Frontiers of Information Technology & Electronic Engineering
www.zju.edu.cn/jzus; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Review:

A systematic review of structured sparse learning*

Lin-bo QIAO[1,2], Bo-feng ZHANG[1], Jin-shu SU[‡1,2], Xi-cheng LU[1,2]

([1]*College of Computer, National University of Defense Technology, Changsha 410073, China*)

([2]*National Laboratory for Parallel and Distributed Processing, National University of Defense Technology, Changsha 410073, China*)

E-mail: qiao.linbo@nudt.edu.cn; bfzhang@nudt.edu.cn; sjs@nudt.edu.cn; xclu@nudt.edu.cn

Received Aug. 21, 2016; Revision accepted Dec. 20, 2016; Crosschecked Mar. 29, 2017

Abstract: High dimensional data arising from diverse scientific research fields and industrial development have led to increased interest in sparse learning due to model parsimony and computational advantage. With the assumption of sparsity, many computational problems can be handled efficiently in practice. Structured sparse learning encodes the structural information of the variables and has been quite successful in numerous research fields. With various types of structures discovered, sorts of structured regularizations have been proposed. These regularizations have greatly improved the efficacy of sparse learning algorithms through the use of specific structural information. In this article, we present a systematic review of structured sparse learning including ideas, formulations, algorithms, and applications. We present these algorithms in the unified framework of minimizing the sum of loss and penalty functions, summarize publicly accessible software implementations, and compare the computational complexity of typical optimization methods to solve structured sparse learning problems. In experiments, we present applications in unsupervised learning, for structured signal recovery and hierarchical image reconstruction, and in supervised learning in the context of a novel graph-guided logistic regression.

Key words: Sparse learning; Structured sparse learning; Structured regularization
http://dx.doi.org/10.1631/FITEE.1601489 **CLC number:** TP391

1 Introduction

Rapid improvement in sensing technologies has created high dimensional data in research fields and industries with many features and a huge number of samples. The increasing data volume becomes a great challenge for contemporary statistical learning algorithms (John Lu, 2010) in various research areas, including high-resolution imaging (Bruckstein *et al.*, 2009), target tracking (Zhang *et al.*, 2012; 2013; 2014; 2015a; 2015b), astronomical data processing (Borne, 2009), genomics (Kim and Xing, 2014), functional and longitudinal processing (Jenatton *et al.*, 2012), and warehouse data analysis in business (Fan *et al.*, 2011). For example, astronomical projects produce more than 10^9 pixels every 20 s and terabytes of data in a single evening (Borne, 2009). Financial data is measured with hundreds of financial instruments and tracked over time with 10^6 trades per second in high-frequency trading (Fan *et al.*, 2011).

It is almost impossible to learn a consistent model with high accuracy, model explicability, and computational efficiency at the same time, unless one assumes that the sample size is much larger than the feature size (Candès and Tao, 2007). However, in high-dimensional settings, the dimension of the feature is often the same as, or even larger than the sample size. Therefore, traditional methods face significant challenges, ranging from theoretical analysis, efficient algorithm design, to model estimation and interpretation. Note that consistent estimators may be obtained if additional assumptions are imposed

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (No. 61303264)

ORCID: Lin-bo QIAO, http://orcid.org/0000-0002-8285-2738 on the traditional models (Negahban *et al.*, 2012). A widely used constraint is that models should be sparse for high-dimensional problems (Bach *et al.*, 2011).

1.1 Sparsity

In a sparse model, only a small number of variables are non-zero among all the variables in the model. The sparsity assumption is typically associated with desired interesting properties such as succinct interpretation, fast evaluation of the model, statistical robustness (sparsity is usually associated with robust statistical performance), and computational advantages, which appeals to a great many researchers.

Sparsity is preferred in learning problems with high-dimensional data. In many applications, though the raw data is high dimensional, the intrinsic dimension is relatively low. For example, in bioinformatics, different high-dimensional genes may belong to the same functional group; in multi-task learning, several estimators are expected to share common types of covariates. In fact, it is now common sense that sparsity is a powerful assumption for contemporary machine learning algorithms (Bach *et al.*, 2012a).

Sparsity is considered to be one of the most significant philosophical and aesthetic principles that have ever existed. It is also known as Occam's razor (Rasmussen and Ghahramani, 2001), "Entities should not be multiplied without necessity", by William Ockham in the 13th century. The parsimony principle has been addressed again and again, which has led to several beautiful results, such as minimal description length (MDL) (John Lu, 2010). Modern sparse learning methods were introduced by Wrinch and Jeffreys (1921), who expressed the sparsity of models in physics as the non-zero number of learning variables. This concept is very close to today's definition of sparsity. Since then, numerous tools (see Mairal *et al.* (2014) and multiple references therein) have been developed in the statistics community to build sparsity-related models, which have greatly improved the explicability of the models, and dramatically decreased the computational cost of the model in the prediction procedure. With the efforts of researchers and engineers, sparse learning has become a popular tool with the development of theoretical frameworks and various efficient algorithms. The theoretical frameworks range from the original idea in underdetermined linear systems (see Bruckstein *et al.* (2009) and references therein), signal processing (Chen and Donoho, 1994) and statistical learning (Tibshirani, 1996), the computational complexity of the ℓ_0-norm regularized problem, the uniqueness of the solution for the ℓ_0-norm regularized optimization (see Donoho and Huo (2001), Tropp *et al.* (2003), Elad (2010), and references therein), the model selection consistency of the convex relaxation from ℓ_0-norm to ℓ_1-norm (Candès *et al.*, 2006; Zhao and Yu, 2006; Candès and Tao, 2007; Candès, 2008; Candès and Recht, 2009; Zhang, 2009), to the statistical analysis of extended algorithms (Elad, 2010; Jenatton, 2011). Proposed algorithms range from greedy algorithms for ℓ_0-norm regularized methods (Tropp, 2004) to convex optimization methods after convex approximation (see Friedman *et al.* (2007), Beck and Teboulle (2009), Jenatton *et al.* (2011), Yang and Yuan (2013), and references therein).

Recently, a line of work has been devoted to the framework of empirical risk minimization with sparsity-inducing regularizations, which is commonly formulated as

$$\min_{\boldsymbol{x}\in\mathcal{X}} \; l(\boldsymbol{x}) + \lambda \cdot r(\boldsymbol{x}), \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^d$ (d is the dimensionality of the features), $\mathcal{X}$ is the feasible domain, $l(\cdot)$ is the empirical loss function, $r(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is a sparsity-regularized function, and λ is a parameter balancing these two terms. Problem (1) accommodates quite a few classic classification and regression models including linear regression obtained by setting $l(\boldsymbol{x}) = \|\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x} - \boldsymbol{b}\|_2/2$, logistic regression obtained by setting $l(\boldsymbol{x}) = \log(1 + \exp(-\boldsymbol{b}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}))$, and linear support vector machine (SVM) obtained by letting $l(\boldsymbol{x}) = \max(0, 1 - \boldsymbol{b}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x})$, where $\boldsymbol{A} \in \mathbb{R}^{d\times N}$ are the data samples and $\boldsymbol{b} \in \mathbb{R}^N$ are the labels of these data samples. For variable selection problems in linear models, sparsity may be directly achieved by adding a penalty function of non-zero elements (Tibshirani, 1996), specifically $r(\boldsymbol{x}) = \|\boldsymbol{x}\|_0$, which is known as the ℓ_0-norm of variable $\boldsymbol{x}$.

In Fig. 1, the signal is recovered through a sparse linear regression algorithm. With the constraint that the basis of the signal is sparse, or the number of bases is relatively small compared to the whole basis, the original basis is determined while the noise is removed.

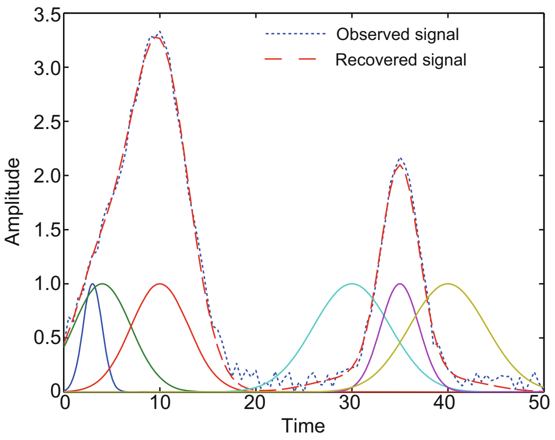

Fig. 1 Signal recovery from observation with noise through a sparse linear regression algorithm. There are six Gaussian distributed bases with means of (3, 4, 10, 30, 35, 40) and variances of (1, 3, 3, 4, 2, 4). The observed signal is a mixture of the second, third, and fourth bases with weights of (1, 3, 2) accordingly. The sparse learning method determines exactly the bases and weights of the original signal (References to color refer to the online version of this figure)

When $r(\cdot)$ is a nonconvex function, it was reported in the literature that nonconvex regularization usually yields a solution with more desirable structural properties. Let us take the ℓ_0-norm regularized least squares problem (i.e., $l(\cdot)$ is a least squares function) as an example. It is well known that such a problem is NP-hard because of its combinatorial nature. To this end, the ℓ_1-norm regularized model was proposed to pursue computational tractability. In spite of computational advantages and successful applications, the ℓ_1 model has some limits in certain scenarios (Candès *et al.*, 2008), since the ℓ_1-norm comes at the price of shifting the resulting estimator by a constant (Fan and Li, 2011) which leads to an over-penalized problem. To circumvent the issues pertaining to the ℓ_1-norm, researchers have managed to impose some nonconvex regularizations on problem (1), which have been proven to be better approximations of the ℓ_0-norm theoretically and computationally. Some nonconvex regularized functions have been widely used in sparse learning (Gong *et al.*, 2013). These nonconvex regularized functions include the ℓ_p-norm ($0 < p < 1$) (Chartrand and Yin, 2008; Foucart and Lai, 2009; Xu *et al.*, 2012; Lai *et al.*, 2013; Wang Y *et al.*, 2013), smoothly clipped absolute deviation (SCAD) (Fan and Li, 2011), log-sum penalty (LSP) (Candès *et al.*, 2008), minimax concave penalty (MCP) (Zhang CH, 2010), and the capped-ℓ_1 penalty (Zhang T, 2010; Zhang, 2013). In Table 1, we present the details of these regularized functions for readers' convenience.

Table 1 Examples of nonconvex penalty $r(\boldsymbol{x})$

Name	$r(\boldsymbol{x})$
LSP	$\gamma\log(1+\lvert\boldsymbol{x}\rvert/\theta)\quad(\theta>0)$
SCAD	$\gamma\int_0^{\lvert\boldsymbol{x}\rvert}\min\left(1,\frac{[\theta\gamma-y]_+}{(\theta-1)\gamma}\right)\mathrm{d}y\quad(\theta>2)$
MCP	$\gamma\int_0^{\lvert\boldsymbol{x}\rvert}\left[1-\frac{y}{\theta\gamma}\right]_+\mathrm{d}y=\begin{cases}\gamma\lvert\boldsymbol{x}\rvert-\frac{\boldsymbol{x}^2}{2\theta}, & \lvert\boldsymbol{x}\rvert\le\theta\gamma,\\ \theta\gamma^2/2, & \lvert\boldsymbol{x}\rvert>\theta\gamma,\\ & (\theta>0)\end{cases}$
Capped-ℓ_1	$\gamma\min(\lvert\boldsymbol{x}\rvert,\theta)\quad(\theta>0)$

1.2 Structured sparsity

Structured sparse learning is commonly used in two situations. First, structured sparse learning is used given the prior knowledge that the model should be structured sparsely. Second, to make the model more interpretable or easier to use in the following procedures, even if the underlying problem does not admit assumed structured sparse solutions, one looks for the best structured sparse approximation.

