When can Multi-Site Datasets be Pooled for Regression? Hypothesis Tests, ℓ_2-consistency and Neuroscience Applications

Hao Henry Zhou [1]   **Yilin Zhang** [1]   **Vamsi K. Ithapu** [1]   **Sterling C. Johnson** [1 2]   **Grace Wahba** [1]   **Vikas Singh** [1]

Abstract

Many studies in biomedical and health sciences involve small sample sizes due to logistic or financial constraints. Often, identifying weak (but scientifically interesting) associations between a set of predictors and a response necessitates pooling datasets from multiple diverse labs or groups. While there is a rich literature in statistical machine learning to address distributional shifts and inference in multi-site datasets, it is less clear *when* such pooling is guaranteed to help (and when it does not) – independent of the inference algorithms we use. In this paper, we present a hypothesis test to answer this question, both for classical and high dimensional linear regression. We precisely identify regimes where pooling datasets across multiple sites is sensible, and how such policy decisions can be made via simple checks executable on each site before any data transfer ever happens. With a focus on Alzheimer's disease studies, we present empirical results showing that in regimes suggested by our analysis, pooling a local dataset with data from an international study improves power.

1. Introduction

In the last two decades, statistical machine learning algorithms for processing massive datasets have been intensively studied for a wide-range of applications in computer vision, biology, chemistry and healthcare (Murdoch & Detsky, 2013; Tarca et al., 2007). While the challenges posed by large scale datasets are compelling, one is often faced with a fairly distinct set of technical issues for studies in biological and health sciences. For instance, a sizable portion of scientific research is carried out by small or medium-sized groups (Fortin & Currie, 2013) supported by modest budgets (Lauer, 2016). Hence, there are logistic/financial constraints on the number of experiments and/or number of participants within a trial, leading to small size datasets. While the analysis may be sufficiently powered to evaluate the *primary* hypothesis of the study/experiment, interesting follow-up scientific questions (often more *nuanced*), come up during the course of the project. These tasks may be underpowered for the sample sizes available. This necessitates efforts to identify similar datasets elsewhere so that the combined sample size of the "pooled" dataset is enough to determine significant associations between a response and a set of predictors, e.g., within linear regression.

Motivating Application. In genomics, funding agencies have invested effort into standardizing/curating data collection across large international projects (ENCODE Project Consortium, 2004). In other disciplines, such as in the study of neurological disorders, heterogeneity in disease etiology, variations in scanners and/or acquisition tools make standardization more difficult. For Alzheimer's disease (AD), a motivation of this work, efforts such as ADNI (Weiner et al., 2015) provide a variety of clinical, imaging and cognitive tests data for $800+$ *older adults*. However, the research focus has now moved to the early stages of disease – as early as *late middle age* – where treatments are expected to be more effective. But (a) the statistical signal at this stage is *weak* and difficult to demonstrate without large sample sizes and (b) such "preclinical" participants are not well represented, even in large ADNI sized studies. Hence, there is a concerted effort in general for smaller standalone projects (focused on a specific disease stage), that can be retrospectively pooled for analysis towards addressing a challenging scientific hypothesis (Jahanshad et al., 2013). Unfortunately, acquisition protocols for various measures across sites are usually different and data are heterogeneous. These issues raise a fundamental technical question. When is it meaningful to pool datasets for estimating a simple statistical model (e.g., linear regression)? When can we guarantee improvements in statistical power, and when are such pooling efforts not worth it? Can we give a hypothesis test and obtain p-values to inform our policies/decisions? While related problems have been stud-

---

[1]University of Wisconsin-Madison [2]William S. Middleton Memorial Veteran's Affairs Hospital. Correspondence to: Hao Zhou <hzhou@stat.wisc.edu>, Vikas Singh <vsingh@biostat.wisc.edu>.

ied in machine learning from an algorithm design perspective, even simple hypothesis tests which can be deployed by a researcher in practice, are currently unavailable. Our goal is to remove this significant limitation.

