KERNEL SELECTION FOR POWER MARKET INFERENCE VIA BLOCK SUCCESSIVE UPPER BOUND MINIMIZATION

Vassilis Kekatos, Yu Zhang, and Georgios Giannakis

Digital Technology Center and ECE Dept., University of Minnesota
Minneapolis, MN 55455, USA
Emails: {kekatos,zhan1220,georgios}@umn.edu

ABSTRACT

Advanced data analytics are undoubtedly needed to enable the envisioned smart grid functionalities. Towards that goal, modern statistical learning tools are developed for day-ahead electricity market inference. Congestion patterns are modeled as rank-one components in the matrix of spatio-temporal prices. The new kernel-based predictor is regularized by the square root of the nuclear norm of the sought matrix. Such a regularizer not only promotes low-rank solutions, but it also facilitates a systematic kernel selection methodology. The non-convex optimization problem involved is efficiently driven to a stationary point following a block successive upper bound minimization approach. Numerical tests on real high-dimensional market data corroborate the interpretative merits and the computational efficiency of the novel method.

Index Terms— Kernel learning; nuclear norm; multi-kernel selection; block successive upper bound minimization.

1. INTRODUCTION

In deregulated electricity markets, an independent system operator (ISO) collects bids submitted by generators and utilities [1]. Compliant with network and reliability constraints, the grid is dispatched in the most economical way. Load patterns and congested transmission lines lead to spatiotemporally-varying energy prices, known as locational marginal prices (LMPs) [2], [3]. Electricity price inference is an important decision making tool for market participants. Further, ISOs recently broadcast price forecasts to proactively relieve congestion [4], and system planners use LMP analytics to identify transmission corridors [5].

Schemes for predicting electricity prices proposed so far include time-series analysis approaches based on auto-regressive (integrated) moving average models and their generalizations; see e.g., [6], [7]. However, these models are confined to linear predictors, whereas markets involve generally nonlinear dependencies. To account for nonlinearities, artificial intelligence approaches, such as fuzzy systems and neural networks, have been also investigated [8], [9], [10], [11]. In [12], market clearance is assumed to be solved as a quadratic program and forecasts are extracted based on the most probable outage combinations. Reviews on electricity price forecasting can be found in [13] and [14].

Different from existing approaches where predictors are trained on a per-LMP basis, a grid-wide kernel-based learning approach is pursued here. Leveraging the price dependence across nodes and hours, market forecasting is cast as a collaborative filtering task [15], [16]. To promote low-rank models, a novel regularizer based on the square root of the nuclear norm of the involved price matrix is introduced. Our analytic results extend kernel selection tools to low-rank multi-task models [17], [18]. The final contribution is an efficient algorithm for solving the non-convex problem involved. Distinct from [19], the solver here minimizes an upper bound of the per block minimizations, hence allowing inference based on market data of even higher dimensions. Forecasting results on the Midwest ISO (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 the transpose and the trace of $\mathbf{A}$, respectively. The ℓ_2-norm of a vector is denoted by $\|\mathbf{a}\|_2$, $\|\mathbf{A}\|_F$ is the Frobenius matrix norm, and $\mathbb{S}_{++}^N$ is the set of $N \times N$ positive definite matrices.

2. PRELIMINARIES ON KERNEL LEARNING

Given pairs $\{(x_n, z_n)\}_{n=1}^N$ of features x_n drawn from a space $\mathcal{X}$ and target values $z_n \in \mathbb{R}$; kernel-based learning aims finding a function $f : \mathcal{X} \to \mathbb{R}$ belonging to the space $\mathcal{H}_\mathcal{K} := \{f(x) = \sum_{n=1}^\infty K(x, x_n)a_n,\ a_n \in \mathbb{R}\}$ defined by a kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. When $K(\cdot,\cdot)$ is a symmetric positive definite function, then $\mathcal{H}_\mathcal{K}$ becomes a reproducing kernel Hilbert space (RKHS) equipped with the norm $\|f\|_\mathcal{K}^2 := \sum_{n=1}^\infty \sum_{m=1}^\infty K(x_n, x_m)a_n a_m < \infty$ [20].