The ℓ_1-norm could be used to induce model parsimony; however, it does not encode the structural information. To encode the structured sparsity, various structured regularizations have been proposed. These regularizations encode the structural information into traditional sparse learning, and they are recognized as structured sparsity inducing norms. Structured sparsity may be achieved by adding explicitly structured regularization. Structured sparsity-inducing norms are natural extensions of the ℓ_0-norm. We could formulate the structured learning problems as

$$\min_{\boldsymbol{x}\in\mathcal{X}}\ l(\boldsymbol{x})+\lambda\cdot R(\boldsymbol{x}), \tag{2}$$

where $R(\cdot)$ is the structured sparsity-inducing regularization, which can be seen as an extension of pure sparsity-inducing penalization. Compared to the sparse learning problem (1), the term $R(\cdot)$ encodes the structured information, which provides a great advantage beyond the traditional sparse learning algorithms to pursue the structured model.

1.3 Aim and scope of this paper

In this article, we will review several structured sparsity-inducing norms (Fig. 2) ranging from grouped sparsity (Fig. 2b), fused sparsity (Fig. 2c), hierarchical sparsity (Fig. 2d), to graphical sparsity (Fig. 2e). This review sheds light on new directions in research fields and engineering problems to take structural information into account.

Structured sparsity has been widely used in practical problems, including model-based compressive sensing (Baraniuk *et al.*, 2010; Asaei *et al.*, 2011a; Duarte and Eldar, 2011; Chen *et al.*, 2014), signal processing (Bach and Jordan, 2006; Asaei *et al.*, 2011; 2014a; 2014b; Najafian, 2016), computer vision (Jenatton *et al.*, 2009; Kim *et al.*, 2013; Chen and Huang, 2014; Karygianni and Frossard, 2014; Xiao *et al.*, 2016), bioinformatics (Wille and Bühlmann, 2006; Zhang SZ *et al.*, 2011; Kim and Xing, 2012), and recommendation systems (Koren *et al.*, 2009; Takacs *et al.*, 2009; Rendle and Schmidt-Thieme, 2010; Zhang ZK *et al.*, 2011). This article focuses mainly on their formulations and algorithms. The models and the algorithms are relatively independent (Xu *et al.*, 2005; Zhang *et al.*, 2006; Xu *et al.*, 2007; Yuan *et al.*, 2015; Hu and Yu, 2016; Xie *et al.*, 2016; Xie and Tong, 2016; Zhu *et al.*, 2016), so is in the privacy research field (Sun *et al.*, 2015a; 2015b; Wu *et al.*, 2016).

Note that there are other books and articles that offer diverse perspectives on sparse learning methods, including Elad (2010) and Mallat (2008) from a perspective of signal processing, and Bach *et al.* (2012b) in a view of optimization. However, this article focuses on intuitive formulations, their variants, and the algorithms to solve them.

1.4 Notations

Vectors are denoted by bold lower-case letters and matrices by upper-case ones. $\|\boldsymbol{x}\|_0$ is the number of non-zero elements in a vector $\boldsymbol{x}$, $\|\boldsymbol{x}\|_1$ is the sum of absolute values of elements in a vector $\boldsymbol{x}$, the ℓ_q-norm of a vector $\boldsymbol{x} \in \mathbb{R}^n$ is defined as $\|\boldsymbol{x}\|_q := (\sum_{i=1}^n |x_i|^q)^{1/q}$ for $q > 0$, and $\|\boldsymbol{x}\|_\infty := \max_{i=1,2,\ldots,n} |x_i|$, where x_i denotes the ith coordinate of $\boldsymbol{x}$. The Frobenius norm of a matrix $\boldsymbol{X} \in \mathbb{R}^{d\times n}$ is defined as $\|\boldsymbol{X}\|_{\mathrm{F}} := (\sum_{i=1}^n \sum_{j=1}^p x_{ij}^2)^{1/2}$, where x_{ij} denotes the entry of $\boldsymbol{X}$ at the ith row and jth column. $\boldsymbol{X}_i$ denotes the ith column of $\boldsymbol{X}$.

2 Comparisons of different structured sparsities and computational complexity of optimization methods

In this section, we summarize all these structured sparsities presented in this article and compare the computational complexity of typical optimization methods of these structured sparse learning problems.

In Table 2, we list formulations of these structured sparse learning problems, their corresponding optimization algorithms, and public available software implementations.

We also compare the convergence rates under convex and strongly convex conditions of typical first-order optimization methods in Table 3. In this article, we focus mainly on the first-order optimization methods. First-order methods typically require access to an objective function's gradient or subgradient. The algorithms typically take the form $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \alpha^t \boldsymbol{g}^t$ for some step sizes α^t and descent direction $\boldsymbol{g}^t$. As such, each iteration takes approximately $O(n)$ time. A comparison between

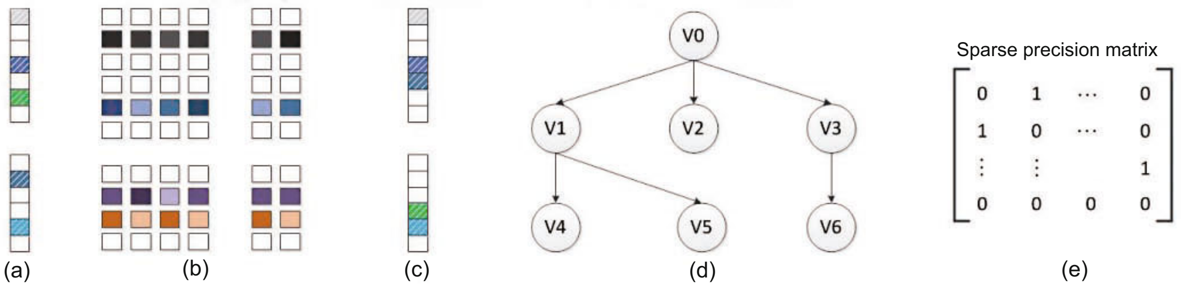

Fig. 2 Illustration of sparsity and its extensions: (a) standard sparsity; (b) grouped sparsity; (c) fused sparsity; (d) hierarchical sparsity; (e) graphical sparsity (References to color refer to the online version of this figure)

Table 2 An overview of formulations, processing algorithms, and software packages for Lasso and its structured extensions

Method	Loss, $l(\boldsymbol{x})$	Regularization, $R(\boldsymbol{x})$	Optimization algorithms	Software packages
Lasso (Tibshirani, 1996)	$\|\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|\|_{\mathrm{F}}^2$	$\lambda\|\boldsymbol{x}\|_1$	Generic QP methods after reformulation (Bach *et al.*, 2012a), alternating direction methods (Boyd *et al.*, 2011), proximal methods (Parikh and Boyd, 2014), block coordinate descent methods (Tseng and Yun, 2009; Wen *et al.*, 2012; Peng *et al.*, 2016), iteratively reweighted methods (Chartrand and Yin, 2008; Lai *et al.*, 2013), working-set and homotopy methods (Bach *et al.*, 2012a)	CVX (Grant and Boyd, 2013), SDPT3 (Toh *et al.*, 2006), YALL1 (Zhang Y *et al.*, 2011), SPGL1 (van den Berg and Friedlander, 2007), SLEP (Liu J *et al.*, 2009), SPAMs (Mairal *et al.*, 2011), SparseLab (Donoho *et al.*, 2007)
Grouped Lasso (Yuan and Lin, 2006)	$\|\|\boldsymbol{y}-\boldsymbol{\beta}_0-\sum\limits_{i=1}^{g}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_i\|\|_{\mathrm{F}}^2$	$\lambda\sum_{i=1}^{g}\sqrt{p_i}\|\|\boldsymbol{\beta}_i\|\|_2$	Generic SOCP methods after reformulation (Bach, 2008b), alternating direction methods (Boyd *et al.*, 2011), proximal methods (Parikh and Boyd, 2014), block coordinate descent methods (Tseng and Yun, 2009; Wen *et al.*, 2012; Peng *et al.*, 2016), iteratively reweighted methods (Chartrand and Yin, 2008; Lai *et al.*, 2013), working-set and homotopy methods (Bach *et al.*, 2012a)	CVX (Grant and Boyd, 2013), SDPT3 (Toh *et al.*, 2006), YALL1 (Zhang Y *et al.*, 2011), SLEP (Liu J *et al.*, 2009), SPAMs (Mairal *et al.*, 2011)
Fused Lasso (Tibshirani *et al.*, 2005)	$\sum\limits_{i=1}^{N}(\boldsymbol{y}_i-\boldsymbol{\beta}_0-\sum\limits_{j=1}^{p}x_{ij}\boldsymbol{\beta}_j)^2$	$\lambda_1\sum_{j=1}^{p}\|\|\boldsymbol{\beta}_j\|\|_1+$ $\lambda_2\sum_{j=1}^{p-1}\|\|\boldsymbol{\beta}_{j+1}-\boldsymbol{\beta}_j\|\|_1$	Generic SDP methods after reformulation (Selesnick and Bayram, 2014), alternating direction methods (Boyd *et al.*, 2011), proximal methods (Parikh and Boyd, 2014), block coordinate descent methods (Tseng and Yun, 2009; Wen *et al.*, 2012; Peng *et al.*, 2016), iteratively reweighted methods (Chartrand and Yin, 2008; Lai *et al.*, 2013), working-set and homotopy methods (Bach *et al.*, 2012a)	CVX (Grant and Boyd, 2013), SDPT3 (Toh *et al.*, 2006), SLEP (Liu J *et al.*, 2009), SPAMs (Mairal *et al.*, 2011)
Hierarchical Lasso (Zhao *et al.*, 2009)	$\|\|\boldsymbol{y}-\boldsymbol{\beta}_0-\sum\limits_{i=1}^{g}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_i\|\|_{\mathrm{F}}^2$	$\lambda\sum_{g\in\mathcal{G}}w_g\|\|\boldsymbol{\alpha}_g\|\|$	Generic SDP methods after reformulation (Bach, 2008a), alternating direction methods (Boyd *et al.*, 2011), proximal methods (Parikh and Boyd, 2014), block coordinate descent methods (Tseng and Yun, 2009; Wen *et al.*, 2012; Peng *et al.*, 2016), iteratively reweighted methods (Chartrand and Yin, 2008; Lai *et al.*, 2013), working-set and homotopy methods (Bach *et al.*, 2012a)	CVX (Grant and Boyd, 2013), SDPT3 (Toh *et al.*, 2006), SLEP (Liu J *et al.*, 2009), SPAMs (Mairal *et al.*, 2011)
Graphical Lasso (Meinshausen and Bühlmann, 2006)	$\log\det\boldsymbol{\Theta}-\mathrm{trace}(\boldsymbol{S}\boldsymbol{\Theta})$	$\lambda\|\|\boldsymbol{\Theta}\|\|_1$	Generic SDP methods after reformulation (Bach, 2008a), alternating direction methods (Boyd *et al.*, 2011), proximal methods (Parikh and Boyd, 2014), block coordinate descent methods (Tseng and Yun, 2009; Wen *et al.*, 2012; Peng *et al.*, 2016), iteratively reweighted methods (Chartrand and Yin, 2008; Lai *et al.*, 2013), working-set and homotopy methods (Bach *et al.*, 2012a)	CVX (Grant and Boyd, 2013), SDPT3 (Toh *et al.*, 2006), SLEP (Liu J *et al.*, 2009), SPAMs (Mairal *et al.*, 2011)