Putting our development in context. The realization that "similar" datasets from multiple sites can be pooled to potentially improve statistical power is not new. With varying empirical success, models tailored to perform regression in multi-site studies (Group, 2002), (Haase et al., 2009), (Klunk et al., 2015) have been proposed, where due to operational reasons, recruitment and data acquisition are distributed over multiple sites, or even countries. When the pooling is being performed *retrospectively* (i.e., after the data has been collected), resolving site-specific confounds, such as distributional shifts or biases in measurements, is essential before estimation/inference of a statistical model. We will *not* develop new algorithms for estimating the distributional mismatch or for performing multi-site regression — rather, our primary goal is to identify the regimes (and give easily computable checks) where this regression task on a pooled dataset is statistically meaningful, assuming that good pre-processing schemes are available. We will present a rigorous yet simple to implement hypothesis test, analyze its behavior, and show extensive experimental evidence (for an important scientific problem). The practitioner is free to use his/her preferred procedure for the "before step" (estimating the distributional shifts).

Contributions: a) Our main result is a hypothesis test to evaluate whether pooling data across multiple sites for regression (before or after correcting for site-specific distributional shifts) can improve the estimation (mean squared error) of the relevant coefficients (while permitting an influence from a set of confounding variables). **b)** We derive analogous results in the high-dimensional setting by leveraging a different set of analysis techniques. Using an existing sparse multi-task Lasso model, we show how the utility of pooling can be evaluated even when the support set of the features (predictors) is not exactly the same across sites using ideas broadly related to high dimensional simultaneous inference (Dezeure et al., 2015). We show ℓ_2-consistency rate, which supports the use of sparse multi-task Lasso when sparsity patterns are not totally identical. **c)** On an important scientific problem of analyzing early Alzheimer's disease (AD) individuals, we provide compelling experimental results showing consistent acceptance rate and statistical power. Via a publicly available software package, this will facilitate many multi-site regression analysis efforts in the short to medium term future.

1.1. Related Work

Meta-analysis approaches. If datasets at multiple different sites *cannot* be shared or pooled, the task of deriving meaningful scientific conclusions from *results of multiple independently conducted analyses* generally falls under the umbrella term of "meta analysis". The literature provides various strategies to cumulate the general findings from analyses on different datasets. But even experts believe that, minor violations of assumptions can lead to misleading scientific conclusions (Greco et al., 2013), and substantial personal judgment (and expertise) is needed to conduct them. It is widely accepted that when the ability to pool the data is an option, simpler schemes may perform better.

Domain adaptation/shift. Separately, the idea of addressing "shift" within datasets has been rigorously studied within statistical machine learning, see (Patel et al., 2015; Li, 2012). For example, domain adaptation, including dataset and covariate shift, seeks to align (the distributions of) multiple datasets to enable follow-up processing (Ben-David & Schuller, 2003). Typically, such algorithms assume a bias in the sampling process, and adopt reweighting as the solution (Huang et al., 2007; Gong et al., 2013). Alternatively, a family of such methods assume that sites (or datasets) differ due to feature distortions (e.g., calibration error), which are resolved, in general, by minimizing some distance measure between appropriate distributions (Baktashmotlagh et al., 2013; Pan et al., 2011; Long et al., 2015; Ganin et al., 2016). In general, these approaches have nice theoretical properties (Ben-David et al., 2010; Cortes & Mohri, 2011; Zhou et al., 2016). However, it is important to note that the domain adaptation literature *focuses on the algorithm itself* – to resolve the distributional site-wise differences. It does *not* address the issue of whether pooling the datasets, after applying the calculated adaptation (i.e., transformation), is beneficial. Our goal in this work is to assess whether multiple datasets can be pooled — either *before* or usually *after* applying the best domain adaptation methods — for improving our estimation of the relevant coefficients within linear regression. We propose a hypothesis test to directly address this question.

The high-dimensional case. High dimensional scenarios, in general, involve predicting a response (e.g., cognitive score) from high dimensional predictors such as image scans (or derived features) and genetic data, which in general, entails Lasso-type formulations unlike the classical regression models. Putting multi-task representation learning (Maurer et al., 2016; Ando & Zhang, 2005; Maurer et al., 2013) together with a sparsity regularizer, we get the multi-task Lasso model (Liu et al., 2009; Kim & Xing, 2010). Although this seems like a suitable model (Chen et al., 2012), it assumes that the multiple tasks (sites here) have an identical active set of predictors. Instead, we find that the sparse multi-task Lasso (Lee et al., 2010), roughly, a multi-task version of sparse group Lasso (Simon et al., 2013; Lee et al., 2010) is a better starting point. There is no theoretical analysis in (Simon et al., 2013); although a ℓ_2-consistency

