Side Information Dependence as a Regularizer for Analyzing Human Brain Conditions across Cognitive Experiments

Shuo Zhou\textsuperscript{1}, Wenwen Li\textsuperscript{2}, Christopher R. Cox\textsuperscript{3}, Haiping Lu\textsuperscript{1,4}
\textsuperscript{1}Department of Computer Science, University of Sheffield, United Kingdom
\textsuperscript{2}Centre for Clinical Brain Sciences, University of Edinburgh, United Kingdom
\textsuperscript{3}Department of Psychology, Louisiana State University, United States
\textsuperscript{4}Sheffield Institute for Translational Neuroscience, United Kingdom
\{szhou20, h.lu\}@sheffield.ac.uk, wenwen.li@ed.ac.uk, chriscox@lsu.edu

Abstract
The increasing of public neuroimaging datasets opens a door to analyzing homogeneous human brain conditions across datasets by transfer learning (TL). However, neuroimaging data are high-dimensional, noisy, and with small sample sizes. It is challenging to learn a robust model for data across different cognitive experiments and subjects. A recent TL approach minimizes domain dependence to learn common cross-domain features, via the Hilbert-Schmidt Independence Criterion (HSIC). Inspired by this approach and the multi-source TL theory, we propose a Side Information Dependence Regularization (SIDeR) learning framework for TL in brain condition decoding. Specifically, SIDeR simultaneously minimizes the empirical risk and the statistical dependence on the domain side information, to reduce the theoretical generalization error bound. We construct 17 brain decoding TL tasks using public neuroimaging data for evaluation. Comprehensive experiments validate the superiority of SIDeR over ten competing methods, particularly an average improvement of 15.6\% on the TL tasks with multi-source experiments.

Introduction
In cognitive neuroscience, neuroimaging can help relate different cognitive functions to patterns of neural activity using functional magnetic resonance imaging (fMRI) (Ogawa et al. 1990). This often takes the form of a classification problem (Cox and Savoy 2003), e.g., distinguishing between brain conditions associated with experimental stimuli. While fMRI produces volumes with the number of voxels in the order of $10^5$, a typical experiment will have on the order of 100 discrete trials. This severely constrains the number of training examples available for the classifier. Moreover, neuroimaging data are noisy and contain a significant amount of physiological, respiratory, and mechanical artifacts, which requires robust modeling against noise (Aydore, Thirion, and Varoquaux 2019).

Transfer learning (TL) is an attractive machine learning scheme that can improve the classification performance on a learning task by leveraging the knowledge from related tasks. The task of interest is called the target domain, while the task(s) to be leveraged is called the source domain (Pan et al. 2011). A TL problem is homogeneous when the feature and label space of the source and target domains are the same, and heterogeneous if they are different.

Transfer learning techniques have been studied in some fMRI applications. Mensch et al. (2017) take a multi-task learning approach to use resting-state data from the Human Connectome Project (Van Essen et al. 2012) to learn a general representation via matrix decomposition and then jointly optimize multiple heterogeneous task-based fMRI classification tasks. Zhang, Chen, and Ramadge (2018) take a matrix factorization approach that relies on shared subjects between datasets to learn better subject factor matrices. Deep models such as autoencoder (Velioglu and Vural 2017; Li, Parikh, and He 2018) and AlexNet (Zhang et al. 2019) pre-trained on generic source data have also been used to represent the target fMRI data for classification.

However, the source data used by the existing fMRI TL studies are independent to the target classification task. While public neuroimaging data from multiple sites, e.g., the OpenNeuro (Gorgolewski et al. 2017), have many similar brain conditions across different cognitive experiments. This enables homogeneous TL studies to leverage the power of overlapping labels across domains. Furthermore, it may potentially offer interpretation/insights from the domain shift perspective for neuroscientists. Here, we make the first attempt, to the best of our knowledge, to investigate homogeneous TL for brain condition decoding.

Homogeneous TL methods are mainly studied in the fields of computer vision (CV) and natural language processing (NLP). They focus on minimizing data distribution mismatch, i.e., making features from the source and target to have as similar distributions as possible, via 1) learning a feature mapping (Pan et al. 2011; Long et al. 2013b); or 2) jointly optimizing the distribution mismatch and classifier parameters (Long et al. 2013a; Wang et al. 2018).

Brain condition decoding presents TL challenges different from those in CV/NLP. fMRI data are generated by brain signals, which are not natural images that human visual system has adapted to interpret. Consequently, fMRI analysis relies heavily on statistics. Furthermore, cognitive stimuli are implemented varying across experiments. Even
the same information can be encoded as different patterns of activity by different brains (Chen et al. 2015). Hence, each subject can be considered as a unique learning task to extract subject-specific features (Rao et al. 2013) in fMRI studies. Additionally, as mentioned before, fMRI data are noisy. Therefore, for TL in brain condition decoding, it can be beneficial to take more domain information, such as experiment designs and subjects, into account to learn a robust model.

Recently, the maximum independence domain adaptation (MIDA) (Yan, Kou, and Zhang 2018) introduces a new domain independence minimization approach to TL. It learns common, cross-domain features by minimizing statistical dependence on auxiliary domain side information, as measured by the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al. 2005). This inspired us to encode different experiment designs and subjects as auxiliary domain covariates for TL in brain condition decoding.

In this paper, we propose a Side Information Dependence Regularization (SIDEr) framework for homogeneous TL in brain condition decoding. The contributions are threefold: (1) We discover the relationship between HSIC and maximum mean discrepancy (MMD) and derive two HSIC-based generalization bounds for and multi-source TL. This enables us to express different experiment designs and subjects as auxiliary domain covariates for TL in brain condition decoding. (2) Under this framework, we construct a simplified HSIC and incorporate the hinge loss that can take unlabeled samples into account following the Manifold Regularization framework formulation (Belkin, Niyogi, and Sindhwani 2006). This gives us the SIDErSVM algorithm. (3) We construct 17 new homogeneous brain decoding TL tasks by identifying datasets with homogeneous brain conditions from public repositories. Experiments on these tasks show the superior performance of SIDEr over ten competing methods.

