Int J Multimed Info Retr (2012) 1:3–15
DOI 10.1007/s13735-012-0001-9

INVITED PAPER

# The heterogeneous feature selection with structural sparsity for multimedia annotation and hashing: a survey

**Fei Wu · Yahong Han · Xiang Liu · Jian Shao ·
Yueting Zhuang · Zhongfei Zhang**

Received: 14 December 2011 / Accepted: 8 January 2012 / Published online: 2 March 2012

**Abstract** There is a rapid growth of the amount of multimedia data from real-world multimedia sharing web sites, such as Flickr and Youtube. These data are usually of high dimensionality, high order, and large scale. Moreover, different types of media data are interrelated everywhere in a complicated and extensive way by *context prior*. It is well known that we can obtain lots of features from multimedia such as images and videos; those high-dimensional features often describe various aspects of characteristics in multimedia. However, the obtained features are often over-complete to describe certain semantics. Therefore, the selection of limited discriminative features for certain semantics is hence crucial to make the understanding of multimedia more interpretable. Furthermore, the effective utilization of intrinsic embedding structures in various features can boost the performance of multimedia retrieval. As a result, the appropriate representation of the latent information hidden in the related features is hence crucial during multimedia understanding. This paper introduces many of the recent efforts in sparsity-based heterogenous feature selection, the representation of the intrinsic latent structure embedded in multimedia, and the related hashing index techniques.

**Keywords** Structural sparsity · Factor decomposition · Latent structure · Hashing index

## 1 Introduction

Natural images and videos can be well approximated by a small subset of elements from an *over-complete* dictionary. The process of choosing a good subset of dictionary elements along with the corresponding coefficients to represent a signal is known as sparse representation [10]. As pointed out in [56], the receptive files of simple cells in mammalian primary visual cortex can be characterized as being spatially localized, oriented, and bandpass (selective to structure at different spatial scales). Therefore, a learning algorithm is crucial to find sparse linear codes for natural scenes. The problem of finding a sparse representation for the data has become an interesting topic recently in computer vision and multimedia retrieval nowadays. The essential challenge to be resolved in sparse representation is to develop an efficient approach with which each original data element could be reconstructed from its corresponding sparse representation.

In this paper, we focus on data mainly on images and videos. The feature selection and hashing of multimedia are the basis for image and video annotation and retrieval. The robust and appropriate techniques for feature selection and hashing can significantly improve the performance of image/video understanding, retrieval, tracking, matching, reconstruction, etc.

This work was done when Z. Zhang was on leave from SUNY Binghamton, USA.

F. Wu · Y. Han · J. Shao · Y. Zhuang
College of Computer Science, Zhejiang University, Hangzhou, China
e-mail: wufei@cs.zju.edu.cn

Y. Han
e-mail: yahong@zju.edu.cn

J. Shao
e-mail: jshao@cs.zju.edu.cn

Y. Zhuang
e-mail: yzhuang@zju.edu.cn

X. Liu · Z. Zhang (✉)
Department of Information Science and Electronic Engineering,
Zhejiang University, Hangzhou, China
e-mail: zhongfei@zju.edu.cn

X. Liu
e-mail: hn_lxa@126.com

It is well known that we can extract high-dimensional features from one given image or video in the real world in different types. These different features can be roughly classified as *Local* (e.g., SIFT, Shape Context and GLOH) versus *Global* (e.g., color, shape and texture) [49], *Dense* (e.g., bag of visual words [24]) versus *Sparse* (e.g., Locality-constrained Linear Coding [73]), *Shadow* versus *Deep* (e.g., Hierarchical Models [58]), and *Multi-scale* (e.g., Spatial Pyramid Matching [40]), *Still* versus *Motion* (e.g., Optical Flow [29]), *Compressed* (e.g., Gabor wavelets [42]) versus and *Uncompressed*. We call these different types of features extracted in the same image or video the heterogeneous features, and the features of the same type the homogeneous features.

Different subsets of heterogenous features have different intrinsic discriminative power to characterize the semantics in multimedia. That is to say, only limited groups of heterogenous features distinguish certain semantics from others. Therefore, the selected visual features for further multimedia processing are usually sparse.

Given high-dimensional heterogeneous features in images and videos, in order to obtain the discriminative features, we often map original features into a subspace to discover their intrinsic structure by dimension reduction such as principal component analysis (PCA), Locally Linear Embedding (LLE), ISOMAP, Laplacian Eigenmap, Local Tangent Space Alignment (LTSA) and Locality Preserving Projections (LPP) [61]. However, it is very hard to discern what original features play an essential role during the semantic understanding in the embedded subspace after the dimension reduction is conducted. As a result, a more *interpretable* approach is necessary for feature selection. That is to say, given the number of extracted over-complete heterogenous features, it is essential to identify the discriminate features for certain semantics.

Motivated by the recent advance in compressed sensing, *sparsity*-based feature selection approaches are developed in computer vision and multimedia retrieval [25,46,48,77,82]. The basic idea of sparsity-based feature selection is to impose a (structural) sparse penalty to select discriminative features. For example, Wright et al. [76] casts the face recognition problem as a liner regression problem with sparse constraints for regression coefficients. To solve the regression problem, Wright et al. [76] reformulate face recognition as an $\ell_1$-norm problem. Cao et al. [11] propose learning different metric kernel functions for different heterogeneous features for image classification. After the introduction of the $\ell_1$-norm at the group level into sparse logistic regression, a heterogeneous feature machines (HFM) is implemented in [11].

For all the above approaches, the $\ell_1$-norm (namely *lasso*, least absolution shrinkage and selection operator) [71] is effectively implemented to make the learning model both sparse and interpretable. However, for the group of features in which the pairwise correlations among them are very high, *lasso* tends to select only one of the pairwise correlated features and cannot induce the group effect. In the "large $p$, small $n$" problem, the "grouped features" situation is an important concern to facilitate a model's interpretability. In order to remedy the deficiency of *lasso*, group *lasso* [87] and elastic net [93] are proposed, respectively. If the structural *priors* embedded in images and videos are appropriately represented, the performance of semantic understanding for images and videos can be boosted. For example, since the extract high-dimensional heterogenous features from images and videos can be naturally divided into disjoint groups of homogeneous features, a *structural grouping sparsity* penalty is proposed in [77] to induce a (*structural*) sparse selection model for the identification of subgroups of homogenous features during image annotation. The motivation in [77] can be illustrated in Fig. 1. After groups of heterogenous features such as color, texture, and shape are extracted from images, the structural grouping sparsity is conducted to set the coefficients ($\boldsymbol{\beta}_i$) of the discriminative feature sets as 1 and the coefficients of other insignificant feature sets as 0. Moreover, the identified subgroup within each selected feature set is further used as the representation of each image. Due to the importance of the introduction of the structural priors into feature selection, Jenatton

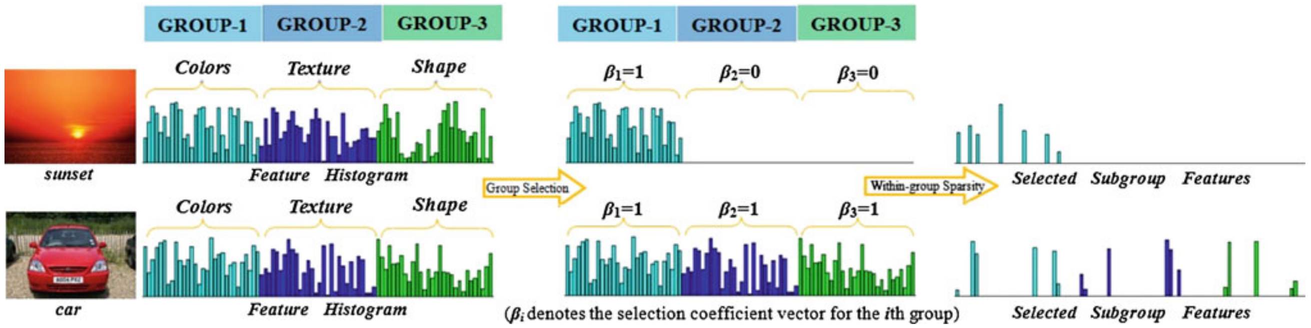

**Fig. 1** The illustration of the high-dimensional heterogeneous feature selection with structural grouping sparsity revised from [77]

et al. recently propose a general definition of the *structured sparsity-inducing norm* in [31,32] to incorporate the prior knowledge or structural constraints to find the suitable linear features. Under the setting of the structured sparsity-inducing norm, *lasso*, group *lasso*, and even the tree-guided group lasso [37] are, respectively, its special cases.

