Online Dictionary Learning with Group Structure Inducing Norms

Zoltán Szabó* SZZOLI@CS.ELTE.HU
*Faculty of Informatics, Eötvös Loránd University, Pázmány P. sétány 1/C, H-1117 Budapest, Hungary

Barnabás Póczos BAPOCZOS@CS.CMU.EDU
School of Computer Science, Carnegie Mellon University, 5000 Forbes Ave, 15213, Pittsburgh, PA, USA

András Lőrincz* ANDRAS.LORINCZ@ELTE.HU

1. Introduction

Thanks to the several successful applications, sparse signal representation has become one of the most actively studied research areas in machine learning. In the *sparse coding* framework one approximates the observations with the linear combination of a few vectors (basis elements) from a *fixed dictionary* (Tropp & Wright, 2010). The general sparse coding problem, i.e., the ℓ_0-norm solution that searches for the least number of basis elements, is NP-hard. To overcome this difficulty, a popular approach is to apply ℓ_p $(0 < p \leq 1)$ relaxations. The $p = 1$ special case, the Lasso problem, has become particularly popular since in this case the relaxation leads to a convex problem.

The traditional form of sparse coding does not take into account any prior information about the structure of hidden representation (also called covariates, or code). However, using *structured sparsity*, that is, forcing different kind of structures (e.g., disjunct groups or trees) on the codes can lead to increased performances in several applications, for example in multiple kernel learning, multi-task learning (a.k.a. transfer learning, joint covariate selection, multiple measurements vector model, simultaneous sparse approximation), feature selection, and compressed sensing (Zhao et al., 2009; Huang & Zhang, 2010; Baraniuk et al., 2010; Bach et al., 2011).

Both dictionary learning and structured sparse coding (when the dictionary is given) are very popular; however, very few works have focused on the combination of these two tasks, i.e., learning *structured dictionaries* by pre-assuming certain structures on the representation (Kavukcuoglu et al., 2009; Jenatton et al., 2010a;b; Mairal et al., 2010b; Rosenblum et al., 2010).

We are interested in structured dictionary learning algorithms that possess the following four properties: (i) They can handle general, overlapping group structures. (ii) The applied regularization can be non-convex and hence allow less restrictive assumptions on the groups' sparsity. (iii) We want online algorithms (Mairal et al., 2010a). Online methods have the advantage over offline ones that they can process more instances in the same amount of time (Bottou & LeCun, 2005), and in many cases this can lead to increased performance. In large systems where the whole dataset does not fit into the memory, online systems can be the only solutions. Online techniques are adaptive: for example in recommender systems when new users appear, we might not want to relearn the dictionary from scratch; we simply want to modify it by the contributions of the new users. (iv) We want an algorithm that can handle missing observations. Using a collaborative filtering example, users usually do not rate every item, and thus some of the possible observations are missing. Several successful structured dictionary learning method have been proposed in the literature; however, to the best of our knowledge, they can possess only two of our four requirements at most.

Our contributions:

• We formulate a general dictionary learning approach, which is (i) online, (ii) enables overlapping group structures with (iii) non-convex group structure inducing regularization, and (iv) handles the partially observable case. We call this problem *online structured dictionary learning* (OSDL).

• We show that several famous structured sparse coding and dictionary learning problems emerge as a special case of OSDL. In particular, we (i) present an application in collaborative filtering where we demonstrate that our algorithm can outperform the state-of-the-art competitors on the Jester (joke recommenda-

ICML-2011 – Structured Sparsity: Learning and Inference Workshop, Bellevue, Washington, USA, 2 July 2011.

tion) dataset, and (ii) we show an illustrative example for finding structured facial components in the color FERET dataset.