Table 3 An overview of optimization methods' computational complexity

Algorithm	Formulation	Convex	Strongly convex	Notes
Subgradient descent	$\min\limits_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$	$O(1/\varepsilon^2)$ (Nesterov, 2004)	$O(1/\varepsilon)$ (Lacoste-Julien *et al.*, 2012)	Cannot be improved upon without further assumptions
Mirror descent	$\min\limits_{\boldsymbol{x}\in\mathcal{C}} f(\boldsymbol{x})$	$O(1/\varepsilon^2)$ (Beck and Teboulle, 2003)	$O(1/\varepsilon)$ (Nemirovski, 2004)	Different parameterizations result in gradient descent and exponentiated gradient descent
Dual averaging	$\min\limits_{\boldsymbol{x}\in\mathcal{C}} f(\boldsymbol{x})$	$O(1/\varepsilon^2)$ (Nesterov, 2009)	$O(\log(1/\varepsilon))$ (Suzuki, 2013)	Cannot be improved upon without further assumptions
Gradient descent	$\min\limits_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$	$O(1/\varepsilon)$ (Nesterov, 2004)	$O(\log(1/\varepsilon))$ (Hazan *et al.*, 2007)	Applicable when $f(\boldsymbol{x})$ is a strongly convex function
Accelerated gradient descent	$\min\limits_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$	$O(1/\sqrt{\varepsilon})$ (Su *et al.*, 2014)	$O(\log(1/\varepsilon))$ (Tseng, 2008)	Applicable when $f(\boldsymbol{x})$ is differentiable. Cannot be improved upon without further assumptions. Has better constants than gradient descent for the strongly convex case
Proximal gradient descent	$\min\limits_{\boldsymbol{x}\in\mathcal{C}} f(\boldsymbol{x})+g(\boldsymbol{x})$	$O(1/\varepsilon)$ (Combettes and Pesquet, 2011)	$O(\log(1/\varepsilon))$ (Suzuki, 2013)	Applicable when $f(\boldsymbol{x})$ is differentiable and $\text{prox}_{\tau t g}(\boldsymbol{x})$ is easily computable
Accelerated proximal gradient descent	$\min\limits_{\boldsymbol{x}\in\mathcal{C}} f(\boldsymbol{x})+g(\boldsymbol{x})$	$O(1/\sqrt{\varepsilon})$ (Mairal, 2013)	$O(\log(1/\varepsilon))$ (Lin *et al.*, 2015)	Applicable when $f(\boldsymbol{x})$ is differentiable and $\text{prox}_{\tau t g}(\boldsymbol{x})$ is easily computable. Has better constants than proximal gradient descent for the strongly convex case
Frank-Wolfe algorithm/conditional gradient algorithm	$\min\limits_{\boldsymbol{x}\in\mathcal{C}} f(\boldsymbol{x})$	$O(1/\varepsilon)$ (Jaggi, 2013)	$O(1/\sqrt{\varepsilon})$ (Garber and Hazan, 2015)	Applicable when $\mathcal{C}$ is bounded and $h_g(\boldsymbol{x}) = \text{argmin}_{\boldsymbol{x}\in\mathcal{C}} < g, \boldsymbol{x} >$ is easily computable. Most useful when $\mathcal{C}$ is a polytope in a very high dimensional space with sparse extrema

The per-iteration costs of all the algorithms are $O(n)$

several first-order methods was provided by Qiao *et al.* (2016a). Higher-order methods are excluded in this article, because some open issues need to be addressed to apply higher-order methods in large-scale machine learning problems.

3 Grouped structured sparsity

In many regression problems, the variables are predefined in groups as prior knowledge in practical situations such as the analysis-of-variance (ANOVA). In this setting, the typical goal is to choose major effects and interactions among variables. For example, in supervised learning problems, to generate the classification label, variables are organized as collections of categorical predictors, and the selection of significant variables corresponds to the selection of groups of variables. To handle these problems, grouped structured sparsity has been proposed, and these extended methods are often called 'grouped Lasso' (Yuan and Lin, 2006).

3.1 Formulations of grouped structured sparsity

The grouped Lasso was studied and generalized by Yuan and Lin (2006). Assume that the number of groups is fixed and finite, and predictors are divided into g groups with p_i as the number of predictors in the ith group G_i, $p=\sum_{i=1}^{g}p_i$, and $G_i\cap G_j=\varnothing$. The grouped Lasso is formulated as

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p}||\boldsymbol{y}-\boldsymbol{\beta}_0-\sum_{i=1}^{g}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_i||_{\mathrm{F}}^2+\lambda\sum_{i=1}^{g}\sqrt{p_i}||\boldsymbol{\beta}_i||_2. \quad (3)$$

The sparsity of the solution depends on the magnitude of the tuning parameter λ, and exploits the non-differentiability of $||\boldsymbol{\beta}_i||_2$ at $\boldsymbol{\beta}_i=\mathbf{0}$. Note that the grouped Lasso estimates and the grouped sparsity pattern converges to the correct patterns in probability (Bach, 2008b).

To obtain sparsity at both group and variables' element-wise levels, Simon *et al.* (2013) proposed a sparse grouped Lasso, which is formulated as

$$\min ||\boldsymbol{y}-\boldsymbol{\beta}_0-\sum_{i=1}^{g}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_i||_{\mathrm{F}}^2+\lambda_1\sum_{i=1}^{g}\sqrt{p_i}||\boldsymbol{\beta}_i||_2 + \lambda_2||\boldsymbol{\beta}||_1, \quad (4)$$

where the second term controls the sparsity at the group level, and the third term controls the sparsity at the variables' element-wise level. When $\lambda_1=0$, problem (4) degenerates to standard Lasso (Tibshirani, 1996). When $\lambda_2=0$, it generates grouped Lasso (problem (3)). Because the grouped Lasso may suffer from estimation inefficiency and selective inconsistency, an adaptive grouped Lasso method (Wang and Leng, 2008) has been proposed as a remedy. Considering the potential non-uniqueness of solutions and high computational costs, a generalized linear model (GLM) (Roth and Fischer, 2008) has been proposed with an active-set algorithm. The number of groups is allowed to grow with the increase of the number of observed data points (Meinshausen and Yu, 2008), and the extension with dynamic group division is also available in Mougeot *et al.* (2013).

3.2 Algorithms for grouped structured sparsity

The grouped Lasso optimization problem (3) can be solved through a coordinate gradient descent algorithm, which is applicable to a broad class of convex loss functions, and convergence of the algorithm is established (see Hong *et al.* (2015) and references therein). The blockwise coordinate descent (BCD) algorithm was also used to solve the problem in Meier *et al.* (2008) and Liu H *et al.* (2009). However, the BCD method may get stuck in ill conditions, and in Vincent and Hansen (2014), a modified BCD algorithm was proposed, which first computes a descent direction at the given point, and then uses a line search to find the next starting point. There are also publicly accessible software implementations that can be used to solve the problem. For example, the SLEP package (Liu J *et al.*, 2009) was used in Xie and Xu (2014). More available implementations are listed in Table 2.

The performance of these algorithms has been comparatively studied in Rakotomamonjy (2011), who concluded that depending on the performance measure, greedy approaches and iterative reweighted algorithms are more efficient in either computational complexity or sparsity recovered.

3.3 Applications of grouped structured sparsity

There are numerous applications to pursue grouped sparsity. In computer vision, each group corresponds to different data sources or data types, and different data sources could be referred to as views. In speech and signal processing, similar groups represent different frequency bands (McAuley *et al.*, 2005). In Vincent and Hansen (2014), grouped Lasso was used in multinomial classification and solved with a coordinate gradient descent method. In document processing, Bengio *et al.* (2009) combined grouped sparse learning with a bag-of-words document representation.

4 Fused structured sparsity

In many real-world applications, coefficients are organized in a specific order and have local constancy. For instance, organizing variables in blocks as prior knowledge may produce a better result. In this setting, it is necessary to extend Lasso to exploit the ordered structure, and the extension is called ‘fussed Lasso’ (Tibshirani *et al.*, 2005).

4.1 Formulations of fused structured sparsity

In regression problems, variables $\boldsymbol{x}$ may have a natural order. Specifically, variables are ordered according to some index variable t, and they have local constancy of the feature profile. These variables are invoked as predictors. To exploit the local constancy of the coefficient, Tibshirani *et al.* (2005) proposed fussed Lasso, which extends the Lasso penalty to take ordering into account. The extended method is also called ‘generalized Lasso’. The fused Lasso can be formulated as

$$\begin{aligned}\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \quad & \sum_{i=1}^{N} \|\boldsymbol{y}_i - \boldsymbol{\beta}_0 - \sum_{j=1}^{p} x_{ij}\boldsymbol{\beta}_j\|_{\mathrm{F}}^2 + \lambda_1 \sum_{j=1}^{p} ||\boldsymbol{\beta}_j||_1 \\ & +\lambda_2 \sum_{j=1}^{p-1} ||\boldsymbol{\beta}_{j+1} - \boldsymbol{\beta}_j||_1, \end{aligned} \tag{5}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ are the variables of the model, and the N pairs $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ are the training data with noise. There are three terms in the formulation: the first term empirically minimizes the training error for a given dataset, the second term encourages sparsity in feature, and the third term is used to penalize the differences of adjacent coefficients. These terms are tuned with parameters λ_1 and λ_2. Note that the fused Lasso makes the assumption that the index t is uniformly distributed (John Lu, 2010), and the third term should be generalized on divided differences $\lambda_2 \sum_{j=1}^{p-1}(\boldsymbol{\beta}_{j+1} - \boldsymbol{\beta}_j)/(t_{j+1} - t_j)$.

A generalized two-dimensional fused Lasso was proposed in Friedman *et al.* (2007). In this model, parameters are laid out in a grid of pixels with a 2D total variation (TV) norm (Rudin *et al.*, 1992), which is usually used in image denoising, image smoothing, and data recovery. The general strategy for two-dimensional fused Lasso can be directly applied in higher-order problems, formulated as a tensor with a higher-order TV-norm as regularization.

4.2 Algorithms to solve fused structured sparsity

The fused Lasso is a strictly convex problem in $\boldsymbol{\beta}$. For 1D fused Lasso, SQOPT (Gill *et al.*, 2008) can be used directly, in which there is a two-phase, active-set algorithm designed for quadratic programming problems with sparse linear constraints. For a class of convex optimization problems, a coordinate descent algorithm was presented in Friedman *et al.* (2007), which is a one-at-a-time coordinate-wise descent algorithm, and can be generalized to solve the 2D fused Lasso or even higher dimensional cases.

To handle large-scale fused Lasso problems, Ye and Xie (2011) proposed an iterative algorithm based on the split Bregman method to solve a class of large-scale fused Lasso problems. Wang LC *et al.* (2013) presented an augmented Lagrangian method (ALM) for general convex loss. Li *et al.* (2014) proposed a fast linearized alternating direction method to solve the general Lasso model, and Qiao *et al.* (2016b; 2016c) improved the method to solve the structed nonconvex problems.

Experiment results in Hoefling (2010) showed that Tibshirani’s and Friedman’s algorithms are clear and fast, and are state-of-the-art methods. The linearized method in Li *et al.* (2014) can be used to handle larger-scale problems.

4.3 Applications of fused structured sparsity

Fused Lasso is validated in protein mass spectroscopy (MS), gene expression, and image smoothing problems. The protein MS, which holds great promise for biomarker identification and proteomics profiling, was also used as a motivating example to demonstrate the efficacy of fused Lasso in Tibshirani *et al.* (2005) and Tibshirani and Wang (2008).

As demonstrated in Huang *et al.* (2005), another important application of fused Lasso is the reconstruction of copy numbers from comparative genomic hybridization (CGH) data arrays. For images with smoothness among pixels, fused Lasso can achieve a rather good performance (Friedman *et al.*, 2007).

5 Hierarchically structured sparsity

In hierarchical sparsity, the variables are organized hierarchically (Xu *et al.*, 2011) or integrated into a tree, and form a union of potentially overlapping groups that were defined previously. The hierarchical sparsity may be achieved through various extended sparsity-inducing norms, and the extension is often called 'hierarchical Lasso' (Zhao *et al.*, 2009).

5.1 Formulations of hierarchically structured sparsity

Hierarchical Lasso assumes that the p variables are assigned to the nodes of a tree T, or a forest. In this setting, if a feature is selected then it implies that all its ancestors in T have already been selected, and if a node is not selected, then its descendants are not selected.

