A Scalable Algorithm for Structured Kernel Feature Selection

Shaogang Ren[1] **Shuai Huang**[2] **John Onofrey**[3] **Xenophon Papademetris** [3] **Xiaoning Qian**[1]
[1]Texas A&M University [2]University of Washington [3]Yale University

Abstract

Kernel methods are powerful tools for nonlinear feature representation. Incorporated with structured LASSO, the kernelized structured LASSO is an effective feature selection approach that can preserve the nonlinear input-output relationships as well as the structured sparseness. But as the data dimension increases, the method can quickly become computationally prohibitive. In this paper we propose a stochastic optimization algorithm that can efficiently address this computational problem on account of the redundant kernel representations of the given data. Experiments on simulation data and PET 3D brain image data show that our method can achieve superior accuracy with less computational cost than existing methods.

1 INTRODUCTION

Feature selection has been one of the important problems to address the infamous curse of dimensionality in applying statistical learning methods to short and fat data with $n/p \ll 1$, where n and p denote the sample size and feature space dimension respectively. Penalized feature selection methods such as the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996) provide one of effective solutions, which typically search for features that are linearly related to the output.

In order to explore potential nonlinear input-output relationships with feature selection, researchers have proposed both parametric and non-parametric methods (Tibshirani, 1996; Tibshirani *et al.*, 2005; Li *et al.*, 2006; Yamada *et al.*, 2014). We focus on non-parametric methods in this paper, specifically, kernel feature selection methods. Kernel methods are arguably among the most popular tools that provide a practical way to capture nonlinear relationships. For example, Quadratic Programming Feature Selection (QPFS) (Rodriguez *et al.*, 2010) solves a quadratic programming problem with quadratic kernelized dependency measures. But with the increasing feature dimension, the Hessian matrix for the quadratic term may become singular and cause computational difficulty. Song *et al.* (2012) proposed a greedy kernel feature selection method with forward feature selection or backward elimination strategies based on Hilbert-Schmidt Independent Criteria (HSIC, Gretton *et al.*, 2005). A related method—Hilbert-Schmidt Feature Selection (HSFS)—proposed in Masaeli *et al.* (2010) can be considered as its continuous relaxation. HSFS was formulated as non-convex optimization problems with only local optimality guarantee from the resulting optimization algorithms. Neither the method in Song *et al.* (2012) nor HSFS can scale up with the feature dimension due to the non-convexity and complexity of their accompanying optimization problems. To address the scalability problem, Sparse Additive Models (SAM) (Ravikumar *et al.*, 2009) have been proposed to efficiently solve kernel feature selection by a back-fitting algorithm (Ravikumar *et al.*, 2009), but it was shown that it may not perform well when features are not additively related. More recently, based on feature vector machines (FVM) (Li *et al.*, 2006), Yamada *et al.* (2014) proposed a high-dimensional kernel feature selection method: HSIC-LASSO, in which the optimization problem can be efficiently solved by the dual augmented Lagrangian (DAL) algorithm (Tomioka *et al.*, 2011).

HSIC-LASSO is a feature-wise kernel method. When studying features from structured data such as images and networks for disease diagnosis, inherent structural and functional relationships among features may need to be integrated in feature selection for better accuracy, reproducibility, and interpretability. Feature-wise kernel selection methods may be further improved with better performance by considering such structural and functional relationships among features, especially when the sample size is limited. Hence, in this paper, we aim to develop such a kernel feature selection method that explicitly imposes structural constraints among selected features. One of such structured penalized feature selection methods is the Fused LASSO (Tibshirani *et al.*, 2005; Xin *et al.*, 2014) in linear regression and classification. The implementation of Fused LASSO for kernel feature selection to capture non-

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38. Copyright 2015 by the authors.

linearity is computationally challenging. When the sample size and feature dimension increase, for example when studying 3-Dimensional brain images, the general batch-based optimization becomes inefficient and even infeasible. To address this computational difficulty, we introduce explicit structural constraints and derive a highly scalable stochastic optimization algorithm for this structured kernel feature selection method that is designed for the classification problems.

In summary, we propose a new structured kernel feature selection method based on the Hilbert-Schmidt Independent Criteria (Gretton *et al.*, 2005) but with explicitly enforced structural constraints to incorporate potential structural and functional relationships among features when they are available. The derived stochastic optimization algorithm is tailored to such a structured kernel feature selection problem and can efficiently solve the problem of very large size, for example for 3D brain images, on account of the redundant kernel representations of the given data. Finally, unlike HSIC-LASSO, which is designed for feature selection and requires separate learning processes for prediction with the selected features, our structured kernel feature selection method is formulated in a supervised learning framework and simultaneously learns the prediction model that can be directly adopted for new data.

The remaining of the paper is organized as follows: Section 2 formulates the structured kernel feature selection problem; Section 3 derives the tailored stochastic optimization algorithm; Section 4 presents and discusses our experimental results with both simulation data and 3D PET brain images; Section 5 provides the discussion on the relationships of our method with the existing kernel feature selection methods in literature; Section 6 concludes the paper and provides future research directions.

2 MODEL FORMULATION

In this section, we present our structured kernel feature selection model for classification.