Notations. $|\cdot|$ denotes the number of elements in a set. $\mathbf{A}_O \in \mathbb{R}^{|O|\times D}$ contains the $O \subseteq \{1,\dots,d\}$ rows of matrix $\mathbf{A} \in \mathbb{R}^{d\times D}$. $\mathbf{I}$ and $\mathbf{0}$ stand for the identity and the null matrices, respectively. For positive numbers p, q, (i) (quasi-)norm ℓ_q of vector $\mathbf{a} \in \mathbb{R}^d$ is $\|\mathbf{a}\|_q = (\sum_{i=1}^d |a_i|^q)^{\frac{1}{q}}$, (ii) $\ell_{p,q}$-norm (group norm) of the same vector is $\|\mathbf{a}\|_{p,q} = \|[\|\mathbf{a}_{P_1}\|_q, \dots, \|\mathbf{a}_{P_K}\|_q]\|_p$, where $\{P_i\}_{i=1}^K$ is a partition of the set $\{1,\dots,d\}$. $S_p^d = \{\mathbf{a} \in \mathbb{R}^d : \|\mathbf{a}\|_p \le 1\}$ is the unit sphere associated with ℓ_p in $\mathbb{R}^d$. For a given set system $\mathcal{G}$, elements of vector $\mathbf{a} \in \mathbb{R}^{|\mathcal{G}|}$ are denoted by a^G, where $G \in \mathcal{G}$, that is $\mathbf{a} = (a^G)_{G\in\mathcal{G}}$. $\Pi_{\mathcal{C}}(\mathbf{x}) = \operatorname{argmin}_{\mathbf{c}\in\mathcal{C}} \|\mathbf{x}-\mathbf{c}\|_2$ denotes the orthogonal projection to the closed and convex set $\mathcal{C} \subseteq \mathbb{R}^d$, where $\mathbf{x} \in \mathbb{R}^d$. $\mathbb{R}^d_+ = \{\mathbf{x} \in \mathbb{R}^d : x_i \ge 0\,(\forall i)\}$. χ stands for the characteristic function.

2. Problem Definition

We define the online structured dictionary learning (OSDL) task as follows. Let the dimension of our observations be denoted by d_x. Assume that in each time instant ($i = 1, 2, \dots$) a set $O_i \subseteq \{1,\dots,d_x\}$ is given, that is, we know which coordinates are observable at time i, and our observation is $\mathbf{x}_{O_i}$. We aim to find a dictionary $\mathbf{D} \in \mathbb{R}^{d_x \times d_\alpha}$ that can approximate the observations $\mathbf{x}_{O_i}$ well from the linear combination of its columns. We assume that the columns of $\mathbf{D}$ belong to a closed, convex, and bounded set $\mathcal{D} = \times_{i=1}^{d_\alpha} \mathcal{D}_i$. To formulate the cost of dictionary $\mathbf{D}$, we first consider a *fixed* time instant i, observation $\mathbf{x}_{O_i}$, dictionary $\mathbf{D}$, and define the hidden representation $\boldsymbol{\alpha}_i$ associated to this triple. Representation $\boldsymbol{\alpha}_i$ is allowed to belong to a closed, convex set $\mathcal{A} \subseteq \mathbb{R}^{d_\alpha}$ ($\boldsymbol{\alpha}_i \in \mathcal{A}$) with certain structural constraints. We express the structural constraint on $\boldsymbol{\alpha}_i$ by making use of a given $\mathcal{G}$ group structure, which is a set system (also called hypergraph) on $\{1,\dots,d_\alpha\}$. We also assume that a set of linear transformations $\{\mathbf{A}^G \in \mathbb{R}^{d_G \times d_\alpha}\}_{G\in\mathcal{G}}$ is given for us. We will use them as parameters to define the structured regularization on the codes. Representation $\boldsymbol{\alpha}$ belonging to a triple $(\mathbf{x}_O, \mathbf{D}, O)$ is defined as the solution of the structured sparse coding task

$$l(\mathbf{x}_O, \mathbf{D}_O) = l_{\mathcal{A},\kappa,\mathcal{G},\{\mathbf{A}^G\}_{G\in\mathcal{G}},\eta}(\mathbf{x}_O, \mathbf{D}_O) \tag{1}$$

$$= \min_{\boldsymbol{\alpha}\in\mathcal{A}} \left[\frac{1}{2}\|\mathbf{x}_O - \mathbf{D}_O\boldsymbol{\alpha}\|_2^2 + \kappa\Omega(\boldsymbol{\alpha})\right], \tag{2}$$