Note that the introduction of the sparsity penalty into the traditional matrix factorization can help achieve a good performance. For example, Kim and Park [38] propose a novel algorithm of sparse NMF to control the degree of sparseness in the nonnegative basis matrix or the nonnegative coefficient matrix. The empirical study shows that the performance can be improved if we impose the sparsity on a factor of NMF by the $\ell_1$-norm minimization into the objective function. Sparse topical coding (STC) is proposed in [92] to discover latent representations of large collections of data by a nonprobabilistic formulation of the topic models. STC can directly control the sparsity of the inferred representations by the conduction of sparsity-inducing regularizers. A hierarchical Bayesian model is developed in [43] to integrate the dictionary learning, sparse coding, and topic modeling for the joint analysis of multiple images and (when present) the associated annotations.

After the discriminative features are selected, we need to represent the intrinsic structures embedded in the heterogenous features. Traditionally, the high-dimensional heterogenous features in images and videos are preferred to being represented merely as concatenated vectors, whose high dimensionality always causes the problem of curse of dimensionality. Besides, as reported in [85], the over-compression problem occurs when the sample vector is very long and the number of training samples is small, which results in a loss of information in the dimension reduction process. At present, many of the representation approaches are proposed such as matrix, tensor, and graph. Tensor is a natural generalization of a vector or a matrix, and has been applied to computer vision, signal processing, and information retrieval [28,45,69]. The tensor algebra defines multilinear operators over a set of vector spaces and captures the high-order information in heterogeneous features. Usually, the traditional graph only models the homogeneous similarity and therefore ignores the high-order relations that are inherent in images and videos. In order to address this drawback of the traditional graph, hypergraph is proposed to represent more complex correlations in images and videos. Hypergraph [5] is a graph in which one edge can connect more than two vertices. This characteristic enables hypergraphs to represent complex and higher-order relations which are difficult to be represented in the traditional undirected or directed graphs. Recently, hypergraphs have been successfully applied to image annotation, image ranking, and music recommendation, and have received considerable attention. For example, spectral clustering is generalized from undirected graphs to hypergraphs in [91], where hypergraph embedding and transductive classification are further developed by spectral hypergraph clustering. Hypergraph spectral learning is utilized in [68] for multi-label classification, where a hypergraph is constructed to exploit the correlation information among different labels. In many real-world applications, the complex spatial–temporal or context in images and videos can be efficiently encoded by a matrix, a tensor, or a graph and then information is lost if the vector representation is used. The interesting issue is whether we can introduce a sparsity penalty into a matrix, a tensor, or a hypergraph to make the representation and learning interpretable. If there is a low-rank structure in a matrix, the penalty of the matrix rank is a good choice to enforce such sparsity. However, a matrix rank is neither continuous nor convex. As a surrogate convex of the nonconvex matrix rank function, the matrix nuclear norm (trace norm, matrix-*lasso*) is specifically employed to encourage the low-rank property. Nuclear norm is defined as the sum of all the singular values as a convex function. The idea of a low-rank matrix is an extension from the concept of “sparse vector” to that of “sparse matrix”. Robust principal component analysis (R-PCA) is proposed in [75] to recover low-rank matrices from corrupted observations by the implementation of the nuclear norm minimization for the low-rank recovery and $\ell_1$-minimization for the error correction. An accelerated R-PCA approach is proposed in [52] for a large-scale image tag transduction under the setting of the nuclear norm. One $\ell_1$-graph is constructed by encoding the overall behavior of the data set in sparse representations in [14].

How to construct an approximate index structure for images and videos with the selected features is essential to the efficient retrieval of a large scale of multimedia. A naive solution to accurately find the relevantly similar examples to a query is to search over all the samples in a database and sort them according to their similarities to the query. However, this becomes prohibitively expensive when the scale of the database is very large. To reduce the complexity of finding the relevant samples for a query, indexing techniques are necessarily required to organize images and videos. However, studies reveal that many of the index structures have an exponential dependency (in space or time or both) upon the number of the dimensions and even a simple brute-force, linear-scan approach may be more efficient than an index-based search in high-dimensional settings [4]. Moreover, an excellent index structure should guarantee that the similarity of two samples in the index space keeps consistent with their similarity in the original data space [59]. Recently, locality-sensitive hashing (LSH) and its variations have been proposed as the indexing approaches for an approximate nearest neighbor search [17,47]. The basic idea in LSH is to use a family of locality preserving hash functions to hash similar data in the high-dimensional space into the same bucket with a higher probability than these for the nonsimilar data.

As shown by semantic hashing [60], LSH could be unstable and lead to an extremely bad result due to its randomized approximate similarity search. Unlike those approaches which randomly project the input data into an embedding space such as LSH, several machine learning approaches are recently developed to generate more compact and approximate binary codewords for data indexing, such as restricted Boltzmann machine (RBM) in semantic hashing [60], parameter sensitive hashing (PSH) in pose estimation [62] and spectral hashing [74]. These approaches attempt to elaborate appropriate hash functions to optimize an underlying hashing objective. Shao et al. [63] introduces the sparse principal component analysis (sparse PCA) and the boosting similarity sensitive hashing (Boosting SSC) into the traditional spectral hashing and calls this approach sparse spectral hashing (SSH).

## 2 Sparsity-based feature selection

### 2.1 Notation and problem formulation

Assume that we have a training set of $n$ labeled samples such as images and videos with $J$ labels (tags) and that the $p$-dimensional heterogenous features can be extracted from each image or video: $\{(\mathbf{x}_i, \mathbf{y}_i) \in \mathbb{R}^p \times \{0,1\}^J : i = 1, 2, \ldots, n\}$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}} \in \mathbb{R}^p$ represents the $p$-dimension feature vector for the $i$th image or video, $p$ represents the dimensionality of features, $\mathbf{y}_i = (y_{i1}, \ldots, y_{iJ})^{\mathrm{T}} \in \{0,1\}^J$ is the corresponding label vector, $y_{ij} = 1$ if the $i$th sample has the $j$th label and $y_{ij} = 0$ otherwise. Unlike the traditional multi-class problem where each sample only belongs to a single category: $\sum_{j=1}^{J} y_{ij} = 1$, in multi-label setting, we relax the constraint to $\sum_{j=1}^{J} y_{ij} \geq 0$. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}}$ be the $n \times p$ training data matrix, and $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^{\mathrm{T}}$ the corresponding $n \times J$ label indicator matrix.

Suppose that the extracted $p$ dimensional heterogenous features are divided into $L$ disjoint groups of homogeneous features, with $p_l$ the number of features in the $l$th group, i.e., $\sum_l^L p_l = p$. For ease of notation, we use a matrix $\mathbf{X}_l \in \mathbb{R}^{n \times p_l}$ to represent the features of the training data corresponding to the $l$th group, with corresponding coefficient vector $\beta_{jl} \in \mathbb{R}^{p_l} (l = 1, 2, \ldots, L)$ for the $j$th label. Let $\boldsymbol{\beta}_j = (\beta_{j1}^{\mathrm{T}}, \ldots, \beta_{jL}^{\mathrm{T}})^{\mathrm{T}}$ be the entire coefficient vector for the $j$th label; we have

$$\mathbf{X}\boldsymbol{\beta}_j = \sum_{l=1}^{L} \mathbf{X}_l \beta_{jl} \tag{1}$$

In the following, we assume that the label indicator matrix $\mathbf{Y}$ is centered and that the feature matrix $\mathbf{X}$ is centered and standardized, namely $\sum_{i=1}^{n} y_{ij} = 0$, $\sum_{i=1}^{n} x_{id} = 0$, and $\sum_{i=1}^{n} x_{id}^2 = 1$, for $j = 1, 2, \ldots, J$ and $d = 1, 2, \ldots, p$. Moreover, we let $||\beta_{jl}||_2^2$ and $||\beta_{jl}||_1$ denote the $\ell_2$-norm and the $\ell_1$-norm of vector $\beta_{jl}$, respectively.