2.1 Structured Kernel Feature Selection

Different from HSIC-LASSO (Yamada *et al.*, 2014), we take the Hinge loss function in our model instead of the least squared loss in HSIC-LASSO since we focus on classification problems in this paper. Without loss of generality, with the input features $X \in \mathbf{R}^{n\times p}$ and output responses $Y \in \{-1,1\}^{n\times 1}$, the penalized kernel feature selection problem can be formulated as follows, with the L_1-norm penalty as typically done in LASSO:

$$\min_{\mathbf{a}} \sum_{m=1}^{n} [n - \bar{L}_m^T(a_0\mathbf{1} + \sum_{i=1}^{p} a_i \bar{K}_m^i)]_+ + \lambda_1 |\mathbf{a}_{1,\dots,p}|_1 \qquad (1)$$

$$+ \lambda_2 \sum_{(i,j)\in E} (a_i - a_j)^2$$

$$s.t. \qquad a_i \geq 0 \quad \forall i \geq 1, \qquad (2)$$

where the first term is the Hinge loss; $\bar{L}_m$ is a n-dimensional vector, corresponding to the mth column of the output kernel matrix $\tilde{L}$; and $\bar{K}_m^i$ corresponds to the mth column of $\tilde{K}^i$, which is the kernel matrix for feature $\mathbf{x}_i$. The structural constraints among candidate features are imposed as quadratic terms of fitting coefficients $\mathbf{a}$ in (1), where E denotes all the available pairwise structural relationships among features. We consider a six-neighborhood-system for 3D images. We note that these quadratic terms can be rewritten in the matrix form with the graph Laplacian based on the feature structural relationships. But for many applications, the Laplacian is highly sparse, and it is not advisable to store and use the Laplacian matrix directly in the algorithm. With the L_1-norm regularization term, the non-negative constraints (2) guarantee that the active features have larger values and non-related features have small values to make the results easily interpretable. As similarly done in Yamada *et al.* (2014), for each feature $\mathbf{x}_i \in X$, we have

$$\tilde{K}^i = HK^iH; \qquad H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T;$$

$$K^i_{k,\ell}(\mathbf{x}_i, \mathbf{x}_i) = exp\left(- \frac{(x_{ki} - x_{\ell i})^2}{2\sigma^2_{\mathbf{x}_i}} \right);$$

$$\bar{K}^i = vec(\tilde{K}^i) \qquad \bar{K}^i_m = \tilde{K}^i_{\bullet,m}.$$

For output responses Y, we adopt the following kernel:

$$\tilde{L} = HYY^TH;$$

$$\bar{L} = vec(\tilde{L}); \qquad \bar{L}_m = \tilde{L}_{\bullet,m}.$$

Note that the output kernel matrix in our model is also different from the one adopted in HSIC-LASSO, which is given as follows:

$$L(y_i, y_j) = \begin{cases} 1/n_{y_i} & \text{if } y_i = y_j \\ 0 & \text{otherwise} \end{cases}$$

$$\tilde{L} = HLH \qquad \bar{L} = vec(\tilde{L}),$$

where n_{y_i} is the number of training samples in class y_i. The proposed kernel on Y discriminates the pairwise sample relationships in the infinite feature space, and also it provides us an approach for label prediction as shown latter in the experiment section. However, it is difficult to get a clear criterion for prediction with the kernel for Y used in HSIC-LASSO when there are different numbers of samples in different classes.

2.2 Interpretation by Hilbert-Schmidt Independent Criteria

The formulated optimization problem in (1) aims to identify predictive features that have large inner-product values between $\bar{L}$ and $\bar{K} = a_0\mathbf{1} + \sum_i \bar{K}^i a_i$ under previously described constraints. By expanding the inner-product $\bar{L}^T\bar{K}$, we have

$$\begin{aligned}\bar{L}^T\bar{K} &= tr(\tilde{L}\tilde{K}) \\ &= a_0 tr(\tilde{L}I) + \sum_i a_i tr(\tilde{L}\tilde{K}^i) \\ &= a_0 tr(\tilde{L}I) + \sum_i a_i HSIC(Y, \mathbf{x}_i).\end{aligned}$$

$HSIC(Y, \mathbf{x}_i) = tr(\tilde{L}\tilde{K}^i)$ is the empirical estimation of Hilbert-Schmidt Independent Criteria (HSIC), which is the same kernel-based independence measure adopted in Song *et al.* (2012) and HSIC-LASSO. As proven in Gretton *et al.* (2005), $HSIC$ always takes non-negative value and is zero if and only if the two variables are independent. When solving the optimization problem (1), the Hinge loss term drives the feature selection for highly correlated features with the output through the HSIC term; thereafter to have larger fitting coefficients a_i's with the non-negative L_1-norm term penalizing less correlated or independent features to have zero coefficients. Finally, with the structural constraints, our new model can robustly recover structurally related groups of features that are responsible for the output, aiming to obtain reproducible and accurate results.

3 STOCHASTIC OPTIMIZATION SOLUTION

In this section, we derive the stochastic optimization algorithm to solve our structured kernel feature selection problem.

3.1 Stochastic Optimization Algorithm

We note that the dimension of $\bar{K}^i$ in (1) is $n^2 \times 1$, and there are p such feature kernel vectors for p features in the problem. When either the sample size or feature dimension is large, many general-purpose first-order optimization algorithms cannot scale up accordingly to solve (1). In order to provide practical and efficient solution algorithms to (1), we develop a stochastic optimization algorithm based on an efficient online algorithm: the dual average method (Xiao, 2010; Yang *et al.*, 2010).

