Kernel Feature Selection via Conditional Covariance Minimization

Jianbo Chen* Mitchell Stern* Martin J. Wainwright Michael I. Jordan

University of California, Berkeley

Abstract

We propose a method for feature selection that employs kernel-based measures of independence to find a subset of covariates that is maximally predictive of the response. Building on past work in kernel dimension reduction, we show how to perform feature selection via a constrained optimization problem involving the trace of the conditional covariance operator. We prove various consistency results for this procedure, and also demonstrate that our method compares favorably with other state-of-the-art algorithms on a variety of synthetic and real data sets.

Formulating Feature Selection

The problem of feature selection:
Given n i.i.d. samples $\{(x_i, y_i) : i = 1, 2, \dots, n\}$ generated from $P_{X,Y}$ together with an integer $m \le d$, select m of the d features $\mathcal{S} = \{X_1, X_2, \dots, X_d\}$ which best predict Y.
Dependence Perspective:
Identify a subset of features $\mathcal{T}$ of size m such that:

$X_{\mathcal{S}\setminus\mathcal{T}}$ is conditionally independent of Y given $X_{\mathcal{T}}$.

Prediction Perspective:
Find the subset of features that minimizes the prediction error:

$$\min_{\mathcal{T}:|\mathcal{T}|\le m} \mathcal{E}_{\mathcal{F}}(X_{\mathcal{T}}) = \min_{\mathcal{T}:|\mathcal{T}|\le m} \inf_{f\in\mathcal{F}_m} \mathbb{E}_{X,Y} L(Y, f(X_{\mathcal{T}})),$$

where $\mathcal{E}_{\mathcal{F}}(X_{\mathcal{T}})$ is the *error of prediction* using only the features in $\mathcal{T}$, $\mathcal{F}$ is a function class from $\mathcal{X}_{\mathcal{T}}$ to $\mathcal{Y}$, and L is a loss.

Conditional Covariance Operator

$(\mathcal{H}_{\mathcal{X}}, k_{\mathcal{X}})$ and $(\mathcal{H}_{\mathcal{Y}}, k_{\mathcal{Y}})$: RKHSs of functions on $\mathcal{X}$ and $\mathcal{Y}$.
(X, Y): a random vector on $\mathcal{X}\times\mathcal{Y}$ with joint distribution $P_{X,Y}$.
Cross-covariance operator: an operator $\Sigma_{YX} : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$ with

$$\langle g, \Sigma_{YX} f\rangle_{\mathcal{H}_{\mathcal{Y}}} = \mathbb{E}_{X,Y}[(f(X) - \mathbb{E}_X[f(X)])(g(Y) - \mathbb{E}_Y[g(Y)])].$$

Conditional covariance operator:

$$\Sigma_{YY|X} = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}.$$

$\Sigma_{YY|X}$ captures **conditional variance:** for $g \in \mathcal{H}_{\mathcal{Y}}$,

$$\langle g, \Sigma_{YY|X} g\rangle_{\mathcal{H}_{\mathcal{Y}}} = \mathbb{E}_X[\mathrm{Var}_{Y|X}[g(Y)|X]].$$

$\Sigma_{YY|X}$ captures **residual error:** for $g \in \mathcal{H}_{\mathcal{Y}}$,

$$\langle g, \Sigma_{YY|X} g\rangle_{\mathcal{H}_{\mathcal{Y}}} = \inf_{f\in\mathcal{H}_{\mathcal{X}}} \mathbb{E}_{X,Y}(g(Y) - f(X))^2.$$

Proposed Method

Feature selection criterion:

$$\min_{\mathcal{T}:|\mathcal{T}|=m} \mathcal{Q}(\mathcal{T}) := \mathrm{Tr}(\Sigma_{YY|X_{\mathcal{T}}}).$$

Property 1. *If $(\mathcal{H}_{\mathcal{X}}, k_{\mathcal{X}})$ is characteristic, then $\mathrm{Tr}(\Sigma_{YY|X}) \le \mathrm{Tr}(\Sigma_{YY|X_{\mathcal{T}}})$ for any $\mathcal{T}$. Moreover, the equality $\mathrm{Tr}(\Sigma_{YY|X}) = \mathrm{Tr}(\Sigma_{YY|X_{\mathcal{T}}})$ holds if and only if $Y \perp\!\!\!\perp X|X_{\mathcal{T}}$.*

Property 2. *The criterion characterizes prediction error:*

$$\mathrm{Tr}(\Sigma_{YY|X_{\mathcal{T}}}) = \mathcal{E}_{\mathcal{F}_m}(X_{\mathcal{T}}) = \inf_{f\in\mathcal{F}_m} \mathbb{E}_{X,Y}(Y - f(X_{\mathcal{T}}))^2.$$

where $\mathcal{F}_m$ is a function space from $\mathbb{R}^m$ to $\mathcal{Y}$ defined from $\mathcal{H}_{\mathcal{X}}$.