Denote $\hat{\beta}(\delta)$ the estimated coefficients obtained by a fitting procedure $\delta$. That is to say, for the $j$th label, we tend to train a regression model $\hat{\beta}_j(\delta)$ with a penalty term as follows to select its corresponding discriminative features:

$$\min_{\hat{\boldsymbol{\beta}}_j} ||\mathbf{Y}_{(:,j)} - \sum_{l=1}^{L} \mathbf{X}_l \hat{\beta}_{jl}||_2^2 + \boldsymbol{\lambda} P(\hat{\boldsymbol{\beta}}_j) \tag{2}$$

where $\mathbf{Y}_{(:,j)} \in (0,1)^{(n \times 1)}$ is the $j$th column of indicator matrix $\mathbf{Y}$ and encodes the label information for the $j$th label, $P(\hat{\boldsymbol{\beta}}_j)$ is the regularizer which imposes structural priors to the high-dimensional features. The trained regression model combines a loss function (measuring the goodness of fit of the model to the data) with a regularized penalty (encouraging the assumed grouping structure). For example, the ridge regression uses the $\ell_2$-norm to avoid overfitting and *lasso* produces sparsity on $\hat{\boldsymbol{\beta}}_j$ by the $\ell_1$-norm. If the estimated coefficients in $\hat{\beta}_{jl}$ for $j$th label are not zero, this means that the $l$th homogeneous features are all selected to make the $j$th label discernible. Simultaneously, homogeneous features may be dropped out for the representation of $j$th label due to their irrelevance. Therefore, we can set up an interpretable model for feature selection.

The solution to $\hat{\beta}_j(\delta)$ can identify all of the discriminative features for each $j$th label; however, the individual conduction of $\hat{\beta}_j(\delta)$ ignores the correlations between labels in the setting of images and videos with multiple labels. The effective utilization of the latent information hidden in the related labels somehow boosts the performance of multi-label annotation. For example, a multiple response regression model, called curds and whey (C&W) is proposed in [9]. Curds and whey sets up the connection between multiple response regressions and canonical correlations. Therefore, the C&W method can be used to boost the performance of multi-label prediction given the prediction results from the regressions of individual labels [77]. Multi-task feature selection (or multi-task feature learning) is an alternative to utilizing the label correlation during feature selection. Argyriou et al. [1] and Obozinski et al. [55] use the $\ell_{1,2}$-norm to regularize the heterogeneous features of different tasks and therefore encourage multiple features to have similar sparsity patterns across tasks (tags).

### 2.2 Lasso and nonnegative garotte

In statistical community, *lasso* [71] is a shrinkage and variable selection method for linear regression, which is a penalized least square method imposing an $\ell_1$-norm penalty to the regression coefficients. Due to the nature of the $\ell_1$-norm penalty, *lasso* continuously shrinks the coefficients toward zero, and achieves its prediction accuracy via the bias–variance trade-off. In signal processing, *lasso* always produces a sparse representation that selects the subset compactly expressing the input signal. In the literature, the *lasso*-based sparse representation methods have been successfully used to solve problems such as face recognition [76] and image classification [57].

In order to select the most discriminative features for the annotation of images by the $j$th tag, *lasso* is defined to train a regression model $\hat{\beta}_j(\delta)$ on the training set of images $\mathbf{X}$ by a $\ell_1$-norm:

$$\min_{\hat{\boldsymbol{\beta}}_j} ||\mathbf{Y}_{(:,j)} - \mathbf{X}\hat{\boldsymbol{\beta}}_j||_2^2 + \lambda||\hat{\boldsymbol{\beta}}_j||_1 \tag{3}$$

where $\lambda > 0$ is the regularized parameter. Due to the nature of the $\ell_1$-norm penalty, by solving (3), most coefficients in the estimated $\hat{\boldsymbol{\beta}}_j$ are shrinked to zero, which could be used to select the discriminative features. It is clear that (3) is an unconstrained convex optimization problem. Many algorithms have been proposed to solve problem (3), such as the quadratic programming methods [71], least angle regression [19] and Gauss-Seidel [65].

It has been shown that the nonnegative matrix factorization (NMF) [41] can learn part-based representation. The nonnegativity constraint makes the representation easy to interpret due to purely additive combinations of nonnegative basis vectors. The model of nonnegative garrote [7] is proposed to solve the following optimization problem

$$\min_{\hat{\boldsymbol{\beta}}_j} ||\mathbf{Y}_{(:,j)} - \mathbf{X}\hat{\boldsymbol{\beta}}_j||_2^2 + \lambda \sum_{l=1}^{p} \hat{\beta}_{jl},$$
$$\text{s.t. } \hat{\beta}_{jl} \geq 0, \quad \forall l \tag{4}$$

where $\lambda > 0$ is the regularized parameter. The nonnegative garrote can be efficiently solved by the classical numerical methods such as the least angle regression (LARS) [19]. Breiman's original implementation [7] to solve (4) is to shrink each ordinary least squares (OLS) estimated coefficient by a nonnegative amount whose sum is subject to an upper bound constraint (the garrote). In the extensive simulation studies, Breiman has shown that the garotte is superior to subset selection and is competitive with ridge regression. Although the motivation of *lasso* comes from the garotte, in overfitting or highly correlated settings, the performance of the garotte deteriorates same as the OLS. In contrast, *lasso* avoids the explicit use of OLS estimates [71].

As mentioned before, Wright et al. [76] introduce $\ell_1$-norm into face recognition and formulates the face recognition as a liner regression with sparse constraints for regression coefficients. However, *lasso* makes the representation unnecessarily additive. This might result in the representation not being interpretable as NMF. Moreover, the class label or discriminant information from the training set is not apparently incorporated during constructing sparse representation, which may limit the ultimate classification accuracy. Liu et al. [46] propose a method for supervised image recognition and refer to it as the nonnegative curds and whey (NNCW). The NNCW procedure consists of two stages. In the first stage, NNCW considers a set of sparse and nonnegative representations of a test image, each of which is a linear combination of the images within a certain class, by solving a set of regression-type NMF problems. In the second stage, NNCW incorporates these representations into a new sparse and nonnegative representation by using the group nonnegative garrote [87]. This procedure is particularly appropriate for discriminant analysis owing to its supervised and nonnegativity nature in sparsity pursuing.

It is natural in group *lasso* to allow the size of each group to grow unbounded, that is, we replace the sum of Euclidean norms with a sum of appropriate Hilbertian norms. Under this setting, several algorithms are proposed to connect multiple kernel learning and group-lasso regularizer together [2]. The composite kernel learning with group structure (CKLGS) is proposed in [86] to select groups of discriminative features. The CKLGS method embeds the nonlinear data with discriminative features into different reproducing kernel Hilbert spaces (RKHS), and then composes these kernels to select groups of discriminative features.

### 2.3 Structural grouping sparsity

If the pairwise correlation between a group of features is very high, lasso tends to individually select only one of the pairwise correlated features and does not induce the group effect. In the "large $p$, small $n$" problem, the "grouped features" situation is an important concern to facilitate a model's interpretability. That is to say, *lasso* is limited in that it treats each input feature independent of each other and hence is incapable of capturing structural priors among heterogenous features. In order to remedy this deficiency of *lasso*, elastic net [93] and group *lasso* [87] are proposed, respectively.

Elastic net [93] generalizes *lasso* to overcome these drawbacks. For any nonnegative $\lambda_1$ and $\lambda_2$, elastic net is defined as a following optimization problem:

$$\min_{\hat{\boldsymbol{\beta}}_j} ||\mathbf{Y}_{(:,j)} - \mathbf{X}\hat{\boldsymbol{\beta}}_j||_2^2 + \lambda||\hat{\boldsymbol{\beta}}_j||_2^2 + \lambda||\hat{\boldsymbol{\beta}}_j||_1 \tag{5}$$

Group *lasso* is proposed by Yuan and Lin [87] by solving the following convex optimization problem:

$$\min_{\hat{\boldsymbol{\beta}}_j} \left\| \mathbf{Y}_{(:,j)} - \sum_{l=1}^{L} \mathbf{X}_l \hat{\beta}_{jl} \right\|_2^2 + \lambda \sum_{l=1}^{L} \sqrt{p_l}||\hat{\beta}_{jl}||_2 \tag{6}$$

where $p$ dimension features are divided into $L$ groups, with $p_l$ the number in group $l$. Note that $||\cdot||_2$ is the *not squared* Euclidean norm. This procedure acts like lasso at the group level: depending on λ, an entire group of features may be dropped out of the model. The key assumption behind the group *lasso* regularizer is that if a few features in one group are important, then most of the features in the same group should also be important. In fact, if the group sizes are all one, (6) reduces to lasso (3).

Yang et al. [83] takes the regions within the same image as a group and proposes spatial group sparse coding (SGSC) for region tagging. In SGSC, the group structure of regions-in-image relationship is incorporated into the sparse reconstruction framework by the group *lasso* penalty. Experimental results show that SGSC achieves a good performance of region tagging by integrating a spatial Gaussian kernel into the group sparse reconstruction.

If there is a linear-ordering (also known as *chain*) in the features, fused *lasso* can be used [70]. For example, in order to remove low-amplitude structures and globally preserve and enhance salient edges, Xu et al. [81] introduces an order penalty into the image smooth based on the mechanism of discretely counting spatial changes.