As the fitting coefficients $\mathbf{a}$ are non-negative, the optimization problem (1) can be rewritten as

$$\begin{aligned}\min_{\mathbf{a}} \; & \sum_{m=1}^{n} [n - \bar{L}_m^T(a_0\mathbf{1} + \sum_{i=1}^{p} a_i \bar{K}_m^i)]_+ + \lambda_1 \sum_{i=1}^{p} a_i \\ & + \lambda_2 \sum_{(i,j)\in E} (a_i - a_j)^2 \qquad (3) \\ s.t. \quad & a_i \geq 0 \quad \forall i \geq 1.\end{aligned}$$

As in the dual average method (Xiao, 2010), the above optimization problem can be considered as two parts: the loss function part, which should be subdifferentiable; and the regularization or constraint part, which should be convex. For our current formulation (3), the objective function in (3) is subdifferentiable and can be directly taken as the loss function part for the dual average optimization. The only constraint term is the non-negative constraints on $\mathbf{a}$. Applying the dual average method (Xiao, 2010), the objective function can be rewritten in each step t for one sample m :

$$\begin{aligned}l_t = & [n - \bar{L}_m^T(a_0\mathbf{1} + \sum_{i=1}^{p} \bar{K}_m^i a_i)]_+ + \lambda_1 \sum_{i=1}^{p} a_i \qquad (4) \\ & + \lambda_2 \sum_{(i,j)\in E} (a_i - a_j)^2.\end{aligned}$$

$\bar{L}_m$ and $\bar{K}_m^i$ can be considered as sample-dependent parts of $\bar{L}$ and $\bar{K}^i$, respectively.

We first compute the subgradient of l_t with respect to fitting coefficients $\mathbf{a}$:

$$\mathbf{g_t}(i) = \begin{cases} \phi(\mathbf{a}), & \text{if } \bar{L}_m^T(a_0\mathbf{1} + \sum_i \bar{K}_m^i a_i) > n; \\ -(\bar{K}_m^i)^T\bar{L}_m + \phi(\mathbf{a}), & \text{otherwise,} \end{cases}$$

$$\phi(\mathbf{a}) = \lambda_1 + 2\lambda_2 \sum_{\{j:(i,j)\in E\}} (a_i - a_j).$$

Here, $\mathbf{g_t}(i)$ gives the ith entry of the subgradient $\mathbf{g_t}$. For a_0, $\bar{K}_m^i$ is $\mathbf{1}$. For the dual average method at step t, we can compute the average subgradient $\bar{\mathbf{g}}_t$:

$$\bar{\mathbf{g}}_t = \frac{t-1}{t}\bar{\mathbf{g}}_{t-1} + \frac{1}{t}\mathbf{g}_t. \qquad (5)$$

According to Xiao (2010), the dual average method requires to solve a modified optimization problem by choosing a simple but strongly convex auxiliary function $h(\mathbf{a})$ as well as a non-decreasing step-size sequence $\{\beta_t\}$. The appropriate choice of the auxiliary function helps make the problem smooth and strongly convex for easier optimization. The appropriate non-decreasing sequence $\{\beta_t\}$ can guarantee fast convergence. For our structured kernel feature selection problem, we need to solve the following optimization problem each step:

$$\min_{\mathbf{a}} \bar{\mathbf{g}}_t^T\mathbf{a} + \frac{\gamma(1+\ln(t))}{t}||\mathbf{a}||^2 \qquad (6)$$

$$s.t. \quad a_i \geq 0, \forall i \geq 1. \qquad (7)$$

Here, we take $h(\mathbf{a}) = ||\mathbf{a}||^2$ as the auxiliary function, which is strongly convex, and $\beta_t = \gamma(1 + \ln(t))$. This auxiliary function $h(\mathbf{a})$ is designed specifically to have an efficient updating rule for solving our original structured kernel feature selection problem (1). Following the derivation of the dual average method in Xiao (2010), we can prove the following theorem that gives the updating rule of our stochastic optimization algorithm.

Theorem 1 With the auxiliary function $h(\mathbf{a}) = ||\mathbf{a}||^2$ and the non-decreasing sequence $\{\beta_t\}$ with $\beta_t = \gamma(1 + \ln(t))$, then the updating rule in each step t for fitting coefficients $\mathbf{a}$ for the problem (1) is:

$$(a_i)_t = \begin{cases} -\frac{t}{2\gamma(1+\ln(t))}\bar{\mathbf{g}}_t(i), & \text{if } i = 0; \\ [-\frac{t}{2\gamma(1+\ln(t))}\bar{\mathbf{g}}_t(i)]_+, & \text{if } i = 1, ..., p. \end{cases}$$

Proof: We can write the Lagrangian of the problem (6) by introducing the Lagrangian multipliers with the non-negative constraint:

$$L(\mathbf{a}, \lambda) = \frac{\gamma(1 + \ln(t))}{t}||\mathbf{a} - (-\frac{t}{2\gamma(1 + \ln(t))}\bar{\mathbf{g}}_t)||_2 \\ - \lambda^T \mathbf{a}_{1,...,p}.$$

We can compute the gradient of the Lagrangian with respect to $\mathbf{a}$ as

$$\bigtriangledown_{\mathbf{a}} L = 2\frac{\gamma(1 + \ln(t))}{t}(\mathbf{a} - (-\frac{t}{2\gamma(1 + \ln(t))}\bar{\mathbf{g}}_t)) - \lambda_{1,...,p}. \tag{8}$$

There is no constraint for a_0. Hence, $a_0 = -\frac{t}{2\gamma(1+\ln(t))}\bar{\mathbf{g}}_t(0)$ does not violate any KKT conditions. For $a_{i:i>0}$, if $-\frac{t}{2\gamma(1+\ln(t))}\bar{\mathbf{g}}_t(i) \geq 0$, we set $a_i = -\frac{t}{2\gamma(1+\ln(t))}\bar{\mathbf{g}}_t(i)$ and $\lambda_i = 0$, and all of the KKT conditions are satisfied. If $-\frac{t}{2\gamma(1+\ln(t))}\bar{\mathbf{g}}_t(i) < 0$, we set $a_i = 0$, and $\lambda_i = \mathbf{g}_t(i)$, so $a_i\lambda_i = 0$ and also $\bigtriangledown_{\mathbf{a}} L(i) = 0$. Therefore, all of the KKT conditions can be met. With the updating rule stated in the theorem, all of the KKT conditions can be satisfied. Finally, as the problem (6) is convex, the updating rule in the theorem provides the optimal solution to (6).