The sought f can be found via the regularization [21], [22]

$$\hat{f}_\mathcal{K} := \arg\min_{f \in \mathcal{H}_\mathcal{K}} \sum_{n=1}^N (z_n - f(x_n))^2 + \mu \|f\|_\mathcal{K}. \quad (1)$$

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.

The regularizer $\|f\|_{\mathcal{K}}$ constraints $f \in \mathcal{H}_{\mathcal{K}}$ 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 [21].

According to the Representer's Theorem, $\hat{f}_{\mathcal{K}}$ admits the form $\hat{f}_{\mathcal{K}}(x) = \sum_{n=1}^{N} K(x, x_n)\hat{a}_n$ for some $\hat{\mathbf{a}} := [\hat{a}_1 \cdots \hat{a}_N]^\top$ [16], [21]. Thus, the functional optimization in (1) is equivalent to the vector optimization

$$\hat{\mathbf{a}} := \arg\min_{\mathbf{a}} \ \|\mathbf{z} - \mathbf{K}\mathbf{a}\|_2^2 + \mu\|\mathbf{a}\|_{\mathbf{K}} \tag{2}$$

where $\mathbf{z} := [z_1 \cdots z_N]^\top$, $\mathbf{K} \in \mathbb{S}_{++}^N$ is the kernel matrix having entries $[\mathbf{K}]_{n,m} := K(x_n, x_m)$, and $\|\mathbf{a}\|_{\mathbf{K}}^2 := \mathbf{a}^\top\mathbf{K}\mathbf{a}$.

The designer is often given candidate kernels $\{K_l\}_{l=1}^L$ and would like to determine which of them provide better inference results. Towards that goal, let the kernel function K in (1) be defined as the convex combination $\sum_{l=1}^L K_l\theta_l$ for some $\theta_l \geq 0$ and $\sum_{l=1}^L \theta_l = 1$. Given $\{(x_n, z_n)\}$ and $\{K_l\}$, kernels $\{K_l\}$ can be selected by minimizing (1) over θ_l's. This double minimization turns out to be equivalent to [23]

$$\{\hat{\mathbf{a}}_l\} := \arg\min_{\{\mathbf{a}_l\}} \ \|\mathbf{z} - \sum_{l=1}^{L} \mathbf{K}_l\mathbf{a}_l\|_2^2 + \mu \sum_{l=1}^{L} \|\mathbf{a}_l\|_{\mathbf{K}_l}. \tag{3}$$

Problem (3) is known to yield many zero $\hat{\mathbf{a}}_l$'s, hence featuring kernel selection. Based on kernel learning, novel models pertinent to power market inference are developed next.

3. PROBLEM FORMULATION

Consider an hourly whole-sale electricity market over a set $\mathcal{N}$ of N pricing nodes indexed by n. The market is considered to be stationary over the T most recent hours comprising the set $\mathcal{T}$. In a day-ahead market, locational marginal prices (LMPs) correspond to the cost of electricity at each node and over one-hour periods for the following day [24]. Viewing market forecasting as an inference problem, hourly LMPs are the target variables. Explanatory variables (features) can be any relevant data: Weather forecasts and load estimates could be utilized as time-related features. The pricing node type (e.g., generator or load) and its geographical location could be node-related features.