**Preliminaries**

**Hilbert-Schmidt Independence Criterion (HSIC)** (HSIC) is a non-parametric criterion for measuring the statistical dependence between two sets $X = \{x_i\}$ and $Y = \{y_i\}$, both with size $n$. HSIC tests whether $\Pr(x) \neq \Pr(y)$ Denoting the empirical HSIC as $\rho_h(X,Y)$, it can be computed via (Gretton et al. 2005)

$$
\rho_h(X,Y) = \text{tr}(KHLH)/(n-1)^2,
$$

(1)

where $K$, $H$, $L \in \mathbb{R}^{n \times n}$, $K_{ij} := k(x_i, x_j)$, $L_{ij} := k(y_i, y_j)$, $k_x(\cdot, \cdot)$ and $k_y(\cdot, \cdot)$ are two kernel functions, e.g., linear, polynomial, or radial basis function (RBF), $H = I - \frac{1}{n}11^T$ is the centering matrix, and $\text{tr}(\cdot)$ is the trace function. HSIC is zero if and only if $X$ and $Y$ are independent. A larger HSIC value suggests stronger statistical dependence.

**Related works.** We summarize the state-of-the-art homogeneous TL methods as the following three approaches.

**Distribution mismatch minimization mapping** is a popular approach to homogeneous TL, which learns a mapping via minimizing the marginal or joint distribution mismatch between a source and a target, e.g., as in Transfer Component Analysis (TCA) (Pan et al. 2011) and Joint Distribution Adaptation (JDA) (Long et al. 2013b). The distribution mismatch is typically measured by the maximum mean discrepancy (MMD) criterion (Borgwardt et al. 2006). TCA also has a semi-supervised version, semi-supervised TCA (SSTCA), that introduces an additional label dependence objective $\rho_h(\phi(X^s), Y)$ to maximize, where $X^s$ denotes labeled data, and $Y$ is a label matrix. **Domain-invariant classifier** is another approach that learns a classifier by optimizing the prediction loss and distribution mismatch jointly. Long et al. (2013a) proposed Adaptation Regularization based Transfer Learning (ARTL) framework by incorporating joint distribution mismatch (as in JDA) into the manifold regularization framework (Belkin, Niyogi, and Sindhwani 2006). Based on ARTL, Wang et al. (2018) proposed Manifold Embedded Distribution Alignment (MEDA) by introducing a trade-off between marginal and conditional distribution mismatch for dynamic transfer.

**Domain dependence minimization mapping.** Yan, Kou, and Zhang (2018) proposed the MIDA method using a new approach that extracts cross-domain features by learning a mapping to minimize the dependence on domain information, e.g., device and time, which is not directly modeled in the previous two approaches. There is also a semi-supervised version of MIDA, i.e., SMIDA, which maximizes the label dependence (as in SSTCA).

**Proposed Method**

This section first defines the transfer learning problem. Then we perform theoretical studies on HSIC to reveal the relationships between HSIC and MMD. This enables us to derive a generalization bound for the multi-source TL setting. Subsequently, the bound motivates the formulation of the Side Information Dependence Regularization (SIDEr) learning framework for homogeneous TL.

**Problem Definition**

In a cognitive experiment, each subject is presented a set of stimuli (conditions) designed by neuroscientists. An experiment typically features one or a few (if repeated) samples per condition per subject. We consider a target dataset (experiment) have both labeled and unlabeled samples, and there are labeled samples with the homogeneous brain conditions that acquired from one or more source experiments, where the experiment designs are different. The objective is to predict the human brain conditions of unlabeled target samples.

The **target** cognitive experiment has $n_t$ fMRI data samples $X_t = [X^s_t, X^u_t] \in \mathbb{R}^{d \times n_t}$ of $m$ brain conditions for classification. $X^s_t \in \mathbb{R}^{d \times \tilde{n}_s}$ and $X^u_t \in \mathbb{R}^{d \times (n_t - \tilde{n}_s)}$ are labeled and unlabeled target data, respectively, $d$ is the number of fMRI features, e.g., voxels.

The **source** consists of data from one or more cognitive experiments with $n_s$ labeled samples $X_s \in \mathbb{R}^{d \times n_s}$ in total, with the same $m$ brain conditions as the target data.

**Domain covariate encoding.** Denote the target and source data jointly as $X = [X_s, X_t] \in \mathbb{R}^{d \times n}$, $n = n_s + n_t$. Each fMRI sample $x_i (i = 1, \cdots, n)$ is collected with a particular experiment implementation $j$ from a particular subject $k$, where $j = 1, \cdots, p$ and $k = 1, \cdots, q$, i.e., there are $p$ unique experiment implementations and $q$ unique...
The empirical MMD can be computed via tracking the MMD between the two domains is estimated to be the difference between the means of the two distributions. Specifically, we use a simple one-hot encoding strategy to encode such domain covariates. Formally, we construct a one-hot experiment implementation covariate matrix \( E \in \mathbb{R}^{n \times p} \), where its \((i, j)\)th element \( e_{i,j} = 1 \) if \( x_i \) is collected from experiment \( j \) and \( e_{i,j} = 0 \) otherwise. Similarly, we construct a one-hot subject covariate matrix \( S \in \mathbb{R}^{n \times q} \), where \( s_{i,k} = 1 \) if \( x_i \) is from subject \( k \) and \( s_{i,k} = 0 \) otherwise. We then obtain the auxiliary domain covariate matrix \( D \in \mathbb{R}^{d \times n} \) by concatenating \( E^T \) and \( S^T \), where \( d = p + q \).