This stochastic optimization algorithm provides an efficient updating rule for our original problem, and this is the key that our method can scale up to high dimensional datasets. Since the objective function in (1) is subdifferentiable, and the constraint set is convex, as shown in Xiao (2010), with a large enough number of samples and iteration steps, the updating rules finally approach to the optimal solution to (1).

The pseudo-code of the final stochastic optimization algorithm is summarized in **Algorithm** 1.

Algorithm 1 Dual Average Algorithm for Structured Kernel Feature Selection

```
Input: Data matrix $X$, Outcome labels $Y$, Feature structural relationship graph $G(V, E)$, a strongly convex auxiliary function $h(\mathbf{a})$, $\lambda_1$, $\lambda_2$.
Initialization: Compute the kernel matrices for $X$ and $Y$; Initialize $\mathbf{a} \in \min_{\mathbf{a}} h(\mathbf{a})$;
repeat
    1 Given the function $l_t$, compute the subgradient on $\mathbf{a}_t$: $\mathbf{g}_t$;
    2 Update the average subgradient $\bar{\mathbf{g}}_t = \frac{t-1}{t}\bar{\mathbf{g}}_{t-1} + \frac{1}{t}\mathbf{g}_t$;
    3 Calculate $\mathbf{a}$ with
      $(a_i)_t = \begin{cases} -\frac{t}{2\gamma(1+\ln(t))}\bar{\mathbf{g}}_t(i) & \text{if } i = 0 \\ [-\frac{t}{2\gamma(1+\ln(t))}\bar{\mathbf{g}}_t(i)]_+ & \text{if } i = 1, ..., p \end{cases}$
until Stopping criteria satisfied
Output: Fitting coefficients $\mathbf{a}$.
```

The required storage of the kernel matrices $\tilde{K^i}$, $i = 1, ..., p$ may take large memory space for high-dimensional datasets. Similar tricks adopted in (Yamada *et al.*, 2014) can be implemented to reduce memory requirements when needed.

3.2 Convergence and Regret Analysis

Following Xiao (2010), we can prove the following theorem:

Theorem 2 With an auxiliary function $h(\mathbf{a}) = ||\mathbf{a}||^2$, and the non-decreasing sequence $\{\beta_t\}$ with $\beta_t = \gamma(1 + \ln(t))$, $\{\mathbf{a}_t\}$ and $\{\mathbf{g}_t\}$ are two sequences generated by **Algorithm** 1. Suppose the optimal solution $\mathbf{a}^*$ to problem (1) satisfies $h(\mathbf{a}^*) \leq D$, for some $D > 0$, and there is a constant G such that $||\mathbf{g}_t||_* \leq G$ for all $t \geq 1$, we have the following properties for **Algorithm** 1:
a) For each $t \geq 1$, the average regret is bounded by

$$R_t(\mathbf{a}) \leq \left(\gamma D^2 + \frac{G^2}{2\gamma}\right)(1 + \ln(t)).$$

b) The sequence of primal variables are bounded by

$$||\mathbf{a}_{t+1} - \mathbf{a}^*|| \leq \\ \frac{2}{\gamma(1 + t + \ln(t))}\left(\left(\gamma D^2 + \frac{G^2}{2\gamma}\right)(1 + \ln(t)) - R_t(\mathbf{a}^*)\right).$$

Also we can have the convergence in the expectation form:
c)

$$\mathbf{E}||\mathbf{a}_{t+1} - \mathbf{a}^*|| \leq \frac{2}{1 + t + \ln(t)}\left(D^2 + \frac{G^2}{2\gamma^2}\right)(1 + \ln(t)).$$

Theorem 2(a) reveals that when $\gamma = \frac{G}{\sqrt{2D}}$, we can have the improved regret bound:

$$R_t(\mathbf{a}) = 2\sqrt{\frac{DG}{\sqrt{2}}}(1 + \ln(t)).$$

From **Theorem 2**(b-c), we can see that our algorithm has a convergence rate of $O(\ln(t)/t)$. A detailed proof of these results can be found in the supplementary file.

4 EXPERIMENTAL RESULTS

We have two sets of experiments to verify the effectiveness and efficiency of our methods on structured high dimensional datasets. The first one is based on simulation experiments using MRI data. The second one is to analyze the 3D PET brain images for Alzheimer's disease (AD) prognosis (Jack *et al.*, 2008; Xin *et al.*, 2014). For these studies, we compare our algorithm with Fused LASSO (Tibshirani *et al.*, 2005; Xin *et al.*, 2014), and HSIC-LASSO. For Fused LASSO we use the recent efficient implementation based on the graph-cut algorithm (Xin *et al.*, 2014) with the same efforts to provide scalable feature selection for 3D brain images.

