

The Fisher-Markov Selector: Fast Selecting Maximally Separable Feature Subset for Multi-Class Classification with Applications to High-Dimensional Data

Qiang Cheng*, Hongbo Zhou, and Jie Cheng

Abstract

Selecting features for multi-class classification is a critically important task for pattern recognition and machine learning applications. Especially challenging is to select an optimal subset of features from high-dimensional data, which typically have much more variables than observations and contain significant noise, missing components, or outliers. Existing methods either cannot handle high-dimensional data efficiently or scalably, or can only obtain local optimum instead of global optimum.

Toward the selection of the globally optimal subset of features efficiently, we introduce a new selector - which we call the Fisher-Markov selector - to identify those features that are the most useful in describing essential differences among the possible groups. Particularly, in this paper, we present a way to represent essential discriminating characteristics together with the sparsity as an optimization objective. With properly identified measures for the sparseness and discriminativeness in possibly high-dimensional settings, we take a systematic approach for optimizing the measures to choose the best feature subset. We use Markov random field optimization techniques to solve the formulated objective functions for simultaneous feature selection.

Our results are non-combinatorial, and they can achieve the exact global optimum of the objective function for some special kernels. The method is fast; particularly, it can be linear in the number of features and quadratic in the number of observations. We apply our procedure to a variety of real-world

The first two authors are with the Computer Science Department, and the third author is with the Electrical and Computer Engineering Department, Southern Illinois University, Carbondale, IL 62901. H. Zhou contributed to the experiment part of this work and J. Cheng contributed to the formulation part. *Contact information: qcheng@cs.siu.edu, Faner Hall, Room 2140, Mail Code 4511, 1000 Faner Drive, Carbondale, IL 62901.

data, including mid-dimensional optical handwritten digit dataset and high-dimensional microarray gene expression datasets. The effectiveness of our method is confirmed by experimental results.

In pattern recognition and from a model selection viewpoint, our procedure says that it is possible to select the most discriminating subset of variables, by solving a very simple unconstrained objective function, which in fact can be obtained with an explicit expression.

Index Terms

Classification, feature subset selection, Fisher's linear discriminant analysis, high dimensional data, kernel, Markov random field.

I. INTRODUCTION

In many important pattern recognition and knowledge discovery tasks, the number of variables or features is high; oftentimes, it may be much higher than the number of observations. Microarray data for instance, often involve thousands of genes, while the number of assays is of the order of hundreds or less. In neurodevelopmental studies with functional MRI (fMRI) images, the number of experiments for the subjects is limited whereas each fMRI image may have tens of thousands of features. For these data, feature selection is often used for dimensionality reduction. It aims at evaluating the importance of each feature and identifying the most important ones, which turns out to be critical to reliable parameter estimation, underlying group structure identification, and classification. With a reduced dimension, the classifier can be more resilient to data overfitting. By eliminating irrelevant features and noise, the time and space efficiencies become significantly better.

Feature selection has several challenges in facing the following problems:

- 1) A small sample size with a large number of features. This has been acknowledged as an important challenge in contemporary statistics and pattern recognition. Selecting the best subset of features is, in general, of combinatorial complexity. Classifying high-dimensional data without dimensionality reduction is difficult and time consuming, if not impossible. The irrelevant variables make no contribution to the classification accuracy of any classifier but, rather, have adverse effects on the performance. With too many irrelevant variables, the classification performance would degrade into that of random guesses, as empirically observed in [1] [2], and theoretically shown in [3].
- 2) Linearly nonseparable classes. These cases can be difficult for many feature selectors. In a high-dimensional space, the difficulty may be alleviated thanks to Cover's theorem [4].

- 3) Noisy features. High-dimensional data usually have significant noise. The noise or irrelevant components can easily lead to data overfitting, especially for any data with a large number of features but a small sample size. The main reason is because, for high-dimensional data, it is easy to find some noise components which may be different for each class given a small number of observations; however, they may not correctly represent the importance of features.

In this paper, we study choosing a desirable subset of features for supervised classification. Even though our main focus is on high-dimensional data, the proposed Fisher-Markov selector can be applied to general data with arbitrary dimensions, as shown with experiments in Section V. The Fisher-Markov selector can be regarded as a filter method, meaning the selected features can be properly used by any available classifiers.

In the same spirit as Fisher's discriminant analysis (FDA) [5], we use the class separability as a criterion to select the best subset of features; that is, we maximize the between-class distance while minimizing the within-class distance - hence the first part of the name. The way we use the class separability, however, is different than the FDA (or its kernelized version KFDA), either the linear discriminant analysis (LDA) or quadratic discriminant analysis (QDA). The FDA is used for extracting features through feature vector transformation (the KFDA does this in a kernel space); in contrast, the Fisher-Markov selector is a general variable (feature) selection strategy. The goal, formulations and optimization techniques of the Fisher-Markov selectors are completely different from those of the FDA (or KFDA). In defining the class separability, we incorporate kernel tricks to map each original input to a higher-dimensional kernel space. By doing so, we may define our feature selection problem in a broad setting, with the aim to subjugate the difficulty of the above-mentioned second challenge when the dimensionality of the input space is insufficiently high. To reduce the adverse effect from noise components, the Fisher-Markov selector employs an l_0 -norm to penalize the complexity. As experimentally shown in Section V, the overfitting problem can be overcome neatly with a principled method. There is of course a huge literature on feature selection, and it is well known that the best subset selection is generally of combinatorial complexity, with which it is literally infeasible for high-dimensional data. By taking advantage of some special kernel functions, the Fisher-Markov selector boils down the general subset selection to seeking an optimal configuration on the Markov random fields (MRF). For the MRF problem, the Fisher-Markov selector then constructs efficient ways to achieve the exact global optimization of the class separability - hence the second part of the name.

In this paper, scalars and vectors are denoted by lower-case letters, and a vector is clearly defined when it first appears. Matrices are denoted by capital letters. Script letters denote spaces or classes, especially,

\mathcal{R}^k denotes the k -dimensional space of real numbers. The operation $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product of two vectors. The transpose of a vector x is denoted by x^T , and its l_q -norm by $\|x\|_q$, $0 \leq q \leq \infty$. We use the notation of $(\cdot)_i$ to denote the i th element of a vector, or $(\cdot)_{ij}$ the ij -th element of a matrix. We use 0 to denote a scalar zero or a vector of zeros whose dimension is clear in the context.

The paper is organized in the following way: We begin by discussing related work in Section II. We formulate a class of new selectors in Section III, and explicitly construct the Fisher-Markov selector in Section IV. Section V introduces experiments demonstrating that our approach is effective in practical applications. And finally, Section VI summarizes our findings and their consequences for variable selection and supervised classification.

II. RELATED WORK

There is a huge literature on variable or feature selection, and many procedures motivated by a wide array of criteria have been proposed over the years. We review in the following only the methods that are closely relevant to our work.

Subset selection methods in discriminant analysis for pattern recognition were initially studied by Dixon and Massey, and Kendall [6] [7]. These and other early work focused on normal-based linear discriminant function (NLDF) for two groups under the assumption of normality and homoscedasticity; see Devijver and Kittler [8], Fukunaga [9], and McLachlan [10] for excellent discussions. Traditional methods considered various selection criteria and optimization techniques. For example, Fu devised mathematical programming approaches to feature selection [11]. The Kullback-Leibler divergence (KLD) [12] and the Bhattacharyya distance [13] were used as selection criteria. Narendra and Fukunaga described the branch and bound algorithm that maximizes a criterion function over all possible subsets of a given size [14]. Rissanen characterized the stochastic complexity of the data by constructing the minimum description length (MDL) principle, and he proposed two methods for variable selection with the MDL principle [15]. These traditional methods are able to handle only large-sample cases, where the number of observations is much larger than that of features. For high-dimensional data, usually they are inapplicable. An exception is Kira and Rendall's Relief [16] which weighed features to maximize a so-called margin. It can be used for high-dimensional data though the results are far from optimal.

Lasso method was proposed by Tibshirani, which is an l_1 penalized squared error minimization problem [17]. It has been shown to be effective in estimating sparse features. More recently, there has been a thrust of research activities in sparse representation of signals - among which are Donoho and Elad [18], Donoho [19], and Candes, Romberg, and Tao [20]. These sparse representation techniques have

a close relationship with feature selection, especially because the l_1 minimization may approximate the sparsest representation of the signal and discard irrelevant variables. In this vein, Candes and Tao have proposed the Dantzig selector that is suitable for high-dimensional data by employing the Lasso-type l_1 regularization technique [21]. It appears to be ineffective for linearly nonseparable data. An interesting model-based approach to feature selection has been constructed by McLachlan, Bean, and Peel for high-dimensional data [22]. Peng, Long, and Ding have proposed a method, mRMR, by combining max-relevance and min-redundancy criteria [23]. Their method uses a greedy optimization of mutual information for univariate feature selection. It handles low- to mid- dimensional data effectively. Wang has presented a feature selection method using a kernel class separability, a lower bound of which is optimized with a gradient descent-type of optimization technique [24]. This method finds local optimum(s) iteratively and is applicable only to stationary kernels.

The above-mentioned work chooses a subset of feature variables by optimizing feature selection criteria, including the probability of error, KLD, NLDF class separations, Bhattacharyya distance, MDL, etc. They select features independently of the subsequent classifiers, which in the literature are referred to as filter methods. Also often used are wrapper methods, which fix a classifier and choose features to maximize the corresponding classification performance. Weston, *et al.*'s method, and Guyon *et al.*'s recursive feature elimination (RFE), e.g., choose features to minimize the classification errors of a support vector machine (SVM) [25] [26]. Wrapper methods choose features in favor of a particular classifier, and they incur higher computational complexities by evaluating the classifier accuracy repeatedly. The feature selector proposed in this paper is mainly a filter method and its resultant subset of features is proper for any classifiers.

Optimizing a variable selection criterion or classifier performance needs a searching strategy or algorithm to search through the space of feature subsets. Globally optimal (in the sense of a chosen criterion) multivariate selection methods usually have a combinatorial complexity [27]. A notable exception is Narendra and Fukunaga's Branch and Bound method whose optimality, however, can only be guaranteed under a monotonicity condition [14]. To alleviate the difficulty, many practically useful, though suboptimal, search strategies have been considered including random search [28] [29], sequential search, etc. The sequential search methods select features one at a time and they are widely used for their efficiency [25] [26] [30]. The selector presented in this paper is an efficient multivariate selection method that may attain or approximate the global optimum of the identified discrimination measure. For instance, the resulting LFS (see Section IV) has a complexity linear in the number of features and quadratic in the number of observations.

The feature selection problem with more than two groups or classes is usually more difficult than those with two groups. The Fisher-Markov selector is built directly for multiple-class classification; that is, the number of classes is greater than or equal to two. Also, there is a close tie between discriminant analysis and multiple regression. Fowlkes, Gnanadesikan, and Kettenring [31] considered the variable selection in three contexts: multiple regression, discriminant analysis, and clustering. This paper only focuses on the problem of feature selection for supervised classification.

III. FORMULATION OF FISHER - MARKOV SELECTOR

In many research fields, feature selection is needed for supervised classification. Specifically, with training data $\{(x_k, y_k)\}_{k=1}^n$, where $x_k \in \mathcal{R}^p$ are p -dimensional feature vectors, and $y_k \in \{\omega_1, \dots, \omega_g\}$ are class labels, the most important features are to be chosen for the most discriminative representation of a multiple-class classification with g classes $\mathcal{C}_i, i = 1, \dots, g$. Each class \mathcal{C}_i has n_i observations. Given a set of new test observations, the selected features are used to predict the (unknown) class label for each observation. This paper focuses on an important situation in which p is much greater than n , though as readers will see later in Sections IV and V the proposed method is applicable to general data including the large-sample case with p less than n .

A. Selection In the Spirit of Fisher

Feature selection methods, especially those independent of the classifiers, need to set a selection criterion. Often used are the KLD [12], Bhattacharyya distance [13], mutual information [23], etc.