The hierarchical Lasso was first presented in Zhao *et al.* (2009). The authors proposed the structured penalty, which is called the composite absolute penalty (CAP). By allowing the groups to overlap, CAP can be used to represent a hierarchy structure among the predictors. In a given group or hierarchical structure, the CAP penalty must be specialized for grouped and hierarchical selection. Assume the grouping is denoted as $G = (G_1, G_2, \ldots, G_K)$, and the norm is denoted as $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_K) \in \mathbb{R}_+^{K+1}$. The CAP penalty with grouping G could be formulated as follows:

$$T_{G,\boldsymbol{\gamma}}(\boldsymbol{\beta}) = \sum_k ||\boldsymbol{\beta}_{G_k}||_{\gamma_k}^{\gamma_0},$$

where $\boldsymbol{\beta}_{G_k} = \{\boldsymbol{\beta}_j | j \in G_k\}$, and the corresponding CAP estimation for the tuning parameter λ is

$$\hat{\boldsymbol{\beta}}_{G,\boldsymbol{\gamma}}(\lambda) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}(L(Z, \boldsymbol{\beta}) + \lambda T_{G,\boldsymbol{\gamma}}(\boldsymbol{\beta})), \quad (6)$$

where $L(Z, \boldsymbol{\beta})$ is the loss function. Specified overlapping patterns corresponding to the given structure can be used for hierarchical variable selection. For piecewise quadratic loss functions, CAP with ℓ_1-norm or ℓ_∞-norms has advantages that their regularization paths are piecewise linear. If $\gamma_i \geq 1$ $(i = 0, 1, \ldots, K)$, then $T(\boldsymbol{\beta})$ is convex. If the loss function $L(\cdot)$ is convex in $\boldsymbol{\beta}$, then the objective function of CAP estimation is convex.

In compressive sensing, Baraniuk *et al.* (2010) presented tree sparsity in the context of sparse wavelet decompositions. The consistency of the sparse estimator of potentially overlapping groups was given in Jacob *et al.* (2009). In sparse coding, Jenatton *et al.* (2010) proposed an extension that the atoms are further assumed to be embedded in a tree, which was achieved using tree-structured sparse regularization norms.

5.2 Algorithms of hierarchically structured sparsity

The BLasso algorithm (Zhao and Yu, 2007), which was derived from a coordinate descent method with a fixed step size applied to the general Lasso loss function, can be used to solve the minimization problem. Zhao *et al.* (2009) extended BLasso and proposed the hiCAP algorithm for hierarchical variables selection, which was valid for the ℓ_2-loss when $\gamma_0 = 1, \gamma_k = \infty$, or a tree-structured hierarchy in graph representation. Considering the formulation's nonseparability and non-smoothness, Chen *et al.* (2012) proposed the smoothing proximal gradient (SPG) method, which combines a smoothing technique with an effective proximal gradient method to solve structured sparse regression problems with a smooth convex loss.

Proximal methods (Parikh and Boyd, 2014), which extend the projection operator to a convex set, have recently been shown effective in solving variational problems. To accelerate the convergence, Mosci *et al.* (2010) added a strictly convex function to the objective function and the experiment results showed that it reduces the number of substantial optimization iterations. After introducing auxiliary variables, Micchelli *et al.* (2013) used an alternating minimization algorithm with a projection procedure to solve the problem and established the theorem of convergence. Combined with an active set strategy, Villa *et al.* (2014) accelerated the proximal method by using a new active set strategy to compute the proximal operator. When the objective function is strongly convex, Xiao and Zhang (2014) proposed a proximal stochastic gradient method iterated in an incremental way, which provides an efficient way to solve large-scale problems.

5.3 Applications of hierarchically structured sparsity

In wavelet decompositions, it is natural to organize them in a tree because of their multi-scale structure, and it benefits image compression and denoising (Baraniuk *et al.*, 2010; Huang *et al.*, 2011). In dictionary learning, Jenatton (2011) used hierarchical dictionary learning in image restoration and performed multi-scale mining of fMRI data for the prediction of simple cognitive tasks (Jenatton *et al.*, 2012). In genetics, Kim and Xing (2010) used it to exploit the tree structure of gene networks for multi-task regression. In topic models, Blei *et al.* (2010) proposed a hierarchical model of latent variables based on Bayesian non-parametric methods to model hierarchies of topics.

6 Graphically structured sparsity

The graph is a powerful data structure in model construction for a lot of machine learning algorithms, such as the graphical model (Wainwright and Jordan, 2008) and high-dimensional model selection (Meinshausen and Bühlmann, 2006). Sparse graphs have a relatively small number of edges, and are economical to use with good interpretability. The problem of estimating sparse graphs may be resolved by 'graphical Lasso' (Banerjee *et al.*, 2008), which is an extension of Lasso on the inverse covariance matrix.

6.1 Formulations of graphically structured sparsity

A graph G consists of a set of vertices N, and an edge set E. In undirected graphical models, each vertex represents a random variable, and each edge represents the dependent relationship between the two vertices. The absence of an edge between two vertices has a special meaning: the corresponding random variables are conditionally independent, given that the rest of the variables are known.

Wermuth (1976) showed that if graph G is Gaussian distributed, then it has the property that conditional independence of vertices corresponds to nonzero entries in the precision matrix. Model selection for undirected Gaussian graphical models is equivalent to selecting non-zero elements in the precision matrix. Dempster (1972) named the problem 'covariance selection'. Considering that the values of the precision matrix are part of continuous variables, it is natural to extend variable selection to edge selection on the graph (see Banerjee *et al.* (2008) and references therein).

Assume that there are N multi-variate normal variables $\boldsymbol{x}_i$ $(i=1,2,\ldots,N)$ with population mean μ and covariance $\boldsymbol{\Sigma}$. The empirical covariance matrix is $\boldsymbol{S}=\sum_{i=1}^{N}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^{\mathrm{T}}/N$, where $\bar{\boldsymbol{x}}$ is the mean of the samples. Without considering the constants, the log-likelihood of the data can be written as $\ell(\boldsymbol{\Theta})=\log\det\boldsymbol{\Theta}-\mathrm{trace}(\boldsymbol{S}\boldsymbol{\Theta})$, where the quantity $\ell(\boldsymbol{\Theta})$ is a convex function of $\boldsymbol{\Theta}$ and the maximum likelihood estimate of $\boldsymbol{\Sigma}$ is $\boldsymbol{S}$. Considering Wishart log-likelihood, there is $\boldsymbol{\Theta}=\boldsymbol{\Sigma}^{-1}$. Then Lagrange constants for all missing edges are formulated as

$$\ell_{\mathrm{c}}(\boldsymbol{\Theta})=\log\det\boldsymbol{\Theta}-\mathrm{trace}(\boldsymbol{S}\boldsymbol{\Theta})-\lambda||\boldsymbol{\Theta}||_1, \qquad (7)$$

where $||\boldsymbol{\Theta}||_1$ is the element-wise ℓ_1-norm of $\boldsymbol{\Theta}$, and term $\lambda||\boldsymbol{\Theta}||_1$ is used as a sparsity-inducing norm on the inverse covariance matrix.

6.2 Algorithms of graphically structured sparsity

The optimization problem (7) is more convenient to solve than the original model selection problem. Banerjee *et al.* (2008) showed that problem (7) is convex and the problem can be solved by optimizing over each row and the corresponding column of $\boldsymbol{W}$ in a block coordinate descent fashion. Concretely, $\boldsymbol{W}$ and $\boldsymbol{S}$ could be partitioned as

$$\boldsymbol{W}=\begin{pmatrix}\boldsymbol{W}_{11} & \boldsymbol{w}_{12}\\ \boldsymbol{w}_{12}^{\mathrm{T}} & w_{22}\end{pmatrix}, \boldsymbol{S}=\begin{pmatrix}\boldsymbol{S}_{11} & \boldsymbol{s}_{12}\\ \boldsymbol{s}_{12}^{\mathrm{T}} & s_{22}\end{pmatrix},$$

and the gradient equation for maximizing problem (7) is $\boldsymbol{\Theta}^{-1}-\boldsymbol{S}-\lambda\cdot\mathrm{sign}(\boldsymbol{\Theta})=\mathbf{0}$. Then, the solution for $\boldsymbol{w}_{12}$ would be given by solving

$$\boldsymbol{w}_{12}=\underset{\boldsymbol{y}}{\mathrm{argmin}}\{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{W}_{11}^{-1}\boldsymbol{y}:||\boldsymbol{y}-\boldsymbol{s}_{12}||_{\infty}\leq\rho\}.$$

Following Banerjee's work, Friedman *et al.* (2008) proposed an algorithm named 'graphical Lasso' using a coordinate descent method. Motivated by the success of convex relaxation for the rank-minimization problem, Chandrasekaran *et al.* (2012) introduced a regularized maximum normal likelihood decomposition framework with a trace norm penalty term, and Ma *et al.* (2013) developed a proximal gradient based alternating direction method of multipliers to solve these problems.

In practice, tuning the parameter is essential for the results. Yuan and Lin (2007) proposed a Bayesian information criterion (BIC) type criterion for the selection of the tuning parameter.

6.3 Applications of graphically structured sparsity

Graphical Lasso has been applied to various research fields, including gene network discovery and social-network data analysis.

Jones and West (2005) applied graphical Lasso to the analysis of gene expression data, which consists of 8408 variables and has roughly a multivariate Gaussian distribution. Friedman *et al.* (2008) used it to analyze a flow cytometry dataset of 11 proteins and 7466 cells, and produced a directed acyclic graph in cell signal data.

In network discovery, Leng and Tang (2012) used it to analyze the U.S. agricultural export data and presented the network and the regions of the US Department of Agriculture export data from 1970 to 2009. Considering that the structure may vary from time to time, Kolar and Xing (2011) proposed that the structure of the undirected graphical model can be consistently estimated in the high-dimensional setting, when the dimensionality of the model is allowed to diverge with the sample size.

7 Experiments

In this section, four numerical experiments are performed to evaluate three of these structured sparse learning methods mentioned in previous sections. The results obtained from these numerical studies are detailed in this section. The first two experiments demonstrate that the structured sparsity-inducing method can recover the signal while the standard sparse learning cannot. The third experiment demonstrates that model parsimony can be obtained through a structured method beyond the standard sparse learning method in wavelet coefficients selection. The fourth experiment demonstrates that the accuracy could be increased with the graphical structure used in logistical regression.

Several publicly accessible software packages are used, including SparseLab (Donoho *et al.*, 2007), SLEP (Liu J *et al.*, 2009), SPAMS (Mairal *et al.*, 2011), YALL1 (Zhang Y *et al.*, 2011), and SPGL1 (van den Berg and Friedlander, 2007). In this article, we use mainly SPAMS and YALL1 to solve the optimization problem. All numerical experiments are conducted with Matlab 7.12.0 on a laptop with an Intel Core I7-4710-MQ, 2.5 GHz CPU, and 4 GB of RAM.

7.1 Measurements

Accuracy, model explainability, and computational complexity are the three most important aspects to consider in machine learning algorithms. The latter two are dominated mainly by model parsimony. Thus, we use prediction accuracy on a test dataset and model parsimony to evaluate the structured sparse learning methods in this study.