4.1 Simulated Active Regions in MRI Images

In this set of experiments, we study the proposed method with a simulation of structural anomalies within MRI anatomical data. From the 1000 Functional Connectomes Project International Neuroimaging Data-Sharing Initiative (Biswal *et al.*, 2010), we randomly selected 200 3D anatomical MRI brain images from healthy subjects. Each image was spatially normalized to a $1mm \times 1mm \times 1mm$ custom, average anatomical template image using a low-dimensional free-form deformation image registration (Rueckert *et al.*, 1999) with $15mm$ control point spacing. For this simulation experiments, we equally partition the total samples into healthy (negative) samples and positive samples by simulating the perturbations from the original images. Considering computation efficiency, only one brain lobe region as shown in Figure 1 is chosen for study. One spherical region within the lobe is randomly perturbed as the active functional area with structural anomaly. Each voxel intensity within the active areas is modified by adding a random value g, which follows a Gaussian distribution, $N(\mu, \sigma^2)$. In our experiments, we take σ as the standard deviation of voxel intensity values of the original image. Among the selected original images without perturbation, the average value of σ is 262.75. We perturb the voxel intensity values in 100 positive samples in the randomly selected single spherical active region with a radius of $r = 4$ voxels. The images in the first row of Figure 1 display three-axis views for one example of an original MRI image. The second and third rows in Figure 1 are the images after perturbation in the active areas at different levels of μ.

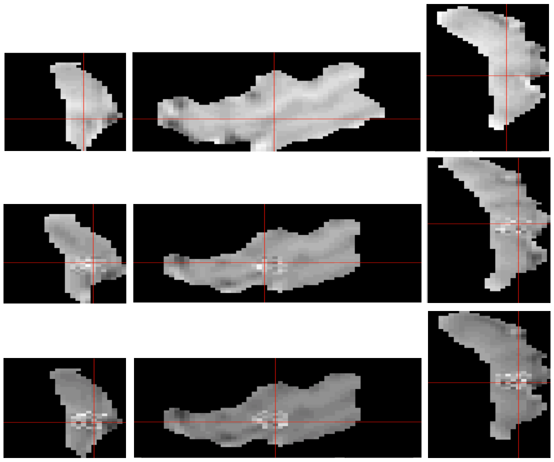

Figure 1: The first row shows one example from the original MRI images; the second row is the corresponding perturbed image at $\mu = 100$; the third row displays the perturbed image at $\mu = 200$.

For Fused LASSO and our method we directly adopt the learned parameters for prediction as both methods are formulated as supervised learning problems. For HSIC-LASSO, kernel SVM (Chang and Lin, 2011) based on the learned features is used for prediction. For the proposed model, we can use the learned parameters to predict the pairwise relationship between the testing sample with all of the training samples. Since it is a binary classification problem, we can use the sign of the accumulated prediction label to determine the final prediction value. As mentioned above, it is difficult to determine the testing sample's relationship with each training sample with HSIC-LASSO, and this forces us to train an additional kernel classifier for HSIC-LASSO. The measure on active region recovery accuracy ACC_{AR} is computed as follows:

$$ACC_{AR} = \frac{2R - ME}{2R},$$

where R denotes the number of voxels in the actual active region; and ME represents the binary voxel-wise matching error between the ground truth active region and the recovered region, which is the number of voxels in both binary images that are not in the overlap region. We take the R active voxels in the recovered region corresponding to the R voxels with the highest average value over all of the positive samples. When the recovered binary functional active region is the same as the ground truth region, $ME = 0$ and thereafter $ACC_{AR} = 1$. When the recovered region does not have any overlap voxel with the ground truth, $ME = 2R$ and hence $ACC_{AR} = 0$.

In this set of experiments, 200 samples are divided into the training set and testing set. The training set contains 50 randomly chosen positive samples and 50 negative ones. The rest of the samples go to the testing set. All of the model

Table 1: Comparison for Simulated MRI Images with Linear Responses

Method	Proposed	FL	HSIC-LASSO
Pred. Accuracy	96%	70 %	69%
Reg. Accuracy	78.1%	33.3%	23.1%
CPU time (sec.)	65.6	431.5	73.7

parameters are learned based on the training set with five-fold cross validation. Since the number of training samples is not large, we use all of training samples in our stochastic algorithm without any subsampling on the training dataset. In this set of simulation experiments, we study all of the three methods on three different types of input-output relationships: linear, additive nonlinear, and non-additive nonlinear.

4.1.1 Linear Response

In this experiment, we compare all of the models based on simulated linear responses from perturbed MRI images with 100 positive samples having the active regions perturbed with random values following $N(\mu, \sigma^2)$ with $\mu = 100$, and the other 100 negative samples from the original MRI images. The output label for each image is directly determined by whether the image is perturbed. The results for the three comparing methods are shown in Table 1, and the recovered regions are shown in Figure 2.

Table 1 shows that our method can achieve higher prediction accuracy as well as higher active region recovery accuracy. Moreover, our algorithm takes fewer computational resources. The results in this experiment show that our method can work robustly even though the active signal is relatively weak. The proposed model and Fused LASSO can get higher ACC values due to the additional structure knowledge of the data that are incorporated in the model formulation. Without the structure constraints, HSIC-LASSO misses many active voxels with the redundancy penalty term in their formulation. This is the reason why the recovered region is sparse and the ACC is low in HSIC-LASSO. We also have tried lower sparse penalty in HSIC-LASSO but it does not significantly change the results. While we note that HSIC-LASSO can achieve similar computing time compared to our proposed method due to the efficiency of their dual augmented Lagrangian (DAL) algorithm. However, HSIC-LASSO does not impose any structural constraints, which is one of bottlenecks for scalability of structured kernel feature selection.