where $l(\mathbf{x}_O, \mathbf{D}_O)$ denotes the loss, $\kappa > 0$, and

$$\Omega(\mathbf{y}) = \Omega_{\mathcal{G},\{\mathbf{A}^G\}_{G\in\mathcal{G}},\eta}(\mathbf{y}) = \|(\|\mathbf{A}^G\mathbf{y}\|_2)_{G\in\mathcal{G}}\|_\eta \tag{3}$$

is the group structure inducing regularizer associated to $\mathcal{G}$ and $\{\mathbf{A}^G\}_{G\in\mathcal{G}}$, and $\eta \in (0,2)$. Here, the first term of (2) is responsible for the quality of approximation on the observed coordinates, and (3) performs regularization defined by the group structure/hypergraph $\mathcal{G}$ and the $\{\mathbf{A}^G\}_{G\in\mathcal{G}}$ linear transformations. The OSDL problem is defined as the minimization of the cost function:

$$\min_{\mathbf{D}\in\mathcal{D}} f_t(\mathbf{D}) := \frac{1}{\sum_{j=1}^t (j/t)^\rho} \sum_{i=1}^t \left(\frac{i}{t}\right)^\rho l(\mathbf{x}_{O_i}, \mathbf{D}_{O_i}), \tag{4}$$

that is, we aim to minimize the average loss of the dictionary, where ρ is a non-negative forgetting rate. If $\rho = 0$, the classical average $f_t(\mathbf{D}) = \frac{1}{t}\sum_{i=1}^t l(\mathbf{x}_{O_i}, \mathbf{D}_{O_i})$ is obtained. When $\eta \le 1$, then for a code vector $\boldsymbol{\alpha}$, the regularizer Ω aims at eliminating the $\mathbf{A}^G\boldsymbol{\alpha}$ terms ($G \in \mathcal{G}$) by making use of the sparsity inducing property of the $\|\cdot\|_\eta$ norm. For $O_i = \{1,\dots,d_x\}$ $(\forall i)$, we get the fully observed OSDL task.

Below we list a few special cases of the OSDL problem:

Special cases for $\mathcal{G}$:

• If $|\mathcal{G}| = d_\alpha$ and $\mathcal{G} = \{\{1\},\{2\},\dots,\{d_\alpha\}\}$, then no dependence is assumed between coordinates α_i, and the problem reduces to the classical task of learning "dictionaries with sparse codes".

• If $|\mathcal{G}| = d_\alpha$ and $\mathcal{G} = \{desc_1,\dots,desc_{d_\alpha}\}$, where $desc_i$ stands for the i^{th} node (α_i) of a tree and its descendants, then we have a tree-structured, hierarchial representation.

• If $|\mathcal{G}| = d_\alpha$, and $\mathcal{G} = \{NN_1,\dots,NN_{d_\alpha}\}$, where NN_i denotes the neighbors of the i^{th} point (α_i) in radius r on a grid, then we obtain a grid representation.

• If $\mathcal{G} = \{\{1\},\dots,\{d_\alpha\},\{1,\dots,d_\alpha\}\}$, then we have an elastic net representation.

• If $\mathcal{G}$ is a partition of $\{1,\dots,d_\alpha\}$, then non-overlapping group structure is obtained.

Special cases for $\{\mathbf{A}^G\}_{G\in\mathcal{G}}$:

• Let (V, E) be a given graph, where V and E denote the set of nodes and edges, respectively. For each $e = (i,j) \in E$, we also introduce $(w_{ij},\ v_{ij})$ weight pairs. Now, if we set $\Omega(\mathbf{y}) = \sum_{e=(i,j)\in E: i<j} w_{ij}|y_i - v_{ij}y_j|$, then we obtain the graph-guided fusion penalty (Chen et al., 2010). The groups $G \in \mathcal{G}$ correspond to the (i,j) pairs, and in this case $\mathbf{A}^G = [w_{ij}, -w_{ij}v_{ij}] \in \mathbb{R}^{1\times 2}$. As a special case, for a chain graph we get the standard fused Lasso penalty by setting the weights to one: $\Omega(\mathbf{y}) = FL(\mathbf{y}) = \sum_{j=1}^{d_\alpha - 1} |y_{j+1} - y_j|$.