Kernel-based predictors could be obviously built on a per-node basis. But locational prices are not independent: they are determined over a transmission grid having capacity and reliability limitations [2], [3]. Leveraging this network-imposed dependence, the price at node n and time t denoted by $p(n,t)$ could be thought of as a function $p : \mathcal{N} \times \mathcal{T} \to \mathbb{R}$ to be inferred. Rigorously, we postulate that $p(n,t)$ belongs to the RKHS $\mathcal{P}$ defined by the product kernel $K_{\otimes}\left((n,t),(n',t')\right) := K(n,n')G(t,t')$, where K and G are kernels evaluated over nodes and hours, respectively. All functions in this RKHS can be expressed as [20], [15]

$$\mathcal{P} = \left\{ p(n,t) = \sum_{r=1}^{R} f_r(n)g_r(t),\ f_r \in \mathcal{H}_K,\ g_r \in \mathcal{H}_G \right\} \tag{4}$$

where $\mathcal{H}_{\mathcal{K}}$ and $\mathcal{H}_{\mathcal{G}}$ are the RKHSs defined respectively by K and G; and R is possibly infinite.

The key presumption here is that $p(n,t)$ is practically the superposition of few $p_r(n,t) := f_r(n)g_r(t)$. At a specific t, usually only a few transmission lines are congested [2], [3]. Each f_r corresponds to the pricing pattern observed whenever a specific congestion scenario occurs. Yet spatial effects are modulated by time. For example, congestion typically occurs during peak demand or high-wind periods. These specifications not only justify using the product kernel $K_{\otimes}$, but they also hint at a relatively small R in (4).

To facilitate parsimonious modeling of $p(n,t)$ using a few $p_r(n,t)$'s, instead of regularizing by $\|p\|_{K_{\otimes}}$, the *trace norm* $\|p\|_*$ is used instead [15], [19]. Upon arranging observed prices and $p(n,t)$'s in the $N \times T$ matrices $\mathbf{Z}$ and $\mathbf{P}$, respectively, market inference is cast as

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

Regularizing by $\|p\|_*$ is known to favor low-rank models [15]. Actually, when $\mathcal{N}$ and $\mathcal{T}$ are Euclidean spaces, $K(n,n') = \delta(n-n')$ and $G(t,t') = \delta(t-t')$ where $\delta(\cdot)$ is the Kronecker delta function; then $p(n,t)$ is the (n,t)-th entry of $\mathbf{P}$ and $\|p\|_*$ is simply its nuclear norm $\|\mathbf{P}\|_*$, i.e., the sum of its singular values [22]. Employing the square root of $\|p\|_*$ in (5) not only inherits this low-rank promoting property, but it facilitates kernel selection and efficient algorithms [19].

Learning the kernels $\mathcal{K}$ and $\mathcal{G}$ is accomplished next by generalizing the multi-kernel learning approach of [17] to the function regularization in (5). In detail, given two sets of candidate kernels, $\{K_l\}_{l=1}^L$ and $\{G_m\}_{m=1}^M$, consider the kernel spaces constructed as the convex hulls, i.e., $\mathcal{K} := \text{conv}\left(\{K_l\}_{l=1}^L\right)$ and $\mathcal{G} := \text{conv}\left(\{G_m\}_{m=1}^M\right)$. Minimizing the outcome of (5) over $\mathcal{K}$ and $\mathcal{G}$ – essentially over the weights of the two convex combinations – provides a disciplined kernel design methodology. The following result asserts that this optimization can be accomplished without even finding the optimal weights.

Theorem 1 ([19])**.** *Consider the function space and the kernel spaces $\mathcal{K}$ and $\mathcal{G}$. Solving the regularization problem*

$$\min_{\mathcal{K},\mathcal{G}} \min_{p \in \mathcal{P}} \ \|\mathbf{Z} - \mathbf{P}\|_F^2 + \mu\sqrt{\|p\|_*}. \tag{6}$$

is equivalent to solving

$$\min_{p \in \mathcal{P}'} \|\mathbf{Z}-\mathbf{P}\|_F^2 + \mu \sum_{l=1}^{L} \sqrt{\sum_{r=1}^{R} \|f_{lr}\|_{\mathcal{K}_l}^2} + \mu \sum_{m=1}^{M} \sqrt{\sum_{r=1}^{R} \|g_{mr}\|_{\mathcal{G}_m}^2} \tag{7}$$