for sparse group lasso is derived in (Chatterjee et al., 2012) using a general proof procedure for M-estimators, it does not take into account the specific sparse group Lasso properties. This makes the result non-informative for sparse group Lasso (much less, sparse multi-task Lasso). Specifically, as we will see shortly, in sparse multi-task Lasso, the joint effects of two penalties induces a special type of asymmetric structure. We show a new result, in the style of Lasso (Meinshausen & Yu, 2009; Liu & Zhang, 2009b), for ℓ_2 convergence rate for this model. It matches known results for Lasso/group Lasso, and identifies regimes where the sparse multi-task (multi-site) setting is advantageous.

Simultaneous High dimensional Inference. Simultaneous high dimensional inference models such as multi sample-splitting and de-biased Lasso is an active research topic in statistics (Dezeure et al., 2015). Multi sample-splitting use half of the dataset for variable selection and the rest for calculating p-values. De-biased Lasso chooses one feature as a response and the others as predictors to estimate a Lasso model; this procedure is repeated for each feature. Estimators from De-biased Lasso asymptotically follow the multi-normal distribution (Dezeure et al., 2016), and using Bonferroni-Holm adjustment produces simultaneous p-values. Such ideas together with the ℓ_2-convergence results for sparse multitask Lasso, will help extend our analysis to the high dimensional setting.

2. Hypothesis Test for Multi-Site Regression

We first describe a simple setting where one seeks to apply standard linear regression to data pooled from multiple sites. For presentation purposes, we will deal with variable selection issues later. Within this setup, we will introduce our main result — a hypothesis test to evaluate statistical power improvements (e.g., mean squared error) when running a regression model on a pooled dataset. We will see that the proposed test is transparent to the use of adaptation algorithms, if any, to pre-process the multi-site data (more details in appendix). In later sections, we will present convergence analysis and extensions to the large p setting. Matrices (vectors/scalars) are upper case (and lower case). $\|.\|_*$ is the nuclear norm.

We first introduce the single-site regression model. Let $X \in \mathbb{R}^{n \times p}$ and $y \in \mathbb{R}^{n \times 1}$ denote the feature matrix of predictors and the response vector respectively. If β corresponds to the coefficient vector (i.e., predictor weights), then the regression model is

$$\min_{\beta} \quad \frac{1}{n}\|y - X\beta\|_2^2 \tag{1}$$

where $y = X\beta^* + \epsilon$ and $\epsilon \sim N(0, \sigma^2)$ I.I.D. if β^* is the true coefficient vector from which y is generated. The mean-squared error (MSE) and ℓ_2-consistency of regres-

sion is well-known. The mean-squared error (MSE) of (1) is $\mathbb{E}\|\hat{\beta} - \beta^*\|_2^2 = \text{tr}\left((X^T X)^{-1}\right)\sigma^2$. If k denotes the number of sites, then one may first apply a domain adaptation scheme to account for the distributional shifts between the k different predictors $\{X_i\}_{i=1}^k$, and then run a regression model. If the underlying "concept" (i.e., predictors and responses relationship) can be assumed to be the same across the different sites, then it is reasonable to impose the *same* β for all sites. For instance, the influence of CSF protein measurements on cognitive scores of an individual may be invariant to demographics. Nonetheless, if the distributional mismatch correction is imperfect, we may define $\Delta\beta_i = \beta_i - \beta^*$ where $i \in \{1, \dots, k\}$ as the residual difference between the site-specific coefficients and the true shared coefficient vector (in the ideal case, we have $\Delta\beta_i = 0$). In the multi-site setting, we can write

$$\min_{\beta} \sum_{i=1}^k \tau_i^2 \|y_i - X_i\beta\|_2^2 \tag{2}$$

where for each site i we have $y_i = X_i\beta^* + X_i\Delta\beta_i + \epsilon_i$ and $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ I.I.D. Here, τ_i is a weighting parameter for each site, if such information is available.