• Let $\nabla \mathbf{y} \in \mathbb{R}^{d_1 \times d_2}$ denote the discrete differential of an image $\mathbf{y} \in \mathbb{R}^{d_1 \times d_2}$ at position $(i,j) \in \{1,\dots,d_1\} \times \{1,\dots,d_2\}$: $(\nabla \mathbf{y})_{ij} = [(\nabla \mathbf{y})^1_{ij}; (\nabla \mathbf{y})^2_{ij}]$, where $(\nabla \mathbf{y})^1_{ij} = (y_{i+1,j} - y_{i,j})\chi_{\{i<d_1\}}$ and $(\nabla \mathbf{y})^2_{ij} = (y_{i,j+1} - y_{i,j})\chi_{\{j<d_2\}}$. Using these notations, the total variation of $\mathbf{y}$ is defined as follows: $\Omega(\mathbf{y}) = \|\mathbf{y}\|_{TV} = \sum_{i=1}^{d_1}\sum_{j=1}^{d_2} \|(\nabla \mathbf{y})_{ij}\|_2$.

Special cases for $\mathcal{D}, \mathcal{A}$:

• $\mathcal{D}_i = S_2^{d_x} \cap \mathbb{R}^{d_x}_+$ $(\forall i)$, $\mathcal{A} = \mathbb{R}^{d_\alpha}_+$: This is the structured non-negative matrix factorization (NMF) problem.

• $\mathcal{D}_i = S_1^{d_x} \cap \mathbb{R}^{d_x}_+$ $(\forall i)$, $\mathcal{A} = \mathbb{R}^{d_\alpha}_+$: This is the structured mixture-of-topics problem.

• Beyond $\mathbb{R}^d$, S_1^d, S_2^d, $S_1^d \cap \mathbb{R}^d_+$, and $S_2^d \cap \mathbb{R}^d_+$, several other constraints can also be motivated for $\mathcal{D}_i$ and $\mathcal{A}$. In the above mentioned examples, the group-norm, elastic net, and fused Lasso constraints have been applied in a "soft" manner, with the help of the Ω regularization. However, we can enforce these constraints in a "hard" way as well: During optimization (Section 3), we can exploit the fact that the projection to the $\mathcal{D}_i$ and $\mathcal{A}$ constraint sets can be computed efficiently. Such constraint sets include, e.g., $\{\mathbf{c} : \|\mathbf{c}\|_{p,q} \leq 1\}$ group norms, the $\{\mathbf{c} : \gamma_1 \|\mathbf{c}\|_1 + \gamma_2 \|\mathbf{c}\|_2^2 \leq 1\}$ elastic net, and the $\{\mathbf{c} : \gamma_1 \|\mathbf{c}\|_1 + \gamma_2 \|\mathbf{c}\|_2^2 + \gamma_3 FL(\mathbf{c}) \leq 1\}$ fused Lasso ($\gamma_1, \gamma_2, \gamma_3 > 0$).

• When applying group norms for both the codes $\boldsymbol{\alpha}$ *and* the dictionary $\mathbf{D}$, we arrive at a double structured dictionary learning scheme.

In sum, the OSDL model provides a unified dictionary learning framework for several actively studied structured sparse coding problems, naturally extends them for partially observable inputs, and allows non-convex regularization as well.