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

Practically solving (7) necessitates transforming the functional to a vector minimization, as pursued next. Note that minimizing (7) 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, each one of the LR functions f_{lr} minimizing (7) can be expressed as

$$f_{lr}(n) = \sum_{n'=1}^{N} K_l(n,n')\beta_{lr,n'}. \tag{8}$$

Upon definining $\boldsymbol{\beta}_{lr} := [\beta_{lr,1} \ \cdots \ \beta_{lr,N}]^\top$, and $\mathbf{f}_{lr} := [f_{lr}(1) \ \cdots \ f_{lr}(N)]^\top$, it holds that $\mathbf{f}_{lr} = \mathbf{K}_l\boldsymbol{\beta}_{lr}$, where $\mathbf{K}_l \in \mathbb{S}_{++}^N$ is the node kernel matrix whose (n,n')-th entry is $K_l(n,n')$. Likewise, each g_{mr} minimizing (7) is written as

$$g_{mr}(t) = \sum_{t'=1}^{T} G_m(t,t')\gamma_{mr,t'}. \tag{9}$$

Similar to $\mathbf{f}_{lr}$, the vector $\mathbf{g}_{mr} := [g_{mr}(1) \ \ldots \ g_{mr}(T)]^\top$ is expressed in terms of the time kernel matrix $\mathbf{G}_m \in \mathbb{S}_{++}^T$ and vector $\boldsymbol{\gamma}_{mr} := [\gamma_{mr,1} \ \cdots \ \gamma_{mr,T}]^\top$, via $\mathbf{g}_{mr} = \mathbf{G}_m\boldsymbol{\gamma}_{mr}$.

Plugging (8)-(9) into the decomposition model dictated by $\mathcal{P}'$ in (7), and after some matrix manipulations yields

$$\mathbf{P} = \sum_{l=1}^{L}\sum_{m=1}^{M} \mathbf{K}_l\mathbf{B}_l\boldsymbol{\Gamma}_m^\top\mathbf{G}_m \tag{10}$$

where $\mathbf{B}_l := [\boldsymbol{\beta}_{l1} \ \cdots \ \boldsymbol{\beta}_{lR}]$ and $\boldsymbol{\Gamma}_m := [\boldsymbol{\gamma}_{m1} \ \cdots \ \boldsymbol{\gamma}_{mR}]$. Using again (8)-(9), the function norms can be written as $\|f_{lr}\|_{\mathcal{K}_l}^2 = \boldsymbol{\beta}_{lr}^\top\mathbf{K}_l\boldsymbol{\beta}_{lr}$ and $\|g_{mr}\|_{\mathcal{G}_m}^2 = \boldsymbol{\gamma}_{mr}^\top\mathbf{G}_m\boldsymbol{\gamma}_{mr}$. Using the properties of the trace operator, it follows that

$$\sum_{r=1}^{R}\|f_{lr}\|_{\mathcal{K}_l}^2 = \|\mathbf{B}_l\|_{\mathbf{K}_l}^2,\ \sum_{r=1}^{R}\|g_{mr}\|_{\mathcal{G}_m}^2 = \|\boldsymbol{\Gamma}_m\|_{\mathbf{G}_m}^2 \tag{11}$$

where $\|\mathbf{X}\|_{\mathbf{B}}^2 := \mathrm{Tr}(\mathbf{X}^\top\mathbf{B}\mathbf{X})$ for any $\mathbf{B} \succ \mathbf{0}$. By (10)-(11), the functional optimization in (7) can be compactly expressed as the non-convex matrix optimization problem

$$\min_{\mathbf{P},\{\mathbf{B}_l\},\{\boldsymbol{\Gamma}_m\}} \|\mathbf{Z}-\mathbf{P}\|_F^2 + \mu\sum_{l=1}^{L}\|\mathbf{B}_l\|_{\mathbf{K}_l} + \mu\sum_{m=1}^{M}\|\boldsymbol{\Gamma}_m\|_{\mathbf{G}_m}$$
$$\text{s.to } \mathbf{P} = \sum_{l=1}^{L}\sum_{m=1}^{M}\mathbf{K}_l\mathbf{B}_l\boldsymbol{\Gamma}_m^\top\mathbf{G}_m. \tag{12}$$

Since (12) admits low-rank minimizers anyway, the column dimension of $\{\mathbf{B}_l\}$ and $\{\boldsymbol{\Gamma}_m\}$ could be possibly restricted to a small R_0. If the $\mathbf{P}$ minimizing (12) 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 [22], [15], [17], [23]. The dimension R will be henceforth set to 20.