Fisher's FDA is well known [8] - [10]. It discriminates the classes by projecting high-dimensional input data onto low-dimensional subspaces with linear transformations, with the goal of maximizing inter-class variations while minimizing intra-class variations. The LDA is simple, theoretically optimal (in a certain sense [10]) and routinely used in practice; for example, in face recognition, bankruptcy prediction, etc. [32]. When the number of observations is insufficient, however, it suffers from a rank deficiency or numerical singularity problem. Some methods have been proposed to handle such numerical difficulty, including Bickel and Levina's naive Bayesian method [33]. Even though Fisher's discriminant analysis performs no feature selection but projection onto subspaces through linear transformations [8] - [10], the spirit of maximizing the class separations can be exploited for the purpose of feature selection. This spirit induces the Fisher-Markov selector's class separability measure (see III-B).

In the spirit of Fisher's class separation, to construct a *maximally separable*, and at the same time, *practically useful* feature selector which is applicable to general (including, e.g., both large-sample and

high-dimensional) datasets, two main questions are yet to be addressed:

- (Q1) Can the selector manage to handle linearly nonseparable and highly noisy datasets with many irrelevant features?
- (Q2) Can the selector's class separation measure be optimized in an algorithmically efficient way?

The first question (Q1) can be approached by using kernel ideas, inspired by [34] - [37] [24], and by using an l_0 -norm. The use of the l_0 -norm is to offer a model selection capability that determines how many features are to be chosen. In response to (Q2), the feature selector must be constructed to admit efficient optimizations. A group of features may lead to the best classification results although each individual feature may not be the best one. To capture the group effect, it appears necessary to seek multivariate features simultaneously instead of sequentially or in a univariate fashion. The multivariate selection is essentially a combinatorial optimization problem, for which exhaustive search becomes infeasible for a large number of features. To overcome this computational difficulty and for simultaneous selection, this paper exploits MRF-based optimization techniques to produce multivariate and fast algorithms.

B. Class Separability Measure

In the original input (or measurement) space, denote the within-, between- (or inter-) class, and total scatter matrices by S_w , S_b , and S_t :

$$S_w = \frac{1}{n} \sum_{j=1}^g \sum_{i=1}^{n_j} (x_i^{(j)} - \mu_j)(x_i^{(j)} - \mu_j)^T, \quad (1)$$

$$S_b = \frac{1}{n} \sum_{j=1}^g n_j (\mu_j - \mu)(\mu_j - \mu)^T, \quad (2)$$

$$S_t = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)(x_i - \mu)^T = S_w + S_b, \quad (3)$$

where $x_i^{(j)}$ denotes the i th observation in class \mathcal{C}_j , and μ_j and μ are sample means for class \mathcal{C}_j and the whole data set respectively; i.e., $\mu_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_i^{(j)}$, and $\mu = \frac{1}{n} \sum_{i=1}^n x_i$.

To measure the class separations, the traces or determinants of these scatter matrices may be used [9]. In this paper, to form a class separation criterion, the traces of these scatter matrices are employed. The above scatter matrices measure the data scattering using means and variances, based on an implicit assumption that the data in each class follow a normal distribution. This assumption works well for large-sample data. To address the linearly nonseparable problem (Q1), kernels will be incorporated into the above scatter matrices. Let $\phi(\cdot)$ be a possibly nonlinear mapping from the input space \mathcal{R}^p to a kernel

space \mathcal{R}^D :

$$\phi : \mathcal{R}^p \rightarrow \mathcal{R}^D \quad (4)$$

$$\phi(x) \rightarrow z. \quad (5)$$

With the kernel trick [35], the inner product in the kernel space becomes $\langle \phi(x_1), \phi(x_2) \rangle = k(x_1, x_2)$, with $k(\cdot, \cdot)$ being some kernel function. After mapping into the kernel space, the scatter matrices can be calculated in the kernel space, where the within-, between-class, and total scatter matrices are denoted by \tilde{S}_w , \tilde{S}_b , and \tilde{S}_t . By simple algebra it is not hard to find the traces of \tilde{S}_w , \tilde{S}_b , and \tilde{S}_t to be

$$\text{Tr}(\tilde{S}_w) = \frac{1}{n} \text{Tr}(K) - \frac{1}{n} \sum_{i=1}^g \frac{1}{n_i} \text{Sum}(K^{(i)}), \quad (6)$$

$$\text{Tr}(\tilde{S}_b) = \frac{1}{n} \sum_{i=1}^g \frac{1}{n_i} \text{Sum}(K^{(i)}) - \frac{1}{n^2} \text{Sum}(K), \quad (7)$$

$$\text{Tr}(\tilde{S}_t) = \frac{1}{n} \text{Tr}(K) - \frac{1}{n^2} \text{Sum}(K), \quad (8)$$

where the operators $\text{Sum}(\cdot)$ and $\text{Tr}(\cdot)$ calculate, respectively, the summation of all elements and the trace of a matrix, and K and $K^{(i)}$ are size $n \times n$ and $n_i \times n_i$ matrices defined by

$$\{K\}_{kl} = k(x_k, x_l), \quad \{K^{(i)}\}_{uv} = k(x_u^{(i)}, x_v^{(i)}), \quad (9)$$

for $k, l \in \{1, \dots, n\}$, $u, v \in \{1, \dots, n_i\}$, and $i = 1, \dots, g$.

The feature selector is denoted by $\alpha = [\alpha_1, \dots, \alpha_p]^T \in \{0, 1\}^p$ with $\alpha_k = 1$ indicating the k th feature is chosen or 0 not-chosen, $k = 1, \dots, p$. The selected features from a feature vector x are given by

$$x(\alpha) = x \odot \alpha, \quad (10)$$

where the operator \odot represents the Hadamard product (also known as the Schur product) which is an entrywise product. With feature selection, K and $K^{(i)}$ become functions of α :

$$\{K(\alpha)\}_{kl} = k(x_k \odot \alpha, x_l \odot \alpha), \quad \{K^{(i)}(\alpha)\}_{uv} = k(x_u^{(i)} \odot \alpha, x_v^{(i)} \odot \alpha), \quad (11)$$

for $k, l \in \{1, \dots, n\}$, $u, v \in \{1, \dots, n_i\}$, and $i = 1, \dots, g$. Hence so do the traces of the scatter matrices, which are denoted by $\text{Tr}(\tilde{S}_w)(\alpha)$, $\text{Tr}(\tilde{S}_b)(\alpha)$, and $\text{Tr}(\tilde{S}_t)(\alpha)$.

The goal of the feature selection is to maximize the class separations for the most discriminative capability of the variables. In the spirit of Fisher, we formulate the following unconstrained optimization for class separability:

$$\text{argmax}_{\alpha \in \{0, 1\}^p} \text{Tr}(\tilde{S}_b)(\alpha) - \gamma \text{Tr}(\tilde{S}_t)(\alpha), \quad (12)$$

where γ is a free parameter whose suitable range of values will be discussed in Section IV. Now we first consider the effect of the signs of γ .

When $\gamma > 0$, by using the Lagrangian multiplier technique [38], it is noted that the unconstrained optimization is equivalent to the following constrained optimization problem:

$$\operatorname{argmax}_{\alpha \in \{0,1\}^p} \operatorname{Tr}(\tilde{S}_b)(\alpha) \quad (13)$$

$$\text{subject to } \operatorname{Tr}(\tilde{S}_t)(\alpha) \leq \text{constant}. \quad (14)$$

It maximizes the inter-class scattering while keeping the total scattering bounded (so the intra-class scattering is automatically minimized); nonetheless, it may be sensitive to spurious features of the data in high-dimensional case.

When $\gamma \leq 0$, the unconstrained optimization in Eq. (12) actually maximizes not only the discriminativeness but the expressiveness jointly. For high-dimensional data where $p \gg n$, there may be many irrelevant or highly noisy components, among which it is easy to find some spurious features that discriminate different groups quite well; for subsequent classifications, however, these spurious features are only misleading. To avoid degrading the generalization accuracy, it is imperative that we avoid spurious features and select those truly discriminative variables that represent the essential information or structure. That is, we should preserve the so-called expressive power of the data. A classical way for obtaining the expressiveness of the data is to use the principal component analysis (PCA), which maximizes the quotient of $\frac{v^T S_t v}{v^T v}$ over nonzero $v \in \mathcal{R}^p$ - Kernel PCA (KPCA) does this by using \tilde{S}_t in the kernel space. Motivated by the KPCA, we use $\operatorname{Tr}(\tilde{S}_t)(\alpha)$ to describe the expressiveness, and we maximize it through our feature selector α . The normalization by $v^T v = \|v\|_2^2$ in the PCA will also have a regularization counterpart in our formulation (which will be $\|\alpha\|_0$ and introduced subsequently). Now the objective function inspired by the KFDA is to maximize $\operatorname{Tr}(\tilde{S}_b)(\alpha) - \gamma_1 \operatorname{Tr}(\tilde{S}_t)(\alpha)$, where $\gamma_1 > 0$ may be a Lagrangian multiplier; and the objective function inspired by the KPDA is to maximize $\operatorname{Tr}(\tilde{S}_t)(\alpha)$. To maximize both objectives jointly, we combine them using some $\gamma_2 > 0$, and we have $\operatorname{Tr}(\tilde{S}_b)(\alpha) - (\gamma_1 - \gamma_2) \operatorname{Tr}(\tilde{S}_t)(\alpha)$. Letting $\gamma = \gamma_1 - \gamma_2$ thus yields Eq. (12). Hence, when $\gamma \leq 0$, by using a nonnegative linear combination of $\operatorname{Tr}(\tilde{S}_b)(\alpha)$ and $\operatorname{Tr}(\tilde{S}_t)(\alpha)$, our feature selector maximizes simultaneously the discriminativeness and expressiveness, useful particularly for high-dimensional data. Our analytical results and experiments have found that $\gamma \leq 0$ often leads to a superior classification performance that is fairly stable for a range of γ values.

Now we are in a position to contrast to the LDA or KFDA to highlight the differences. The LDA computes a linear transformation, often in the form of a matrix, to project onto a subspace of the original

space (KFDA does this in the kernel space), whereas our goal is to select variables using $\alpha \in \{0, 1\}^p$, which is a highly nonlinear process. In terms of optimization strategies, the LDA or KFDA often uses generalized eigenvalue techniques [38], whereas our selectors make use of the MRF techniques that will be specified in Section IV.

In the presence of a large number of irrelevant or noise variables, jointly optimizing discriminativeness and expressiveness can help but it alone is still insufficient: The classifier's performance may still be degraded as observed empirically by Fidler, *et al.* [39]. The reason is that overfitting becomes overly easy. To avoid detrimental effect of overfitting, model selection has to be explicitly considered. There is a huge literature on model selection, and many procedures have been proposed such as Akaike's AIC [40], Schwarz's BIC (also called Schwarz Criterion) [41], and Foster and George's canonical selection procedure [42]. Foster and George's procedure makes use of an l_0 norm, which is the number of nonzero elements of a predictor for multiple regression problems. The l_0 norm has also been used by Weston *et al.* [43]. They formed essentially an l_0 -norm SVM, rather than the l_2 - [35] [37] or l_1 -norm SVM. The optimization of the l_0 -norm SVM, however, is not convex and can only be performed approximately. To handle the noise robustness issue raised in (Q1), the l_0 norm is utilized in our feature selector to regularize the discrimination measure. Hence we obtain the following feature selection criterion,

$$\operatorname{argmax}_{\alpha \in \{0,1\}^p} \{ \operatorname{Tr}(\tilde{S}_b)(\alpha) - \gamma \operatorname{Tr}(\tilde{S}_t)(\alpha) - \beta \|\alpha\|_0 \}, \quad (15)$$

where β is a constant factor for the l_0 norm whose feasible values will be discussed in Section IV.

More explicitly, plugging the expressions for the scatter matrices into Eq. (15), we have the following feature selector,

$$(FS) \quad \operatorname{argmax}_{\alpha \in \{0,1\}^p} \frac{1}{n} \left[\sum_{i=1}^g \frac{1}{n_i} \operatorname{Sum}(K^{(i)}(\alpha)) - \gamma \operatorname{Tr}(K(\alpha)) + \frac{\gamma - 1}{n} \operatorname{Sum}(K(\alpha)) \right] - \beta \|\alpha\|_0. \quad (16)$$

The formulated feature selection process (FS), in general, is a combinatorial optimization problem for many kernel functions, which is computationally feasible only when p is not large. For some special kernel functions, surprisingly, global optimum can be obtained efficiently for large p . To address the computational efficiency issue raised in (Q2) and to obtain global optimum solutions, we consider these special kernel functions for selecting maximally separable feature subset via (FS).