The heterogenous features in images and videos are *naturally* grouped. For example, color and shape, respectively, discern the aspects of visual characteristics. That is to say, it is convenient to select discriminative features from high-dimensional heterogeneous features by performing feature selection at a group level. However, the group *lasso* does not yield sparsity within a group. That is, if the selection coefficients of a group is nonzero, the selection coefficient of each feature within that group will all be nonzero.

In order to utilize the structure priors between heterogeneous and homogeneous features for image annotation, Wu et al. [77] proposes a framework of multi-label boosting by the selection of heterogeneous features with structural grouping sparsity (**MtBGS**). MtBGS formulates the multi-label image annotation problem as a multiple response regression model with a structural grouping penalty. A benefit of performing multi-label image annotation via regression is the ability to introduce penalties. Many of the penalties can be introduced into the regression model for a better prediction. Hastie et al. [27] proposes the penalized discriminant analysis (PDA) to tackle problems of overfitting in situations of large numbers of highly correlated predictors (features). PDA introduces a quadratic penalty with a symmetric and positive definite matrix $\Omega$ into the objective function. Elastic net [93] is proposed to conduct automatic variable selection and group selection of the correlated variables simultaneously by imposing both $\ell_1$ and $\ell_2$-norm penalties. Furthermore, motivated by elastic net, Clemmensen et al. [15] extended PDA to sparse discriminant analysis (SDA).

The basic motivation of imposing structural grouping penalty in MtBGS is to perform heterogeneous feature group selection and subgroup identification within homogeneous features simultaneously. As we know, some subgroups of features in high-dimensional heterogenous features have a discriminative power for predicting certain labels of a given image.

For each label $j$ and its corresponding indicator vector, the regression model of **MtBGS** is defined as follows:

$$\min_{\hat{\boldsymbol{\beta}}_j} \left\| \mathbf{Y}_{(:,j)} - \sum_{l=1}^{L} \mathbf{X}_l \hat{\beta}_{jl} \right\|_2^2 + \lambda_1 \sum_{l=1}^{L} ||\hat{\beta}_{jl}||_2 + \lambda_2 ||\hat{\boldsymbol{\beta}}_j||_1 \tag{7}$$

where $\lambda_1 \sum_{l=1}^{L} ||\hat{\beta}_{jl}||_2 + \lambda_2 ||\hat{\boldsymbol{\beta}}_j||_1$ is the regularizer $P(\hat{\boldsymbol{\beta}}_j)$ in (2) and is called the *structural grouping penalty* in [77].

Let $\hat{\boldsymbol{\beta}}_j$ be the solution to (7); we predict the probability $\hat{\mathbf{y}}_u$ that unlabeled images $\mathbf{X}^u$ belong to the $j$th label as follows:

$$\hat{\mathbf{y}}_u = \mathbf{X}^u \hat{\boldsymbol{\beta}}_j \tag{8}$$

Unlike group *lasso*, the above structural grouping penalty in (7) not only selects the groups of heterogeneous features, but also identifies the subgroup of homogeneous features within each selected group.

Note that when $\lambda_1 = 0$, (7) reduces to the traditional *lasso* under the multi-label learning setting, and $\lambda_2 = 0$ for the group *lasso* [87].

As stated before, for problems where the heterogeneous features lie in a high-dimensional space with a sparsity structure and only a few common important features are shared by labels (tasks), regularized regression methods have been proposed to recover the shared sparsity structure across tasks. According to [9], if the labels are correlated we may be able to obtain an accurate prediction. In order to take advantage of correlations between the labels to boost multi-label annotation, MtBGS utilizes the curds and whey (C&W) [9] method to boost the annotation performance.

In order to tackle problems of overfitting in situations of large numbers of highly correlated predictors, Hastie et al. [27] introduce a quadratic penalty with a symmetric and positive definite matrix $\Omega$ into the objective function. Taking into account the ability of *elastic net* which simultaneously conducts automatic variable selection and group selection of correlated variables, Clemmensen et al. [15] formulate (single-task) MLDA as SDA by imposing both $\ell_1$ and $\ell_2$ norm regularization. Han et al. [25] extends single-task SDA to the multi-task problem with a method called multi-task sparse discriminant analysis (MtSDA). MtSDA uses a quadratic optimization approach for prediction of the multiple labels. In SDA, the identity matrix is commonly used as the penalty matrix. MtSDA introduces a large class of equicorrelation matrices with the identity matrix as a special case and indicates that an equicorrelation matrix has a grouping effect under some conditions.

### 2.4 Structured sparsity-inducing norm

Jenatton et al. propose a general definition of structured sparsity-inducing norm in [31,32], based on which many sparsity penalties, such as lasso, group lasso, and even the tree-guided group lasso [37], may be instantiated.

**Definition 1** (*Structured sparsity-inducing norm*) Given a $p$-dimensional feature vector $\mathbf{x}$, let us assume that the set of groups of features $\mathcal{G} = \{g_1, \ldots, g_{|\mathcal{G}|}\}$ is defined as a subset of the power set of $\{1, \ldots, p\}$; the structured sparsity-inducing norm $\Omega(\mathbf{x})$ is defined as

$$\Omega(\mathbf{x}) \equiv \sum_{g\in\mathcal{G}} w_g ||\mathbf{x}_g||_2$$

where $\mathbf{x}_g \in \mathbb{R}^{|g|}$ is the sub-vector of $\mathbf{x}$ for the input feature index in group $g$, and $w_g$ is the predefined weight for group $g$.

In Definition 1, if we ignore weight $w_g$ and let $\mathcal{G}$ be the set of singleton, i.e., $\mathcal{G} = \{\{1\}, \{2\}, \ldots, \{p\}\}$, $\Omega(\mathbf{x})$ is instantiated to be an $\ell_1$-norm of vector $\mathbf{x}$.

### 2.5 Tree and graph-guided sparsity

In a typical setting, the input features lie in a high-dimensional space, and one is interested in selecting a small number of features that influence the annotation output. In order to handle more general structures such as tree or graph, various models that further extend group *lasso* and fused *lasso* are proposed [13,37]. Tree-guided group *lasso* [37] is a multi-task sparse feature selection method. The penalty of tree-guided group *lasso* is imposed on the output direction of the coefficient matrix $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J)$, with the goal of integrating the correlations among multiple labeled tags into the process of sparse feature selection. Tree-guided group *lasso* is formulated as to solve the following regularized regression model:

$$\min_{\hat{\mathbf{B}}} ||\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}||_2^2 + \gamma \sum_{d=1}^{p} \sum_{v\in\mathcal{G}_\mathbf{T}} w_v ||\hat{\mathbf{B}}_v^d||_2 \quad (9)$$

For ease of notation, in (9), we let $\mathbf{B}^d = \mathbf{B}_{(d,:)}$. We call $\sum_{v\in\mathcal{G}_\mathbf{T}} w_v ||\mathbf{B}_v^d||_2$ the penalty of tree-guided group *lasso*. Specifically, $\sum_{g\in\mathcal{G}_\mathbf{T}} w_v ||\mathbf{B}_v^d||_2$ is a special example of $\Omega(\mathbf{B}^d)$ in Definition 1, when a set of groups $\mathcal{G}_\mathbf{T}$ is induced from a tree structure $\mathbf{T}$ that is defined on vector $\mathbf{B}^d$. For the details of definitions of $w_v$ and $\mathbf{T}$, refer to [37].

Furthermore, let us assume that the structure of the $p$-dimensional features for each image and video $\mathbf{x}_i$ is available as a graph $G$ with a set of nodes $V = \{1, 2, \ldots, p\}$ and a set of edges $E$. Let $w_{ml} \geq 0$ denote the weight of the edge $e = (m, l) \in E$, corresponding to the correlation between two features for nodes $m$ and $l$. With $w_{ml} \geq 0$, we only consider the positively correlated features. In order to integrate graph $G$ into the process of structural feature selection and guide the regularization process, a penalty of graph-guided fusion (G$^2$F) [13] $\Omega_G(\boldsymbol{\beta})$ is imposed, and the graph-guided feature selection framework is taken as follows:

$$\min_{\hat{\beta}} \frac{1}{2}||\mathbf{Y}_{(:,j)} - \mathbf{X}\hat{\boldsymbol{\beta}}||_2^2 + \gamma\Omega_G(\hat{\boldsymbol{\beta}}) + \lambda||\hat{\boldsymbol{\beta}}||_1 \quad (10)$$

where the G$^2$F penalty $\Omega_G(\hat{\boldsymbol{\beta}})$ is defined as [13]:

$$\Omega_G(\hat{\boldsymbol{\beta}}) = \sum_{e=(m,l)\in E, m<l} w_{ml} |\hat{\beta}_m - \hat{\beta}_l| \quad (11)$$

where $\hat{\beta}_m$ and $\hat{\beta}_l$ are estimated coefficients in $\hat{\boldsymbol{\beta}}$ corresponding to the selection coefficients of the $m$th and $l$th features, respectively. The weight $w_{ml}$ measures the fusion penalty for each edge $e = (m, l)$ such that $\hat{\beta}_m$ and $\hat{\beta}_l$ are for highly correlated features with a larger $w_{ml}$ receiving a greater fusion effect. Therefore, the graph-guided fusion penalty in (11) encourages highly correlated features corresponding to a densely connected sub-network in $G$ to be jointly selected as the relevant features.