3. Optimization

In this section we briefly summarize our proposed method for solving the OSDL problem. The optimization of cost function (4) is equivalent to the joint optimization of dictionary $\mathbf{D}$ and representation $\{\boldsymbol{\alpha}_i\}_{i=1}^t$, i.e., the minimization of $\arg\min_{\mathbf{D}\in\mathcal{D}, \{\boldsymbol{\alpha}_i \in \mathcal{A}\}_{i=1}^t} f_t(\mathbf{D}, \{\boldsymbol{\alpha}_i\}_{i=1}^t)$, where

$$f_t = \frac{1}{\sum_{j=1}^t (j/t)^\rho} \sum_{i=1}^t \left(\frac{i}{t}\right)^\rho \left[\frac{1}{2}\|\mathbf{x}_{O_i} - \mathbf{D}_{O_i}\boldsymbol{\alpha}_i\|_2^2 + \kappa\Omega(\boldsymbol{\alpha}_i)\right].$$

We optimize $\mathbf{D}$ online in an alternating manner by using the sequential observations $\mathbf{x}_{O_i}$. We use the actual dictionary estimation $\mathbf{D}_{t-1}$ and sample $\mathbf{x}_{O_t}$ to optimize (2) for representation $\boldsymbol{\alpha}_t$. For the estimated representations $\{\boldsymbol{\alpha}_i\}_{i=1}^t$, we derive our dictionary estimation $\mathbf{D}_t$ from the quadratic optimization problem

$$\hat{f}_t(\mathbf{D}_t) = \min_{\mathbf{D}\in\mathcal{D}} f_t(\mathbf{D}, \{\boldsymbol{\alpha}_i\}_{i=1}^t). \quad (5)$$

3.1. Representation Optimization ($\boldsymbol{\alpha}$).

Using the variational properties of $\|\cdot\|_\eta$, one can show that the solution $\boldsymbol{\alpha}$ of the following optimization task is equal to the solution of (2): $\arg\min_{\boldsymbol{\alpha}\in\mathcal{A}, \mathbf{z}\in\mathbb{R}_+^{|\mathcal{G}|}} J(\boldsymbol{\alpha}, \mathbf{z})$, where

$$J(\boldsymbol{\alpha}, \mathbf{z}) = \frac{1}{2}\|\mathbf{x}_{O_t} - (\mathbf{D}_{t-1})_{O_t}\boldsymbol{\alpha}\|_2^2 + \kappa\frac{1}{2}\left(\boldsymbol{\alpha}^T\mathbf{H}\boldsymbol{\alpha} + \|\mathbf{z}\|_\beta\right),$$

and $\mathbf{H} = \mathbf{H}(\mathbf{z}) = \sum_{G\in\mathcal{G}} (\mathbf{A}^G)^T\mathbf{A}^G/z^G$. The optimization of $J(\boldsymbol{\alpha}, \mathbf{z})$ can be carried out by iterative alternating steps. One can minimize the quadratic cost function on the convex set $\mathcal{A}$ for given $\mathbf{z}$ with standard solvers. For fixed $\boldsymbol{\alpha}$, $\mathbf{z} = (z^G)_{G\in\mathcal{G}}$ can be calculated as follows: $z^G = \|\mathbf{A}^G\boldsymbol{\alpha}\|_2^{2-\eta}\|(\|\mathbf{A}^G\boldsymbol{\alpha}\|_2)_{G\in\mathcal{G}}\|_\eta^{\eta-1}$.

3.2. Dictionary Optimization (D).

We use the block-coordinate descent method for the optimization of $\mathbf{D}$: we optimize columns $\mathbf{d}_j$ in $\mathbf{D}$ one-by-one by keeping the other columns $(\mathbf{d}_i, i \neq j)$ fixed. For a given j, $\hat{f}_t$ is quadratic in $\mathbf{d}_j$. We find the minimum by solving $\frac{\partial \hat{f}_t}{\partial \mathbf{d}_j}(\mathbf{u}_j) = \mathbf{0}$, and then we project this solution to the constraint set $\mathcal{D}_j$ ($\mathbf{d}_j \leftarrow \Pi_{\mathcal{D}_j}(\mathbf{u}_j)$). One can show by differentiation that $\mathbf{u}_j$ satisfies the

$$\mathbf{C}_{j,t}\mathbf{u}_j = \mathbf{b}_{j,t} - \mathbf{e}_{j,t} + \mathbf{C}_{j,t}\mathbf{d}_j \quad (\mathbf{C}_{j,t} \in \mathbb{R}^{d_x \times d_x}) \quad (6)$$