IV. FEATURE SUBSET SELECTION USING FISHER-MARKOV SELECTOR

To obtain efficient and global optimization for (FS) for large p , some special kernels will be considered including particularly the polynomial kernels. We consider the polynomial kernel [35] - [37]

$$k(x_1, x_2) = (1 + \langle x_1, x_2 \rangle)^d, \quad (17)$$

where d is a degree parameter; or alternatively,

$$k'(x_1, x_2) = (\langle x_1, x_2 \rangle)^d. \quad (18)$$

A. Fisher-Markov Selector with Linear Polynomial Kernel

Incorporating the feature selector α , with $d = 1$, the kernel becomes

$$k_1(x_1, x_2)(\alpha) = 1 + \langle x_1 \odot \alpha, x_2 \odot \alpha \rangle = 1 + \sum_{i=1}^p x_{1i}x_{2i}\alpha_i; \quad \text{or,}$$

$$k'_1(x_1, x_2)(\alpha) = \sum_{i=1}^p x_{1i}x_{2i}\alpha_i.$$

Now plugging $k_1(\cdot, \cdot)$ (or $k'_1(\cdot, \cdot)$) into Eq. (15) or (16), and defining θ_j to be

$$\theta_j = \frac{1}{n} \sum_{i=1}^g \frac{1}{n_i} \sum_{u,v=1}^{n_i} x_{uj}^{(i)} x_{vj}^{(i)} - \frac{\gamma}{n} \sum_{i=1}^n x_{ij}^2 + \frac{\gamma-1}{n^2} \sum_{u,v=1}^n x_{uj} x_{vj}, \quad (19)$$

we have the following special case of (FS):

Proposition IV.1 *With a linear polynomial kernel of $d = 1$, and θ_j defined in Eq. (19), the Fisher-Markov feature selector (FS), which maximizes the class separations, turns out to be*

$$(LFS) \quad \operatorname{argmax}_{\alpha \in \{0,1\}^p} \sum_{j=1}^p (\theta_j - \beta) \alpha_j.$$

For given β and γ , the feature selector (LFS) has an optimum $\alpha^* \in \{0,1\}^p$ such that

$$\theta_j > \beta \iff \alpha_j^* = 1. \quad (20)$$

The computational complexity for obtaining α^* with the given training data is $O(n^2p)$.

The (LFS) is a special case of Markov random field (MRF) problem with binary labels, in which there is no pairwise interaction term. For this particular maximization problem, α^* is obviously the unique global optimum solution. When β and γ are given, the Fisher-Markov selector has a computational complexity linear in p and quadratic in n . For high-dimensional data with $p \gg n$, for instance gene expression levels, the number of features may be thousands or tens of thousands while the size of a training sample is at the order of hundred. The (LFS) is fast in terms of the number of features.

The coefficient θ_j indicates the importance of the j th feature. The greater the value of θ_j , the more important the corresponding feature. As observed in Section V, the Fisher-Markov selector is pretty insensitive to the value of γ as long as it is in the feasible range of values that will be discussed later. The value of β , however, is important in determining the number of chosen features. It is clear from the

above Proposition IV.1 that the regularization factor β turns out to be a global threshold. This paper uses cross validation during training to set a proper value for β .

Now some performance analysis on the (LFS) is in order. It is usually true that the observations x_u and x_v are statistically independent when $u \neq v$. Let's first consider simple cases where each feature variable follows a Gaussian distribution. If the j th variable contains no class discrimination information, it behaves like random noise across the classes. That is, $x_{uj}^{(i)}$ are independently and identically distributed (i.i.d.) Gaussian white noise, $x_{uj}^{(i)} \sim i.i.d. N(0, \sigma_j^2)$ for any $u = 1, \dots, n_i$ and $i = 1, \dots, g$. On the other hand, if the j th variable is an important feature for classification, it does contain class discrimination information, and thus we model it by $x_{uj}^{(i)} \sim i.i.d. N(\mu_{i,j}, \sigma_j^2)$. Then we have the following property of the (LFS):

Proposition IV.2 *Using the (LFS), if the j th variable contains no class discrimination information, then*

$$\theta_j \sim \frac{\sigma_j^2}{n} [\chi_g^2 - \gamma \chi_n^2 + (\gamma - 1) \chi_1^2]; \quad (21)$$

if the j th variable is an important feature for classification, then

$$\begin{aligned} \theta_j \sim & (1 - \gamma) \left[\sum_{i=1}^g \lambda_i \mu_{i,j}^2 - \left(\sum_{i=1}^g \lambda_i \mu_{i,j} \right)^2 \right] \\ & + 2\sigma_j \sqrt{\sum_{i=1}^g \left(\frac{\lambda_i}{n} + \gamma^2 \lambda_i^2 \right) \mu_{i,j}^2 + \frac{(\gamma - 1)^2}{n} \left(\sum_{i=1}^g \lambda_i \mu_{i,j} \right)^2} G_1 + \frac{\sigma_j^2}{n} [\chi_g^2 - \gamma \chi_n^2 + (\gamma - 1) \chi_1^2]; \quad (22) \end{aligned}$$

where χ_k^2 is a Chi-square distribution with k degrees of freedom (d.o.g.), G_1 is a standard normal distribution, and λ_i is defined to be n_i/n , thus $\sum_{i=1}^g \lambda_i = 1$.

Proof: The proof is obtained from direct calculations of Gaussian random variables. When the j th feature variable contains no discrimination information, $\frac{1}{\sqrt{n_i}} \sum_{u=1}^{n_i} x_{uj}^{(i)}$ is Gaussian distributed with mean zero and variance σ_j^2 . It follows that $\left(\frac{1}{\sqrt{n_i}} \sum_{u=1}^{n_i} x_{uj}^{(i)} \right)^2$ has a distribution of $\sigma_j^2 \chi_1^2$. Thus, the first additive term in θ_j has a distribution of $\frac{\sigma_j^2}{n} \chi_g^2$. Similarly, we can calculate the second and third terms in θ_j and hence Eq. (21).

When the j th variable is an important feature for classification, $\frac{1}{\sqrt{n_i}} \sum_{u=1}^{n_i} x_{uj}^{(i)} = \sqrt{n_i} \mu_{i,j} + \sigma_j G_1$. Then we have the first term in θ_j as

$$\sum_{i=1}^g \lambda_i \left(\frac{1}{n_i} \sum_{u=1}^{n_i} x_{uj}^{(i)} \right)^2 = \sum_{i=1}^g \lambda_i (\mu_{i,j}^2 + \frac{2\sigma_j}{\sqrt{n_i}} G_1 + \frac{\sigma_j^2}{n_i} \chi_1^2) = \sum_{i=1}^g \lambda_i \mu_{i,j}^2 + 2 \frac{\sigma_j}{\sqrt{n}} \sqrt{\sum_{i=1}^g \lambda_i \mu_{i,j}^2} G_1 + \frac{\sigma_j^2}{n} \chi_g^2.$$

Similarly, we can derive the expression for the second and third terms in θ_j , and hence Eq. (22). \blacksquare

Though the above proposition is derived based on Gaussian distributions, the conclusion is still true for nonGaussian distributions in the asymptotic regime, even when the observations have certain correlations - e.g., with the Lindeberg-Feller condition [44], or when a stationary random sequence or a random field satisfies some mixing conditions [45] [46] - provided the central limit theorem (CLT) holds. When the total number of observations n becomes large and n_i/n tends to a constant λ_i , $\frac{1}{\sqrt{n_i}} \sum_{u=1}^{n_i} x_{uj}^{(i)}$ follows approximately a Gaussian distribution as a consequence of the CLT. And the rest of the proof of Proposition IV.2 remains the same.

As a consequence of the above proposition, the mean and variance of θ_j are, in the absence of discrimination information,

$$E(\theta_j) = -\frac{\sigma_j^2}{n}(\gamma n + 1 - g - \gamma); \quad \text{Var}(\theta_j) = \frac{2\sigma_j^4}{n^2}(g + n\gamma^2 + (\gamma - 1)^2); \quad (23)$$

and in the presence of discrimination information for classification,

$$E(\theta_j) = (1 - \gamma) \left[\sum_{i=1}^g \lambda_i \mu_{i,j}^2 - \left(\sum_{i=1}^g \lambda_i \mu_{i,j} \right)^2 \right] - \frac{\sigma_j^2}{n}(\gamma n + 1 - g - \gamma); \quad (24)$$

$$\text{Var}(\theta_j) = 4\sigma_j^2 \left[\sum_{i=1}^g \left(\frac{\lambda_i}{n} + \gamma^2 \lambda_i^2 \right) \mu_{i,j}^2 + \frac{(\gamma - 1)^2}{n} \left(\sum_{i=1}^g \lambda_i \mu_{i,j} \right)^2 \right] + \frac{2\sigma_j^4}{n^2}(g + n\gamma^2 + (\gamma - 1)^2). \quad (25)$$

A feature variable will be selected only when it contributes to distinguishing the classes. The means, in the absence or presence of discrimination information, have to be separated sufficiently far apart with respect to the variances. Specifically, we require that the difference between the means be larger than κ times the standard deviations, usually with $\kappa \geq 2$. This leads to a separation condition for the j th feature to be selected as stated in the following corollary:

Corollary IV.3 *A condition for the j th feature variable to be chosen by the (LFS) is that $|1 - \gamma| U$ is larger than both $\kappa \frac{\sqrt{2\sigma_j^2}}{n} \sqrt{g + n\gamma^2 + (\gamma - 1)^2}$ and $2\kappa\sigma_j \left[\frac{U}{n} + \gamma^2 W^2 + \frac{1+(\gamma-1)^2}{n} \left(\sum_{i=1}^g \lambda_i \mu_{i,j} \right)^2 + \frac{\sigma_j^2}{2n^2} (g + n\gamma^2 + (\gamma - 1)^2) \right]^{1/2}$, where $U = \left[\sum_{i=1}^g \lambda_i \mu_{i,j}^2 - \left(\sum_{i=1}^g \lambda_i \mu_{i,j} \right)^2 \right]$ and $W = \sqrt{\sum_{i=1}^g \lambda_i^2 \mu_{i,j}^2}$.*

In the above, the term U is always nonnegative due to the Cauchy-Schwartz inequality. Corollary IV.3 can be used to design a proper γ that separates well the discriminative from non-discriminative features.

In the case that $\sigma_j^4 \gamma^2 = o(n)$, $1 + (\gamma - 1)^2 \left(\sum_{i=1}^g \lambda_i \mu_{i,j} \right)^2 = o(n)$, the separation condition essentially becomes

$$|1 - \gamma| U > 2\kappa\sigma_j |\gamma| W. \quad (26)$$

Then the feasible range of γ values can be chosen to be $L < \gamma \leq U/(U + 2\kappa\sigma_j W)$ where L is a lower bound for γ with $L = U/(U - 2\kappa\sigma_j W) < 0$ if $U < 2\kappa\sigma_j W$; or $L = -\infty$ if $U \geq 2\kappa\sigma_j W$.

The above condition provides insights into the feasible range of γ values and why oftentimes a negative γ may lead to a better performance than a positive one, as pointed out in Section III. This condition may also guide us to choose proper values of β , which needs to be chosen to cut off from the feature distributions in the absence of class discriminative information. As empirically observed in our experiments, there is usually a large range of γ values satisfying the above condition, and the (LFS) is quite insensitive to γ values. This paper uses the cross validation to choose a suitable β given a classifier.

The procedures of the LFS are illustrated in Algorithm 1 with chosen γ and β .

Algorithm 1 Algorithm for LFS: Fisher-Markov Selector Using Polynomial Kernel with $d = 1$.