The prediction accuracy on the test dataset is formulated as

$$\mathrm{Acc}_{\text{test}} = \frac{\mathrm{TP}+\mathrm{TN}}{P+N}, \tag{8}$$

where TP is true positive with hit, and TN is true negative with correct rejection. We propose a parsimony ratio to measure model parsimony, which is formulated as

$$\mathrm{PR} = \frac{\text{Number of variables in the model}}{\text{Number of variables in the orginal model}}. \tag{9}$$

7.2 Signal recovery with fused structured sparsity

To show the capability of the fused structure for solving fused structured problems, we conduct the following tests. First, we generate the regression coefficient $\hat{\boldsymbol{x}} \in \mathbb{R}^n$ for $n = 1000$ as

$$\hat{\boldsymbol{x}} = \begin{cases} r_1, & j = 1, 2, \cdots, 100, \\ r_2, & j = 201, 202, \cdots, 300, \\ r_3, & j = 401, 402, \cdots, 500, \\ r_4, & j = 601, 602, \cdots, 700, \\ 0, & \text{else}, \end{cases}$$

where scalers r_1, r_2, r_3, and r_4 are randomly generated and uniformly distributed on (0, 1). The plot of $\hat{\boldsymbol{x}}$ is shown in Fig. 3a. The entries of matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ with $m = 500$ and $n = 1000$ are drawn from standard distribution $\mathcal{N}(0,1)$. Observations $\boldsymbol{b} \in \mathbb{R}^m$

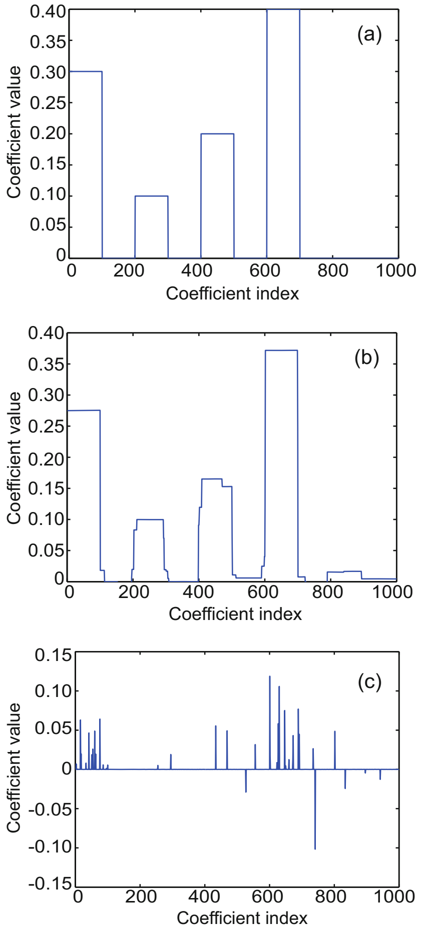

Fig. 3 Coefficients recovered with fused structure regularization: (a) original coefficients; (b) coefficients recovered with the fused sparse model (problem (5)); (c) coefficients recovered with sparse learning (without the fused structure regularized term)

are then created as the signs of $A\hat{\boldsymbol{x}}+\boldsymbol{e}$, where $\boldsymbol{e} \in \mathbb{R}^m$ is a random vector with distribution $\mathcal{N}(0, 0.05)$. Parameters in the fused model, specifically problem (5), are setted as $\lambda_1 = 5 \times 10^{-4}$, $\lambda_2 = 5 \times 10^{-2}$, while in the sparse model (problem (1)), $\lambda_1 = 5 \times 10^{-2}$.

Fig. 3b shows that the fused model can preserve the natural ordering rather well. Fig. 3c shows that the sparse model presents a sparse solution, but cannot preserve the natural ordering. This example demonstrates that fused structured sparse learning surpasses sparse learning when the original data has a fused structure.

7.3 Signal recovery improved by grouped structured sparsity

To show the capability of the grouped structure for solving grouped structured problems, we conduct the following tests. We consider solving problem (3) with weighted groups. First, we create a random $m \times n$ encoding matrix, and perform scaling by normalizing the rows of the encoding matrix. Then, we generate groups with the desired number of unique groups, and a weight is determined for each group. Next, the observations are generated through the encoding matrix and grouped sparse vector, followed by a Gaussian noise added to the observation.

The original signal is presented in Fig. 4a. Fig. 4b shows that the grouped model can recover the signal rather well. Fig. 4c shows that the sparse model without the grouped regularization term presents a sparse solution. A sparse method without the grouped regularized term could not recover the original signal, although the data fitting loss is relatively low. This example demonstrates that grouped structured sparse learning surpasses sparse learning when the original data has a grouped structure.

7.4 Model parsimony gained by hierarchically structured sparsity

To show the capability of the hierarchical structure, we conduct the following tests. First, an image with five rectangles is randomly generated. Then, we apply ordinary linear sampling to measure the 4096 wavelet coefficients directly. Compared to the linear structure, we also perform a hierarchical structure that has 1152 wavelet coefficients.

Fig. 5b shows that the linear model can reconstruct images well with $\|\boldsymbol{x}_{\text{lin}} - \boldsymbol{x}_0\|_{\text{F}}^2/\|\boldsymbol{x}_0\|_{\text{F}}^2 = 0.3567$. Fig. 5c shows that the hierarchical model can greatly reduce the number of coefficients, specifically from 4096 to 1152, with a model parsimony rate of 1152/4095=28.1%, while the relative error is comparable, $\|\boldsymbol{x}_{\text{hie}} - \boldsymbol{x}_0\|_{\text{F}}^2/\|\boldsymbol{x}_0\|_{\text{F}}^2 = 0.3568$. This example demonstrates that hierarchically structured sparse learning can greatly improve model parsimony beyond sparse learning when the original data is structured hierarchically.

7.5 Prediction accuracy improved by graphically structured sparsity

To show the ability of the graphical structure to solve graphically structured problems, we conduct the following tests. The experiment is conducted on the binary classification datasets: ‘20

Newsgroups ’ (www.cs.nyu.edu/~roweis/data.html). The 20 Newsgroups dataset is a collection of approximately 20 000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. The 20 Newsgroups collection has become a popular dataset for experiments in text applications of machine learning techniques, such as text classification and text clustering. Here, the first 100 words are selected and we use 80% of the samples for training and 20% for testing. To reduce statistical variability, experimental results are averaged over 10 repetitions. First, we use graphical Lasso to generate the graphical relationship. Then, we propose a novel graph-guided method to do classification, and compare the average predication accuracy of the standard classifier without a graphical structure penalty.

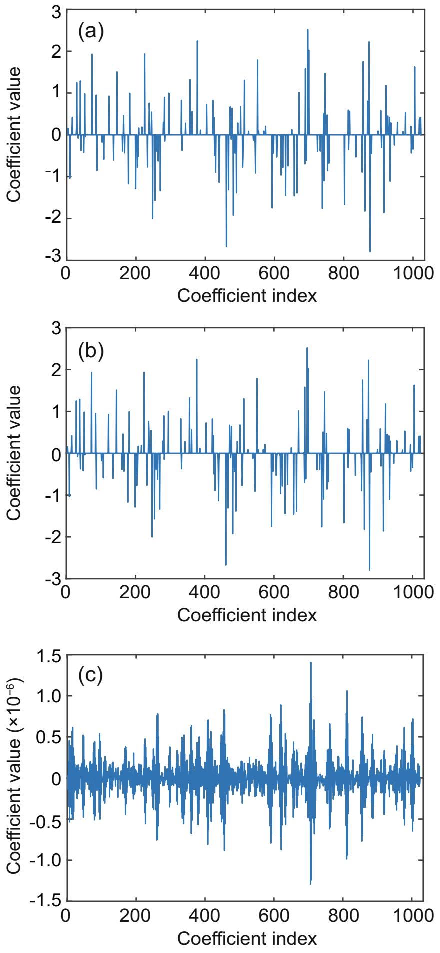

Fig. 4 Coefficients recovered with grouped structure regularization: (a) original coefficients; (b) coefficients recovered with the grouped sparse model (problem (5)); (c) coefficients recovered with sparse learning (without the grouped structure regularized term)

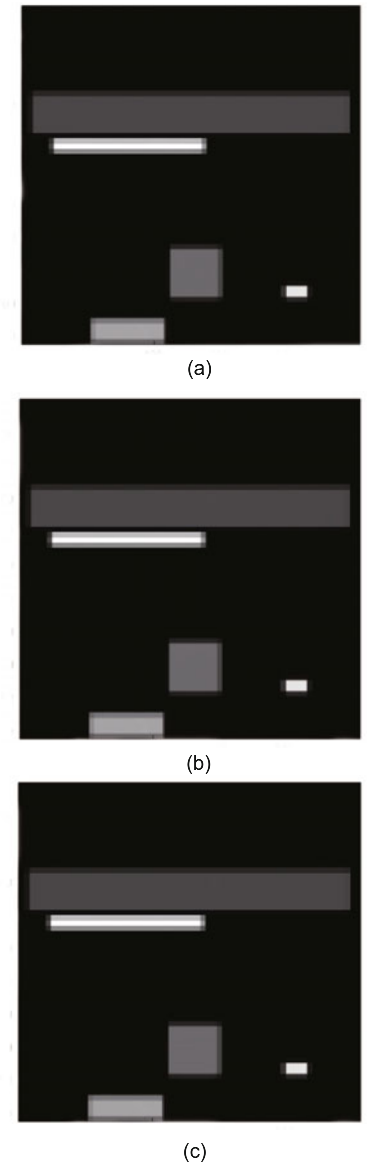

Fig. 5 Images reconstructed with hierarchically structured sparsity: (a) original image; (b) linear reconstruction from 4096 samples ($\|\boldsymbol{x}_{\rm lin}-\boldsymbol{x}_0\|_{\rm F}^2/\|\boldsymbol{x}_0\|_{\rm F}^2=0.3567$); (c) hierarchically structured reconstruction from 1152 samples ($\|\boldsymbol{x}_{\rm hie}-\boldsymbol{x}_0\|_{\rm F}^2/\|\boldsymbol{x}_0\|_{\rm F}^2=0.3568$ and the model parsimony rate is 1152/4096=28.1%)

In the second step, we propose a novel graph-guided logistic regression approach, which is formulated as

$$\min_{\boldsymbol{x}}\ l(\boldsymbol{x})+\frac{\gamma}{2}\|\boldsymbol{x}\|_2^2+\lambda\|\boldsymbol{F}\boldsymbol{x}\|_1, \qquad (10)$$

where $l(\boldsymbol{x})=\sum_{i=1}^{N}l(\boldsymbol{x},\xi_i)/N$, $l(\boldsymbol{x},\xi_i)$ is the logistic loss on data sample $\boldsymbol{\xi}_i$, and $\lambda>0$ is a regularization parameter. Furthermore, $\boldsymbol{F}$ is the penalty matrix promoting the desired graphical structure of $\boldsymbol{x}$.

Specifically, $\boldsymbol{F}$ in problem (10) is generated by sparse inverse covariance selection (Scheinberg *et al.*, 2010). We observe that problem (10) can be solved by the stochastic primal dual hybrid gradient method proposed in our prior work (Qiao *et al.*, 2016a).

Fig. 6a shows that the graphical model can generate the graphical relationship among words very well. In Fig. 6b, we see that graphical logistic regression could greatly improve the prediction accuracy beyond standard logistic regression. This example shows that a graphically structured model can greatly improve prediction accuracy.

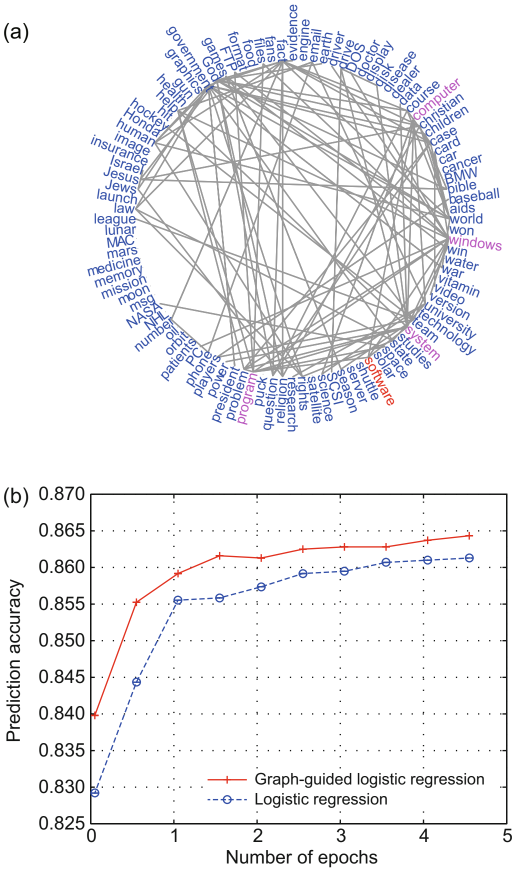

Fig. 6 Classification with graphically structured sparsity: (a) graphical relationship among words; (b) average predication accuracy of graph-guided logistic regression and standard logistic regression

8 Conclusions and discussion

Structured sparse learning methods incorporate specific structure information with sparse learning methods, and have been used in various fields. In this article, we reviewed the development of the theory, formulations, algorithms, and applications of the latest structured sparse learning methods, including grouped structured sparsity, fused structured sparsity, hierarchically structured sparsity, and graphically structured sparsity. For each type of structured sparsity, we presented the original formulation and its variations and the mathematical motivation of these methods, addressed the algorithms for solving these problems, and discussed the fact that applications with prior knowledge lead to improved explicability of the sparse estimations and/or increased prediction performance in related research fields.