Remark 1. Having solved (12), price forecasts can be issued not only for $t \notin \mathcal{T}$, but also for new nodes $n \notin \mathcal{N}$. This is an important feature when dealing with markets having seasonal pricing models: e.g., MISO updates its commercial grid quarterly by adding, removing, and redefining nodes.

4. BSUM SOLVER

A block-coordinate descent (BCD) solver of (12) was developed in [19]. That BCD solver partitioned optimization variables into blocks $\{\mathbf{B}_1,\ldots,\mathbf{B}_L,\boldsymbol{\Gamma}_1,\ldots,\boldsymbol{\Gamma}_M\}$. By cyclically iterating over blocks, per block minimizations were carried out *exactly* while retaining the rest of the variables fixed. To handle market data of even higher dimensions, a block successive upper minimization (BSUM) solver is devised next.

Consider minimizing (12) over a specific block $\mathbf{B}_l$, while all other variables are maintained to their most recent values $\{\hat{\mathbf{B}}_{l'}\}_{l'\neq l}$, $\{\hat{\boldsymbol{\Gamma}}_m\}_{m=1}^M$. Upon rearranging terms in (12), $\mathbf{B}_l$ can be updated as

$$\hat{\mathbf{B}}_l = \arg\min_{\mathbf{B}_l} \|\mathbf{Z}_l^B - \mathbf{K}_l\mathbf{B}_l\mathbf{H}^\top\|_F^2 + \mu\|\mathbf{B}_l\|_{\mathbf{K}_l} \tag{13}$$

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

$$\hat{\boldsymbol{\Gamma}}_m = \arg\min_{\boldsymbol{\Gamma}_m} \|\mathbf{Z}_m^\Gamma - \mathbf{F}\boldsymbol{\Gamma}_m^\top\mathbf{G}_m\|_F^2 + \mu\|\boldsymbol{\Gamma}_m\|_{\mathbf{G}_m} \tag{14}$$

where $\mathbf{F} := \sum_{l=1}^{L}\mathbf{K}_l\hat{\mathbf{B}}_l$ and $\mathbf{Z}_m^\Gamma := \mathbf{Z} - \sum_{m'\neq m}\mathbf{F}\boldsymbol{\Gamma}_{m'}^\top\mathbf{G}_{m'}$. Problems (13)-(14) exhibit the same canonical convex form:

$$\min_{\mathbf{X}} \|\mathbf{A} - \mathbf{B}\mathbf{X}\mathbf{C}^\top\|_F^2 + \mu\|\mathbf{X}\|_{\mathbf{B}} \tag{15}$$

for an $\mathbf{A} \in \mathbb{R}^{d_1\times d_3}$, $\mathbf{B} \in \mathbb{S}_{++}^{d_1}$, and $\mathbf{C} \in \mathbb{R}^{d_3\times d_2}$. Let $h(\mathbf{X})$ denote the cost function in (15). Instead of directly minimizing $h(\mathbf{X})$, the new BSUM solver successively minimizes a function $u(\mathbf{X};\hat{\mathbf{X}})$, constructed at the most recent update $\hat{\mathbf{X}}$. The function $u(\mathbf{X};\hat{\mathbf{X}})$ should satisfy $h(\mathbf{X}) \leq u(\mathbf{X};\hat{\mathbf{X}})$ and $h(\hat{\mathbf{X}}) = u(\hat{\mathbf{X}};\hat{\mathbf{X}})$ for all $\mathbf{X},\hat{\mathbf{X}}$ [25].