- 1: **Input:** A data matrix of training examples $[x_1, \dots, x_n] \in \mathcal{R}^{p \times n}$ for g classes, a vector of class labels of the training examples $y = [y_1, \dots, y_n]$, where $y_k \in \{\omega_1, \dots, \omega_g\}$, $k = 1, \dots, n$.
 - 2: Compute the Markov coefficients θ_j by Eq. (19).
 - 3: Solve the MRF maximization problem of (LFS) in Proposition IV.1 by Eq. (20).
 - 4: **Output:** Estimated feature selector of α^* .
-

B. Fisher-Markov Selector with Quadratic Polynomial Kernel

With $d = 2$, the polynomial kernel in Eq. (17) or (18) as a function of α becomes

$$k_2(x_1, x_2)(\alpha) = (1 + \langle x_1 \odot \alpha, x_2 \odot \alpha \rangle)^2 = (1 + \sum_{i=1}^p x_{1i} x_{2i} \alpha_i)^2, \quad \text{or} \quad (27)$$

$$k'_2(x_1, x_2)(\alpha) = (\sum_{i=1}^p x_{1i} x_{2i} \alpha_i)^2. \quad (28)$$

Now plugging $\frac{1}{2}k_2(x_1, x_2)(\alpha)$ (or $\frac{1}{2}k'_2(x_1, x_2)(\alpha)$) into Eq. (15) or (16), and defining θ_{jl} to be

$$\theta_{jl} = \frac{1}{n} \sum_{i=1}^g \frac{1}{n_i} \sum_{u,v=1}^{n_i} x_{uj}^{(i)} x_{vj}^{(i)} x_{ul}^{(i)} x_{vl}^{(i)} - \frac{\gamma}{n} \sum_{i=1}^n x_{ij}^2 x_{il}^2 + \frac{\gamma-1}{n^2} \sum_{u,v=1}^n x_{uj} x_{vj} x_{ul} x_{vl}, \quad 1 \leq j, l \leq p, \quad (29)$$

we obtain the following special case of (FS):

Proposition IV.4 *With a quadratic polynomial kernel $\frac{1}{2}k_2(x_1, x_2)(\alpha)$, the Fisher-Markov feature selector (FS) turns out to be*

$$(QFS) \quad \operatorname{argmax}_{\alpha \in \{0,1\}^p} \sum_{j=1}^p (\theta_j - \beta) \alpha_j + \frac{1}{2} \sum_{j=1}^p \sum_{l=1}^p \theta_{jl} \alpha_j \alpha_l.$$

Alternatively, with a quadratic polynomial kernel $\frac{1}{2}k'_2(x_1, x_2)(\alpha)$, it is

$$(QFS') \quad \operatorname{argmax}_{\alpha \in \{0,1\}^p} \sum_{j=1}^p -\beta \alpha_j + \frac{1}{2} \sum_{j=1}^p \sum_{l=1}^p \theta_{jl} \alpha_j \alpha_l.$$

Here θ_j is defined in Eq. (19) and θ_{jl} in Eq. (29). An exact global optimum can be obtained efficiently by optimizing the above binary-label MRF if and only if $\theta_{jl} \geq 0$ for all $1 \leq j < l \leq p$.

Clearly, the (QFS) is an extension to the (LFS) with additional pair-wise interaction terms, whereas the (QFS) and (QFS') have the same pair-wise interactions. Both (QFS) and (QFS') seek optimal configurations for binary-label MRFs. A few comments to optimizing the MRFs are in order now. Ever since Geman and Geman's seminal work [47], the MRFs have been widely used for vision, image processing and other tasks [48] [49]. It is generally hard to find a global optimum and the optimization can be even NP-hard [50]. In a seminal work [51] Picard and Ratliff computed a maximum-flow/minimum-cut on a graph to optimize the MRF energy. Ishikawa [52] showed that convexity of the difference of pairwise labels is a sufficient and necessary condition (even for more than two labels). Kolmogorov and Zabih [53] gave a sufficient and necessary condition for optimizing binary MRF with pairwise interactions via s-t minimum-cut. For the MRFs that do not satisfy the conditions, a significant progress has been made recently with mainly two types of methods to approximate the global optimum, those based on graph-cuts [52] [54], and those based on belief propagation (also called message passing) [55] [56].

Proof: For the problem of (QFS), it is easily seen that

$$\max_{\alpha \in \{0,1\}^p} \sum_{j=1}^p (\theta_j - \beta) \alpha_j + \frac{1}{2} \sum_{j=1}^p \sum_{l=1}^p \theta_{jl} \alpha_j \alpha_l \quad (30)$$

$$= \min_{\alpha \in \{0,1\}^p} \sum_{j=1}^p (\beta - \theta_j - 1/4 \sum_{k=1}^p (\theta_{jk} + \theta_{kj})) \alpha_j + \frac{1}{4} \sum_{j=1}^p \sum_{l=1}^p \theta_{jl} (\alpha_j - \alpha_l)^2. \quad (31)$$

For the problem of (QFS'), the same pair-wise terms are obtained. As each pair-wise interaction term is a convex function in terms of $(\alpha_j - \alpha_l)$ for binary MRF energy minimization problem, invoking Ishikawa's result [52], the conclusion immediately follows. Alternatively, this can be shown by using Kolmogorov and Zabih's result [53]. Defining $E^{jl}(\alpha_j, \alpha_l) = -\theta_{jl} \alpha_j \alpha_l$, then the sufficient and necessary condition for binary MRF energy minimization, which is the so-called regularity condition in [53], becomes $E^{jl}(0, 0) + E^{jl}(1, 1) \leq E^{jl}(0, 1) + E^{jl}(1, 0)$ for $j < l$; that is, $\theta_{jl} \geq 0$. The exact solution to the MRF can be sought using, e.g., the minimum cut algorithm specified in [53]. ■

In the following, we provide a sufficient and necessary condition, in terms of γ , for attaining exact global optimal solutions:

Corollary IV.5 *An equivalent condition, both sufficient and necessary, for guaranteeing an exact global*

optimum for solving (QFS) or (QFS') is

$$\gamma \leq \frac{1}{B_{jl}} \left[n \sum_{i=1}^g \frac{1}{n_i} \left(\sum_{u=1}^{n_i} x_{uj}^{(i)} x_{ul}^{(i)} \right)^2 - \left(\sum_{u=1}^n x_{uj} x_{ul} \right)^2 \right], \quad \text{for } 1 \leq j < l \leq p, \quad (32)$$

with a term B_{jl} defined as

$$B_{jl} = \left[n \sum_{i=1}^n x_{ij}^2 x_{il}^2 - \left(\sum_{u=1}^n x_{uj} x_{ul} \right)^2 \right].$$

If $B_{jl} = 0$ for some (j, l) , then $1/B_{jl}$ is defined to be $+\infty$.

In the above, $B_{jl} \geq 0$ for any pair of (j, l) due to the Cauchy-Schwartz inequality, and $B_{jl} = 0$ if and only if $x_{ij} x_{il}$ is a constant for all i , $1 \leq i \leq n$. Given a pair of indexes j, l , the complexity for calculating a single θ_{jl} is $O(n^2)$, and thus the complexity for calculating all θ_{jl} is $O(n^2 p^2)$. The MRF solver has a complexity only in terms of p . For traditional large-sample problems when $p \ll n$, (QFS) or (QFS') may be solved with an efficient MRF solver for an optimal (or near-optimal) subset of features. In high-dimensional settings where $p \gg n$, the complexity of solving (QFS) or (QFS') may be prohibitive and thus it is recommended that (LFS) be used instead.

In the case that $p < n$, when the pairwise interaction terms satisfy the sufficient and necessary condition, e.g., by choosing proper γ values, graph-cut method can be utilized to find the exact, globally optimal configuration to (QFS) or (QFS'). When there is at least a pairwise interaction term that does not satisfy the condition, however, it has been shown for binary MRFs the optimization is NP-hard [53]. In this case, globally optimum solutions to the MRFs can still be approximated efficiently by using either a graph-cut method or a message passing algorithm [55].

The procedures of the QFS are illustrated in Algorithm 1 with chosen γ and β .

Algorithm 2 Algorithm for QFS: Fisher-Markov Selector Using Polynomial Kernel with $d = 2$.

- 1: **Input:** A data matrix of training examples $[x_1, \dots, x_n] \in \mathcal{R}^{p \times n}$ for g classes, a vector of class labels of the training examples $y = [y_1, \dots, y_n]$, where $y_k \in \{\omega_1, \dots, \omega_g\}$, $k = 1, \dots, n$.
 - 2: Compute the Markov coefficients θ_j by Eq. (19) and θ_{jl} by Eq. (29).
 - 3: Solve the MRF maximization problem of (QFS) (or (QFS')) in Proposition (IV.4) by using an MRF solver, e.g., the one specified in [53].
 - 4: **Output:** Estimated feature selector of α^* .
-

In a similar vein to obtaining (LFS), (QFS) and (QFS'), as we apply the polynomial kernel with $d = 3$, an MRF with triple-wise interaction terms can be constructed. The resulting optimization problem is an

extension of (QFS) or (LFS). However, the computation of all θ_{jkl} would incur $O(n^2p^3)$ complexity, and thus it is not suitable for applications with high-dimensional data. The derivations are nonetheless similar and the sufficient and necessary conditions for exactly maximizing the triple-wise MRF can also be studied in a similar way to (QFS).

V. EXPERIMENTS

A. Experiment Design

To verify the validity and efficiency of our proposed methods, we perform experiments on a number of synthetic and real-world datasets. We make comprehensive comparisons with several state-of-the-art feature selection methods such as mRMR [23], SVM-RFE [26], l_0 -norm SVMs [43] and Random Forests [29]. We use standard RFE and l_0 -SVM provided by the Spider package¹, the Random Forest package of Breiman², as well as the mRMR package³. Four types of classifiers are used in our experiments, Naive Bayesian (NB), C4.5, SVM, and K-SVM (SVM with a RBF kernel). Both NB and C4.5 are the Weka implementation⁴. For SVM and K-SVM, we use multiple-class LibSVM (extended from binary classification with a one-versus-all approach) in the MatlabArsenal⁵ package. We will denote our LFS Fisher-Markov selector by MRF (d=1) or simply MRF, and QFS by MRF (d=2) in the following sections. Also, we will denote the l_0 -SVM feature selection method by L0.

Table I shows the characteristics of the datasets used in our experiments. The smallest real-world one is Iris [5], which has four features; the largest one, Prostate Cancer, has 12600 features. As a comprehensive evaluation, we consider various datasets: those from different applications ranging from optical handwritten digits recognition to gene analysis; those with different numbers of classes ranging from 2 to 10 classes; and those with a small training size and a large number of features. All data and programs in our experiments are available online⁶.

¹downloaded from <http://www.kyb.tuebingen.mpg.de/bs/people/spider/main.html>

²downloaded from <http://www.stat.berkeley.edu/users/breiman/RandomForests>

³downloaded from <http://penglab.janelia.org/proj/mRMR/>

⁴downloaded from <http://www.cs.waikato.ac.nz/ml/weka/>

⁵downloaded from <http://www.informedia.cs.cmu.edu/yanrong/MATLABArsenal/MATLABArsenal.htm>

⁶available at <http://www.cs.siu.edu/~qcheng/featureselection/>

TABLE I
CHARACTERISTICS OF THE DATASETS USED IN OUR EXPERIMENTS.

Dataset	Dim.	Class	Training	Test	Classifier	Comments
Synthetic-1	3	2	924	308	SVM	For visual inspection
Synthetic-2	52	2	750	250	SVM	Linearly non-separable
Iris	4	3	100	50	C4.5	Classic, small size
Wine	13	3	134	45	SVM, C4.5, NB	Classic, middle size
Optical Pen	64	10	1347	450	SVM, C4.5, NB	Classic, middle size, multi-class
Leukemia-S3	7070	2	54	18	SVM, C4.5, NB	High-dimensional
Nci9	9712	9	45	16	SVM, C4.5, NB	High-dimensional, multi-class
Prostate Cancer	12600	2	102	34	SVM, C4.5, NB	High-dimensional
Lung Cancer	12533	2	32	149	SVM, C4.5, NB	High-dimensional, small training set

B. Evaluation Criteria and Statistical Significance Test

For a given dataset and a given classifier, we adopt the following criteria to compare these five feature selection methods:

- 1) If a feature selection method can produce a smaller classification error rate than the other four methods, then its performance is better than the other four.
- 2) If more than one feature selection method can give rise to the same error rate, the one using the smallest number of features in achieving this error rate is the best.