**Multi-source view of brain decoding.** As mentioned in Sec. Introduction, each cognitive experiment can be designed differently, and each subject can encode the stimulus differently, i.e., \( P(x \mid E_i) \neq P(x \mid E_j) \), and \( P(x \mid S_i) \neq P(x \mid S_j) \). \( E \) and \( S \) denote an experiment and a subject, respectively. Traditional TL methods consider two different datasets (experiments in brain decoding) as different domains. If we also consider each subject as a domain as in (Rao et al. 2013), then each unique experiment-subject combination is a domain, i.e., brain decoding TL tasks is essentially a multi-source transfer problem. Therefore, in the following, we study HSIC in the multi-source TL setting.

**Theoretical Studies on HSIC**

The domain dependence of the data can be computed via \( \rho_h(x, D) \), using the domain covariate matrix \( D \) defined above. Now we show when using a one-hot encoding and a linear kernel for \( D \) in HSIC, we can derive an equivalence between HSIC and MMD, as shown in the following lemma. Based on this lemma, we can derive generalization bounds for HSIC-based TL and formulate our new framework.

**Lemma 1. HSIC is proportional to MMD when there are only two discrete domains, i.e., with a degenerated one-hot domain covariate vector, e.g., \( d_0 = [0 \cdots 0 1 \cdots 1] \in \mathbb{R}^n \), and linear kernel is used for \( d_0 \) in HSIC.**

**Proof.** The MMD between the two domains is

\[
\text{MMD}(X_s, X_t) = \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} x_{s_i} - \frac{1}{n_t} \sum_{i=1}^{n_t} x_{t_i} \right\|^2 \tag{2}
\]

where \( \mathcal{H}_k \) denotes a reproducing kernel Hilbert space (RKHS). The empirical MMD can be computed via \( \text{tr}(KL') \) (Pan et al. 2011), where \( K = k(X, X) \in \mathbb{R}^{n \times n} \), \( X = [X_s, X_t] \in \mathbb{R}^{d \times n} \), and \( L' \in \mathbb{R}^{n \times n} \) is defined as

\[
L'_{ij} = \begin{cases} \frac{1}{n_s} & \text{if } x_i, x_j \in X_s, \\ \frac{1}{n_t} & \text{if } x_i, x_j \in X_t, \\ 0 & \text{otherwise}. \end{cases} \tag{3}
\]

By Eq. (1), \( \rho_h(x, d_0) = \text{tr}(KLH)/\left(n - 1\right)^2 \), where \( K = k(X, X) \) is exactly the same kernel matrix as in the MMD, and \( L = d_0 \cdot d_0 \), i.e., \( L_{i,j} = 1 \), if \( x_i, x_j \in X_t \), and otherwise \( L_{i,j} = 0 \). Let \( \tilde{L} = HLH \), resulting in

\[
\tilde{L}_{ij} = \begin{cases} \frac{1}{n_s^2} & \text{if } x_i, x_j \in X_s, \\ \frac{1}{n_t^2} & \text{if } x_i, x_j \in X_t, \\ 0 & \text{otherwise}. \end{cases} \tag{4}
\]

By comparing Eq. (3) and Eq. (4), we have

\[
\text{MMD}(X_s, X_t) = u \rho_h(x, d_0), \tag{5}
\]

where \( u = \frac{\left(n^2 - (n - 1)^2\right)}{\left(n - 1\right)^2} \). For a learning task, \( u \) is a constant. This completes the proof.

**Generalization bound.** For the multi-source setting, we define a domain \( j \) as the samples \( X_j \) drawn from a distribution \( D_j \) on the inputs \( X \) and a labeling function \( f_j : X \to \{0, 1\} \). A hypothesis \( f : X \to \{0, 1\} \). We consider a classifier trained on \( J \) \((J \geq 1)\) distinct source domains \( X_s = [X_1, \ldots, X_j] \), with a total of \( n_s \) samples and a domain weight vector \( \alpha = [\alpha_1, \ldots, \alpha_J] \), where \( \sum_{j=1}^{J} \alpha_j = 1 \), and derive the bounds on its generalization performance on a target domain, i.e., \( \epsilon(f) \) or \( \epsilon(f, f_t) \).

By the proof of Theorems 4 and 5 in (Ben-David et al. 2010), \( \epsilon(f) \leq \epsilon_s(f) + d_{\Delta(\mathcal{H})}(x_s, x_t) + \lambda^* + \Omega \), where \( d_{\Delta(\mathcal{H})}(x_s, x_t) = 2 \sup_{f,f' \in \mathcal{H}} |\epsilon_s(f, f') - \epsilon(f, f')| \) is the empirical symmetric \( H \)-divergence, \( \lambda^* \) is the risk of an ideal joint hypothesis, and \( \Omega \) denotes the complexity of hypothesis space. Here we derive the following lemma for \( \lambda^* \) and then an HSIC-based bound for the multi-source setting.

**Lemma 2.** Let \( H \) be a hypothesis space, \( \alpha_{j+1} = 1 \), and \( D_{j+1} = D_i \), for \( j \in \{1, \ldots, J\} \), let \( X_j \) be samples drawn from \( D_i \) with domain weight \( \alpha_j \) and labeling function \( f_j \), let \( f^* = \arg \min_{f \in H} \sum_{j=1}^{J} \alpha_j \epsilon_j(f) \) be the ideal joint hypothesis, \( \lambda^* = \sum_{j=1}^{J+1} \alpha_j \epsilon_j(f^*) \), then

\[
\lambda^* \leq \frac{1}{2} \sum_{j=1}^{J+1} \alpha_j \left(d_{\Delta(\mathcal{H})}(X_j, X_{\alpha_j})\right) + \Omega, \tag{6}
\]

where \( X_\alpha = [X_1, \ldots, j, X_j] \), \( X_{\alpha,j} \) denotes exclude \( X_j \) from \( X_\alpha \), and \( \Omega = \sum_{j=1}^{J+1} \alpha_j E_{x \sim D_{\alpha_j}} [f^*(x) - f_j(x)] + \Omega \).