Our main goal is to test if the combined regression improves the estimation for a single site. We can pose this question in terms of improvements in the mean squared error (MSE). Hence, W.L.O.G. using site 1 as the reference, we set $\tau_1 = 1$ in (2) and consider $\beta^* = \beta_1$,

$$\min_{\beta} \|y_1 - X_1\beta\|_2^2 + \sum_{i=2}^k \tau_i^2 \|y_i - X_i\beta\|_2^2 \tag{3}$$

Clearly, when the sample size is not large enough, the multi-site formulation in (3) may reduce variance significantly, because of the averaging effect in the objective function, while increasing the bias by a little. This reduces the Mean Squared Error (MSE), see Figure 1. Note that while traditionally, the unbiasedness property was desirable, an extensive body of literature on ridge regression suggests that the quantity of interest should really be $\mathbb{E}\|\hat{\beta} -$

Figure 1. β_1 and β_2 are 1st and 2nd site coefficients. After combination, β_1's bias increases but variance reduces, resulting in a smaller MSE.

$\beta^*\|_2^2$ (Hoerl & Kennard, 1970; James & Stein, 1961). These ideas are nicely studied within papers devoted to the "bias-variance" trade-off. Similar to these results, we will focus on the mean squared error because the asymptotic consistency properties that come with an unbiased estimator are not meaningful here anyway — the key reason we want to pool datasets in the first place is because of small sample sizes. We now provide a result showing how the tuning parameters $\tau_2, \dots, \tau_k$ can be chosen.
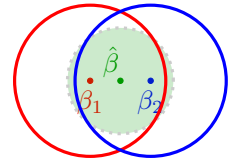
Theorem 2.1 $\tau_i = \frac{\sigma_1}{\sigma_i}$ *achieves the smallest variance in* $\hat{\beta}$.

Remarks: This result follows from observing that the each site's contribution is inversely proportional to site-specific noise level, σ_i. We will show that this choice of τ_is also leads to a simple mechanism to setup a hypothesis test.

2.1. Sharing all βs

In the specification above, the estimates of β_i across all k sites are restricted to be the same. Without this constraint, (3) is equivalent to fitting a regression *separately* on each site. So, a natural question is whether this constraint improves estimation. To evaluate whether MSE is reduced, we first need to quantify the change in the bias and variance of (3) compared to (1). To do so, we introduce a few notations. Let n_i be the sample size of site i, and let $\hat{\beta}_i$ denote the regression estimate from a specific site i. We have $\Delta\hat{\beta}_i = \tilde{\beta}_i - \hat{\beta}_1$. We define the length kp vector $\Delta\beta^T$ as $\Delta\beta^T = (\Delta\beta_2^T, ..., \Delta\beta_k^T)$ (similarly for $\Delta\hat{\beta}^T$). We use $\hat{\Sigma}_i$ for the sample covariance matrix of the data (predictors) from the site i and $G \in \mathbb{R}^{(k-1)p \times (k-1)p}$ for the covariance matrix of $\Delta\hat{\beta}$, where $G_{ii} = (n_1\hat{\Sigma}_1)^{-1} + (n_i\tau_i^2\hat{\Sigma}_i)^{-1}$ and $G_{ij} = (n_1\hat{\Sigma}_1)^{-1}$ whenever $i \neq j$.

Let the difference in bias and variance between the single site model in (1) and the multi-site model in (3) be $Bias_\beta$ and Var_β respectively. Let $\hat{\Sigma}_2^k = \sum_{i=2}^k n_i\tau_i^2\hat{\Sigma}_i$ and $\hat{\Sigma}_1^k = n_1\hat{\Sigma}_1 + \hat{\Sigma}_2^k$. We have,

Lemma 2.2 *For model* (3)*, we have*

$$\frac{\|Bias_\beta\|_2^2}{\|G^{-1/2}\Delta\beta\|_2^2} \leq \|(\hat{\Sigma}_1^k)^{-2}(\hat{\Sigma}_2^k(n_1\hat{\Sigma}_1)^{-1}\hat{\Sigma}_2^k + \hat{\Sigma}_2^k)\|_*, \quad (4)$$

$$Var_\beta = \sigma_1^2 \left\|(n_1\hat{\Sigma}_1)^{-1} - (n_1\hat{\Sigma}_1 + \hat{\Sigma}_2^k)^{-1}\right\|_*. \quad (5)$$

Remarks: The above result bounds the increase in bias and the reduction in variance (see discussion of Figure 1). Since our goal is to test MSE reduction — in principle, we can use bootstrapping to calculate MSE approximately. This procedure has a significant computational footprint. Instead, (4) (from a one-step Cauchy-Schwartz inequality), gives a *sufficient condition for MSE reduction* as shown below.