Note that, if $w_{ml} = 1$ for all $e = (m, l)$, the penalty definition in (11) reduces to:

$$\Omega_G(\hat{\boldsymbol{\beta}}) = \sum_{e=(m,l)\in E, m<l} |\hat{\beta}_m - \hat{\beta}_l| \quad (12)$$

The standard fused *lasso* [70] penalty $\sum_{j=1}^{J-1} |\hat{\beta}_{j+1} - \hat{\beta}_j|$ is a special case of (12). Furthermore, if the edge set $E$ consists of edges of pairs of regions, i.e., graph $G$ is defined to be a full connected graph, the G$^2$F penalty in (11) is instantiated to be the grouping pursuit penalty [64].

### 2.6 Sparsity constrained tensor factorization

As introduced before, nuclear norm is recently proposed to discover the low-rank structure in a matrix and is denoted as the matrix-*lasso*. Unlike a matrix, a tensor is a multidimensional array. More formally, an $N$-way or $N$th order tensor is an element of the tensor product of $N$ vector spaces, each of which has its own coordinate system. Tensors include vectors and matrices as the first-order and the second-order special cases, respectively.

Many data in signal processing, computer vision, and multimedia retrieval can be naturally represented as a tensor (i.e., multi-way arrays). Due to the ability of release of the over-fitting problem in vector-based learning, Tao et al. propose a supervised tensor learning (STL) framework in [69]. Based on STL and its alternating projection optimization procedure, the generalization of support vector machines (SVM) is extended to support tensor machines (STM). Wu et al. [79] introduces a higher-order tensor framework for video analysis, which represents image frame, audio and text in video shots as data points by the three-order tensor (*Tensor*Shot). A transductive support tensor machine (TSTM) is then developed to learn and classify tensorshots in videos. Since the TensorShot dimension reduction method discovers the intrinsic tensorshot manifold before classification, the dimension-reduced tensorshots of training and test data sets show that TSTM not only is a natural extension of TSVM in tensor space, but also has a more powerful classification capability.

Tensor factorization and decomposition have several advantages over the traditional two-order matrix factorizations even when most of the data are missing. Moreover, tensor decomposition and factorization explicitly exploit the high-order structures that are lost when the regular matrix factorization approaches such as PCA, SVD, NMF, and ICA are directly implemented to tensors.

Two of the most commonly used tensor decompositions are the Tucker decomposition [72] and CANDECOMP/PARAFAC (CANonical DECOMPosition or PARAllel FACtors model, abbreviated as CP) [12,26]. Both Tucker decomposition and CP are often considered as higher-order generalizations of the matrix singular value decomposition (SVD) or PCA.

The Tucker decomposition is a form of higher-order PCA and decomposes a tensor into a core tensor multiplied (or transformed) by a matrix along each mode. The CP decomposition factorizes a tensor into a sum of component rank-one tensors. An interesting property of tensor decompositions by CP is the uniqueness under a weak condition. However, Tucker decompositions are not unique [39].

When data or signals are inherently represented by nonnegative numbers, imposing nonnegativity constraints to tensor decomposition is shown to provide a physically meaningful interpretation. The block principal pivoting method is developed to solve nonnegativity constrained least squares (NNLS) problems for computing a low-rank nonnegative CP decomposition in [36].

## 3 Multimedia spectral hashing with sparsity

The summarization of multimedia (images and videos) by much more compact sets of binary bits is of strong interest to many multimedia processing applications. The summaries, or *hashes*, can be used as a content identification to efficiently query similar images or videos in a database. Multimedia hashing is usually implemented in two steps: first, an intermediate code is obtained by the extraction of the representative features from images and videos; second, this intermediate code is quantized by a vector quantization to generate a binary code. In general, these two steps are independent of each other.

The hashing algorithms seek compact binary codes of data points, so that the Hamming distance between codewords correlates with a semantic similarity such as the semantic hashing [60]. As one of the representative hashing approaches, the spectral hashing [74] is formulated as the problem of the semantic hashing as a form of graph partitioning and is designed as an efficient eigenvector solution to graph partition to generate the best binary code for data indexing. Spectral hashing finds a projection from Euclidean space to Hamming space and guarantees that data points close in Euclidean space are mapped to similar binary codewords. In order to avoid NP problems, spectral relaxation is implemented in spectral hashing to obtain a number of eigenvectors with the minimal eigenvalues from a Laplacian matrix by the PCA. However, the traditional PCA suffers from the fact that each principal component is a linear combination of all the original variables; thus, it is often difficult to interpret the results [94].

Assume that we have a collection of $n$ $p$-dimensional data points $\{(\mathbf{x}_i) \in \mathbb{R}^p : i = 1, 2, \ldots, n\}$, where $\mathbf{x}_i$ represents the feature vector for the $i$th data point, and $p$ represents the dimension of the features from the training data. $\Theta$ is an efficient indexing function to map each $x_i$ from the $p$-dimensional Euclidean space to the $k$-dimensional Hamming space $y_i$; we define $\Theta$ as follows:

$$\Theta : \mathbf{x}_i \in \mathbb{R}^p \rightarrow y_i \in \{-1, 1\}^k \tag{13}$$

The indexing function $\Theta$ defined above has the following characteristics: (1) $\Theta$ is a semantic hashing function. That is to say, if the distance between the $i$th data point $x_i$ and the $j$th data point $x_j$ is small in terms of the Euclidean distance in $\mathbb{R}^p$ space, their distance in terms of the Hamming distance in $\{-1, 1\}^k$ space is also small; (2) $\Theta$ tends to generate a compact binary code $y_i$. Only a small number of bits are required to code the whole data set. Additionally, the coding is expected to be *structure preserving*, which means that only a limited subset of features is chosen to index the original data and to preserve the intrinsic structure hidden in the images and videos. The $n$ $p$-dimensional data are written as $\mathbf{X} \in \mathbb{R}^{n\times p}$ and their corresponding $n$ $k$-dimensional binary codes are written as $\mathbf{Y} \in \{-1, 1\}^{n\times k}$.

Spectral hashing considers such requirements as a particular problem of thresholding a subset of eigenvectors of the Laplacian graph as follows [74]:

$$\begin{aligned} \text{minimize:}\ & \sum_{ij} \mathbf{W}(i, j) \parallel y_i - y_j \parallel^2 \\ \text{subject to:}\ & y_i \in \{-1, 1\}^k \\ & \sum_i y_i = 0 \\ & \frac{1}{n} \sum_i y_i y_i^{\mathrm{T}} = \mathbf{I} \end{aligned} \tag{14}$$

where $\mathbf{W} \in R^{N\times N}$ is the similarity matrix and $\mathbf{W}(i,j) = \exp(-\parallel x_i - x_j \parallel^2 /\epsilon^2)$. $y_i \in \{-1,1\}^k$ guarantees that the indexing code is binary; $\sum_i y_i = 0$ guarantees that each bit has 50% to be $-1$ or 1 when we choose 0 as a threshold; and $\frac{1}{n}\sum_i y_i y_i^{\mathrm{T}} = \mathbf{I}$ guarantees that the bits are uncorrelated to each other.

Though finding the solution to (14) is an NP-complete problem, by introducing the Laplacian matrix and removing the constraint $y_i \in \{-1,1\}^m$, [74] turns it into a graph partition so that the solution is simply the $m$ eigenvectors of Laplacian matrix $\mathbf{L}$ with the minimal eigenvalue, where $\mathbf{L} = \mathbf{D} - \mathbf{W}$ and $\mathbf{D}$ is a diagonal matrix with its entries $\mathbf{D}(i,i) = \Sigma_j^n \mathbf{W}(i,j)$. As a result, the solution to (14) is transformed into Laplacian eigenmap dimension reduction and thus PCA is directly used in spectral hashing [74].