4.1.2 Additive Nonlinear Response

In this experiment, we set $\mu = 200$ for perturbations. Among 200 original images, 150 are chosen to be perturbed by adding random values following $N(\mu, \sigma^2)$ to the corresponding voxels in the selected active regions. In addition, in order to create a nonlinear response model, not all of these samples are labelled as positive samples. We divide the voxels within the active regions into four groups: $V1, V2, V3, V4$ according to the spacial order in the image. Then we compute a nonlinear response value $\psi = \sum_{\forall v1 \in V1, v2 \in V2, v3 \in V3, v4 \in V4} sin(v1) + exp(v2/c1) + v3/c2 + (v4/c3)^2$, where $c1 = 2000, c2 = 1500$, and $c3 = 1500$ are constants in this experiment. All the perturbed images are ranked in an ascending order of ψ values. The top 100 samples are considered as positive samples while the other 100 samples are labelled as health (or negative) samples.

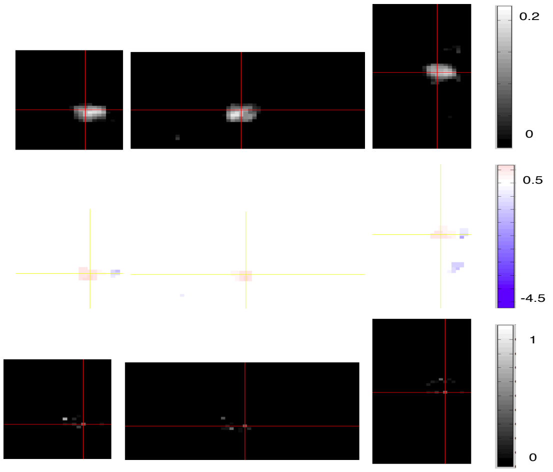

Figure 2: Active regions recovered by the proposed method, Fused LASSO and HSIC-LASSO for simulated MRI images with linear responses.

The results for this experiment are presented in Table 2. In the supplementary file, Figure 1 illustrates the recovered regions by three methods. It is clear that our proposed model takes lead in both accuracies and speed. The high prediction accuracy compared to the Fused LASSO is due to the kernel method in our model for incorporating potential nonlinear input-output relationships. By enforcing structural constraints, our structured kernel feature selection also performs superior to HSIC-LASSO. It is interesting to note that the Fused LASSO can achieve higher ACC for active region recovery compared to HSIC-LASSO because of the incorporated spacial structures. However, Fused LASSO takes much more computing time than the other two methods due to the incorporated non-smooth structure constraints even with the fast proximal and graph-cut algorithms implemented in Xin *et al.* (2014).

Based on these simulation experiments, our structured kernel feature selection with the dual average stochastic optimization algorithm can robustly recover potential active function regions, accurately predict output responses, and scale better with both the sample size and feature dimen-

Table 2: Comparison for Simulated MRI Images with Additive Nonlinear Responses

Method	Proposed	FL	HSIC-LASSO
Pred. Accuracy	94%	62 %	65%
Reg. Accuracy	74.5%	64.5%	27.9%
CPU time (sec.)	62.1	414.3	80.5

Table 3: Comparison for Simulated MRI Images with Non-additive Nonlinear Responses

Method	Proposed	FL	HSIC-LASSO
Pred. Accuracy	75%	69 %	60%
Reg. Accuracy	70.9%	28.0%	0.0%
CPU time (sec.)	69.5	2230.4	89.9

sion compared to the other existing feature selection methods.

4.1.3 Non-additive Nonlinear Response

In this experiment, the simulation data is generated in a similar way as in the previous experiments. But this time we randomly choose the voxels in the four groups, and the nonlinear response value $\psi = \sum_{\forall v1 \in V1, v2 \in V2, v3 \in V3, v4 \in V4} v1 \times exp(v2/c1)/c2 + (v3/c3)^2 \times v4$, where $c1 = 2000$, $c2 = 6200$ and $c3 = 1500$. Similarly, top ranked 100 perturbed images in the ascending order of ψ are set as positive samples and the remaining 100 images are negative samples.

The results of this experiment for prediction accuracies, active region recovery accuracies, and computational time are given in Table 3. In the supplementary file, Figure 2 displays the recovered regions by three methods. As visualized in the figures, our method is much more robust than the other two methods. For non-additive and nonlinear responses, the objective function is more complicated, and Fused LASSO and HSIC-LASSO take longer time to reach to the optimal values. The computational time for the Fused LASSO has increased dramatically. The possible reason is that as the problem becomes complicated, the line search step in the proximal algorithm in the Fused LASSO takes much longer time. In this experiment, HSIC-LASSO fails to identify any responsive voxels inside the active region due to the lack of structural constraints in their formulation.

The results in this set of experiments show that our model can recover active function regions in high dimensional structured data, even when the response signal is weak and complicated.

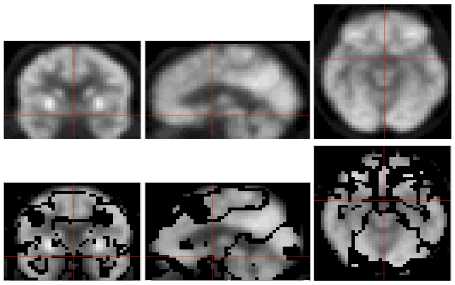

Figure 3: The first row displays the mean image of the original PET images in three-axis views and the second row shows the corresponding mean image after preprocessing.