For high dimensional datasets it is usually unnecessary to run through all features to arrive at meaningful conclusions. It is often sufficient to take the number of features from 1 to a maximum number of 60 (or 120) in our experiments. To measure the mean and variance of the performance, we take a multi-fold random split approach with the random splits usually repeated 20 times. Besides the comparison on any single dataset, we also conduct statistical significance tests over multiple datasets; on the recommendation from [57], we use the Wilcoxon signed rank test to compare MRF with the other four methods.

C. Parameter Setting for MRF

We conduct experiments to examine the parameter settings of the proposed MRF ($d=1$) method. Regardless of the setting for classifiers, there are two parameters β and γ for MRF. The parameter β controls the number of features to be selected; a proper value of β can be found during the training stage. To investigate the relationship between γ and the performance of the MRF method, we design

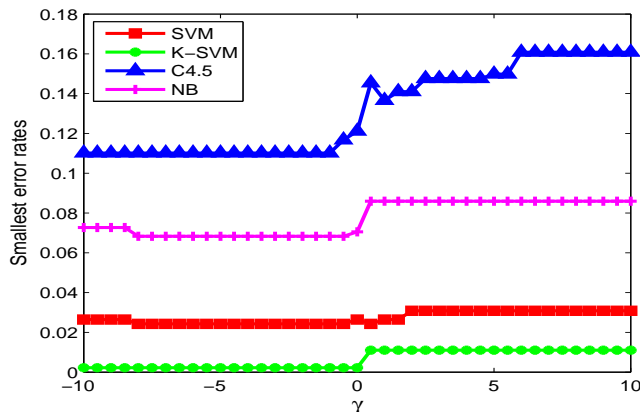


Fig. 1. The relationship between γ and the smallest classification error rates by using MRF method on Optical Pen dataset. Four different classifiers are used.

experiment using the Optical Pen dataset. We let γ vary from -10 through 10 with a step size of 0.5 . For each γ value, we run MRF with the number of selected features ranging from 1 up to 64 . The smallest classification errors attained by four classifiers are shown in Figure 1. It illustrates that the performance of MRF is quite stable for a large range of γ values as discussed in Section IV-A.

D. Summary of the Results

A summary of comparison results is reported here for these five methods with four classifiers; and further details will be provided subsequently. We have four tables: 1) Table II is a summary of performance evaluation for various feature selection methods for several datasets; 2) Table III is a summary of their computational efficiency; 3) Table IV summarizes the multi-fold random split performance evaluation; and 4) Table V shows the results from a statistical significance test over multiple datasets.

From Table II, we observe that MRF attains the smallest classification error rates (underlined results) over all six datasets; and especially, on high dimensional datasets such as Prostate Cancer and Lung Cancer, MRF outperforms the other four methods. Other methods also produce good results on some datasets, but they are more time consuming than MRF, as shown in Table III for time efficiency comparisons.

From the multi-fold random split test results in Table IV, we can see that MRF often outperforms the other four methods. On several datasets, other method(s) may also achieve the lowest average error rates (those underlined in Table IV); in this case, MRF has smaller variances.

Finally, we conduct the Wilcoxon signed rank test according to the recommendation from [57]. The rank test is between MRF and each of the other four methods based on the results in Tables II and IV.

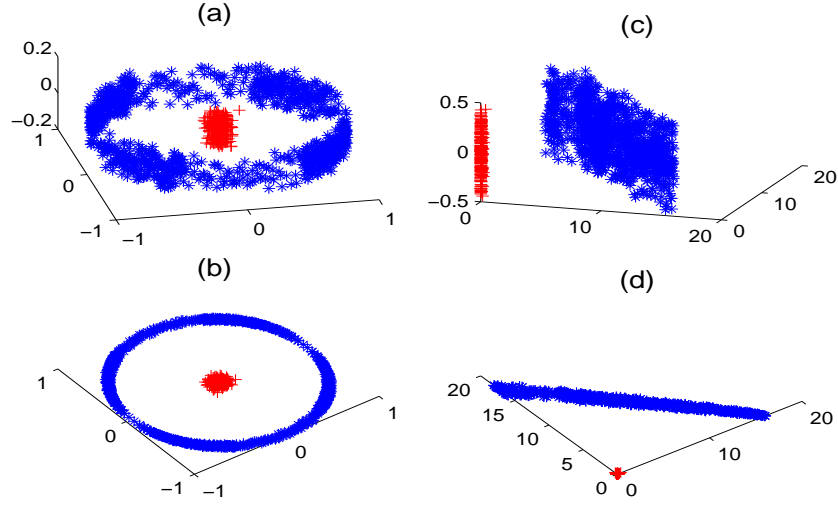


Fig. 2. Illustration of the data points of Synthetic-1 dataset in the original and transformed spaces. The red ‘+’ represents the points of the first class and blue ‘*’ of the second class. (a) is a 3-dimensional view of the original space, and (b) is the projection of (a) onto the X-Y plane. (c) is the data distribution in the transformed space using the kernel function given in Eq. (33), and (d) is the projection of (c) onto the X-Y plane.

The testing results in terms of p -values are shown in Table V; and it is evident the differences between MRF and the other four methods are statistically significant.

E. Experiments on Synthetic datasets

1) *Three-dimensional linearly non-separable problem:* The first synthetic dataset, Synthetic-1, is mainly designed for visual inspection in a 3-dimensional Euclidean space. On the X-Y plane, we first generate two classes of data. For the first class, we draw 982 points from a uniform distribution on a unit circle centered at $(0, 0)$; for the second class, we draw 250 points from a Gaussian distribution with a variance of 0.2 and a mean of $(0, 0)$. These two classes are linearly separable in a transformed space induced by the following kernel:

$$K(x, y) = (x_1^2 - 0.001x_1)(y_1^2 - 0.001y_1) + (x_2^2 - 0.001x_2)(y_2^2 - 0.001y_2). \quad (33)$$

Subsequently, for every two-dimensional point (x, y) , a third dimension z is added as noise from a uniform distribution in $[-0.2, 0.2]$. Figure 2 illustrates the data points in the original and transformed spaces.

TABLE II

PERFORMANCE EVALUATION ON SOME REAL-WORLD DATASETS. THE VALUES IN THE 4TH TO 8TH COLUMNS ARE THE BEST PERFORMANCES (THE SMALLEST ERROR RATE, IN PERCENTAGE) ACHIEVED BY MRF, RANDOM FOREST, mRMR, RFE AND L0 RESPECTIVELY BY USING DIFFERENT NUMBERS OF FEATURES FROM 1 UP TO $Max\#$ (GIVEN IN THE 2ND COLUMN). FOR EACH DATASET THE PERFORMANCE OF RANDOM FOREST IS AVERAGED OVER 10 RANDOM RUNS (THE OTHER FOUR METHODS ARE DETERMINISTIC). FOR EACH CLASSIFIER, THE BOLD RESULTS ARE THE BEST AMONG THESE FIVE METHODS. FOR EACH DATASET, THE UNDERLINED RESULTS ARE THE BEST OVER ALL CLASSIFIERS AND ALL FIVE METHODS.

Dataset	Max # selected features	Classifier	MRF	Random Forest	mRMR	RFE	L0
Wine	10	SVM	<u>0.00</u>	0.00	0.00	2.22%	2.22%
		K-SVM	2.22%	0.00	0.00	2.22%	2.22%
		NB	<u>0.00</u>	2.22%	0.00	2.22%	0.00
		C4.5	4.44%	5.11%	6.67%	6.67 %	8.89%
Optical Pen	40	SVM	2.42%	2.22%	2.64%	6.61%	7.49%
		K-SVM	<u>0.22%</u>	0.57%	0.66%	1.76%	2.64%
		NB	6.83%	7.22%	7.27%	14.32%	12.78%
		C4.5	11.67%	15.04%	13.66%	16.52 %	18.94%
Leukemia-S3	60	SVM	<u>0.00</u>	7.89%	0.00	5.26%	5.26%
		K-SVM	0.00	6.32%	5.26%	5.26%	5.26%
		NB	<u>0.00</u>	6.32%	5.26%	0.00	0.00
		C4.5	5.26%	11.58%	5.26%	5.26 %	5.26%
NCI9	60	SVM	<u>36.84%</u>	43.68%	<u>36.84%</u>	47.37%	42.11%
		K-SVM	<u>36.84%</u>	48.95%	<u>36.84%</u>	47.37%	<u>36.84%</u>
		NB	47.37%	56.84%	42.11%	57.89%	47.37%
		C4.5	52.63%	64.74%	52.63%	73.68 %	63.16%
Prostate Cancer	60	SVM	<u>0.00</u>	32.06%	61.77%	2.94%	8.82%
		K-SVM	26.47%	37.35%	73.53%	26.47%	26.47%
		NB	26.47%	54.12%	26.47%	26.47%	26.47%
		C4.5	23.53%	38.53%	26.47%	26.471 %	26.47%
Lung Cancer	60	SVM	<u>0.67%</u>	1.34%	1.34%	2.68%	3.36%
		K-SVM	6.04%	4.63%	13.42%	8.05%	10.07%
		NB	2.01%	2.82%	12.75%	8.05%	8.05%
		C4.5	9.40%	6.64%	24.83%	9.40 %	9.40%

TABLE III

COMPUTATIONAL EFFICIENCY EVALUATION (IN SECONDS) ON SEVERAL REAL-WORLD DATASETS. THE TIME COST OF RANDOM FOREST IS AVERAGED OVER 10 RUNS. IT SHOULD BE NOTED THAT mRMR IS NOT SCALABLE AND ITS TIME COST INCREASES QUICKLY WITH RESPECT TO THE NUMBER OF SELECTED FEATURES.

Dataset	Selected features	MRF	Random Forest	mRMR	RFE	L0
Leukemia-S3	60	0.09	2.84	5.22	0.57	0.34
NCI9	60	0.08	3.45	5.50	0.84	0.27
Prostate Cancer	60	0.45	10.25	294.62	1.62	0.48
Lung Cancer	60	0.09	2.26	170.21	1.31	0.20

We perform tests using MRF (d=1) and MRF (d=2) methods on this dataset. For each class, three quarters of the data are uniformly sampled for training and the rest are used for testing. Since the third dimension z is independent of the class label, the meaningful features in this trivial problem are the first two variables. The coefficient distribution is illustrated in Figure 3 (a), from which we can readily determine the importance order (y,x,z) for this dataset. Figure 3 (b) and (c) show classification error rates of MRF (d=1) and mRMR in the original space respectively. With exactly the same settings, MRF (d=1) correctly selects (y), (y,x), (y,x,z) while mRMR selects (y), (y,z), (y,z,x). It shows that mRMR cannot work in the original space for this linearly non-separable problem.

We further perform tests using MRF (d=2) in the original space, and the results are shown in Figure 4. This method correctly selects features in the order of (y), (y,x), and (y,x,z).

For other types of kernels, directly solving Eq. (16) might not be computationally efficient or even feasible when p is large. In this case, a potentially useful approach might be to use Taylor or other types of expansions to approximate or bound the objective function with linear or quadratic polynomial kernels, and we leave this as a line of future research due to the space limit.

2) *52-dimensional linearly non-separable problem*: To further verify the performance of our proposed MRF method in higher-dimensional situations, we perform experiments on a set of 52-dimensional synthetic data, Synthetic-2, originally used by Weston *et al.* [43]. This linearly non-separable classification problem consists of 1000 examples assigned to two classes: $y = 1$ and $y = -1$; each example e_i admits 52 features, $e_i = (x_{i_1}, x_{i_2}, \dots, x_{i_{52}})$. The data are generated in the following way:

Step 1. The probability of $y = 1$ or -1 is equal.