**Proof.** For any \( i \in \{1, \ldots, J+1\} \),

\[
\epsilon_i(f^*) = \epsilon_i(f^*) + \alpha_i \epsilon_i(f^*, f_i) - \alpha_i \epsilon_i(f^*, f_i) \\
\leq |\epsilon_i(f^*, f_i) - \epsilon_i(f^*, f_i)| + |\epsilon_i(f^*, f_i)| \\
\leq \frac{1}{2} d_{\Delta(\mathcal{H})}(D_i, D_{\alpha_i}) + \alpha_i |\epsilon_i(f^*, f_i)| \\
\leq \frac{1}{2} d_{\Delta(\mathcal{H})}(X_i, X_{\alpha_i}) + \Omega_i \\
+ E_{x \sim D_{\alpha_i}} |f^*(x) - f_i(x)|.
\]

Sum over \( \epsilon_i(f^*), i \in \{1, \ldots, J+1\} \), completes the proof.

**Theorem 1 (Multi-source).** Let \( H \) be a hypothesis space, \( \alpha_{J+1} = 1 \), and \( D_{J+1} = D_i \), for \( j \in \{1, \ldots, J\} \), let \( X_j \) be labeled samples of size \( n_j \) drawn from \( D_j \) with domain weight \( \alpha_j \) and labeling function \( f_j \). Let \( D \) be a one-hot domain covariate matrix, then for \( f \in H \):

\[
\epsilon(f) \leq \epsilon_s(f) + \frac{3}{5} \rho_h(x, D U) + \Omega, \tag{8}
\]

where \( \epsilon_s(f) \) is the empirical risk of \( f \) on the source data, \( U = \text{diag}(\mathbf{u}) \), \( \text{diag}(\cdot) \) is the diagonal function, \( \mathbf{u} \in \mathbb{R}^n \) is a vector, \( u_i = \alpha_i n_i^2 (n - 1)^2 / (n - 2)(n - 1+\Omega_i) \), if \( x_i \in X_j, i = 1, \ldots, n, \) and \( \Omega = \sum_{j=1}^{J+1} \alpha_j E_{x \sim D_{\alpha_j}} |f^*(x) - f_j(x)| + \Omega \).
Proof. By the theoretical results in (Ben-David et al. 2010) mentioned above and Lemma 2, we have
\[
\epsilon_{\ell}(f) \leq \hat{\epsilon}_{s}(f) + \frac{1}{2} \sum_{j=1}^{J+1} \alpha_j \left( d \Delta \mathcal{H}(X_j, X_{\alpha \setminus j}) \right) + \Omega \\
+ \frac{1}{2} \sum_{j=1}^{J+1} \alpha_j \left( \hat{d} \Delta \mathcal{H}(X_j, X_{\alpha \setminus j}) \right) + \Omega
\]
(9)
\[
\leq \hat{\epsilon}_{s}(f) + \frac{3}{2} \sum_{j=1}^{J+1} \alpha_j \left( \hat{d} \Delta \mathcal{H}(X_j, X_{\alpha \setminus j}) \right) + \Omega.
\]
Empirical $\mathcal{H}$-divergence can be estimated by MMD as in the existing TL studies, e.g., (Long et al. 2013a). Hence, by Lemma 1, we have
\[
\epsilon_{\ell}(f) \leq \hat{\epsilon}_{s}(f) + \frac{3}{2} \sum_{j=1}^{J+1} u_j \text{tr}(\mathbf{K} \mathbf{H} \mathbf{J}) + \Omega
\]
(10)
where $L_j = d_j^T d_j$, $d_j \in \mathbb{R}^n$ is the $j$th row of $D$, e.g., $d_1 = [1 \cdots 1 0 \cdots 0]$. This completes the proof. \qed

The Framework

Our ultimate goal is to learn a classifier for the unlabeled target data. From Theorem 1, the bound of $\epsilon_{\ell}(f)$ can be decreased by simultaneously minimizing 1) the empirical error on labeled data, and 2) the dependence on domain covariates. This observation enables us to propose a new Side Information Dependence Regularization (SIDeR) learning framework that optimizes these two objectives. Here, we follow the manifold regularizer framework that can take unlabeled samples into account and formulate SIDeR as
\[
\min_f \mathcal{L}(f(X^t), Y^t) + \sigma \|f\|^2_K + \lambda \rho_h(f(X), D),
\]
(11)
where $\sigma, \lambda \geq 0$ are hyper-parameters, $X^t \in \mathbb{R}^{d \times n}$ denotes all labeled samples, $f(\cdot)$ is the decision function of a classifier, $\|f\|^2_K$ is the Tikhonov regularization term, and $Y$ denotes training labels. For each term in SIDeR framework, $\mathcal{L}(f(X^t), Y^t)$ minimizes the empirical risk, $\|f\|^2_K$ minimizes the model complexity, and $\rho_h(f(X), D)$ minimizes the domain dependence in the label decision space.

Connection to existing methods. SIDeR minimizes prediction error and domain dependence simultaneously, and therefore it can also be viewed as combining the virtues from both domain-invariant classifier methods and domain dependence minimization mapping. We summarize the relationship between SIDeR and related methods as follows.

SIDeR vs. ARTL. By Lemma 1, ARTL without manifold regularization and conditional distribution mismatch is equivalent to SIDeR with the degenerated domain covariate matrix $D_0$. However, SIDeR can model multiple sources and domain covariates, making it more flexible than ARTL. Moreover, it is easier to extend SIDeR to leverage the rich continuous side information in public neuroimaging dataset, such as subjects’ age, IQ, and handiness score. For the same reason, TCA is equivalent to MIDA with $d_0$.

\textbf{Algorithm 1} Side Information Dependence Regularization (SIDeR) with SVM Loss

\textbf{Input:} Input data matrix $X \in \mathbb{R}^{d \times n}$ (first $n$ samples are labeled), label vector $y \in \mathbb{R}^n$, and domain covariates.

\textbf{Hyper-parameters:} Penalty $C$, trade-off parameter $\lambda$, kernel function $k_x(\cdot, \cdot)$ and corresponding hyper-parameters.