Empirical estimate (with a linear kernel on Y):

$$\min_{|\mathcal{T}|=m} \widehat{\mathcal{Q}}^{(n)}(\mathcal{T}) := \mathrm{Tr}(\mathbf{Y}^T(G_{X_{\mathcal{T}}} + n\varepsilon_n I_n)^{-1}\mathbf{Y}),$$

where

$$G_{X_{\mathcal{T}}} = (I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T)K_{X_{\mathcal{T}}}(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T),$$
$$K_{X_{\mathcal{T}}} = (k_X(x_i^{\mathcal{T}}, x_j^{\mathcal{T}}))_{n\times n},$$
$$\mathbf{Y} \in \mathbb{R}^{n\times k},$$

and $x^{\mathcal{T}} \in \mathbb{R}^d$ is a vector with $x_i^{\mathcal{T}} = x_i$ if $i \in \mathcal{T}$ or 0 otherwise.

Theorem 1. [Feature Selection Consistency] *Define the set of all optimal feature subsets to be $A = \mathrm{argmin}_{|\mathcal{T}|\le m} Q(\mathcal{T})$, and let $\widehat{\mathcal{T}}^{(n)} \in \mathrm{argmin}_{|\mathcal{T}|\le m} \widehat{\mathcal{Q}}^{(n)}(\mathcal{T})$ be a global optimum of the empirical estimate. If $\varepsilon_n \to 0$ and $\varepsilon_n n \to \infty$ as $n \to \infty$, we have*

$$P(\widehat{\mathcal{T}}^{(n)} \in A) \to 1.$$

Optimization

We relax the initial NP-hard formulation to obtain:

$$\begin{aligned} \min_{w} \quad & \mathbf{y}^T(G_{w\odot X} + n\varepsilon_n I_n)^{-1}\mathbf{y} \\ \text{subject to} \quad & 0 \le w_i \le 1, \ i = 1, \dots, d, \\ & \mathbf{1}^T w \le m. \end{aligned}$$

where $\odot$ is the Hadamard product.

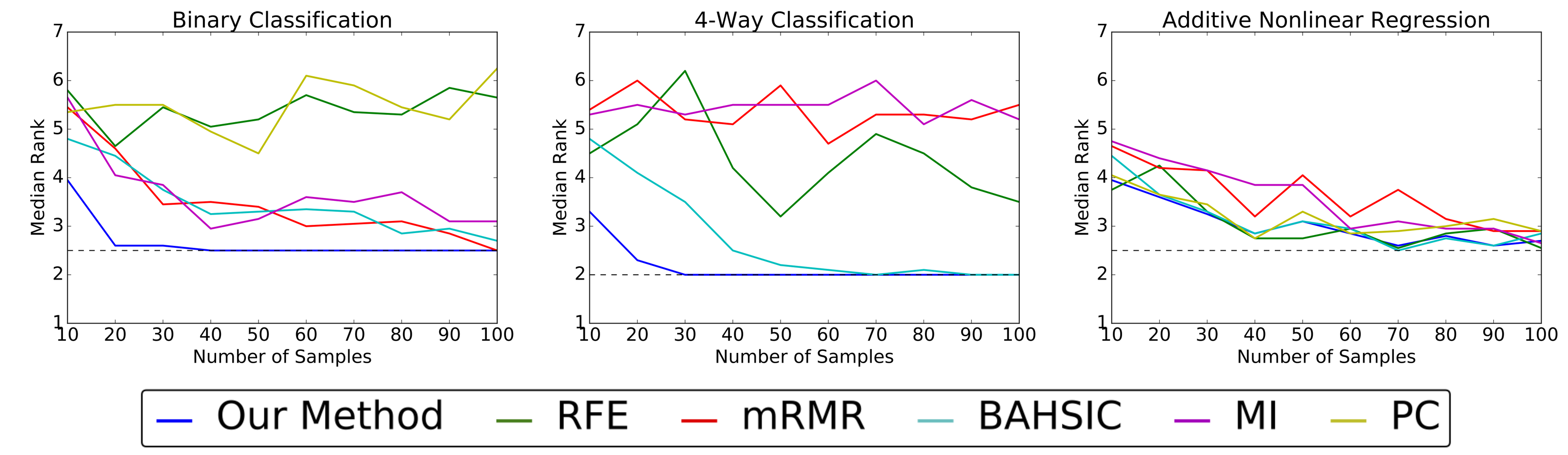
We may further use a kernel approximation $G_w \approx V_w V_w^T$:

$$(G_{w\odot X} + n\varepsilon_n I_n)^{-1} \approx \frac{1}{\varepsilon_n n}(I - V_w(V_w^T V_w + \varepsilon_n n I_D)^{-1} V_w^T).$$

Both objectives are optimized using projected gradient descent.