Step 2. If $y = -1$ then $(x_{i_1}, x_{i_2})^T$ are drawn from $N(\mu_1, \Sigma)$ or $N(\mu_2, \Sigma)$ with equal probabilities,

TABLE IV

RANDOM SPLITS PERFORMANCE EVALUATION ON REAL-WORLD DATASETS (PROSTATE CANCER AND LUNG CANCER ARE NOT INCLUDED HERE AS THEY HAVE INDEPENDENT TESTING DATASETS). THE 3RD TO 7TH COLUMNS ARE THE BEST PERFORMANCES (THE ENTRY $a(b)\%$: $a\%$ = THE AVERAGE OF THE SMALLEST ERROR RATES, AND $b\%$ = VARIANCE (BOTH IN PERCENTAGE) ACHIEVED BY USING DIFFERENT NUMBERS OF FEATURES FROM 1 UPTO $Max\#$. EACH CLASS OF EACH DATASET IS RANDOMLY DIVIDED INTO $\#Folds$, WITH ONE FOLD FOR TESTING AND THE REST FOR TRAINING. FOR EACH CLASSIFIER, THE BOLD VALUES ARE THE BEST RESULTS AMONG THESE FIVE METHODS. FOR EACH DATASET, THE UNDERLINED VALUES ARE THE BEST RESULTS OVER ALL CLASSIFIERS AND FIVE METHODS. FOR LARGE DATASETS, WE TERMINATED MRMR AFTER RUNNING OVER THREE HOURS: THE MRMR IS NOT SCALABLE AS SHOWN IN TABLE III.

dataset/Max #/#Folds	Classifier	MRF	Random Forest	mRMR	RFE	L0
Iris/2/10	SVM	4.67(0.19)%	5.33(0.35)%	5.33(0.39)%	24.10(1.20)%	24.00(0.90)%
	K-SVM	<u>1.33(0.16)%</u>	2.67(0.21)%	2.67(0.21)%	21.33(1.30)%	21.33(1.00)%
	NB	<u>1.33(0.10)%</u>	<u>1.33(0.37)%</u>	<u>1.33(0.37)%</u>	23.33(0.84) %	23.33(0.82)%
	C4.5	2.00(0.10)%	2.67(0.16)%	3.33(0.16)%	28.67(0.95) %	28.67(0.96)%
Wine/10/10	SVM	1.84(0.23)%	1.32(0.22)%	2.11(0.20)%	1.58(0.27)%	1.84(0.20)%
	K-SVM	<u>0.79(0.39)%</u>	<u>0.79(0.42)%</u>	1.58(0.37)%	1.32(0.35)%	1.58(0.39)%
	NB	5.53(0.20)%	2.11(0.23)%	5.53(0.21)%	2.37(0.22) %	2.11(0.21)%
	C4.5	7.37(0.31)%	7.37 (0.22)%	8.68(0.29)%	7.89(0.25) %	8.42(0.18)%
Optical Pen/40/4	SVM	3.94(0.04)%	4.19(0.03)%	4.12(0.03)%	8.13(0.04)%	7.82(0.03)%
	K-SVM	<u>0.89(0.04)%</u>	1.34(0.04)%	1.30(0.04)%	3.35(0.04)%	3.72(0.04)%
	NB	9.27(0.02)%	9.10(0.02)%	9.05(0.02)%	16.15(0.02) %	13.99(0.03)%
	C4.5	14.01(0.03)%	13.77(0.04)%	13.88(0.03)%	18.00(0.02)%	19.05(0.03)%
Leukemia-S3/60/4	SVM	<u>2.11(0.17)%</u>	6.32(0.55)%	-	<u>2.11(0.21)%</u>	2.37(0.19)%
	K-SVM	2.37(0.44)%	5.00(0.88)%	-	2.37(0.55)%	2.37(0.50)%
	NB	2.89(0.18)%	5.00(0.59)%	-	8.16(0.26) %	2.37(0.20)%
	C4.5	4.74(0.22)%	11.32(0.56)%	-	2.37(0.21)%	6.05(0.17) %
NCI9/60/4	SVM	48.68(0.85)%	53.62(0.89)%	-	66.31(0.87)%	65(0.76)%
	K-SVM	55.00(0.87)%	60.26(0.88)%	-	69.47(0.64)%	68.16(0.85)%
	NB	66.58(1.04)%	64.47(0.78)%	-	69.47(0.57) %	70.00(0.60)%
	C4.5	70.26(1.04)%	69.74(1.01)%	-	77.63(0.85)%	74.74(0.95) %

TABLE V

STATISTICAL SIGNIFICANCE TEST USING THE WILCOXON SIGNED RANK TEST [57] OVER MULTIPLE DATASETS. THE RANK TEST IS PERFORMED BETWEEN MRF AND EACH OF THE OTHER FOUR METHODS BASED ON TABLES II AND IV.

Pair of methods	(MRF, Random Forest)	(MRF, mRMR)	(MRF, RFE)	(MRF, L0)
p -values of Wilcoxon test	0.0043	0.0062	<0.0001E-4	<0.0001E-4

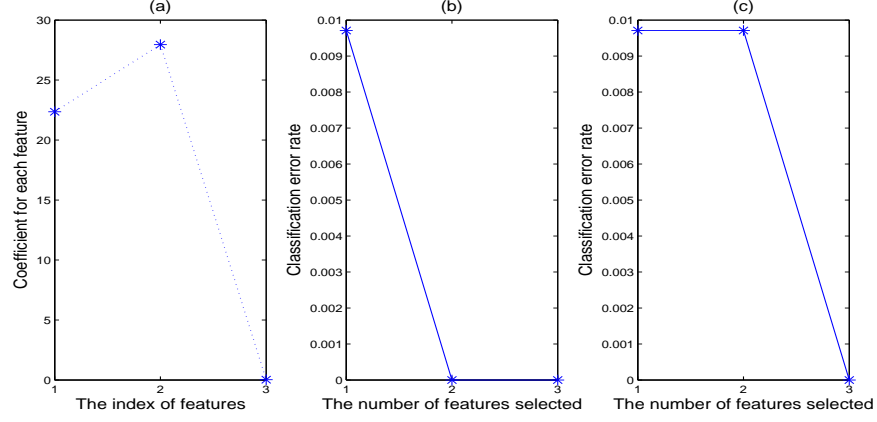


Fig. 3. The MRF ($d=1$) coefficient distribution and the comparison between MRF ($d=1$) and mRMR on Synthetic-1 dataset. (a) is the distribution of the coefficients θ_j using MRF ($d=1$) method ($\gamma = -0.5$), and the coefficient values of θ_j indicate the importance of the j th feature. (b) is the classification error rate by using MRF ($d=1$). (c) is the classification error by using mRMR. A two-class linear SVM classifier is used in this experiment.

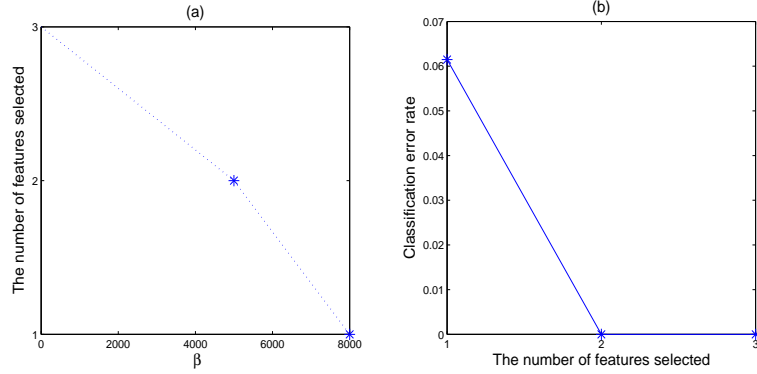


Fig. 4. The performance of MRF ($d=2$) on Synthetic-1 dataset. (a) The relationship between the number of selected features and parameter β with $\gamma = -0.5$. (b) the classification error rate with respect to the number of selected features. A two-class SVM with a polynomial kernel of order 2 is used as a classifier in this experiment.

$\mu_1 = (-\frac{3}{4}, -3)^T$, $\mu_2 = (\frac{3}{4}, 3)^T$, and $\Sigma = I$; if $y = 1$ then $(x_{i_1}, x_{i_2})^T$ are drawn from $N(\mu_3, \Sigma)$ or $N(\mu_4, \Sigma)$ with equal probabilities, $\mu_3 = (3, -3)^T$, and $\mu_4 = (-3, 3)^T$.

Step 3. The rest features x_{i_j} are noise randomly generated from $N(0, 20)$, $j \in \{3, \dots, 52\}$, $i \in \{1, \dots, 1000\}$.

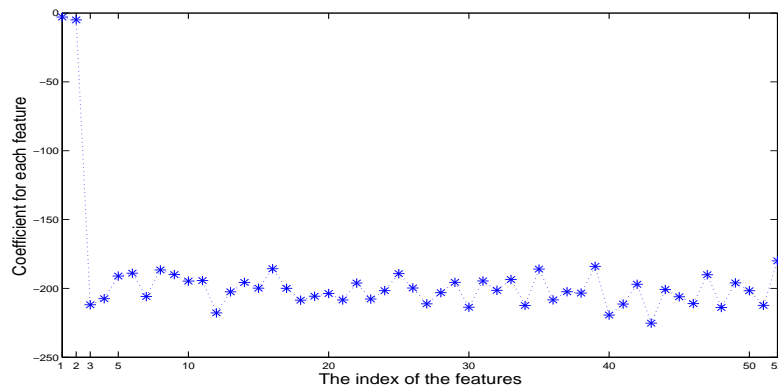


Fig. 5. The distribution of the coefficients θ_j for Synthetic-2 dataset by using MRF with $\gamma = -0.5$. The first two features admit much higher coefficients than the other 52 noise features.

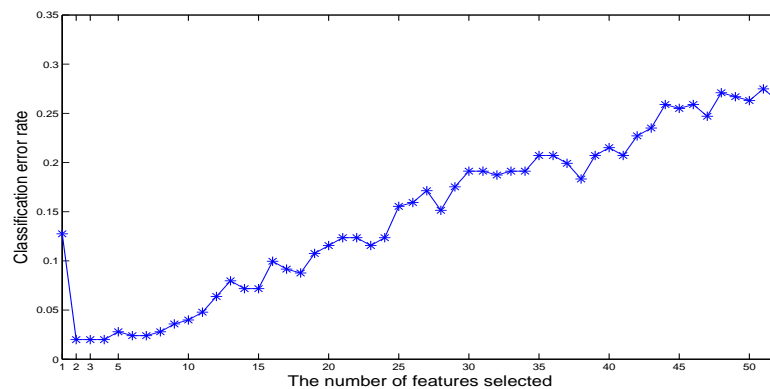


Fig. 6. MRF ($d=1$) classification error rates w.r.t the number of selected features for Synthetic-2 dataset. SVM with a polynomial kernel of order 2 is used. When noise features are included, the classification error rates go higher. When only two informative features are selected (x_{i_1} and x_{i_2}), we have the lowest classification error rate.

For each class, we uniformly sample three quarters of the data as training examples and the rest as test examples. The distribution of the coefficients θ_j is shown in Figure 5. From this distribution, we can clearly determine the (most) important features x_{i_1} and x_{i_2} ; and the other features are distinct from these two. The classification error rates with respect to (w.r.t.) the number of features are illustrated in Figure 6. It becomes evident that more features may not always induce lower classification errors. Actually, more incorrectly selected features may result in larger classification errors. In Figure 7, we also report the results by using MRF ($d=2$) which works well on this dataset.

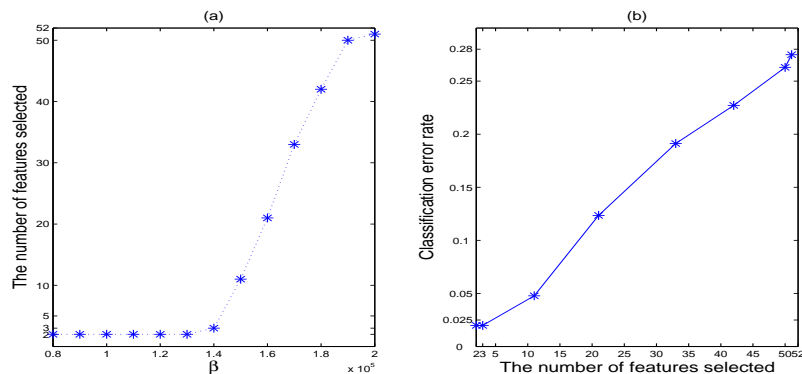


Fig. 7. Performance evaluation of using MRF ($d=2$) on Synthetic-2 dataset. (a) is the selected number of features w.r.t. the value of β with $\gamma = -0.5$. (b) is the classification error rate w.r.t the number of the selected features. SVM with a polynomial kernel of order 2 is used. When only two features are selected (x_{i_1} and x_{i_2}), we have the lowest classification error rate in figure (b).