linear equation system, where

$$\mathbf{C}_{j,t} = \sum_{i=1}^t \left(\frac{i}{t}\right)^\rho \boldsymbol{\Delta}_i\alpha_{i,j}^2, \;\; \mathbf{e}_{j,t} = \sum_{i=1}^t \left(\frac{i}{t}\right)^\rho \boldsymbol{\Delta}_i\mathbf{D}\boldsymbol{\alpha}_i\alpha_{i,j},$$

$$\mathbf{B}_t = \sum_{i=1}^t \left(\frac{i}{t}\right)^\rho \boldsymbol{\Delta}_i\mathbf{x}_i\boldsymbol{\alpha}_i^T = [\mathbf{b}_{1,t}, \dots, \mathbf{b}_{d_\alpha,t}], \quad (7)$$

matrices $\mathbf{C}_{j,t}$ are diagonal, $\mathbf{e}_{j,t} \in \mathbb{R}^{d_x}$, $\mathbf{B}_t \in \mathbb{R}^{d_x \times d_\alpha}$, $\boldsymbol{\Delta}_i \in \mathbb{R}^{d_x \times d_x}$ is the diagonal matrix representation of the O_i set (for $j \in O_i$ the j^{th} diagonal is 1 and is 0 otherwise). It is sufficient to update statistics $\{\{\mathbf{C}_{j,t}\}_{j=1}^{d_\alpha}, \mathbf{B}_t, \{\mathbf{e}_{j,t}\}_{j=1}^{d_\alpha}\}$ online for the optimization of $\hat{f}_t$, which can be done exactly for $\mathbf{C}_{j,t}$ and $\mathbf{B}_t$:

$$\mathbf{C}_{j,t} = \gamma_t\mathbf{C}_{j,t-1} + \boldsymbol{\Delta}_t\alpha_{tj}^2, \quad \mathbf{B}_t = \gamma_t\mathbf{B}_{t-1} + \boldsymbol{\Delta}_t\mathbf{x}_t\boldsymbol{\alpha}_t^T,$$

where $\gamma_t = \left(1 - \frac{1}{t}\right)^\rho$ and the recursions are initialized by (i) $\mathbf{C}_{j,0} = \mathbf{0}$, $\mathbf{B}_0 = \mathbf{0}$ for $\rho = 0$ and (ii) in an arbitrary way for $\rho > 0$. According to numerical experiences, $\mathbf{e}_{j,t} = \gamma_t\mathbf{e}_{j,t-1} + \boldsymbol{\Delta}_t\mathbf{D}_t\boldsymbol{\alpha}_t\alpha_{t,j}$ is a good approximation for $\mathbf{e}_{j,t}$ with the actual estimation $\mathbf{D}_t$ and with initialization $\mathbf{e}_{j,0} = \mathbf{0}$.

4. Illustration

In this section we demonstrate the applicability of the proposed OSDL approach on (i) structured NMF, and (ii) collaborative filtering problems.

4.1. Online Structured NMF on Faces

It has been shown on the CBCL database that dictionary vectors of the offline NMF method can be interpreted as face components. However, to the best of our knowledge, there is no existing NMF algorithm as of yet which could handle general $\mathcal{G}$ group structures in an online fashion. Our OSDL method is able to do that, can also cope with only partially observed inputs, and can be extended with non-convex sparsity-inducing norms. We illustrate our approach on the color FERET dataset, which is a large scale 140×120 face dataset. These images were the observations for our ODSL method ($\mathbf{x}_i$, $d_x = 49,140 = 140 \times 120 \times 3$ minus some masking at the bottom corners). The group structure $\mathcal{G}$ was chosen to be hierarchical; we applied a full, 8-level binary tree ($d_\alpha = 255$), η was set to 0.5 and κ was $\frac{1}{2^{10.5}}$. The optimized $\mathbf{D}$ dictionary is shown in Fig. 1. We can observe that the proposed algorithm is able to naturally develop and hierarchically organize the elements of the dictionary, and the colors are separated as well. This example demonstrates that our method can be used for large scale problems where the dimension of the observations is about $50,000$.