Table 4: Comparison on Pet 3D Brain Images

Method	Proposed	FL	HSIC-LASSO
Pred. Accuracy	95.0%	85.9 %	87.9%
CPU time (sec.)	163.5	2786.2	187.9

4.2 PET 3D Brain Images

In this section, we test the proposed method on a 3D positron emission tomography (PET) dataset, which is collected from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Jack *et al.*, 2008). We collected 95 Alzheimer's disease (AD) patients and 102 healthy subjects in this set of experiments. With the affine transformation and subsequent non-linear warping algorithm (Friston *et al.*, 1995) in the SPM MATLAB toolbox, each image was spatially normalized to the Montreal Neurological Institute (MNI) template (Fonov *et al.*, 2011). The data was resampled and the resolution was reduced to $4mm \times 4mm \times 4mm$ to save computation time. Student's t-test was used to remove the voxels that do not differ significantly between patients and healthy people. Furthermore, the voxels with very small intensity values are also removed to reduce computational cost. Figure 3 shows the mean image before and after pre-processing.

The dataset is divided into two sets: the training set contains 51 healthy people and 47 patients, the testing set has 51 healthy people and 48 patients. The parameters are learned by five-fold cross validation on the training data set according to the prediction accuracy. Table 4 provides the performance comparison for the three comparing methods. We can see that our method again performs much better on prediction than the other two approaches. Figure 4 gives the predicted active regions by three models. We use the voxel-wise average intensity values of the healthy brain images as the reference background, and then we add in the learned voxel-wise fitting coefficient weights by the three models on the background. We can that see our method can recover multiple coherent regions.

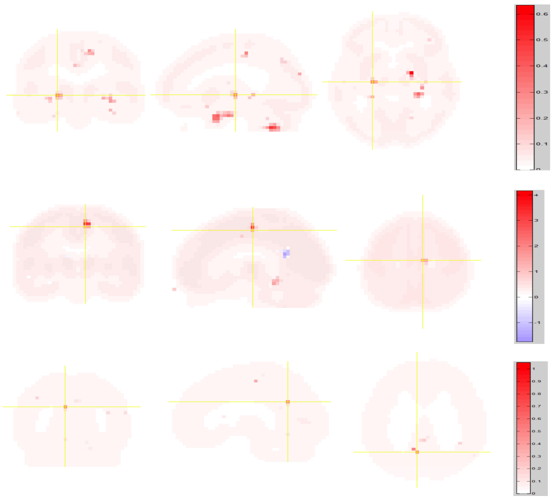

Figure 4: Active regions recovered by the proposed method, Fused LASSO and HSIC-LASSO for PET 3D brain images.

5 RELATED WORK

Although based on the same HSIC, our structured kernel feature selection method is quite different from the existing kernel feature selection methods in addition to our explicitly enforced structural constraints in the formulation (3). For example, the formulation for the Hilbert-Schmidt Feature Selection (HSFS) (Masaeli *et al.*, 2010) is as follows:

$$\min_{W \in R^{P \times P}} -HSIC(WX, Y) + \lambda \sum_{i=1}^{P} ||w_i||_\infty,$$

where $W = [w_1, ..., w_d]$ is a transformation matrix. Limited-memory BFGS (L-BFGS) algorithm (Nocedal and Wright, 2003) can be used to solve the problem. One limitation of HSFS is that the objective function is non-convex. Hence, with different starting points for optimization, we may get different solutions. In addition, the estimation of the transformation matrix with p^2 variables is computationally expensive, especially when we have a large number of candidate features as witnessed in our 3D image analysis problems.

Our model is also quite different from other feature-wise nonlinear methods, including HSIC, FVM, HSIC-LASSO (Cortes *et al.*, 2012; Li *et al.*, 2006; Yamada *et al.*, 2014). In Yamada *et al.*(2014), they propose to minimize the following objective function:

$$\frac{1}{2}||\bar{L} - \sum_{k=1}^{p} \bar{K}||^2 = \frac{1}{2} HSIC(Y, Y) - \sum_{i} a_i HSIC(Y, X_{\bullet i})$$
$$+ \frac{1}{2} \sum_{ij} a_i a_j HSIC(X_{\bullet i}, X_{\bullet j}).$$

With the last term, their methods aim to eliminate the correlated redundant features. In this paper we are trying to identify all of the features that have potential predictive power to the output, which has more reasonable applications such as in identifying functional regions inside human brain images for neurodegenerative disease prognosis and diagnosis. In addition, the least squared loss function adopted in these methods may give degenerated results when solving binary classification problems as the kernel matrix $\bar{L}$ on output Y will degenerate to a bi-value matrix.

To summarize, our structured kernel feature selection problem is specifically designed for classification with the Hinge loss function, which can be represented by HSIC terms as we have shown earlier. Enforcing that related features should be selected together as they have higher probability in similarly correlating the output, our structured kernel feature selection can get more robust feature selection results. In addition to the differences in formulations, we derive a tailored stochastic optimization algorithm so that the proposed method can be implemented to efficiently solve feature selection and active region recovery when we have big and high-dimensional data such as 3D brain images in our experiments.

6 CONCLUSIONS

In this paper we propose a new kernel feature selection model for binary classification problems. Based on Hilbert-Schmidt Independent Criteria, with the structure knowledge among features incorporated into the objective function, our model can effectively and robustly identify the active regions related to the outcome of interest. Our method can scale up to large-scale data problems with the efficient stochastic algorithm based on the dual average method. Experimental results on both simulation data and real-world 3D image data have verified the effectiveness and efficiency of the proposed method. Our structured formulation for kernel feature selection together with the accompanying stochastic optimization method provides a practical approach for large-scale structured data feature selection and active function region recovery from 3D brain images. Our model can be further improved with the less memory techniques (Yamada *et al.*, 2014) and faster stochastic methods (Xiao, 2010), which will be our future research directions.