F. Experiments on Real-world Datasets with Small or Medium Feature Size

1) *Iris*: As one of the classical examples, Fisher’s Iris dataset [5] has 3 different classes: Setosa, Versicolor, and Virginica. Each observation consists of 4 ordered features: “Sepal Length”, “Sepal Width”, “Petal Length” and “Petal Width”. Nine tenths of the observations in each class are used as training examples and the rest as test examples. The distribution of coefficients θ_j is illustrated in Fig. 8. As seen from this figure, the third feature, “Petal Length”, has the highest MRF coefficient value. While there is no big difference among the rest three features, we can still obtain an importance order: “Petal Width” > “Sepal Length” > “Sepal Width” (varying γ values in the feasible ranges can highlight the difference). These results are in accordance with previous conclusions on Iris dataset in the literature.

We compare several feature selection methods on this dataset and the results are illustrated in Fig. 9. In the same settings and using C4.5 as a classifier, both mRMR and our MRF methods can correctly determine the importance of the features, while L0 and RFE fail on this dataset.

2) *Optical Pen*: To test its performance on multi-class datasets, we perform experiments on the Optical Pen dataset from the UCI repository [58]. This middle-sized dataset (64 features) has 10 different classes, one class for one digit [59]. We uniformly sample three quarters of observations of each class as training examples and the rest as test examples. Figure 10 shows the coefficient distribution for MRF method. From Figure 11 we can see MRF achieves the lowest classification error rate when 39 features are selected. We also perform experiments on this dataset using different classifiers, with the comprehensive

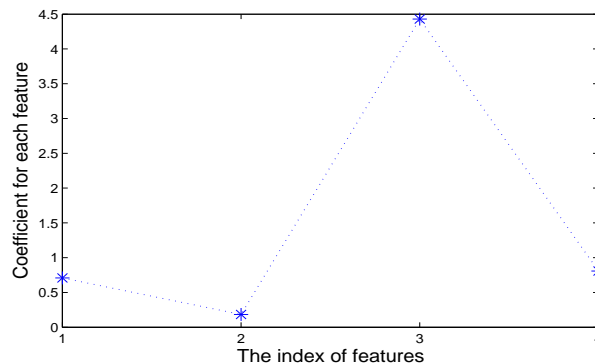


Fig. 8. The distribution of the coefficients θ_j for Iris data by using MRF with $\gamma = -0.5$. From this figure, we can readily determine the importance of features: The third feature, “Petal Width”, has the highest coefficient and thus is of the greatest importance.

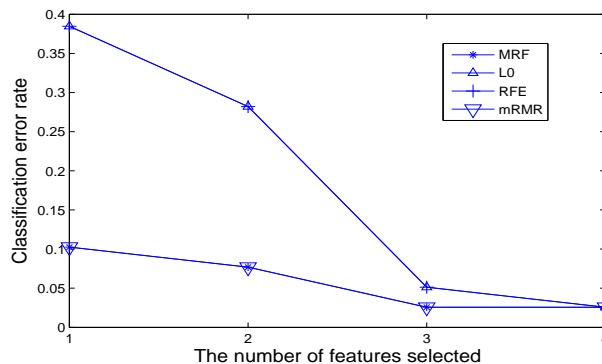


Fig. 9. Comparison of mRMR, RFE, L0, and MRF on Iris data. L0 and RFE overlap in this figure, and so do mRMR and MRF. Both L0 and RFE fail in this example, whereas mRMR and MRF produce the correct feature subsets. C4.5 is used as the classifier in this example.

comparison results provided in Tables II and IV.

G. Experiments on Real-world High-dimensional datasets

1) *Leukemia*: Leukemia-S3⁷, originally by Golub *et al.* [60], has 7070 features. We run MRF and the other four methods to select a maximum of 120 features, and the classification results using a linear SVM classifier are illustrated in Figure 12. It is clear that a proper feature selection method is vital for this dataset as only a handful of features are meaningful for class discrimination. Moreover, with more

⁷Dataset can be downloaded from <http://penglab.janelia.org/proj/mRMR/>

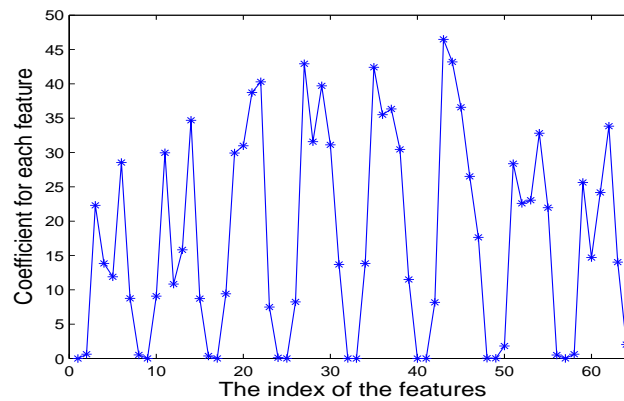


Fig. 10. Coefficient distribution of θ_j for Optical Pen dataset by using MRF with $\gamma = -0.5$.

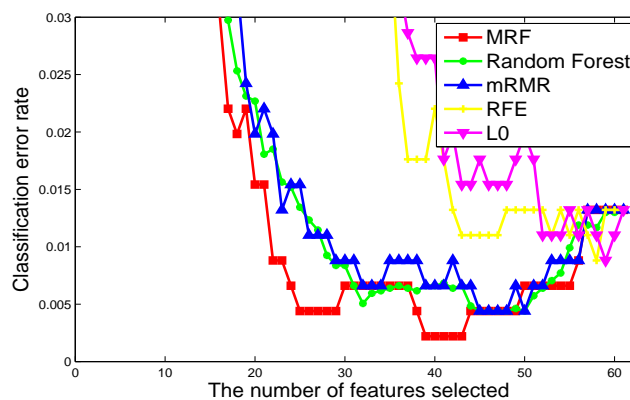


Fig. 11. Comparison of mRMR, Random Forest (100 trees), MRF, RFE and L0 on Optical Pen dataset. SVM with RBF kernel is used as the classifier. MRF achieves the lowest classification error rate when about 39 features are selected.

noise features included, the classification error rates go higher for all feature selection methods. The comprehensive comparison results are listed in Tables II, III, and IV.

2) *NCI9*: *NCI9*⁸ [61] [62] has 60 observations, 9703 features, and 9 types of cancers. The training set is generated by uniformly sampling three quarters of the observations for each class, and the rest is served as a test set. This dataset is challenging and the lowest classification error rate (by selecting up to 60 features) is 36.84% by using MRF and mRMR with a linear SVM classifier. For these five different methods and a linear SVM, the classification error rates are shown in Figures 13. The comprehensive comparison results are shown in Tables II, III and IV.

⁸Dataset can be downloaded from <http://penglab.janelia.org/proj/mRMR/>

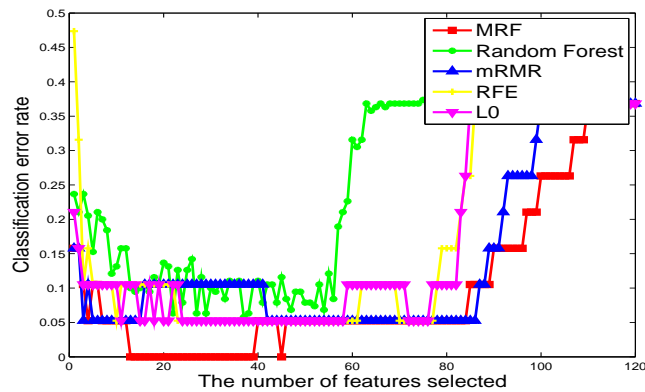


Fig. 12. Comparison of mRMR, Random Forest (100 trees), RFE, LO and MRF on Leukemia-S3 data. A linear SVM classifier is used. MRF outperforms the other four methods. Including more noise features will lead to larger classification error rates for all five feature selection methods.

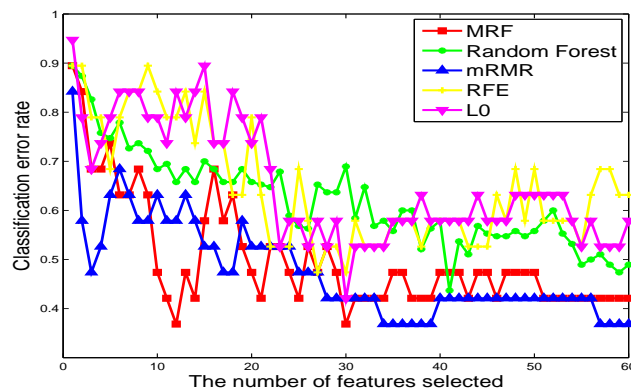


Fig. 13. Comparison of mRMR, Random Forest (100 trees), RFE, LO and MRF on NCI9 data. A linear SVM classifier is used. Both MRF and mRMR methods achieve the lowest error rate, but MRF requires only 12 features while mRMR requires more than 34 features.

3) *Prostate Cancer*: Prostate Cancer dataset [63] [64] has 12600 features. We use the data from [63] as a training dataset and those from [64] as an independent testing dataset. The coefficient distribution of θ_j is shown in Figure 14. From this figure, we can see that there are only a handful of distinct features (around 20). We perform classification tests on the first 60 most significant features. The resulting error rates are shown in Figure 15. MRF outperforms the other four methods. The comprehensive comparison results are provided in Tables II and III.

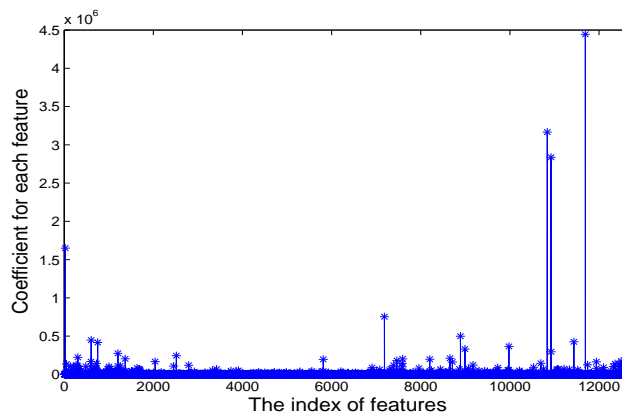


Fig. 14. Coefficient distribution of θ_j for Prostate Cancer data ($\gamma = -0.5$). There are only a handful of (around 20) distinct features.

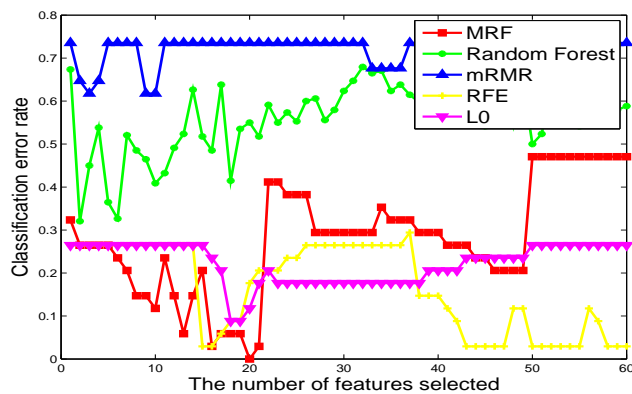


Fig. 15. Comparison of mRMR, Random Forest (100 trees), RFE, L0 and MRF on Prostate Cancer data. The first 60 most important features are selected for each method. A linear SVM classifier is used. MRF outperforms all other four methods.

4) *Lung Cancer*: The coefficient distribution of θ_j Lung Cancer dataset⁹ (more than 12000 features) is shown in Figure 16. The classification error rates are shown in Figure 17. MRF achieves the best performance on this dataset. The comprehensive comparison results using different classifiers are given in Tables II and III.