Theorem 2.3 a) *Model* (3) *has smaller MSE of* $\hat{\beta}$ *than model* (1) *whenever*

$$H_0 : \|G^{-1/2}\Delta\beta\|_2^2 \leq \sigma_1^2. \quad (6)$$

b) *Further, we have the following test statistic,*

$$\left\|\frac{G^{-1/2}\Delta\hat{\beta}}{\sigma_1}\right\|_2^2 \sim \chi_{(k-1)*p}^2 \left(\left\|\frac{G^{-1/2}\Delta\beta}{\sigma_1}\right\|_2^2\right), \quad (7)$$

where $\|G^{-1/2}\Delta\beta/\sigma_1\|_2$ *is called a "condition value".*

Remarks: This is our main test result. Although σ_i is typically unknown, it can be easily replaced using its site-specific estimation. Theorem 2.3 implies that we can conduct a non-central χ^2 distribution test based on the statistic. Also, (6) shows that the non-central χ^2 distribution, which the test statistics will follow, has a non-central parameter smaller than 1 when the sufficient condition H_0 holds. Meanwhile, in obtaining the (surprisingly simple) sufficient condition H_0, no other arbitrary assumption is needed except the application of Cauchy-Schwartz inequality. From a practical perspective, Theorem 2.3 implies that the sites, in fact, do *not* even need to share the full dataset to assess whether pooling will be useful. Instead, the test only requires *very high-level* information such as $\hat{\beta}_i$, $\hat{\Sigma}_i$, σ_i and n_i for all participating sites – which can be transferred very cheaply with no additional cost of data storage, or privacy implications. The following result deals with the special case where we have two participating sites.

Corollary 2.4 *For the case where we have two participating sites, the condition* (6) *from Theorem 2.3 reduces to*

$$H_0 : \Delta\beta^T((n_1\hat{\Sigma}_1)^{-1} + (n_2\tau_2^2\hat{\Sigma}_2)^{-1})^{-1}\Delta\beta \leq \sigma_1^2. \quad (8)$$

Remarks: The left side above relates to the Mahalanobis distance between β_1, β_2 with covariance $(n_1\hat{\Sigma}_1)^{-1} + (n_2\tau_2^2\hat{\Sigma}_2)^{-1}$, implying that the test statistic is a type of a normalized metric between the two regression models.

2.2. Sharing a subset of βs

In numerous pooling scenarios, we are faced with certain systemic differences in the way predictors and responses associate across sites. For example, socio-economic status may (or may not) have a significant association with a health outcome (response) depending on the country of the study (e.g., due to insurance coverage policies). Unlike in Section 2.1, we now relax the restriction that all coefficients are the same across sites, see Fig. 2. The model in (3) will now include another design matrix of predictors $Z \in R^{n \times q}$ and corresponding coefficients γ_i for each site i,

$$\min_{\beta,\gamma} \quad \sum_{i=1}^k \tau_i^2\|y_i - X_i\beta - Z_i\gamma_i\|_2^2 \quad (9)$$

$$y_i = X_i\beta^* + X_i\Delta\beta_i + Z_i\gamma_i^* + \epsilon_i, \ \tau_1 = 1 \quad (10)$$

Our goal is still to evaluate whether the MSE of β reduces. We do not take into account the MSE change in γ because they correspond to site-specific variables. For estimation, $\hat{\beta}$ can first be computed from (9). Treating it as a fixed entity now, $\hat{\gamma}_i$ can be computed using y_i and Z_i on each site independently. Clearly, if $\hat{\beta}$ is close to the "true" β^*, it will also enable a better estimation of site-specific variables. It turns out that, if $\hat{\Sigma}_i$s are replaced by the conditional covariance, the analysis from Section 2.1 still holds for this case.