\textbf{Output:} Coefficient vector $w$.

1: Encode domain covariates into a matrix $D \in \mathbb{R}^{d \times n}$ with one-hot encoding;
2: Construct matrix $\tilde{Y} \in \mathbb{R}^{n \times n}$, where $\tilde{Y}_{i,i} = y_i$, and the rest are zeros, identity matrix $I$, and centering matrix $\mathbf{H}$; 3: Construct kernel matrices $K = k_x(X, X), \mathbf{L} = D^T D$; 4: Learn the optimal Lagrange multipliers $\alpha^*$ by solving the QP problem of Eq. (14); 5: Compute $w = (I + \lambda \mathbf{H} \mathbf{L} \mathbf{H} K)^{-1} \tilde{Y} \alpha^*$. 6: return Coefficient vector $w$.

SIDeR vs. SMIDA. We can also view SIDeR as 1) replacing the label dependence term $\rho_h(\phi(X), Y)$ in SMIDA with the prediction loss, and 2) learning a mapping to a one-dimensional classification space (i.e., a line) rather than a low-dimensional subspace.

\textbf{Proposed Algorithm}

\textbf{Simplified HSIC.} In SIDeR framework, we aim to optimize the domain dependence in the decision space. If we view the coefficient vector $w$ as a classifier-based feature mapping, this mapping projects input features to a one-dimensional space (i.e., a line), where the projected values represent the decision scores. Following the principle of dependence minimization, we aim to learn a domain-independent classifier by minimizing the dependence of the decision scores (projected values) on domain side information, i.e., experiment implementations and subjects. By the Representer Theorem (Schölkopf, Herbrich, and Smola 2001), we can simplify the HSIC $\rho_h(f(X), D)$ to the following version
\[
\rho_{sh}(f(X), D) = \text{tr}((w^T K)^T (w^T K) H L H K) = w^T K H L H K w.
\]
where $L = D^T D$ (linear kernel) according to Lemma 1.

\textbf{SIDeR with SVM loss.} We can plug in any loss function for the first term in SIDeR of Eq. (11), such as the square loss, logistic loss, or hinge loss. In this paper, we consider only binary classification with $y \in \mathbb{R}^n$, $y_i \in \{-1, 1\}$, $i = 1, \cdots, n$, i.e., decoding $m = 2$ brain conditions. Here we choose hinge loss, which is robust to binary classification problems, for empirical risk minimization as in support vector machines (SVMs). We define $f(X) = w^T \phi(X)$, $\phi$ is a linear or non-linear kernel mapping, $w$ is a coefficient vector, $y = \text{sgn}(f(x))$, where $\text{sgn}(\cdot)$ is the sign function that extracts the sign of a real number, i.e., $(1 \text{ or } -1)$. Using the Representer Theorem again, we have $f(x) = \sum_{i=1}^n w_i k_x(x_i, x_i)$, and therefore $f(x_i) = \sum_{i=1}^n w_i k_x(x_i, x_i)$.
The computational complexity of HSIC is $O(n^2 + d^2)$ when linear kernel is used, i.e., $K = X^\top X$, $L = D^\top D$ (Gretton et al. 2005). In brain decoding problems, $d \gg n$ and $d \gg d$, so the overall computational complexity of HSIC is $O(nd^2)$. However, HSIC only needs to be computed once. The complexity of solving the quadratic programming problem for Eq. (14) is $O(n^3)$.

### Experiments

This section evaluates SIDeR against ten competing methods on 17 TL tasks in brain decoding.

#### Experimental Setup

**Dataset selection.** We selected six datasets (A to F) that are most meaningful from psychological perspective from the public OpenfMRI repository, as summarized in Table 1. Each dataset is from an experiment. Subjects from the same accession number ($ds \times \times \times$) are the same and there is no overlapping subject between accession numbers. There are two brain conditions selected from each dataset, with each as a class and having the same number of samples. Thus, we have binary classification problems that discriminate between brain conditions in an experiment.

**Preprocessing.** Each sample was preprocessed using FSL (Jenkinson et al. 2012) with the protocol in (Poldrack et al. 2013) to obtain the Z-score statistical parametric map (SPM) (Friston et al. 1994; 1998) of size $91 \times 109 \times 91$, which is then reduced to a vector of size $228 \times 546$ by masking the voxels outside of the brain.

**Seventeen TL tasks.** We constructed 17 TL problems with increasing psychological difficulty, as determined by discrepancy across experimental paradigms, subjects involved, cognitive control demands (e.g., inhibiting a planned response or ignoring a distracting or misleading stimulus), and complexity or modality of response. Table 2 summarizes how each pair of tasks relate on these dimensions. We denote source and target experiments as $S$ and $T$ and classify a positive condition against a negative condition. We define 17 TL tasks ($S \to T$) as listed in the first columns of Tables 3 and 4 with the three classification problems below:

- **“Successful stop” vs “Unsuccessful stop”:** Twelve single-source experiment TL tasks.
- **“Congruent correct” vs “Incongruent correct”:** Two single-source experiment TL tasks.

#### Table 1: Information on the OpenfMRI data used. ‘Exp’ indexes the six cognitive experiments A–F. #AC is the accession number of an OpenfMRI project, where the same group of subjects are used in each project and there is no overlapping subject between projects. #Sub indicate the number of unique subjects for each dataset. Each of the six experiments has two brain conditions to classify. Each subject in each experiment contributed two positive and two negative brain condition samples, respectively.