⁹downloaded from <http://www.chestsurg.org>

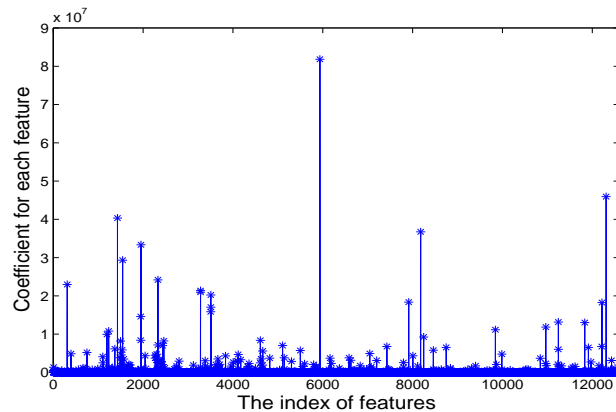


Fig. 16. Coefficient distribution of θ_j for Lung Cancer data ($\gamma = -0.5$). There are only a small fraction of highly discriminative features.

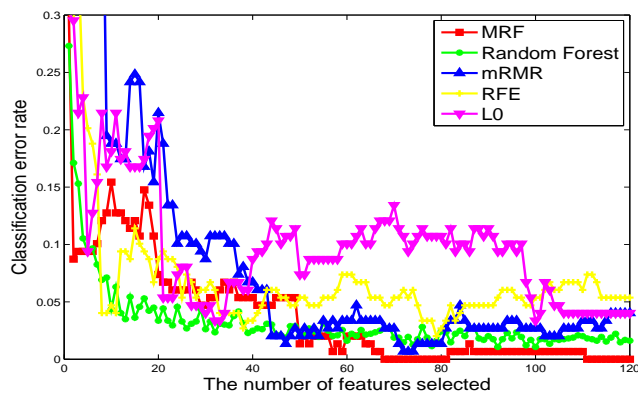


Fig. 17. Comparison of mRMR, Random Forest (100 trees), RFE, L0, and MRF on Lung Cancer data. The first 60 most important features are selected for each method. A linear SVM classifier is used. MRF outperforms the other four methods.

VI. CONCLUSIONS

In this paper, we study selecting the maximally separable subset of features for supervised classifications from among all subsets of variables. A class of new feature selectors is formulated in the spirit of Fisher's class separation criterion of maximizing between-class separability and minimizing within-class variations. The formulation addresses particularly three challenges encountered in pattern recognition and classifications: A small sample size with a large number of features, linearly non-separable classes, and noisy features. By choosing specific kernel functions we construct the Fisher-Markov selectors that boil down the general subset selection problem (simultaneous selection) to seeking optimal configurations on

the Markov random fields (MRF). For the consequent MRF problem, the Fisher-Markov selectors admit efficient algorithms to attain the global optimum of the class separability criterion without fitting the noise components. The resulting multivariate feature selectors are capable of selecting features simultaneously and have efficient computational complexity; for example, the LFS is linear in the number of features and quadratic in the number of observations. The Fisher-Markov selectors can be applied to general data with arbitrary dimensions and multiple classes, and the selected feature subset is useful for general classifiers. We have conducted extensive experiments on synthetic data as well as standard real-world datasets, and the experimental results have confirmed the effectiveness of the Fisher-Markov selectors.

REFERENCES

- [1] Domingos, P. and Pazzani, M. (1997). "On the optimality of the simple Bayesian classifier under zero-one loss," *Machine Learning*, 29:103-130.
- [2] Dudoit, S., Fridlyand, J., and Speed, T. (2002). "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of American Statistical Association*, 97: 77-87.
- [3] Fan, J. and Fan, Y. (2008). "High dimensional classification using features annealed independence rules," *Ann. Statistics.*, 36: 2232-2260.
- [4] Cover, T.M. (1965). "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Trans. on Electronic Computers*, EC-14:326-334.
- [5] Fisher, R.A. (1936). "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, 7:197-188.
- [6] Dixon, W.J. and Massey, F.J. (1957). *Introduction to Statistical Analysis*. Second Ed. New York: McGraw-Hill.
- [7] Kendall, M.G. (1957). *A Course in Multivariate Analysis*. London: Griffin.
- [8] Devijver, P.A. and Kittler, J. (1982). *Pattern Recognition: A Statistical Approach*. London: Prentice-Hall.
- [9] Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*. New York: Academic Press, second edition.
- [10] McLachlan, G.J. (2004). *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley.
- [11] Fu, K.-S. (1968). *Sequential Methods in Pattern Recognition and Machine Learning*. New York: Academic Press.
- [12] Fu, K.-S., Min, P.J., and Li, T.J. (1970). "Feature selection in pattern recognition," *IEEE Trans. Syst. Science Cybern.*, SSC-6: 33-39.
- [13] Chen, C.H. (1975). "On a class of computationally efficient feature selection criteria," *Pattern Recognition*, 7: 87-94.
- [14] Narendra, P. and Fukunaga, K. (1977). "A branch and bound algorithm for feature subset selection," *IEEE Trans. Computer*, C-26(9): 917-922.
- [15] Rissanen, J. (1989). *Stochastic complexity in Statistical Inquiry*. Singapore: World Scientific Publishing Company.
- [16] Kira, K. and Rendall, L.A. (1992). "A practical approach to feature selection," *Proc. Int. Conf. Machine Learning*, 249-256.
- [17] Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. Ser. B*, 58:267-288.
- [18] Donoho, D.L. and Elad, M. (2003). "Optimally sparse representation in general (nonorthogonal) dictionaries via l_1 minimization," *Proc. Natl. Acad. Sci. UDA*, 100:2197-2202.
- [19] Donoho, D.L. (2006). "Compressed sensing," *IEEE Trans. Inform. Theory*. 52: 1289-1306.
- [20] Candes, E.J., Romberg, J. and Tao, T. (2006). "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure Appl. Math.* 59: 1207-1223.

- [21] Candes, E. and Tao, T. (2007) "The Dantzig selector: statistical estimation when p is much larger than n ," *Annals of Statistics*, 35(6): 2313-2351.
- [22] McLachlan, G.J., Bean, R.W., and Peel, D. (2002). "A mixture model-based approach to the clustering of microarray expression data," *Bioinformatics*, 18:413-422.
- [23] Peng, H., Long, F., and Ding, C. (2005). "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226-1238.
- [24] Wang, L. (2008). "Feature selection with kernel class separability," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 30(9): 1534-1546.
- [25] Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T. and Vapnik, V. (2000). "Feature selection for SVMs," *Advances in Neural Information Processing Systems (NIPS'00)*, Leen, T.K., Dietterich, T.G. and Tresp, V., eds, 668-674.
- [26] Guyon, I., Weston, J., Barnhill, S., Vapnik, V. (2002). "Gene selection for cancer classification using support vector machines," *Machine Learning*, 46(1-3):389-422.
- [27] Webb, A. (2002). *Statistical Pattern Recognition*. 2nd ed, West Sussex, UK: Wiley.
- [28] Liu, H. and Motoda, H. (1999). *Feature Selection for Knowledge Discovery and Data Mining*. London: Kluwer Academic Publishers.
- [29] Breiman, L. (2001). "Random forests," *Machine Learning*, 45(1):5-32.
- [30] Koller, D. and Sahami, M. (1996). "Toward optimal feature and subset selection problem," *Proc. Int. Conf. Machine Learning*. Morgan Kaufman, 284-292, Bari, Italy.
- [31] Fowlkes, E.B., Gnanadesikan, R., and Kettnering, J.R. (1987). "Variable selection in clustering and other contexts," *Design, Data, and Analysis*, C.L. Mallows (Ed.). New York: Wiley, 13-34.
- [32] Duda, R.O., Hart, P.E., and Stork, D.G. (2000). *Pattern Classification*. 2nd ed, Wiley-Interscience.
- [33] Bickel, P. and Levina, E. (2004). "Some theory of Fisher's linear discriminant function, 'naive Bayes', and some alternatives where there are many more variables than observations," *Bernoulli*, 10: 989-1010.
- [34] Mika, S., Ratsch, G., and Muller, K.-R. (2001). "A mathematical programming approach to the Kernel Fisher algorithm," *Advances in Neural Information Processing Systems*, 13:591-597.
- [35] Vapnik, V.N. (1998). *Statistical Learning Theory*. New York: Wiley.
- [36] Scholkopf, B. and Smola, A.J. (2002). *Learning with Kernels*. Cambridge, MA:MIT Press.
- [37] Vapnik, V.N. (1999). *The Nature of Statistical Learning Theory*. New York:Springer.
- [38] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. New York: Cambridge University Press.
- [39] Fidler, S. Slocaj, D., and Leonardis, A. (2006). "Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(3): 337-350.
- [40] Akaike, H. (1974). "A new look at the statistical model identification," *IEEE Trans. Automatic Control*, 19:716-723.
- [41] Schwarz, G. (1978). "Estimating the dimension of a model," *Ann. Statist.* 6: 361-379.
- [42] Foster, D.P. and George, E.I. (1994). "The risk inflation criterion for multiple regression," *Ann. Statist.* 22: 1947-1975.
- [43] Weston, J., Elisseeff, A., Schlkopf, B., and Tipping, M.E. (2003). "Use of the zero-norm with linear models and kernel methods," *Journal of Machine Learning Research*, 3:1439-1461.
- [44] Greenwood, P.E. and Shirayev, A.N. (1985). *Contiguity and the Statistical Invariance Principle*. New York: Gordon and Breach.
- [45] Rosenblatt, M. (2000). *Gaussian and Non-Gaussian Linear Time Series and Random Fields*. Springer.
- [46] Bosq, D. (1998). *Nonparametric Statistics for Stochastic Processes*. Springer.

- [47] Geman S. and Geman D. (1984). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6(6): 721-741.
- [48] Winkler, G. (2006). *Image Analysis, Random Fields and Dynamic Monte Carlo Methods. A Mathematical Introduction*. 3rd edition, Springer-Verlag.
- [49] Dai, S., Baker, S., and Kang, S.B. (2009). "An MRF-based deinterlacing algorithm with exemplar-based refinement," *IEEE Trans. on Image Processing*, 18(4): 956-968.
- [50] Hochbaum, D.S. (2001). "An efficient algorithm for image segmentation, Markov random fields and related problems," *Journal of ACM*, 48(2): 686-701.
- [51] Picard, J.P. and Ratliff, H.D. (1975). "Minimum cuts and related problem," *Networks*, 5:357-370.
- [52] Ishikawa, H. (2003). "Exact optimization for Markov random fields with convex priors," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(10): 1333-1336.
- [53] Kolmogorov, V. and Zabih, R. (2004). "What energy can be minimized via graph cuts?" *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(2): 147-159.
- [54] Boykov, Y., Veksler, O. and Zabih, R. (2001). "Fast approximate energy minimization via graph cuts," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(11): 1222-1239.
- [55] Wainwright, M., Jaakkola, T. and Willsky, A. (2005). "MAP estimation via agreement on (hyper)trees: message-passing and linear programming," *IEEE Trans. Information Theory*, 51(11): 3697-3717.
- [56] Yedidia, J., Freeman, W. and Weiss, Y. (2004). "Constructing free energy approximations and generalized belief propagation algorithms," *IEEE Trans. Information Theory*, 51: 2282-2312.
- [57] Demsar, J. (2006). "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, 7: 1-30.
- [58] Blake, C.L., Newman, D.J., Hettich, S., and Merz, C.J. (1998). *UCI repository of machine learning databases*. URL <http://www.ics.uci.edu/mllearn/MLRepository.html>.
- [59] Garris, M.D., et al. (1994). *NIST Form-Based Handprint Recognition System, NISTIR 5469*.
- [60] Golub, T., et al. (1999). "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*: 286, 531-537. The data is available at <http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi>
- [61] Ross, D.T., et al. (2000). "Systematic variation in gene expression patterns in human cancer cell lines," *Nature Genetics*, 24(3): 227-234.
- [62] Scherf, U., et al. (2000). "A cDNA microarray gene expression database for the molecular pharmacology of cancer," *Nature Genetics*, 24(3): 236-244.
- [63] Singh, D., et al. (2002). "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, 1:203-209. The data is available at <http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi>
- [64] Welsh, J.B., et al. (2001). "Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer," *Cancer Research*, 61: 5974-5978.