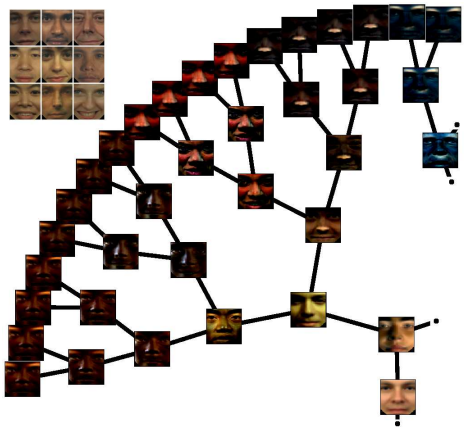

Figure 1. Illustration of the online learned structured NMF dictionary. Upper left corner: training samples.

4.2. Collaborative Filtering

The OSDL approach can also be used for solving the online collaborative filtering problem by simply setting the t^{th} user's known ratings to be the observations ($\mathbf{x}_{O_t}$). We have chosen the Jester, joke recommendation dataset for the illustrations, which is a standard benchmark for CF. To the best of our knowledge, the top results on this database are RMSE (root mean square error) = 4.1123 based only on neighbor information and RMSE = 4.1229 using an unstructured dictionary learning model. Our extensive numerical experiments demonstrate that using toroid (hexagonal grid) and hierarchical group structures increase performance; our OSDL method achieved RMSE = 4.0774 on this problem.

Acknowledgments. The research was partly supported by the Department of Energy (grant number DESC0002607). The Project is supported by the European Union and co-financed by the European Social Fund (grant agreements no. TÁMOP 4.2.1/B-09/1/KMR-2010-0003 and KMOP-1.1.2-08/1-2008-0002).

References

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. *Optimization for Machine Learning*, chapter Convex optimization with sparsity-inducing norms. MIT Press, 2011.

Baraniuk, R., Cevher, V., Duarte, M., and Hegde, C. Model-based compressive sensing. *IEEE T. Inform. Theory*, 56:1982 – 2001, 2010.

Bottou, L. and LeCun, Y. On-line learning for very large data sets. *Appl. Stoch. Model. Bus. - Stat. Learn.*, 21 (2):137–151, 2005.

Chen, X., Lin, Q., Kim, S., Carbonell, J., and Xing, E. An efficient proximal gradient method for general structured sparse learning. Technical report, 2010. http://arxiv.org/abs/1005.4717.

Huang, J. and Zhang, T. The benefit of group sparsity. *Ann. Stat.*, 38(4):1978–2004, 2010.

Jenatton, R., Mairal, J., Obozinski, G., and Bach, F. Proximal methods for sparse hierarchical dictionary learning. In *ICML*, pp. 487–494, 2010a.

Jenatton, R., Obozinski, G., and Bach, F. Structured sparse principal component analysis. *J. Mach. Learn. Res.:W&CP*, 9:366–373, 2010b.

Kavukcuoglu, K., Ranzato, M.'A., Fergus, R., and LeCun, Y. Learning invariant features through topographic filter maps. In *CVPR*, pp. 1605–1612, 2009.

Mairal, J., Bach, F., Ponce, J., and Sapiro, G. Online learning for matrix factorization and sparse coding. *J. Mach. Learn. Res.*, 11:10–60, 2010a.

Mairal, J., Jenatton, R., Obozinski, G., and Bach, F. Network flow algorithms for structured sparsity. In *NIPS*, pp. 1558–1566, 2010b.

Rosenblum, K., Z.-Manor, L., and Eldar, Y. Dictionary optimization for block-sparse representations. In *AAAI Fall Symp. on Manifold Learning*, 2010.

Tropp, J. and Wright, S. Computational methods for sparse solution of linear inverse problems. *IEEE special issue on Applications of sparse representation and compressive sensing*, 98(6):948–958, 2010.

Zhao, P., Rocha, G., and Yu, B. The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Stat.*, 37(6A):3468–3497, 2009.