To derive a computationally convenient upper bound $u(\hat{\mathbf{X}};\hat{\mathbf{X}})$, let $h_1(\mathbf{X})$ denote the first summand of $h(\mathbf{X})$ and consider its Taylor expansion at $\hat{\mathbf{X}}$, which yields $h_1(\mathbf{X}) = h_1(\hat{\mathbf{X}}) - 2\,\mathrm{Tr}\left[(\mathbf{X}-\hat{\mathbf{X}})^\top\mathbf{B}^\top(\mathbf{A}-\mathbf{B}\hat{\mathbf{X}}\mathbf{C}^\top)\mathbf{C}\right] + \|\mathbf{B}(\mathbf{X}-\hat{\mathbf{X}})\mathbf{C}^T\|_F^2$. By upper bounding the third summand in this expansion, $h(\mathbf{X})$ can be upper bounded by $u(\mathbf{X};\hat{\mathbf{X}}) := \|\mathbf{A}-\mathbf{B}\hat{\mathbf{X}}\mathbf{C}^\top\|_F^2 - 2\,\mathrm{Tr}\left[(\mathbf{X}-\hat{\mathbf{X}})^\top\mathbf{B}^\top(\mathbf{A}-\mathbf{B}\hat{\mathbf{X}}\mathbf{C}^\top)\mathbf{C}\right] + \lambda_{\max}(\mathbf{C}^\top\mathbf{C})\lambda_{\max}(\mathbf{B})\|\mathbf{X}-\hat{\mathbf{X}}\|_{\mathbf{B}}^2 + \mu\|\mathbf{X}\|_{\mathbf{B}}$. Further, after ignoring constant terms and completing the squares, minimizing $u(\mathbf{X};\hat{\mathbf{X}})$ can be shown to be equivalent to

$$\min_{\mathbf{X}} \|\mathbf{X}-\bar{\mathbf{X}}\|_{\mathbf{B}}^2 + \frac{\mu}{\lambda_{\max}(\mathbf{C}^\top\mathbf{C})\lambda_{\max}(\mathbf{B})}\|\mathbf{X}\|_{\mathbf{B}} \tag{16}$$

Algorithm 1 BSUM algorithm for solving (12)

Randomly initialize $\{\hat{\mathbf{B}}_l\}_{l=1}^L$ and $\{\hat{\mathbf{\Gamma}}_m\}_{m=1}^M$
$\mathbf{F} = \sum_{l=1}^L \mathbf{K}_l\hat{\mathbf{B}}_l$;$\mathbf{H} = \sum_{m=1}^M \mathbf{G}_m\hat{\mathbf{\Gamma}}_m$; $\mathbf{R} = \mathbf{Z} - \mathbf{F}\mathbf{H}^\top$
repeat
  Compute $\mathbf{H} = \sum_{m=1}^M \mathbf{G}_m\hat{\mathbf{\Gamma}}_m$ and $\lambda_{\max}(\mathbf{H}^\top\mathbf{H})$
  for $l = 1 \to L$ **do**
    $\bar{\mathbf{B}}_l = \hat{\mathbf{B}}_l + \frac{1}{\lambda_{\max}(\mathbf{H}^\top\mathbf{H})\lambda_{\max}(\mathbf{K}_l)}\mathbf{K}_l\mathbf{R}\mathbf{H}$
    $\mathbf{R} = \mathbf{R} + \mathbf{K}_l\hat{\mathbf{B}}_l\mathbf{H}^\top$
    $\hat{\mathbf{B}}_l = \mathsf{S}(\bar{\mathbf{B}}_l; \mathbf{K}_l, \mathbf{H})$
    $\mathbf{R} = \mathbf{R} - \mathbf{K}_l\hat{\mathbf{B}}_l\mathbf{H}^\top$
  end for
  Compute $\mathbf{F} = \sum_{l=1}^L \mathbf{K}_l\hat{\mathbf{B}}_l$ and $\lambda_{\max}(\mathbf{F}^\top\mathbf{F})$
  for $m = 1 \to M$ **do**
    $\bar{\mathbf{\Gamma}}_m = \hat{\mathbf{\Gamma}}_m + \frac{1}{\lambda_{\max}(\mathbf{F}^\top\mathbf{F})\lambda_{\max}(\mathbf{G}_m)}\mathbf{G}_m\mathbf{R}^\top\mathbf{F}$
    $\mathbf{R} = \mathbf{R} + \mathbf{F}\hat{\mathbf{\Gamma}}_m^\top\mathbf{G}_m$
    $\hat{\mathbf{\Gamma}}_m = \mathsf{S}(\bar{\mathbf{\Gamma}}_m; \mathbf{G}_m, \mathbf{F})$
    $\mathbf{R} = \mathbf{R} - \mathbf{F}\hat{\mathbf{\Gamma}}_m^\top\mathbf{G}_m$
  end for
until convergence.