However, PCA suffers from the fact that it is difficult to interpret the result of dimension reduction due to the lack of sparseness of the principal vectors, i.e., all the data coordinates participate in the linear combination [94]. Usually, the codebook is often *over-complete*. Intuitively, given a data point, only a limited subset of features is sufficient to represent this data point. Take an image as an example. Color-related visual words could be salient features to represent an image of rainbow and shape-related visual words are better to distinguish an image of a car from others.

Motivated by the nature of the grouping effect in *elastic net*, sparse PCA in [94] transforms PCA as a *nonconvex* regression-type optimization problem via the *elastic net* penalty to estimate PCs with sparse loadings. Due to its nonconvex solution of sparse PCA in [94], a *convex relaxation* method is developed [16] to achieve a globally optimal solution to sparse PCA. In essence, the convex relaxation method is a semi-definite program. Since sparse PCA maximizes the variance explained by a particular linear combination of the input variables and constrains the number of nonzero coefficients in this combination, Shao et al. [63] introduces the idea of sparse PCA [94] into spectral hashing for data indexing and calls this approach as SSH. The proposed SSH not only achieves dimensionality reduction, but also reduces the number of the explicitly used features during indexing. In order to resolve the *out-of-sample* problem, same as the parameter-sensitive hashing [62], here SSH introduces boosting similarity sensitive coding (Boost SSC) into SSH in order to find a more practical threshold for the quantization of the binary code.

As discussed before, The structural information is of great significance in many applications. For example, due to the importance of local features in face recognition, NMF is used to learn the facial parts of the images and shows a better performance than other holistic representation methods such as PCA and vector quantization. It is a quite interesting issue to embed structural *prior* into the compact binary codes by the replacement of sparse PCA with the structured sparse principal component analysis [32] in SSH.

## 4 Cross-media analysis and retrieval

As an example to showcase the applications, there are huge collections of heterogeneous media data from microblog, mobile phone, social networking Web sites, and news media Web sites. These heterogeneous media data are integrated together to reflect social behaviors. Different from the traditional structural and nonstructural data, these heterogeneous media data are referred to as *cross-media* with three properties: (1) *Cross-modality*: heterogeneous features are obtained from data in different modalities; (2) *Cross-domain*: heterogeneous features may be obtained from different domains (e.g., from both target domain and auxiliary domain for problems such as topic modeling, multimedia annotation); (3) *Cross space*: the virtual world (cyberspace) and the real-world (reality) complement each other [23]. Here, sparse representation also plays an important role [78]. For example, all of the heterogenous features from different views can be unified as a consensus representation for the cross-media semantics, and factorized into a latent spaces with a structured sparsity that can be exploited to simultaneously learn a low-dimensional latent space [33]; traditional canonical correlation analysis (CCA) can be extended to sparse CCA and therefore learn the multi-modal correlations of media objects [78]; graphical *lasso* can be applied to discover the network community [22]. Moreover, different from the traditional content-based single media retrieval systems, in content-based cross-media retrieval system, multimedia objects are retrieved uniformly. The query examples and retrieval results do not need to be of the same media type. For example, users can query images by submitting either an audio example or an image example in a cross-media retrieval system [84].

## 5 Computational issues

### 5.1 Complexity

In principle, the $\ell_1$-norm [71] and the structured sparsity-inducing norm [32] penalized sparse feature selection problems can be solved by the generic optimization solvers. For example, the sparse penalized problems are first posed as a second-order cone programming or a quadratic programming [71] formulation and then solved by the interior-point methods. However, such approaches are expensive even for the problems of a moderate size.

Recently, inspired by Nesterov's method [53] and the fast iterative shrinkage-thresholding algorithm (FISTA) [3], the first-order gradient approach has been widely used to solve optimization problems with a convex loss function (e.g., least-square loss) and nonsmooth penalty (e.g., the $\ell_1$-norm). It has been shown that the first-order gradient approach can achieve the optimal convergence rate $O(\frac{1}{\sqrt{\epsilon}})$ for a desired accuracy $\epsilon$. FISTA only deals with relatively simple and well-separated nonsmooth penalties, such as $\ell_1$-norm, group *lasso* penalty. Though the pathwise coordinate descent [21] method has been widely applied to solve many complex sparsity penalties, this method may get stuck and does not converge to the exact solution for certain nonseparated penalties, such as the fused *lasso* [70] penalty.

In order to efficiently solve the complex sparsity penalized problems, e.g., graph-guided fusion [13] and tree-structured groups [30] penalties, many proximal gradient methods [13,30] are developed. The main challenges are to find an approximation of the nonseparable and nonsmooth structured sparsity-inducing norm. The smoothing proximal gradient (SPG) [13] method is an efficient proximal gradient method for general structured sparse feature selection methods. According to the smoothing method in [54], SPG first finds a separable and smooth approximation of $\Omega(\mathbf{x})$, and then solves this transformed simple $\ell_1$-norm penalized sparse learning problem by the FISTA approach. It has been proven that SPG achieves a convergence rate of $O(\frac{1}{\epsilon})$ for a desired accuracy $\epsilon$ [13], which is faster than the subgradient method with a convergence rate of $O(\frac{1}{\epsilon^2})$. The gap between $O(\frac{1}{\epsilon})$ and $O(\frac{1}{\sqrt{\epsilon}})$ [3] is due to the approximation of the nonseparable and nonsmooth structured sparsity-inducing norm penalty.

The most common theoretical approach to understanding the behavior of the algorithms is the worst-case analysis. However, there are many algorithms that work exceedingly well in practice, but are known to perform poorly in the worst-case analysis or lack a good worst-case analysis according to the theory of the smoothed analysis [67]. In Tibshirani's original paper [71], he has found that the model selection problem with $\ell_1$-norm usually can be solved with the iteration number within the range of $(0.5p, 0.75p)$ in practice. Take the algorithm of MtBGS [77], for example, the runtime performance of the regression model with the structural grouping sparsity is also very efficient when implemented as the *cyclic coordinate descent* method. From the description of the coordinate descent by the Gauss–Seidel method, we see that for a complete cycle through all the coordinates, it takes $O(k)$ operations, where $k$ is the number of the nonzero elements when the sparsity of the data is considered. Thus the complexity of the regression model with the structural grouping sparsity is roughly $O(p \times n)$.

### 5.2 Consistent selection and nonconvex relaxation

Given high-dimensional heterogenous features and their corresponding semantics, the question is whether there is a true model and whether the true model can select all the features for the representation of their semantics; that is to say whether the selected features are *consistent* with the data. For example, after $p$-dimensional heterogenous features $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}} \in \mathbb{R}^p$ are extracted from the $i$th image or video, assume a *true model* (*an oracle*) is found to select all the discriminate features for the $i$th image or video and the coefficients of the selected features are denoted as $A = \{m : \beta^*_m \neq 0\}$ and$|A| = p_0 < p$. If $\hat{\beta}(\delta)$ is the estimated coefficients produced by a fitting procedure $\delta$, the question is whether the selection result by $\delta$ is the same as that of the result by the true model. According to [20,95], $\delta$ is called an *oracle* procedure if $\hat{\beta}(\delta)$ has the following oracle properties: $\delta$ can identify the correct subset, $\{m : \hat{\beta}_j \neq 0\} = A$; and $\delta$ has the optimal estimate rate.

When we talk about the inconsistent selection, it means that the correct sparse subset of the relevant variables cannot be identified asymptotically with a large probability. Without loss of generality, assume that the first $q$ elements of vector $\hat{\beta}(\delta)$ are nonzeroes and that others are zeros. Let $\hat{\beta}(\delta)_{(1)} = (\hat{\beta}(\delta)_1, \ldots, \hat{\beta}(\delta)_q)$ and $\hat{\beta}(\delta)_{(2)} = (\hat{\beta}(\delta)_{q+1}, \ldots, \hat{\beta}(\delta)_p)$, then $\hat{\beta}(\delta)_{(1)} \neq 0$ element-wise and $\hat{\beta}(\delta)_{(2)} = 0$. Recent work [50,88,90,95] has given some conditions for a consistent selection in *lasso*. It has been shown that in the classical case when $p$ and $q$ are fixed, a simple condition, called the *Irrepresentable Condition* (IC) on the generated covariance matrices, is necessary and sufficient for the model selection consistency by Lasso. An *Elastic Irrepresentable Condition* (EIC) is given in [34] to show that Elastic Net can consistently select the true model if and only if EIC is satisfied. One of the consistency conditions of group *lasso* is given in [44].