<table>
<thead>
<tr>
<th>Exp</th>
<th>#AC</th>
<th>Exp Description</th>
<th>#Sub</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>ds007</td>
<td>Stop signal with spoken pseudo word naming (Xue, Aron, and Poldrack 2008)</td>
<td>20</td>
</tr>
<tr>
<td>B</td>
<td>ds007</td>
<td>Stop signal with spoken letter naming (Xue, Aron, and Poldrack 2008)</td>
<td>20</td>
</tr>
<tr>
<td>C</td>
<td>ds007</td>
<td>Stop signal with manual response (Xue, Aron, and Poldrack 2008)</td>
<td>20</td>
</tr>
<tr>
<td>D</td>
<td>ds008</td>
<td>Conditional stop signal (Aron et al. 2007)</td>
<td>13</td>
</tr>
<tr>
<td>E</td>
<td>ds101</td>
<td>Simon task [Unpublished]</td>
<td>21</td>
</tr>
<tr>
<td>F</td>
<td>ds102</td>
<td>Flanker task (Kelly et al. 2008)</td>
<td>26</td>
</tr>
</tbody>
</table>

#### Table 2: Domain differences from the psychological perspective. A=B means when A is used as target experiment, B will be used as source experiment, and vice versa.

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Paradigm</th>
<th>Subjects</th>
<th>Control</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>A=B</td>
<td>Same</td>
<td>Same</td>
<td>Same</td>
<td>Different</td>
</tr>
<tr>
<td>A=C</td>
<td>Same</td>
<td>Same</td>
<td>Different</td>
<td>Different</td>
</tr>
<tr>
<td>B=C</td>
<td>Same</td>
<td>Same</td>
<td>Different</td>
<td>Different</td>
</tr>
<tr>
<td>C=D</td>
<td>Same</td>
<td>Different</td>
<td>Similar</td>
<td>Different</td>
</tr>
<tr>
<td>A=D</td>
<td>Same</td>
<td>Different</td>
<td>Different</td>
<td>Different</td>
</tr>
<tr>
<td>B=D</td>
<td>Same</td>
<td>Different</td>
<td>Different</td>
<td>Different</td>
</tr>
<tr>
<td>E=F</td>
<td>Different</td>
<td>Different</td>
<td>Similar</td>
<td>Different</td>
</tr>
</tbody>
</table>
Table 3: Classification accuracy in percentage for 14 single-source experiment TL tasks (mean ± standard deviation over ten replications of cross-validation). ‘Avg’ is the average over the 14 tasks. Each task is denoted as Source→Target experiment, e.g., A→B means A is the source and B is the target. Subscript r denotes the RBF kernel gives better results, and for those without subscript r, the linear kernel gives better results. The best result for each task is in **bold**, and the second best is underlined.

<table>
<thead>
<tr>
<th>SVM′</th>
<th>PCA′</th>
<th>PCA++r′</th>
<th>TCA</th>
<th>SSTCA-</th>
<th>JDA</th>
<th>ARSVM</th>
<th>MEDA</th>
<th>MIDA</th>
<th>SMIDA</th>
<th>SIDER_{SVM}</th>
</tr>
</thead>
<tbody>
<tr>
<td>A→B</td>
<td>63.4±2.6</td>
<td>62.5±4.4</td>
<td>59.7±4.5</td>
<td>55.0±4.3</td>
<td>48.0±3.9</td>
<td>63.7±2.8</td>
<td>65.7±2.2</td>
<td>57.8±5.3</td>
<td>64.5±5.0</td>
<td>54.9±2.8</td>
</tr>
<tr>
<td>B→A</td>
<td>59.2±4.0</td>
<td>60.4±5.7</td>
<td>60.8±3.1</td>
<td>59.8±3.2</td>
<td>50.4±4.6</td>
<td>54.4±1.8</td>
<td>64.0±2.8</td>
<td>62.5±4.5</td>
<td>71.2±3.5</td>
<td>52.4±1.5</td>
</tr>
<tr>
<td>A→C</td>
<td>68.4±2.6</td>
<td>66.9±6.3</td>
<td>74.0±5.9</td>
<td>70.6±6.3</td>
<td>48.0±4.6</td>
<td>58.4±3.1</td>
<td>70.9±2.7</td>
<td>70.1±2.1</td>
<td>78.4±2.9</td>
<td>57.5±2.1</td>
</tr>
<tr>
<td>C→A</td>
<td>59.2±4.0</td>
<td>60.4±5.7</td>
<td>67.6±5.2</td>
<td>63.6±4.0</td>
<td>51.8±3.4</td>
<td>53.7±3.6</td>
<td>59.5±3.5</td>
<td>57.9±3.1</td>
<td>70.6±4.0</td>
<td>52.4±1.3</td>
</tr>
<tr>
<td>B→C</td>
<td>68.4±2.6</td>
<td>66.9±6.3</td>
<td>78.9±3.4</td>
<td>86.4±5.3</td>
<td>49.9±3.6</td>
<td>63.1±5.2</td>
<td>73.5±2.1</td>
<td>73.6±3.6</td>
<td>80.1±4.4</td>
<td>58.5±2.5</td>
</tr>
<tr>
<td>C→B</td>
<td>63.4±2.6</td>
<td>62.5±4.4</td>
<td>66.6±4.0</td>
<td>75.3±4.0</td>
<td>52.6±4.4</td>
<td>54.1±3.5</td>
<td>62.5±2.1</td>
<td>57.0±2.1</td>
<td>73.2±3.3</td>
<td>58.4±2.4</td>
</tr>
<tr>
<td>C→D</td>
<td>74.6±3.2</td>
<td>79.6±8.2</td>
<td>68.9±4.9</td>
<td>74.5±5.3</td>
<td>44.8±2.6</td>
<td>71.7±2.3</td>
<td>81.3±4.0</td>
<td>64.6±3.4</td>
<td>74.2±6.6</td>
<td>66.5±2.8</td>
</tr>
<tr>
<td>D→C</td>
<td>68.4±2.6</td>
<td>66.9±6.3</td>
<td>75.0±1.4</td>
<td>71.9±4.6</td>
<td>56.0±5.3</td>
<td>58.0±2.4</td>
<td>70.9±2.6</td>
<td>62.8±4.1</td>
<td>74.4±4.2</td>
<td>61.4±3.6</td>
</tr>
<tr>
<td>A→D</td>
<td>74.6±3.2</td>
<td>78.7±2.6</td>
<td>57.9±4.3</td>
<td>54.0±3.2</td>
<td>53.7±3.2</td>
<td>64.2±9.9</td>
<td>78.5±1.9</td>
<td>52.9±4.5</td>
<td>63.3±6.0</td>
<td>72.7±4.6</td>
</tr>
<tr>
<td>D→A</td>
<td>59.2±4.0</td>
<td>60.4±5.7</td>
<td>63.7±5.3</td>
<td>58.0±4.8</td>
<td>50.0±3.1</td>
<td>55.5±1.9</td>
<td>59.2±3.4</td>
<td>58.4±3.6</td>
<td>50.6±2.4</td>
<td>48.1±1.7</td>
</tr>
<tr>
<td>B→D</td>
<td>74.6±3.2</td>
<td>79.6±8.2</td>
<td>68.5±2.1</td>
<td>67.9±2.6</td>
<td>60.2±4.0</td>
<td>62.5±4.7</td>
<td>85.6±2.3</td>
<td>55.4±4.7</td>
<td>66.9±4.2</td>
<td>78.7±3.2</td>
</tr>
<tr>
<td>D→B</td>
<td>63.4±2.6</td>
<td>65.2±4.4</td>
<td>60.5±5.6</td>
<td>50.1±5.4</td>
<td>54.0±4.0</td>
<td>54.7±3.1</td>
<td>61.4±2.4</td>
<td>60.4±5.7</td>
<td>64.7±3.6</td>
<td>55.5±2.3</td>
</tr>
<tr>
<td>E→F</td>
<td>66.9±1.7</td>
<td>67.5±6.2</td>
<td>51.4±1.8</td>
<td>52.3±3.2</td>
<td>56.0±2.3</td>
<td>50.5±2.2</td>
<td>62.5±2.0</td>
<td>60.2±2.2</td>
<td>66.4±3.0</td>
<td>61.5±3.8</td>
</tr>
<tr>
<td>F→E</td>
<td>53.3±2.8</td>
<td>51.8±3.9</td>
<td>51.3±2.4</td>
<td>49.3±1.9</td>
<td>52.1±2.1</td>
<td>50.2±1.7</td>
<td>49.9±2.2</td>
<td>52.7±1.5</td>
<td><strong>53.7±4.5</strong></td>
<td>49.2±1.7</td>
</tr>
<tr>
<td>Avg</td>
<td>65.5±3.0</td>
<td>66.5±5.7</td>
<td>64.2±3.8</td>
<td>63.4±4.1</td>
<td>51.9±3.6</td>
<td>58.2±3.4</td>
<td>67.5±2.6</td>
<td>60.6±3.7</td>
<td>68.0±4.2</td>
<td>59.1±2.6</td>
</tr>
</tbody>
</table>