where $\bar{\mathbf{X}} := \hat{\mathbf{X}} + \frac{1}{\lambda_{\max}(\mathbf{C}^\top\mathbf{C})\lambda_{\max}(\mathbf{B})}\mathbf{B}(\mathbf{A} - \mathbf{B}\hat{\mathbf{X}}\mathbf{C}^\top)\mathbf{C}$. Interestingly, the minimizer of (16) is provided in closed-form

$$\mathsf{S}(\bar{\mathbf{X}}; \mathbf{B}, \mathbf{C}) := \bar{\mathbf{X}}\left[1 - \frac{\mu}{2\lambda_{\max}(\mathbf{C}^\top\mathbf{C})\lambda_{\max}(\mathbf{B})\|\bar{\mathbf{X}}\|_{\mathbf{B}}}\right]_+$$

where $[a]_+ := \max\{0, a\}$. The above update reveals that depending on the value of μ, many of the $\{\mathbf{B}_l\}$ and $\{\mathbf{\Gamma}_m\}$ will be zero matrices, thus effecting kernel selection. The BSUM solver is tabulated as Alg. 1. Notice that finding the maximum eigenvalues of $\mathbf{H}^\top\mathbf{H}$ and $\mathbf{F}^\top\mathbf{F}$ involve only $\mathcal{O}(R^2)$ operations. BSUM iterates are guaranteed to converge to a stationary point of (12) [25].

5. NUMERICAL TESTS

The proposed low-rank multi-kernel learning approach was tested using real data from the MISO market. Day-ahead hourly LMPs were collected across $N = 1,732$ nodes for the period June 1 to August 31, 2012. Two pools of $K = 5$ nodal and $L = 5$ temporal kernels were constructed as briefly outlined next; see [19] for details. Kernels $\mathbf{K}_1$ and $\mathbf{K}_2$ were selected as Laplacian kernels of a surrogate of the nodal connecivity graph; $\mathbf{K}_3$ as a Gaussian kernel; $\mathbf{K}_4$ as the identity matrix; and $\mathbf{K}_5$ as the sample covariance of historical prices. Regarding temporal kernels, several features were utilized including yesterday's same-hour LMPs; load, and weather forecasts; as well as categorical features such as hour of the day and day of the week. 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. Parameter μ is tuned via cross-validation over the first two weeks.

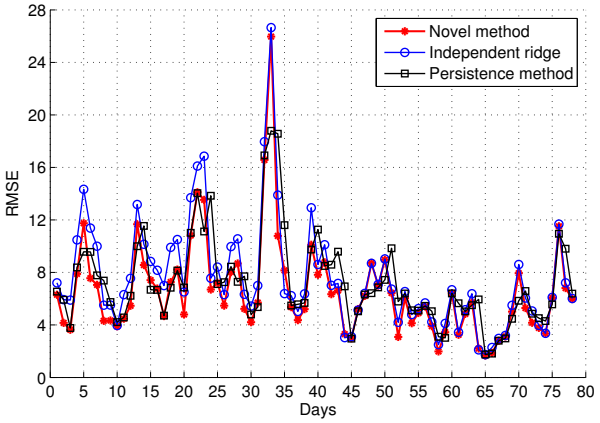

Fig. 1. RMSE comparison of forecasting methods. The RMSEs averaged across 78 evaluation days are 6.53 (red), 7.55 (blue), and 7.20 (black).