Many of penalty regularizers can be developed for feature selection; however, Fan and Li [20] provide a deep insight into how to choose a penalty function. In their analysis, they encourage choosing penalty functions satisfying certain mathematical conditions such that the resulting penalized likelihood estimate possesses the properties of sparsity, continuity and unbiasedness. These mathematical conditions imply that the penalty function must be singular at the origin and *nonconvex* over $(0, \infty)$. Accordingly, a number of nonconvex relaxation approaches, such as the smoothly clipped absolute deviation (SCAD) penalty [20] and the minimax concave (MC) penalty [89], have been proposed. Shi et al. [66] treats the MC penalty as a nonconvex relaxation of the $l_0$ penalty for dictionary learning and achieves a robust and sparse representation.

### 5.3 Stability of selection

Two key issues in the design of multimedia learning algorithms are bias and variance, and one needs to find a trade-off between them. Therefore, besides a good accuracy of sparse learning algorithms, we also desire the property of a low variance, or a high *stability*. In a broad sense, stability means that an algorithm is well posed, so that given two very similar data sets, the algorithm's performance varies little [80]. In the landmark work about stability in [6], stability is explored based on the sensitivity analysis, which aims at determining how much the variation of the data can influence the performance of a learning algorithm. Two sources of the instability come from the sampling mechanism used to generate the input data and the noise in the input data, respectively. The former is mainly investigated by the sampling randomness, for example using the re-sampling methods [77] of cross validation, jackknife [51], and bootstrap [18], whereas the latter is usually referred to as the perturbation analysis.

In [6], stability is taken as an avenue for proving the generalization performance of an algorithm. More specifically, stability is a principled way of establishing bounds on the difference between the empirical and the generalization errors. In statistical learning theory, Vapnik–Chervonenkis dimension (VC-dimension) is a measure of the capacity or complexity of a statistical classification algorithm. It has been proven that an algorithm having a search space of a finite VC-dimension is stable in the sense that its stability is bounded by its VC-dimension [35]. Therefore, Bousquet and Elisseeff [6] use the stability to derive the generalization error bound based on the empirical error and the leave-one-out error.

*lasso* [71] is known to have the stability problems [8]. Although its predictive performance is not disastrous, the selected predictor may vary a lot. Typically, given two correlated variables, *lasso* only selects one of the two, at random. In [80], Xu et al. prove that sparsity and stability are at odds with each other. They show that sparse algorithms are not stable: if an algorithm encourages sparsity, e.g., *lasso*, then its sensitivity to small perturbations of the input data remains bounded away from zero. Based on the uniform stability [6] properties, they have proven that a sparse algorithm can have nonunique optimal solutions and is therefore ill-posed.

Breiman [8] has shown that the unstable linear regression process can be stabilized by perturbing the data, getting a new predictor sequence and then averaging over many such predictor sequences. Thus, how to develop a stable feature selection model by the perturbing technique is an open problem. Furthermore, Xu et al. also prove that the stability bound of the elastic net [93] coincides with that of an $\ell_2$ regularization algorithm and thus has the uniform stability. Consequently, it is interesting to explore the stability of the recent proposed sparse models, such as group lasso [87], SDA [15], and even the structured sparsity penalized regression models.

## 6 Conclusion

This paper surveys some recent research work on heterogeneous feature selection, representation, and hashing for images and videos after the introduction of sparsity constraints. The utilization of sparsity in images and videos make multimedia understanding and retrieval interpretable. However, how to define the intrinsic spatial–temporal structure in images and videos and then to apply an appropriate sparse penalty is still an open problem for multimedia research.

**Acknowledgments** This work was supported in part by National Basic Research Program of China (2012CB316400), NSFC (No. 90920303, 61070068) and the Fundamental Research Funds for the Central Universities. Zhongfei Zhang was also supported in part by US NSF (IIS-0812114, CCF-1017828). Yahong Han was also supported in part by Scholarship Award for Excellent Doctoral Student granted by the Ministry of Education, China.

## References

1. Argyriou A, Evgeniou T, Pontil M (2007) Multi-task feature learning. In: Advances in neural information processing systems (NIPS)
2. Bach FR (2008) Consistency of the group Lasso and multiple kernel learning. J Mach Learn Res 9:1179–1225
3. Beck A, Teboulle M (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J Imaging Sci 2(1):193–202
4. Berchtold S, Bohm C, Jagadish HV, Kriegel HP, Sander J (2000) Independent quantization: an index compression technique for high-dimensional data spaces. In: Proceedings of international conference on data engineering (ICDE), pp 577–588
5. Berge C (1973) Graphs and hypergraphs. North-Holland, Amsterdam
6. Bousquet O, Elisseeff A (2002) Stability and generalization. J Mach Learn Res 2:499–526
7. Breiman L (1995) Better subset regression using the nonnegative garrote. Technometrics 373–384
8. Breiman L (1996) Heuristics of instability and stabilization in model selection. Ann Stat 24(6):2350–2383
9. Breiman L, Friedman J (1997) Predicting multivariate responses in multiple linear regression. J R Stat Soc Ser B (Methodological) 59(1):3–54
10. Candes EJ, Donoho DL (2002) New tight frames of curvelets and optimal representations of objects with piecewise C2 singularities. In: Communications on pure and applied mathematics, pp 219–266
11. Cao L, Luo J, Liang F, Huang T (2009) Heterogeneous feature machines for visual recognition. In: Proceedings of the IEEE internation conference on computer vision (ICCV), pp 1095–1102
12. Carroll JD, Chang JJ (1970) Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition. Psychometrika 35(3):283–319
13. Chen X, Lin Q, Kim S, Carbonell J, Xing E (2010) Efficient proximal gradient method for general structured sparse learning. Preprint, arXiv:1005.4717
14. Cheng B, Yang J, Yan S, Fu Y, Huang T (2010) Learning with $l_1$-graph for image analysis. IEEE Trans Image Process (TIP) 19(4):858–866
15. Clemmensen L, Hastie T, Ersbøll B (2008) Sparse discriminant analysis. http://www-stat.stanford.edu/hastie/Papers/

