power grids: A signal processing perspective," *IEEE Signal Process. Mag.*, vol. 30, no. 5, pp. 107–128, Sep. 2013.

[10] A. Gómez-Expósito, A. J. Conejo, and C. Canizares, Eds., *Electric Energy Systems, Analysis and Operation*. Boca Raton, FL: CRC Press, 2009.

[11] M. Gonen and E. Alpaydin, "Multiple kernel learning algorithms," *J. Machine Learning Res.*, vol. 12, pp. 2211–2268, Sep. 2011.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, 2009.

[13] V. Kekatos, Y. Zhang, and G. B. Giannakis, "Electricity market forecasting via low-rank multi-kernel learning," *IEEE J. Sel. Topics Signal Process.*, Oct. 2013 (submitted). [Online]. Available: http://arxiv.org/abs/1310.0865

[14] D. Kirschen and G. Strbac, *Power System Economics*. West Sussex, England: Wiley, 2010.

[15] E. D. Kolaczyk, *Statistical Analysis of Network Data, Methods and Models*. New York, NY: Springer, 2010.

[16] V. Koltchinskii and M. Yuan, "Sparsity in multiple kernel learning," *The Annals of Statistics*, vol. 38, no. 6, pp. 3660–3695, 2010.

[17] M. Mardani, G. Mateos, and G. B. Giannakis, "Decentralized sparsity-regularized rank minimization: Algorithms and applications," *IEEE Trans. Signal Process.*, vol. 61, no. 21, pp. 5374–5388, Nov. 2013.

[18] C. Michelli and M. Pontil, "Learning the kernel function via regularization," *J. Machine Learning Res.*, vol. 6, pp. 1099–1125, Sep. 2005.

[19] A. L. Ott, "Experience with PJM market operation, system design, and implementation," *IEEE Trans. Power Syst.*, vol. 18, no. 2, pp. 528–534, May 2003.

[20] A. T. Puig, A. Wiesel, G. Fleury, and A. H. Hero, "Multidimensional shrinkage-thresholding operator and group LASSO penalties," *IEEE Signal Process. Lett.*, vol. 18, no. 6, pp. 363–366, Jun. 2011.

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

[22] P. Tseng, "Convergence of block coordinate descent method for nondifferentiable minimization," *Journal on Optimization Theory and Applications*, vol. 109, pp. 475–494, Jun. 2001.

[23] U.S. Department of Energy, "National Electric Transmission Congestion Study," 2012. [Online]. Available: http://energy.gov/oe/services/electricity-policy-coordination-and-implementation/transmission-planning/2012-national

[24] L. Wu and M. Shahidehpour, "A hybrid model for day-ahead price forecasting," *IEEE Trans. Power Syst.*, vol. 25, no. 3, pp. 1519–1530, Aug. 2010.

[25] Q. Zhou, L. Tesfatsion, and C.-C. Liu, "Short-term congestion forecasting in wholesale power markets," *IEEE Trans. Power Syst.*, vol. 26, no. 4, pp. 2185–2196, Nov. 2011.