Acknowledgements

This work was partially supported by Awards #1447235 and #1244068 from the National Science Foundation; as well as Award R21DK092845 from the National Institute Of Diabetes And Digestive And Kidney Diseases, National Institutes of Health.

References

R. Tibshirani, Regression shrinkage and selection via the LASSO, Journal of the Royal Statistical Society, vol. 58, pp. 267-288, 1996.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, Sparsity and smoothness via the fused LASSO, Journal of the Royal Statistical Society Series B, pp. 91-108, 2005.

F. Li, Y. Yang, and E. P. Xing, From LASSO regression to feature vector machine, in NIPS, 2006.

M. Yamada, W. Jitkrittum, L. Sigal, E. P. Xing, and M. Sugiyama, High-dimensional feature selection by feature-wise kernelized LASSO, Neural Computation, vol. 26, pp. 185-207, 2014.

I. Rodriguez-Lujan, R. Huerta, C. Elkan, and C. S. Cruz, Quadratic programming feature selection, Journal of Machine Learning Research, vol. 11, pp. 1491-1516, 2010.

L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt, Feature selection via dependence maximization, Journal of Machine Learning Research, vol. 13, pp. 1393-1434, 2012.

A. Gretton, O. Bousquet, A. Smola, and B. Schlkopf, Measuring statistical dependence with Hilbert-Schmidt norms, Algorithmic Learning Theory, vol. 3734, pp. 63-77, 2005.

M. Masaeli, G. Fung, and J. G. Dy, From transformation-based dimensionality reduction to feature selection, in ICML, 2010.

P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman, Sparse additive models, Journal of Machine Learning Research, vol. 71, pp. 1009-1030, 2009.

R. Tomioka, T. Suzuki, and M. Sugiyama, Superlinear convergence of dual augmented lagrangian algorithm for sparsity regularized estimation, Journal of Machine Learning Research, vol. 12, pp. 1537-1586, 2011.

B. Xin, Y. Kawahara, Y. Wang, and W. Gao, Efficient generalized fused LASSO with its application to the diagnosis of Alzheimers disease, in AAAI, 2014.

L. Xiao, Dual averaging methods for regularized stochastic learning and online optimization, Journal of Machine Learning Research, pp. 2543-2596, 2010.

H. Yang, Z. Xu, I. King, and M. R. Lyu, Online learning for group LASSO, in ICML, 2010.

C. Jack, M. Bernstein, N. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. Britson, J. Whitwell, C. Ward, A. Dale, J. Felmlee, J. Gunter, D. Hill, R. Killiany, N. Schuff, S. Fox-Bosetti, C. Lin, C. Studholme, C. DeCarli, K. Gunnar, H. Ward, G. Metzger, K. Scott, R. Mallozzi, D. Blezek, J. Levy, J. Debbins, A. Fleisher, M. Albert, R. Green, G. Bartzokis, G. Glover, J. Mugler, and M. Weiner, The Alzheimers disease neuroimaging initiative (ADNI): MRI methods, J Magn Reson Imaging, vol. 27, pp. 685-691, 2008.

B. B. Biswal, M. Mennes, X.-N. Zuo, S. Gohel, C. Kelly, S. M. Smith, C. F. Beckmann, J. S. Adelstein, R. L. Buckner, S. Colcombe, A.-M. Dogonowski, M. Ernst, D. Fair, M. Hampson, M. J. Hoptman, J. S. Hyde, V. J. Kiviniemi, R. Kotter, S.- J. Li, C.-P. Lin, M. J. Lowe, C. Mackay, D. J. Mad- den, K. H. Madsen, D. S. Margulies, H. S. Mayberg, K. McMahon, C. S. Monk, S. H. Mostofsky, B. J. Nagel, J. J. Pekar, S. J. Peltier, S. E. Petersen, V. Riedl, S. A. R. B. Rombouts, B. Rypma, B. L. Schlaggar, S. Schmidt, R. D. Seidler, G. J. Siegle, C. Sorg, G. J. Teng, J. Veijola, A. Villringer, M. Walter, L. Wang, X.-C. Weng, S. Whitfield-Gabrieli, P. Williamson, C. Windischberger, Y.-F. Zang, H.-Y. Zhang, F. X. Castellanos, and M. P. Milham, Toward discovery science of human brain function, Proceedings of the National Academy of Sciences, vol. 107, no. 10, pp. 4734-4739, 2010.

D. Rueckert, L. Sonoda, C. Hayes, D. Hill, M. Leach, and D. Hawkes, Nonrigid registration using freeform deformations: Application to breast MR images, Medical Imaging, IEEE Transactions on, vol. 18, no. 8, pp. 712-721, 1999.

C.-C. Chang and C.-J. Lin, Libsvm : a library for support vector machines, ACM Transactions on Intelligent Systems and Technology, vol. 2, no. 27, pp. 1-27, 2011.

K. J. Friston, J. Ashburner, C. D. Frith, J.-B. Poline, J. D. Heather, and R. S. J. Frackowiak, Spatial registration and normalization of images, Human Brain Mapping, pp. 165-189, 1995.

V. Fonov, A. Evans, K. Botteron, C. Almli, R. McKinstry, D. Collins, and Brain Development Cooperative Group, Unbiased average age-appropriate atlases for pediatric studies, NeuroImage, vol. 54, no. 1, pp. 317-323, 2011.

J. Nocedal and S. J. Wright, Numerical Optimization. Springer Press, 2003.

C. Cortes, M. Mohri, and A. Rostamizadeh, Algorithms for learning kernels based on centered alignment, Journal of Machine Learning Research, vol. 13, pp. 795-828, 2012.