Experiments have been conducted to demonstrate the advantage of structured sparse learning algorithms beyond standard sparse learning methods. These experiments demonstrated that the structured sparsity-inducing method could achieve better performance than the standard sparse learning method. We also proposed a novel graph-guided logistic regression method to demonstrate the efficacy of the graphical structure. However, the experiment results on super computers (Yang *et al.*, 2010; 2011) are expected and power efficient algorithms (Lai *et al.*, 2015; 2016) and algorithms for new infrastructures (Chen *et al.*, 2016) are still required.

Though structured sparse learning methods have shown great success from scientific research fields to industrial engineering, there are still many issues to be addressed:

1. Online learning algorithms for structured sparsity problems. Most current structured sparsity is optimized in a batch way. In real-world applications, the training data volume may be huge or be given in a sequential way. Online learning is a better choice to address these problems.

2. Efficient algorithms for non-convex models. Today, most structured models are solved through a convex approximation of the original formulation. However, in statistics, it was reported in the literature that non-convex regularization usually yields a solution with more desirable structural properties, for example, the ℓ_0-norm regularized least squares problem (i.e., $l(\cdot)$ is a least squares function). Efficient non-convex algorithms are needed for these strict models.

3. Specific structure inducing regularization. Many structure-inducing regularizations have been proposed, and many of them have been applied in a wide range of fields. For specifically structured

problems, we still need new regularization to induce the specific structure.

References

Asaei, A., Bourlard, H., Cevher, V., 2011a. Model-based compressive sensing for multi-party distant speech recognition. Proc. ICASSP, p.4600-4603. http://dx.doi.org/10.1109/ICASSP.2011.5947379

Asaei, A., Taghizadeh, M.J., Bourlard, H., *et al.*, 2011b. Multi-party speech recovery exploiting structured sparsity models. Proc. Conf. on Int. Speech Communication Association, p.192-195.

Asaei, A., Bourlard, H., Taghizadeh, M.J., *et al.*, 2014a. Model-based sparse component analysis for reverberant speech localization. Proc. ICASSP, p.1439-1443 http://dx.doi.org/10.1109/ICASSP.2014.6853835

Asaei, A., Golbabaee, M., Bourlard, H., *et al.*, 2014b. Structured sparsity models for reverberant speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.*, **22**(3):620-633. http://dx.doi.org/10.1109/TASLP.2013.2297012

Bach, F., 2008a. Consistency of trace norm minimization. *J. Mach. Learn. Res.*, **9**:1019-1048.

Bach, F., 2008b. Consistency of the group Lasso and multiple kernel learning. *J. Mach. Learn. Res.*, **9**:1179-1225.

Bach, F., Jenatton, R., Mairal, J., *et al.*, 2011. Convex optimization with sparsity-inducing norms. *In*: Sra, S., Nowozin, S., Wright, S.J. (Eds.), Optimization for Machine Learning. MIT Press, Cambridge, p.1-35.

Bach, F., Jenatton, R., Mairal, J., *et al.*, 2012a. Optimization with sparsity-inducing penalties. *Found. Trends Mach. Learn.*, **4**(1):1-106. http://dx.doi.org/10.1561/2200000015

Bach, F., Jenatton, R., Mairal, J., *et al.*, 2012b. Structured sparsity through convex optimization. *Stat. Sci.*, **27**(4):450-468. http://dx.doi.org/10.1214/12-STS394

Bach, F., Jordan, M.I., 2006. Learning spectral clustering, with application to speech separation. *J. Mach. Learn. Res.*, **7**:1963-2001.

Banerjee, O., El Ghaoui, L., d'Aspremont, A., 2008. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.*, **9**:485-516.

Baraniuk, R.G., Cevher, V., Duarte, M.F., *et al.*, 2010. Model-based compressive sensing. *IEEE Trans. Inform. Theory*, **56**(4):1982-2001. http://dx.doi.org/10.1109/Tit.2010.2040894

Beck, A., Teboulle, M., 2003. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.*, **31**(3):167-175. http://dx.doi.org/10.1016/S0167-6377(02)00231-6

Beck, A., Teboulle, M., 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imag. Sci.*, **2**(1):183-202. http://dx.doi.org/10.1137/080716542

Bengio, S., Pereira, F., Singer, Y., *et al.*, 2009. Group sparse coding. Proc. NIPS, p.82-89.

Blei, D.M., Griffiths, T.L., Jordan, M.I., 2010. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *J. ACM*, **57**(2):7. http://doi.acm.org/10.1145/1667053.1667056

Borne, K., 2009. Scientific data mining in astronomy. arXiv:0911.0505.

Boyd, S., Parikh, N., Chu, E., *et al.*, 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, **3**(1):1-122. http://dx.doi.org/10.1561/2200000016

Bruckstein, A.M., Donoho, D.L., Elad, M., 2009. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Rev.*, **51**(1):34-81. http://dx.doi.org/10.1137/060657704

Candès, E., Tao, T., 2007. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Stat.*, **35**(6):2313-2351. http://dx.doi.org/10.1214/009053606000001523

Candès, E.J., 2008. The restricted isometry property and its implications for compressed sensing. *Comput. Rend. Math.*, **346**(9-10):589-592. http://dx.doi.org/10.1016/j.crma.2008.03.014

Candès, E.J., Recht, B., 2009. Exact matrix completion via convex optimization. *Found. Comput. Math.*, **9**(6):717-772. http://dx.doi.org/10.1007/s10208-009-9045-5

Candès, E.J., Romberg, J.K., Tao, T., 2006. Stable signal recovery from incomplete and inaccurate measurements. *Commun. Pure Appl. Math.*, **59**(8):1207-1223. http://dx.doi.org/10.1002/Cpa.20124

Candès, E.J., Wakin, M.B., Boyd, S.P., 2008. Enhancing sparsity by reweighted ℓ_1 minimization. *J. Four. Anal. Appl.*, **14**(5):877-905. http://dx.doi.org/10.1007/s00041-008-9045-x

Chandrasekaran, V., Parrilo, P.A., Willsky, A.S., 2012. Latent variable graphical model selection via convex optimization. *Ann. Stat.*, **40**(4):1935-1967. http://dx.doi.org/10.1214/11-AOS949

Chartrand, R., Yin, W.T., 2008. Iteratively reweighted algorithms for compressive sensing. Proc. ICASSP, p.3869-3872. http://dx.doi.org/10.1109/Icassp.2008.4518498

Chen, C., Huang, J.Z., 2014. Exploiting the wavelet structure in compressed sensing MRI. *Magn. Reson. Imag.*, **32**(10):1377-1389. http://dx.doi.org/10.1016/j.mri.2014.07.016

Chen, C., Li, Y.Q., Huang, J.Z., 2014. Forest sparsity for multi-channel compressive sensing. *IEEE Trans. Signal Process.*, **62**(11):2803-2813. http://dx.doi.org/10.1109/TSP.2014.2318138

Chen, H.Y., Sun, Z.G., Yi, F., *et al.*, 2016. BufferBank storage: an economic, scalable and universally usable in-network storage model for streaming data applications. *Sci. China Inform. Sci.*, **59**(1):1-15. http://dx.doi.org/10.1007/s11432-015-5299-5

Chen, S., Donoho, D., 1994. Basis pursuit. Proc. Asilomar Conf. on Signals, Systems and Computers, p.41-44.

Chen, X., Lin, Q.H., Kim, S., *et al.*, 2012. Smoothing proximal gradient method for general structured sparse regression. *Ann. Appl. Stat.*, **6**(2):719-752. http://dx.doi.org/10.1214/11-AOAS514

Combettes, P.L., Pesquet, J.C., 2011. Proximal splitting methods in signal processing. *In*: Bauschke, H.H., Burachik, R.S., Combettes, P.L., *et al.* (Eds.), Fixed-Point Algorithms for Inverse Problems in Science and Engineering. Springer, Berlin, p.185-212. http://dx.doi.org/10.1007/978-1-4419-9569-8_10

Dempster, A.P., 1972. Covariance selection. *Biometrics*, **28**:157-175.

Donoho, D.L., Huo, X., 2001. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory*, **47**(7):2845-2862. http://dx.doi.org/10.1109/18.959265

Donoho, D.L, Drori, I., Stodden, V.C, *et al.*, 2007. SparseLab. http://sparselab.stanford.edu/

Duarte, M.F., Eldar, Y.C., 2011. Structured compressed sensing: from theory to applications. *IEEE Trans. Signal Process.*, **59**(9):4053-4085. http://dx.doi.org/10.1109/TSP.2011.2161982

Elad, M., 2010. Sparse and Redundant Representations: from Theory to Applications in Signal and Image Processing. Springer, Berlin. http://dx.doi.org/10.1007/978-1-4419-7011-4

Fan, J.Q., Li, R.Z., 2011. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.*, **96**(456):1348-1360. http://dx.doi.org/10.1198/016214501753382273

Fan, J.Q., Lv, J.C., Qi, L., 2011. Sparse high-dimensional models in economics. *Ann. Rev. Econ.*, **3**:291-317. http://dx.doi.org/10.1146/annurev-economics-061109-080451

Foucart, S., Lai, M.J., 2009. Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$. *Appl. Comput. Harmon. Anal.*, **26**(3):395-407. http://dx.doi.org/10.1016/j.acha.2008.09.001

Friedman, J., Hastie, T., Höfling, H., *et al.*, 2007. Pathwise coordinate optimization. *Ann. Appl. Stat.*, **1**(2):302-332. http://dx.doi.org/10.1214/07-Aoas131

Friedman, J., Hastie, T., Tibshirani, R., 2008. Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, **9**(3):432-441. http://dx.doi.org/10.1093/biostatistics/kxm045

Garber, D., Hazan, E., 2015. Faster rates for the Frank-Wolfe method over strongly-convex sets. Proc. ICML, p.541-549.

Gill, P.E., Murray, W., Saunders, M.A., 2008. User's Guide for SQOPT Version 7: Software for Large-Scale Linear and Quadratic Programming. http://www-leland.stanford.edu/group/SOL/guides/sqdoc7.pdf

Gong, P.H., Zhang, C.S., Lu, Z.S., *et al.*, 2013. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. Proc. ICML, p.37-45.

Grant, M., Boyd, S., 2013. CVX: Matlab Software for Disciplined Convex Programming. Version 2.0 Beta. http://cvxr.com/cvx/

Hazan, E., Agarwal, A., Kale, S., 2007. Logarithmic regret algorithms for online convex optimization. *Mach. Learn.*, **69**(2):169-192. http://dx.doi.org/10.1007/s10994-007-5016-8

Hoefling, H., 2010. A path algorithm for the fused Lasso signal approximator. *J. Comput. Graph. Stat.*, **19**(4):984-1006. http://dx.doi.org/10.1198/jcgs.2010.09208

Hong, M.Y., Razaviyayn, M., Luo, Z.Q., *et al.*, 2015. A unified algorithmic framework for block-structured optimization involving big data. arXiv:1511.02746.

Hu, T.C., Yu, J.H., 2016. Max-margin based Bayesian classifier. *Front. Inform. Technol. Electron. Eng.*, **17**(10): 973-981. http://dx.doi.org/10.1631/FITEE.1601078

Huang, J.Z., Zhang, T., Metaxas, D., 2011. Learning with structured sparsity. *J. Mach. Learn. Res.*, **12**:3371-3412.

Huang, T., Wu, B.L., Lizardi, P., *et al.*, 2005. Detection of DNA copy number alterations using penalized least squares regression. *Bioinformatics*, **21**(20):3811-3817. https://doi.org/10.1093/bioinformatics/bti646

Jacob, L., Obozinski, G., Vert, J.P., 2009. Group Lasso with overlap and graph Lasso. Proc. ICML, p.433-440. http://dx.doi.org/10.1145/1553374.1553431

Jaggi, M., 2013. Revisiting Frank-Wolfe: projection-free sparse convex optimization. Proc. ICML, p.427-435.