- **“Successful stop” vs “Unsuccessful stop”**: Three multi-source experiment TL tasks.

Ten methods compared. We evaluate SIDER_{SVM} against ten methods: three simple baselines 1) SVM′, 2) PCA′, and 3) PCA++r′; and seven state-of-the-art TL methods discussed in Preliminary: 4) TCA (Pan et al., 2011), 5) SSTCA (Pan et al., 2011), 6) JDA (Long et al., 2013b), 7) ARSVM (Long et al., 2013a) of ARTL, 8) MEDA (Wang et al., 2018), 9) MIDA (Yan, Kou, and Zhang, 2018), and 10) SMIDA (Yan, Kou, and Zhang, 2018). SVM′ and PCA′ use only the target data while PCA++r′ uses both source and target data. For multi-source experiments TL, the multiple source experiments are used as one single source domain by MMD based methods, i.e., TCA, SSTCA, JDA, and ARSVM, PCA, TCA, SSTCA, JDA, MIDA, and SMIDA only learn a feature mapping so they use SVM as the classifier. Both linear and RBF kernels were studied for such SVM classifiers, and also for ARSVM, MEDA and SIDER_{SVM}. We will report the result from the best performing kernel.

We performed 10 × 5-fold cross-validation on the target domain. For each split, the target training samples and all source samples (except for SVM′ and PCA′) were used for training, with the target test samples for testing. On each training set, the optimal hyper-parameters for all methods are determined using the search strategy in (Pan et al., 2011) with 10 further random splits (20% for validation, 80% for training). To compare the difference between MMD and HSIC as a regularizer, manifold regularization and conditional distribution mismatch of ARSVM are not considered. Sensitivity studies for SIDER are provided later in this section, which will validate SIDER can offer stable performance for a wide range of hyper-parameter settings.

**Results and Discussions**

Tables 3 and 4 summarize the decoding accuracy of the 17 TL tasks with both the mean and standard deviation. The best result for each task is in **bold** and the second best is underlined. We have five key observations:

- On the whole, SIDER_{SVM} outperformed all the comparing methods. From results of using experiment A or B as target domain, we can observe the performance gain decreases from easy to difficult tasks. This indicates a plausible correlation between the TL improvements and the transfer difficulties from easy to hard tasks.

SIDER_{SVM} outperformed the best existing method (ARSVM) by **15.6%** (83.2% vs. 67.6%) in TL tasks with multi-source experiments (Table 4). On the other hand, SIDER_{SVM} obtained lower accuracy on A→B→C compared to B→C, and the rest results show using multiple source experiments is better. Thus, source selections can influence the transfer performance. If there is no clear preference of a particular source dataset, transfer with multiple sources is preferred in our opinion.

- MIDA and SIDER_{SVM} outperformed the corresponding MMD-counterparts TCA and ARSVM, respectively. Based on Theorem 1, this confirmed that making use of multiple domain side information (experiments and subjects) is beneficial in brain decoding.

- SIDER_{SVM} outperformed SMIDA, and ARSVM outperformed SSTCA, which indicates prediction loss (hinge loss) is more robust than variance preserving and label dependence maximization in brain decoding problems.