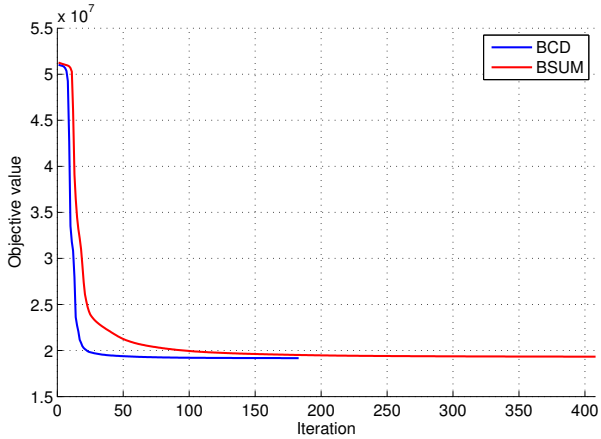

Fig. 2. Objective value convergence for BSUM and BCD in [19]. BSUM terminates in 6.5 minutes (408 iterates), while BCD in 32.3 minutes (183 iterates).

Several factors not captured by the publicly available features used here can severely affect the market. To account for this non-stationarity, the designed day-ahead predictors depend on market data only from the previous week. The forecasting performance is illustrated in Fig. 1. Three approaches were tested: (i) the novel multi-kernel learning method; (ii) a per-node ridge regression predictor; and (iii) the persistence forecast which repeats yesterday's prices. Clearly, the novel method attains almost consistently the lowest root mean-square error (RMSE). Even though $R = 20$, the $\{\mathbf{B}_l\}$ and $\{\mathbf{\Gamma}_m\}$ minimizing (12) had rank 10. Out of the 10 kernels, $\mathbf{K}_4$ and $\mathbf{G}_1$ were consistently not selected.

Figure 2 compares the objective convergence of the BSUM solver developed here, and the BCD solver of [19]. Both solvers were randomly initialized at the same point. BSUM required more iterations than BCD to reach the same cost value. Albeit, BSUM iterations are computationally more efficient than those of BCD; thus, BSUM is nearly five times faster than BCD in terms of total running time.

6. REFERENCES

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

[2] 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.

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

[4] 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

[5] 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

[6] J. Contreras, R. Espinola, F. J. Nogales, and A. J. Conejo, "ARIMA models to predict next-day electricity prices," *IEEE Trans. Power Syst.*, vol. 18, no. 3, pp. 1014–1020, Aug. 2003.

[7] R. C. Garcia, J. Contreras, M. van Akkeren, and J. B. C. Garcia, "A GARCH forecasting model to predict day-ahead electricity prices," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 867–874, May 2005.

[8] L. Zhang, P. B. Luh, and K. Kasiviswanathan, "Energy clearing price prediction and confidence interval estimation with cascaded neural network," *IEEE Trans. Power Syst.*, vol. 18, no. 1, pp. 99–105, Feb. 2003.

[9] A. M. Gonzalez, A. M. S. Roque, and J. G. Gonzalez, "Modeling and forecasting electricity prices with input/output hidden Markov models," *IEEE Trans. Power Syst.*, vol. 20, no. 1, pp. 13–24, Feb. 2005.

[10] G. Li, C.-C. Liu, C. Mattson, and J. Lawarree, "Day-ahead electricity price forecasting in a grid environment," *IEEE Trans. Power Syst.*, vol. 22, no. 1, pp. 266–274, Feb. 2007.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

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

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

[19] 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

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

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

[22] 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.

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

[24] 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.

[25] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, 2013.