Jenatton, R., 2011. Structured Sparsity-Inducing Norms: Statistical and Algorithmic Properties with Applications to Neuroimaging. PhD Thesis, École Normale Supérieure de Cachan, Cachan, France.

Jenatton, R., Obozinski, G., Bach, F., 2009. Structured sparse principal component analysis. Proc. AISTATS, p.366-373.

Jenatton, R., Mairal, J., Bach, F.R., *et al.*, 2010. Proximal methods for sparse hierarchical dictionary learning. Proc. ICML, p.487-494.

Jenatton, R., Mairal, J., Obozinski, G., *et al.*, 2011. Proximal methods for hierarchical sparse coding. *J. Mach. Learn. Res.*, **12**:2297-2334.

Jenatton, R., Gramfort, A., Michel, V., *et al.*, 2012. Multiscale mining of fMRI data with hierarchical structured sparsity. *SIAM J. Imag. Sci.*, **5**(3):835-856. http://dx.doi.org/10.1137/110832380

John Lu, Z.Q., 2010. The elements of statistical learning: data mining, inference, and prediction. *J. R. Stat. Soc. A*, **173**(3):693-694. http://dx.doi.org/10.1111/j.1467-985X.2010.00646_6.x

Jones, B., West, M., 2005. Covariance decomposition in undirected Gaussian graphical models. *Biometrika*, **92**(4): 779-786. https://doi.org/10.1093/biomet/92.4.779

Karygianni, S., Frossard, P., 2014. Structured sparse coding for image denoising or pattern detection. Proc. ICASSP, p.3533-3537. http://dx.doi.org/10.1109/ICASSP.2014.6854258

Kim, B.S., Park, J.Y., Gilbert, A.C., *et al.*, 2013. Hierarchical classification of images by sparse approximation. *Image Vis. Comput.*, **31**(12):982-991. http://dx.doi.org/10.1016/j.imavis.2013.10.005

Kim, S., Xing, E.P., 2010. Tree-guided group Lasso for multi-task regression with structured sparsity. Proc. ICML, p.543-550.

Kim, S., Xing, E.P., 2012. Tree-guided group Lasso for multi-response regression with structured sparsity, with an application to eQTL mapping. *Ann. Appl. Stat.*, **6**(3):1095-1117. http://dx.doi.org/10.1214/12-Aoas549

Kim, S., Xing, E.P., 2014. Exploiting genome structure in association analysis. *J. Comput. Biol.*, **21**(4):345-360. http://dx.doi.org/10.1089/cmb.2009.0224

Kolar, M., Xing, E.P., 2011. On time varying undirected graphs. Proc. AISTATS, p.407-415.

Koren, Y., Bell, R., Volinsky, C., 2009. Matrix factorization techniques for recommender systems. *Computer*, **42**(8):30-37. http://dx.doi.org/10.1109/MC.2009.263

Lacoste-Julien, S., Schmidt, M., Bach, F., 2012. A simpler approach to obtaining an $O(1/t)$ convergence rate for the projected stochastic subgradient method. arXiv:1212.2002.

Lai, M.J., Xu, Y.Y., Yin, W.T., 2013. Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization. *SIAM J. Numer. Anal.*, **51**(2):927-957. http://dx.doi.org/10.1137/110840364

Lai, Z.Q., Lam, K.T., Wang, C.L., *et al.*, 2015. Latency-aware DVFS for efficient power state transitions on many-core architectures. *J. Supercomput.*, **71**(7):2720-2747. http://dx.doi.org/10.1007/s11227-015-1415-y

Lai, Z.Q., Lam, K.T., Wang, C.L., *et al.*, 2016. PoweRock: power modeling and flexible dynamic power management for many-core architectures. *IEEE Syst. J.*, in press. http://dx.doi.org/10.1109/JSYST.2015.2499307

Leng, C.L., Tang, C.Y., 2012. Sparse matrix graphical models. *J. Am. Stat. Assoc.*, **107**(499):1187-1200. http://dx.doi.org/10.1080/01621459.2012.706133

Li, X.X., Mo, L.L., Yuan, X.M., *et al.*, 2014. Linearized alternating direction method of multipliers for sparse group and fused Lasso models. *Comput. Stat. Data Anal.*, **79**:203-221. http://dx.doi.org/10.1016/j.csda.2014.05.017

Lin, H.Z., Mairal, J.L., Harchaoui, Z., 2015. A universal catalyst for first-order optimization. Proc. NIPS, p.3384-3392.

Liu, H., Palatucci, M., Zhang, J., 2009. Blockwise coordinate descent procedures for the multi-task Lasso, with applications to neural semantic basis discovery. Proc. ICML, p.649-656. http://dx.doi.org/10.1145/1553374.1553458

Liu, J., Ji, S., Ye, J., 2009. SLEP: Sparse Learning with Efficient Projections. http://www.public.asu.edu/~jye02/Software/SLEP

Ma, S.Q., Xue, L.Z., Zou, H., 2013. Alternating direction methods for latent variable Gaussian graphical model selection. *Neur. Comput.*, **25**(8):2172-2198. http://dx.doi.org/10.1162/NECO_a_00379

Mairal, J., 2013. Optimization with first-order surrogate functions. Proc. ICML, p.783-791.

Mairal, J., Bach, F., Ponce, J., *et al.*, 2011. SPAMS: SPArse Modeling Software. http://spams-devel.gforge.inria.fr/

Mairal, J., Bach, F., Ponce, J., 2014. Sparse modeling for image and vision processing. *Found. Trends Comput. Graph. Vis.*, **8**(2-3):85-283. http://dx.doi.org/10.1561/0600000058

Mallat, S., 2008. A Wavelet Tour of Signal Processing: the Sparse Way (3rd Ed.). Elsevier/Academic Press, Amsterdam.

McAuley, J., Ming, J., Stewart, D., *et al.*, 2005. Subband correlation and robust speech recognition. *IEEE Trans. Speech Audio Process.*, **13**(5):956-964. http://dx.doi.org/10.1109/TSA.2005.851952

Meier, L., van de Geer, S., Bühlmann, P., 2008. The group Lasso for logistic regression. *J. R. Stat. Soc. B*, **70**(1):53-71. http://dx.doi.org/10.1111/j.1467-9868.2007.00627.x

Meinshausen, N., Bühlmann, P., 2006. High-dimensional graphs and variable selection with the Lasso. *Ann. Stat.*, **34**(3):1436-1462. http://dx.doi.org/10.1214/009053606000000281

Meinshausen, N., Yu, B., 2008. Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Stat.*, **37**(1):246-270. http://dx.doi.org/10.1214/07-AOS582

Micchelli, C.A., Morales, J.M., Pontil, M., 2013. Regularizers for structured sparsity. *Adv. Comput. Math.*, **38**(3):455-489. http://dx.doi.org/10.1007/s10444-011-9245-9

Mosci, S., Rosasco, L., Santoro, M., *et al.*, 2010. Solving structured sparsity regularization with proximal methods. *LNCS*, **6322**:418-433. http://dx.doi.org/10.1007/978-3-642-15883-4_27

Mougeot, M., Picard, D., Tribouley, K., 2013. Grouping strategies and thresholding for high dimensional linear models. *J. Stat. Plan. Infer.*, **143**(9):1417-1438. http://dx.doi.org/10.1016/j.jspi.2013.03.001

Najafian, M., 2016. Acoustic Model Selection for Recognition of Regional Accented Speech. PhD Thesis, University of Birmingham, Birmingham, UK.

Negahban, S.N., Ravikumar, P., Wainwright, M.J., *et al.*, 2012. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Stat. Sci.*, **27**(4):538-557. http://dx.doi.org/10.1214/12-Sts400

Nemirovski, A., 2004. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM J. Optim.*, **15**(1):229-251. http://dx.doi.org/10.1137/S1052623403425629

Nesterov, Y., 2004. Introductory Lectures on Convex Optimization: a Basic Course. Springer Science & Business Media. http://dx.doi.org/10.1007/978-1-4419-8853-9

Nesterov, Y., 2009. Primal-dual subgradient methods for convex problems. *Math. Program.*, **120**(1):221-259. http://dx.doi.org/10.1007/s10107-007-0149-x

Parikh, N., Boyd, S., 2014. Proximal algorithms. *Found. Trends Optim.*, **1**(3):127-239. http://dx.doi.org/10.1561/2400000003

Peng, Z.M., Wu, T.Y., Xu, Y.Y., *et al.*, 2016. Coordinate friendly structures, algorithms and applications. arXiv:1601.00863.

Qiao, L.B., Lin, T.Y., Jiang, Y.G., *et al.*, 2016a. On stochastic primal-dual hybrid gradient approach for compositely regularized minimization. Proc. European Conf. on Artificial Intelligence, p.167-174. http://dx.doi.org/10.3233/978-1-61499-672-9-167

Qiao, L.B., Zhang, B.F., Su, J.S., *et al.*, 2016b. Linearized alternating direction method of multipliers for constrained nonconvex regularized optimization. Proc. Asian Conf. on Machine Learning, p.97-109.

Qiao, L.B., Zhang, B.F., Zhuang, L., *et al.*, 2016c. An efficient algorithm for tensor principal component analysis via proximal linearized alternating direction method of multipliers. Proc. Int. Conf. on Advanced Cloud and Big Data, p.283-288. http://dx.doi.org/10.1109/CBD.2016.056

Rakotomamonjy, A., 2011. Surveying and comparing simultaneous sparse approximation (or group-Lasso) algorithms. *Signal Process.*, **91**(7):1505-1526. http://dx.doi.org/10.1016/j.sigpro.2011.01.012

Rasmussen, C.E., Ghahramani, Z., 2001. Occam's razor. Proc. NIPS, p.294-300.

Rendle, S., Schmidt-Thieme, L., 2010. Pairwise interaction tensor factorization for personalized tag recommendation. Proc. 3rd ACM Int. Conf. on Web Wearch and Data Mining, p.81-90. http://dx.doi.org/10.1145/1718487.1718498

Roth, V., Fischer, B., 2008. The group-Lasso for generalized linear models: uniqueness of solutions and efficient algorithms. Proc. ICML, p.848-855. http://dx.doi.org/10.1145/1390156.1390263

Rudin, L.I., Osher, S., Fatemi, E., 1992. Nonlinear total variation based noise removal algorithms. *Phys. D*, **60**(1-4):259-268. http://dx.doi.org/10.1016/0167-2789(92)90242-F

Scheinberg, K., Ma, S., Goldfarb, D., 2010. Sparse inverse covariance selection via alternating linearization methods. Proc. NIPS, p.2101-2109.

Selesnick, I.W., Bayram, I., 2014. Sparse signal estimation by maximally sparse convex optimization. *IEEE Trans. Signal Process.*, **62**(5):1078-1092. http://dx.doi.org/10.1109/TSP.2014.2298839

Simon, N., Friedman, J., Hastie, T., *et al.*, 2013. A sparse-group Lasso. *J. Comput. Graph. Stat.*, **22**(2):231-245. http://dx.doi.org/10.1080/10618600.2012.681250

Su, W.J., Boyd, S., Candès, E., 2014. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. Proc. NIPS, p.2510-2518.

Sun, Y.P., Chen, S.H., Han, B., *et al.*, 2015a. A novel location privacy mining threat in vehicular Internet access service. *LNCS*, **9204**:467-476. http://dx.doi.org/10.1007/978-3-319-21837-3_46

Sun, Y.P., Zhang, B.F., Zhao, B.K., *et al.*, 2015b. Mix-zones optimal deployment for protecting location privacy in VANET. *Peer-to-Peer Netw. Appl.*, **8**(6):1108-1121. http://dx.doi.org/10.1007/s12083-014-0269-z

Suzuki, T.J., 2013. Dual averaging and proximal gradient descent for online alternating direction multiplier method. Proc. ICML, p.392-400.