- In Table 3, TCA and MIDA outperformed the corresponding semi-supervised version SSTCA and SMIDA. However, their performance were close when more source samples were used (Table 4). This observation indicates that label dependence is more susceptible to overfitting. Additionally, the performance of methods with conditional distribution alignment, i.e., JDA and MEDA, were inferior to marginal distribution alignment methods, i.e., TCA and ARSVM. In summary, using label dependence or conditional distribution alignment may lead to inferior performance in small sample size brain decoding TL tasks.
Hyper-parameter sensitivity. We evaluated the sensitivity of SIDeR$_{SVM}$ with linear kernel against hyper-parameters $C$ and $\lambda$ under five-fold cross validation. Figure 1a shows the sensitivity against $C \in [10^{-3}, 10^3]$ when fixing $\lambda = 1$. We can observe that the accuracy stays stable when $C \leq 1$, and shows a trend of decreasing when $C \in [10^0, 10^3]$. Since a smaller value of $C$ can lead to a larger SVM classification margin, we expect a classifier with a larger margin to generalize better and have higher prediction accuracy. Figure 1b shows the sensitivity against $\lambda \in \{0, 0.01, 0.1, 1, 10, 100\}$ when fixing $C = 1$. We can observe that the prediction accuracy of SIDeR$_{SVM}$ kept almost constant when $\lambda > 0$ and it was not sensitive to $\lambda$. When $\lambda = 0$, i.e., without minimizing domain dependence, SIDeR$_{SVM}$ becomes a standard kernel SVM and the performance dropped significantly.

Sensitivity to sample size. Figure 1c shows the sensitivity of SIDeR$_{SVM}$ with respect to labeled target training sample size for single/multi-source experiment transfer, averaged over fourteen/three single/multi-source experiment transfer tasks over 20 random splits, respectively. Multi-source experiment transfer obtained better performance, especially when the labeled target training size is smaller.

Convergence. Figure 1d illustrates the primal and dual cost of SIDeR$_{SVM}$ on the TL task A→B. We can observe that the costs converged within ten iterations.

Model visualization. We visualize the three models with the best performance on the learning task A&B→C, which are SVM$^t$, ARSVM, and SIDeR$_{SVM}$, using the Python package Nilearn (Abraham et al. 2014). Figure 2 depicts the learned weights averaged over 5-fold cross validation in the voxel space. We can observe that without training with source data, SVM$^t$ (Fig. 2a) highlight different areas compared to ARSVM (Fig. 2b) and SIDeR$_{SVM}$ (Fig. 2c), which identified some common clusters around cingulate gyrus (shaded in blue). These clusters in Fig. 2c are clearer and less noisy, suggesting that SIDeR$_{SVM}$ has identified more coherent brain functional areas.

Table 4: Classification accuracy in percentage for three multi-source experiment TL tasks.

<table>
<thead>
<tr>
<th>Task</th>
<th>SVM$^t$</th>
<th>PCA$^t$</th>
<th>PCA$^{t-t}$</th>
<th>TCA</th>
<th>SSTCA</th>
<th>JDA</th>
<th>ARSVM</th>
<th>MEDA</th>
<th>MIDA</th>
<th>SMIDA</th>
<th>SIDeR$_{SVM}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A&amp;B→C</td>
<td>59.2±5.0</td>
<td>60.4±5.7</td>
<td>51.5±1.5</td>
<td>52.1±1.1</td>
<td>52.1±3.4</td>
<td>51.7±2.8</td>
<td>63.2±2.5</td>
<td>52.6±1.4</td>
<td>51.4±1.8</td>
<td>55.4±2.4</td>
<td>80.3±3.1</td>
</tr>
<tr>
<td>A&amp;C→B</td>
<td>63.4±2.6</td>
<td>62.5±4.4</td>
<td>53.7±2.1</td>
<td>54.5±3.1</td>
<td>52.2±3.2</td>
<td>53.8±3.3</td>
<td>67.0±1.6</td>
<td>52.6±1.3</td>
<td>61.7±2.2</td>
<td>60.4±2.2</td>
<td>79.5±2.0</td>
</tr>
<tr>
<td>A&amp;B→C</td>
<td>68.4±2.6</td>
<td>66.9±6.3</td>
<td>50.6±2.5</td>
<td>52.1±1.1</td>
<td>56.5±3.5</td>
<td>52.0±3.1</td>
<td>71.9±2.6</td>
<td>57.0±0.8</td>
<td>65.4±2.2</td>
<td>60.5±3.2</td>
<td>89.9±1.7</td>
</tr>
<tr>
<td>Avg</td>
<td>63.7±3.1</td>
<td>62.6±5.5</td>
<td>52.0±2.0</td>
<td>52.9±1.8</td>
<td>53.6±3.4</td>
<td>52.5±3.0</td>
<td>67.6±2.2</td>
<td>54.1±2.6</td>
<td>59.5±2.0</td>
<td>58.8±2.6</td>
<td>83.2±2.3</td>
</tr>
</tbody>
</table>

Figure 1: The sensitivity of the classification accuracy with respect to hyper-parameters $C$, $\lambda$, and labeled target training data size, and the convergence study for SIDeR$_{SVM}$ with linear kernel.

Figure 2: Visualization of the top 1% learned voxel weights in magnitude for three classifiers averaged over 5-fold cross validation on A&B→C.

Conclusion

In this paper, we proposed Side Information Dependence Regularization (SIDeR) learning framework for TL in analyzing human brain conditions across subjects and experiments. We incorporated the SVM loss into SIDeR to simultaneously minimize the empirical prediction risk and the dependence on domain side information measured by a simplified HSIC. We evaluated SIDeR against SVM, PCA and seven state-of-the-art TL methods on seventeen TL tasks. Experimental results showed the superior overall performance of SIDeR over other methods, particularly on multi-source experiment transfer, with a 15.6% improvement. This confirmed the benefits of leveraging domain side information and HSIC in TL for brain condition decoding.

Acknowledgments

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/R014507/1 and the European Research Council (ERC) grant GAP: 670428 - BRAIN2MIND_NEUROCOMP.
References