16. d'Aspremont A, El Ghaoui L, Jordan MI, Lanckriet GRG (2004) A direct formulation for sparse PCA using semidefinite programming. In: Advances in neural information processing systems (NIPS)
17. Datar M, Immorlica N, Indyk P, Mirrokni VS (2004) Locality-sensitive hashing scheme based on p-stable distributions. In: Proceedings of annual symposium on computational geometry, pp 253–262
18. Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7(1):1–26
19. Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. Ann Stat 32(2):407–499
20. Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96(456):1348–1360
21. Friedman J, Hastie T, Holger H, Tibshirani R (2007) Pathwise coordinate optimization. In: Annals of applied statistics, pp 302–332
22. Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9(3)
23. Gewin V (2011) Self-reflection. Nature 471:667–669
24. Grauman K, Darrell T (2005) Pyramid match kernels: discriminative classification with sets of image features. In: Proceedings of international conference on computer vision (ICCV), pp 1458-1465
25. Han Y, Wu F, Jia J, Zhuang Y , Yu. B (2010) Multi-task sparse discriminant analysis (MtSDA) with overlapping categories. In: Proceedings of the AAAI conference on artificial intelligence (AAAI), pp 469-474
26. Harshman RA (1970) Foundations of the PARAFAC procedure: models and conditions for an "explanatory" multimodal factor analysis. University of California at Los Angeles
27. Hastie T, Buja A, Tibshirani R (1995) Penalized discriminant analysis. Ann Stat 23(1):73–102
28. He X, Niyogi P (2006) Tensor subspace analysis. In: Advances in neural information processing systems (NIPS)
29. Irani M (1999) Multi-frame optical flow estimation using subspace constraints. In: Proceedings of international conference on computer vision (ICCV), pp 623–633
30. Jenatton R , Julien M, Guillaume O, Bach F (2010) Proximal methods for sparse hierarchical dictionary learning. In: Proceedings of the 27th international conference on machine learning (ICML)
31. Jenatton R, Audibert JY , Bach F (2009) Structured variable selection with sparsity-inducing norms. Preprint, arXiv:0904.3523
32. Jenatton R, Obozinski G, Bach F (2010) Structured sparse principal component analysis. In: Proceedings of international conference on artificial intelligence and statistics (AISTATS)
33. Jia Y, Salzmann M, Darrell T (2010) Factorized latent spaces with structured sparsity. In: Advances in neural information processing systems (NIPS)
34. Jia J, Yu B (2010) On model selection consistency of the elastic net when $p \gg n$. Stat Sin 20:595–611
35. Kearns M, Ron D (1999) Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. Neural Comput 11(6): 1427–1453
36. Kim J, Park H (2011) Fast nonnegative matrix factorization: an active-set-like method and comparisons. SIAM J Sci Comput
37. Kim S, Xing EP (2010) Tree-guided group lasso for multi-task regression with structured sparsity. In: Proceedings of the 27th international conference on machine learning (ICML)
38. Kim H, Park H (2007) Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinforma/Comput Appl Biosci 23:1495–1502
39. Kolda TG, Bader BW (2009) Tensor decompositions and applications. SIAM Rev 51(3):455–500
40. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition (CVPR)
41. Lee D, Seung H (1999) Learning the parts of objects by nonnegative matrix factorization. Nature 401(6755):788–791
42. Lee TS (1996) Image representation using 2D Gabor wavelets. IEEE Trans Pattern Anal Mach Intell 18(10):959–971
43. Li L, Zhou M, Sapiro G, Carin L (2011) On the integration of topic modeling and dictionary learning. In: Proceedings of international conferences on machine learning (ICML)
44. Liu H, Zhang J (2009) Estimation consistency of the group lasso and its applications. In: Proceedings of the twelfth international conference on artificial intelligence and statistics (AISTATS)
45. Liu N, Zhang B, Yan J, Chen J (2005) Text representation: from vector to tensor. In: Proceedings of international conferences on data mining (ICDM)
46. Liu Y, Wu F, Zhang Z, Zhuang Y , Yan S (2010) Sparse representation using nonnegative curds and whey. In: Proceedings of computer vision and pattern recognition (CVPR), pp 3578–3585
47. Lv Q, Josephson W, Wang Z, Charikar M., Li K (2007) Multi-probe LSH: efficient indexing for high-dimensional similarity search. In: Proceedings of international conference on very large data bases (VLDB), pp 950–961
48. Ma ZG, Yang Y, Nie FP, Uijlings J, Sebe N (2011) Exploiting the entire feature space with sparsity for automatic image annotation. In: Proceedings of the ACM multimedia (ACM MM)
49. Maron O, Ratan A.L (1998) Multiple-instance learning for natural scene classification. In: Proceedings of the international conference on machine learning (ICML), pp 341–349
50. Meinshausen N, Yu B (2009) Lasso-type recovery of sparse representations for high-dimensional data. Ann Stat 37(1):246–270
51. Miller RG (1974) The jackknife—a review. Biometrika 6(1):1–15
52. Mu Y, Dong J, Yuan X, Yan S (2011) Accelerated low-rank visual recovery by random projection. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)
53. Nesterov Y (2007) Gradient methods for minimizing composite objective function. Technical report, Universit catholique de Louvain. Center for Operations Research and Econometrics (CORE)
54. Nesterov Y (2005) Smooth minimization of non-smooth functions. Math Program 103(1):127–152
55. Obozinski G, Taskar B, Jordan MI (2006) Multi-task feature selection. Technical report, Statistics Department, UC Berkeley
56. Olshausen B, Field D (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381:607–609
57. Quattoni, A, Collins, M, Darrell, T (2008) Transfer learning for image classification with sparse prototype representations. In: Proceedings of computer vision and pattern recognition (CVPR), pp 1-8
58. Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. Nat Neurosci 2(11):1019–1025
59. Salakhutdinov R, Hinton GE (2007) Learning a nonlinear embedding by preserving class neighbourhood structure. In: AI and statistics
60. Salakhutdinov R, Hinton G (2009) Semantic hashing. Int J Approx Reason 50(7):969–978
61. Saul L, Weinberger K, Sha F, Ham J, Lee D (2006) Spectral methods for dimensionality reduction. In: Semisupervised learning, pp 293-308
62. Shakhnarovich G, Viola P, Darrell T (2003) Fast pose estimation with parameter-sensitive hashing. In: Proceedings of IEEE international conference on computer vision (ICCV), pp 750-757
63. Shao J, Wu F, Ouyang C, Zhang X (2011) Sparse spectral hashing. In: Pattern recognition letters

64. Shen X, Huang HC (2010) Grouping pursuit through a regularization solution surface. J Am Stat Assoc 105(490):727–739
65. Shevade S, Keerthi S (2008) A simple and efficient algorithm for gene selection using sparse logistic regression. Bioinformatics 19(17):2246–2253
66. Shi J, Ren X, Dai G, Wang J, Zhang Z (2011) A non-convex relaxation approach to sparse dictionary learning. In: Proceedings of computer vision and pattern recognition (CVPR)
67. Spielman D, Teng S (2009) Smoothed analysis: an attempt to explain the behavior of algorithms in practice. Commun ACM 52(10):76–84
68. Sun L, Ji S, Ye J (2008) Hypergraph spectral learning for multi-label classification. In: Proceeding of the 14th ACM SIGKDD international conference on knowledge discovery and data mining (SIGKDD), pp 668-676
69. Tao D, Li X, Wu X, Hu W, Maybox J (2005) Supervised tensor learning. In: Proceedings of IEEE conference on data mining (ICDM)
70. Tibshirani R, Saunders M (2005) Sparsity and smoothness via the fused lasso. J R Stat Soc Ser B (Statistical Methodology) 67(1):91–108
71. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B (Statistical Methodology) 58(1):267–288
72. Tucker LR (1996) Some mathematical notes on three-mode factor analysis. Psychometrika 31(3):279–311
73. Wang J, Yang J, Yu K, Lv F (2005) Locality-constrained linear coding for image classification. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition (CVPR), pp 3360-3367
74. Weiss Y, Torralba A, Fergus R (2009) Spectral hashing. In: Advances in neural information processing systems (NIPS), pp 1753-1760
75. Wright J, Ganesh A, Rao S , Ma Y (2009) Exact recovery of corrupted low-rank matrices. Robust principal component analysis. In: Advances in neural information processing systems (NIPS)
76. Wright J, Yang A, Ganesh A, Sastry S, Ma Y (2009) Robust face recognition via sparse representation. IEEE Trans Pattern Anal Mach intell 31(2):210–227
77. Wu F, Han Y, Tian Q, Zhuang Y (2010) Multi-label boosting for image annotation by structural grouping sparsity. In: Proceedings of the 2010 ACM international conference on multimedia (ACM MM), pp 15–24
78. Wu F, Zhang H, Zhuang Y (2006) Learning semantic correlations for cross-media retrieval. In: Proceedings of IEEE international conference on image processing (ICIP), pp 1465-1468
79. Wu F, Liu Y, Zhuang Y (2009) Tensor-based transductive learning for multimodality video semantic concept detection. IEEE Trans Multimed 11(5):868–878
80. Xu H, Mannor S, Caramanis C (2011) Sparse algorithms are not stable: a no-free-lunch theorem. In: IEEE transactions on pattern analysis and machine intelligence
81. Xu L, Lu C, Xu Y, Jia J (2011) Image smoothing via L0 gradient minimization. ACM Trans Graph (SIGGRAPH, Asia 2011) 30(6)
82. Yang Y, Shen HY, Ma ZG, Huang Z, Zhou XF (2011) L21-norm regularized discriminative feature selection for unsupervised learning. In: International joint conferences on artificial intelligence (IJCAI)
83. Yang Y, Yang Y, Huang Z, Shen HT, Nie F (2011) Tag localization with spatial correlations and joint group sparsity. In: Proceedings of computer vision and pattern recognition (CVPR), pp 881–888
84. Yang Y, Wu F, Xu D, Zhuang Y, Chia LT (2010) Cross-media retrieval using query dependent search methods. Pattern Recognit 43(8):2927–2936
85. Yu H, Bennamoun M (2006) 1D-PCA, 2D-PCA to nD-PCA. In: Proceedings of international conference on pattern recognition (ICPR), pp 181–184
86. Yuan Y, Wu F, Zhuang Y, Shao J (2011) Image annotation by composite kernel learning with group structure. In: Proceedings of ACM conference on multimedia (ACM MM)
87. Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B (Methodological) 68(1):49–67
88. Yuan M, Lin Y (2007) On the non-negative garrotte estimator. J R Stat Soc Ser B (Statistical Methodology) 69(2):143–161
89. Zhang CH (2010) Nearly unbiased variable selection under minimax concave penalty. Ann Stat 38(2):894–942
90. Zhao P, Yu B (2006) On model selection consistency of Lasso. J Mach Learn Res 7:2541–2563
91. Zhou D, Huang J, Schölkopf B (2006) Learning with hypergraphs: clustering, classification, and embedding. In: Advances in neural information processing systems (NIPS), pp 1601–1608
92. Zhu J, Xing EP (2011) Sparse topical coding. In: Proceedings of the 27th international conference on uncertainty in artificial intelligence (UAI)
93. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Ser B (Statistical Methodology) 67(2):301–320
94. Zou H, Hastie T, Tibshirani R (2006) Sparse principal component analysis. J Comput Graph Stat 15(2):265–286
95. Zou H (2006) The adaptive lasso and its oracle properties. J Am Stat Assoc 101(476):1418–1429