Takacs, G., Pilaszy, I., Nemeth, B., *et al.*, 2009. Scalable collaborative filtering approaches for large recommender systems. *J. Mach. Learn. Res.*, **10**:623-656.

Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. B*, **58**(1):267-288.

Tibshirani, R., Wang, P., 2008. Spatial smoothing and hot spot detection for CGH data using the fused Lasso. *Biostatistics*, **9**(1):18-29. http://dx.doi.org/10.1093/biostatistics/kxm013

Tibshirani, R., Saunders, M., Rosset, S., *et al.*, 2005. Sparsity and smoothness via the fused Lasso. *J. R. Stat. Soc. B*, **67**(1):91-108. http://dx.doi.org/10.1111/j.1467-9868.2005.00490.x

Toh, K., Todd, M.J., Tütüncü, R.H., 2006. SDPT3 Version 4.0: a Matlab Software for Semidefinite-Quadratic-Linear Programming. http://www.math.nus.edu.sg/~mattohkc/sdpt3.html

Tropp, J.A., 2004. Greed is good: algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, **50**(10):2231-2242. http://dx.doi.org/10.1109/Tit.2004.834793

Tropp, J.A., Gilbert, A.C., Muthukrishnan, S., *et al.*, 2003. Improved sparse approximation over quasi-incoherent dictionaries. Proc. Int. Conf. on Image Processing, p.37-40. http://dx.doi.org/10.1109/ICIP.2003.1246892

Tseng, P., 2008. On Accelerated Proximal Gradient Methods for Convex-Concave Optimization. http://www.mit.edu/~dimitrib/PTseng/papers/apgm.pdf

Tseng, P., Yun, S., 2009. A coordinate gradient descent method for nonsmooth separable minimization. *Math. Program.*, **117**(1):387-423. http://dx.doi.org/10.1007/s10107-007-0170-0

van den Berg, E., Friedlander, M.P., 2007. SPGL1: a Solver for Large-Scale Sparse Reconstruction. http://www.cs.ubc.ca/labs/scl/spgl1

Villa, S., Rosasco, L., Mosci, S., *et al.*, 2014. Proximal methods for the latent group Lasso penalty. *Compt. Optim. Appl.*, **58**(2):381-407. http://dx.doi.org/10.1007/s10589-013-9628-6

Vincent, M., Hansen, N.R., 2014. Sparse group Lasso and high dimensional multinomial classification. *Comput. Stat. Data Anal.*, **71**:771-786. http://dx.doi.org/10.1016/j.csda.2013.06.004

Wainwright, M.J., Jordan, M.I., 2008. Graphical models, exponential families, and variational inference. *Found. Trend. Mach. Learn.*, **1**(1-2):1-305. http://dx.doi.org/10.1561/2200000001

Wang, H.S., Leng, C.L., 2008. A note on adaptive group Lasso. *Comput. Stat. Data Anal.*, **52**(12):5277-5286. http://dx.doi.org/10.1016/j.csda.2008.05.006

Wang, L.C., You, Y., Lian, H., 2013. A simple and efficient algorithm for fused Lasso signal approximator with convex loss function. *Comput. Stat.*, **28**(4):1699-1714. http://dx.doi.org/10.1007/s00180-012-0373-6

Wang, Y., Wang, J.J., Xu, Z.B., 2013. On recovery of block-sparse signals via mixed ℓ_2/ℓ_q $(0 < q \leq 1)$ norm minimization. *EURASIP J. Adv. Signal Process.*, **2013**: 1-17. http://dx.doi.org/10.1186/1687-6180-2013-76

Wen, Z., Goldfarb, D., Scheinberg, K., 2012. Block coordinate descent methods for semidefinite programming. *In*: Anjos, M.F., Lasserre, J.B. (Eds.), Handbook on Semidefinite, Conic and Polynomial Optimization. Springer US, Boston, p.533-564. http://dx.doi.org/10.1007/978-1-4614-0769-0_19

Wermuth, N., 1976. Analogies between multiplicative models for contingency tables and covariance selection. *Biometrics*, **32**:95-108.

Wille, A., Bühlmann, P., 2006. Low-order conditional independence graphs for inferring genetic networks. *Stat. Appl. Genet. Mol. Biol.*, **5**(1). http://dx.doi.org/10.2202/1544-6115.1170

Wrinch, D., Jeffreys, H., 1921. On certain fundamental principles of scientific inquiry. *Phil. Mag.*, **42**(249):369-390. http://dx.doi.org/10.1080/14786442108633773

Wu, Y.L., Lu, X.C., Su, J.S., *et al.*, 2016. An efficient searchable encryption against keyword guessing attacks for sharable electronic medical records in cloud-based system. *J. Med. Syst.*, **40**:258. http://dx.doi.org/10.1007/s10916-016-0609-z

Xiao, J.J., Qiao, L.B., Stolkin, R., *et al.*, 2016. Distractor-supported single target tracking in extremely cluttered scenes. *LNCS*, **9908**:121-136. http://dx.doi.org/10.1007/978-3-319-46493-0_8

Xiao, L., Zhang, T., 2014. A proximal stochastic gradient method with progressive variance reduction. *SIAM J. Optim.*, **24**(4):2057-2075. http://dx.doi.org/10.1137/140961791

Xie, H., Tong, R.F., 2016. Image meshing via hierarchical optimization. *Front. Inform. Technol. Electron. Eng.*, **17**(1):32-40. http://dx.doi.org/10.1631/FITEE.1500171

Xie, Y.C., Huang, H., Hu, Y., *et al.*, 2016. Applications of advanced control methods in spacecrafts: progress, challenges, and future prospects. *Front. Inform. Technol. Electron. Eng.*, **17**(9):841-861. http://dx.doi.org/10.1631/FITEE.1601063

Xie, Z.X., Xu, Y., 2014. Sparse group Lasso based uncertain feature selection. *Int. J. Mach. Learn. Cybern.*, **5**(2):201-210. http://dx.doi.org/10.1007/s13042-013-0156-6

Xu, X., Zhang, B.F., Zhong, Q.X., 2005. Text categorization using SVMs with Rocchio ensemble for Internet information classification. *LNCS*, **3619**:1022-1031. http://dx.doi.org/10.1007/11534310_107

Xu, X., Hu, D.W., Lu, X.C., 2007. Kernel-based least squares policy iteration for reinforcement learning. *IEEE Trans. Neur. Netw.*, **18**(4):973-992. http://dx.doi.org/10.1109/tnn.2007.899161

Xu, X., Liu, C.M., Yang, S.X., *et al.*, 2011. Hierarchical approximate policy iteration with binary-tree state space decomposition. *IEEE Trans. Neur. Netw.*, **22**(12):1863-1877. http://dx.doi.org/10.1109/tnn.2011.2168422

Xu, Z., Chang, X., Xu, F., *et al.*, 2012. $L_{1/2}$ regularization: a thresholding representation theory and a fast solver. *IEEE Trans. Neur. Netw. Learn. Syst.*, **23**(7):1013-1027. http://dx.doi.org/10.1109/TNNLS.2012.2197412

Yang, J.F., Yuan, X.M., 2013. Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization. *Math. Comput.*, **82**:301-329. http://dx.doi.org/10.1090/S0025-5718-2012-02598-1

Yang, X.J., Liao, X.K., Xu, W.X., *et al.*, 2010. Th-1: China's first petaflop supercomputer. *Front. Comput. Sci. China*, **4**(4):445-455. http://dx.doi.org/10.1007/s11704-010-0383-x

Yang, X.J., Liao, X.K., Lu, K., *et al.*, 2011. The TianHe-1A supercomputer: its hardware and software. *J. Comput. Sci. Technol.*, **26**(3):344-351. http://dx.doi.org/10.1007/s11390-011-1137-4

Ye, G.B., Xie, X.H., 2011. Split Bregman method for large scale fused Lasso. *Comput. Stat. Data Anal.*, **55**(4):1552-1569. http://dx.doi.org/10.1016/j.csda.2010.10.021

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. B*, **68**(1):49-67. http://dx.doi.org/10.1111/j.1467-9868.2005.00532.x

Yuan, M., Lin, Y., 2007. Model selection and estimation in the Gaussian graphical model. *Biometrika*, **94**(1):19-35. http://dx.doi.org/10.1093/biomet/asm018

Yuan, M., Yang, B.X., Ma, Y.D., *et al.*, 2015. Multi-scale UDCT dictionary learning based highly undersampled MR image reconstruction using patch-based constraint splitting augmented Lagrangian shrinkage algorithm. *Front. Inform. Technol. Electron. Eng.*, **16**(12):1069-1087. http://dx.doi.org/10.1631/FITEE.1400423

Zhang, B.F., Su, J.S., Xu, X., 2006. A class-incremental learning method for multi-class support vector machines in text classification. Proc. ICMLC, p.2581-2585. http://dx.doi.org/10.1109/ICMLC.2006.258853

Zhang, C.H., 2010. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.*, **38**(2):894-942. http://dx.doi.org/10.1214/09-AOS729

Zhang, S.Z., Wang, K., Chen, B.L., *et al.*, 2011. A new framework for co-clustering of gene expression data. *LNCS*, **7036**:1-12. http://dx.doi.org/10.1007/978-3-642-24855-9_1

Zhang, T., 2009. Some sharp performance bounds for least squares regression with L_1 regularization. *Ann. Stat.*, **37**(5A):2109-2144. http://dx.doi.org/10.1214/08-AOS659

Zhang, T., 2010. Analysis of multi-stage convex relaxation for sparse regularization. *J. Mach. Learn. Res.*, **11**:1081-1107.

Zhang, T., 2013. Multi-stage convex relaxation for feature selection. *Bernoulli*, **19**(5B):2277-2293. http://dx.doi.org/10.3150/12-BEJ452

Zhang, T.Z., Ghanem, B., Liu, S., *et al.*, 2012. Robust visual tracking via multi-task sparse learning. Proc. CVPR, p.2042-2049. http://dx.doi.org/10.1109/CVPR.2012.6247908

Zhang, T.Z., Ghanem, B., Liu, S., *et al.*, 2013. Robust visual tracking via structured multi-task sparse learning. *Int. J. Comput. Vis.*, **101**(2):367-383. http://dx.doi.org/10.1007/s11263-012-0582-z

Zhang, T.Z., Jia, K., Xu, C.S., *et al.*, 2014. Partial occlusion handling for visual tracking via robust part matching. Proc. CVPR, p.1258-1265. http://dx.doi.org/10.1109/CVPR.2014.164

Zhang, T.Z., Liu, S., Ahuja, N., *et al.*, 2015a. Robust visual tracking via consistent low-rank sparse learning. *Int. J. Comput. Vis.*, **111**(2):171-190. http://dx.doi.org/10.1007/s11263-014-0738-0

Zhang, T.Z., Liu, S., Xu, C.S., *et al.*, 2015b. Structural sparse tracking. Proc. CVPR, p.150-158. http://dx.doi.org/10.1109/CVPR.2015.7298610

Zhang, Y., Yang, J., Yin, W., 2011. YALL1: Your Algorithms for L1. http://yall1.blogs.rice.edu

Zhang, Z.K., Zhou, T., Zhang, Y.C., 2011. Tag-aware recommender systems: a state-of-the-art survey. *J. Comput. Sci. Technol.*, **26**:767-777. http://dx.doi.org/10.1007/s11390-011-0176-1

Zhao, P., Yu, B., 2006. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, **7**:2541-2563.

Zhao, P., Yu, B., 2007. Stagewise Lasso. *J. Mach. Learn. Res.*, **8**:2701-2726.

Zhao, P., Rocha, G., Yu, B., 2009. The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Stat.*, **37**(6a):3468-3497. http://dx.doi.org/10.1214/07-Aos584

Zhu, Y.T., Zhao, Y.B., Liu, J., *et al.*, 2016. Low complexity robust adaptive beamforming for general-rank signal model with positive semidefinite constraint. *Front. Inform. Technol. Electron. Eng.*, **17**(11):1245-1252. http://dx.doi.org/10.1631/FITEE.1601112