Synthetic Experiments

Synthetic data sets: binary classification, 4-way classification, additive nonlinear regression.
Other algorithms: Recursive feature elimination (RFE), Minimum Redundancy Maximum Relevance (mRMR), BAHSIC, mutual information (MI) and Pearson's correlation (PC).
Evaluation: Median rank assigned to true features.



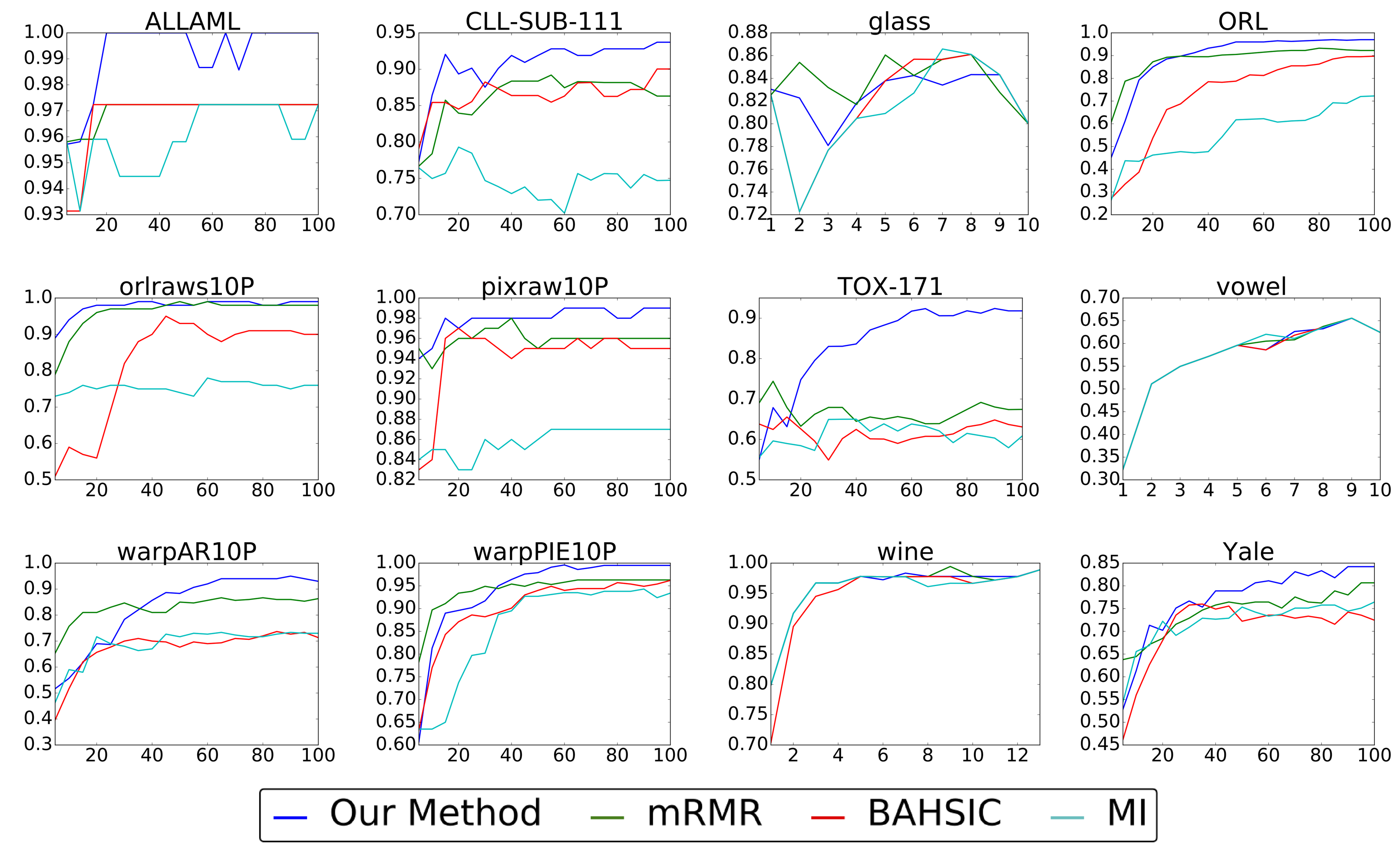
Plots of median rank vs. number of samples

Real-world Experiments

Summary of data sets:

	ALLAML	CLL-SUB-111	glass	ORL	orlraws10P	pixraw10P
Samples	72	111	214	400	100	100
Features	7,129	11,340	10	1,024	10,304	10,000
Classes	2	3	6	40	10	10
	TOX-171	vowel	warpAR10P	warpPIE10P	wine	Yale
Samples	171	990	130	210	178	165
Features	5,784	10	2,400	2,420	13	1,024
Classes	4	11	10	10	3	15

Other nonlinear algorithms: mRMR, BAHSIC, and MI.
Evaluation: Accuracy of a kernel SVM on selected features.



Plots of accuracy vs. number of selected features