Specifically, let $\hat{\Sigma}_{ab_i}$ be the sample covariance matrix between features a and b from some site i. We have,

Theorem 2.5 *Analysis in Section 2.1 holds for β in (9) by replacing $\hat{\Sigma}_i$ with $\tilde{\Sigma}_i = \hat{\Sigma}_{xx_i} - \hat{\Sigma}_{xz_i}(\hat{\Sigma}_{zz_i})^{-1}\hat{\Sigma}_{zx_i}$*

Remarks: The test now allows evaluating power improvements focused only on the subset of coefficients that is shared and permits site-specific confounds. For example, we can test which subset of parameters might benefit from parameter estimation on pooled data from multiple sites.

3. Pooling in High Dimensional Regression

We now describe our analysis of pooling multi-site data in the high-dimensional setting where $p \gg n$. The challenge here is that *variable section* has to be a first order concern. In classical regression, ℓ_2 consistency properties are well known and so our focus in Section 2 was devoted to deriving sufficient conditions for the hypothesis test. In other words, imposing the same β across sites works in (3) because we understand its consistency. In contrast, here, one cannot enforce a shared β for all sites *before* the active set of predictors within each site are selected — directly imposing the same β leads to a loss of ℓ_2-consistency, making follow-up analysis problematic. Therefore, once a suitable model for high-dimensional multi-site regression is chosen, the first requirement is to characterize its consistency.

We start with the multi-task Lasso (a special case of group Lasso) (Liu et al., 2009), where the authors show that the strategy selects better explanatory features compared to separately fitting Lasso on each site. But this algorithm underperforms when the sparsity pattern of the predictors is not identical across sites, so we use a recent variant called sparse multi-task Lasso (Lee et al., 2010) – essentially substituting "sites" for "tasks". The sparse *multi-site* Lasso in $p \gg n$ setting (p is the number of predictors) is given as

$$\hat{B}^\lambda = \arg\min_\beta \sum_{i=1}^{k}\|y_i - X_i\beta_i\|_2^2 + \lambda\Lambda(B) \quad (11)$$

$$\Lambda(B) = \alpha\sum_{j=1}^{p}\|\beta^j\|_1 + (1-\alpha)\sqrt{k}\sum_{j=1}^{p}\|\beta^j\|_2, \quad (12)$$

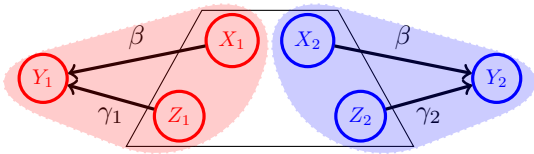where λ is the Lasso regularization parameter. Here, $B \in$

Figure 2. X and Z influence the response Y. After adjustment, X_1 and X_2 may be close requiring same β. However, Z_1 and Z_2 may differ a lot, and we need different γ_1 and γ_2.

$R^{k\times p}$ is a matrix where the ith row gives the coefficients from i^{th} site (k sites in total). Also, β_i with subscript denotes the ith row (site) of B, we use β^j with superscript to give the j-th column (coefficients) of B. The hyperparameter $\alpha \in [0,1]$ balances the two penalties (and will be used shortly); a larger α weighs the ℓ_1 penalty more and a smaller α puts more weight on the grouping. Similar to a Lasso-based regularization parameter, λ here will produce a solution path (to select coefficients) for a given α. We first address the consistency behavior of the sparse multi-site Lasso in (11), which was not known in the literature.

3.1. ℓ_2 consistency

Our analysis of (11) is related to known results for Lasso (Meinshausen & Yu, 2009) and the group Lasso (Liu & Zhang, 2009a). Recall that $X_1, \ldots, X_k$ are the data matrices from k sites. We define $\bar{n} = \max_{i=1}^{k}\{n_i\}$ and $C = \bar{n}^{-1}\text{DIAG}(X_1^T X_1, ..., X_k^T X_k)$ where $\text{DIAG}(A, B)$ corresponds to constructing a block-diagonal matrix with A and B as blocks on the diagonal. We require the following useful properties of C ($\|\cdot\|_0$ denotes ℓ_0-norm).

Definition 3.1 *The m-sparse minimal and maximal eigenvalues of C, denoted by $\phi_{\min}(m)$ and $\phi_{\max}(m)$, are*

$$\min_{\nu:\|\nu\|_0\leq\lceil m\rceil}\frac{\nu^T C\nu}{\nu^T\nu} \quad and \quad \max_{\nu:\|\nu\|_0\leq\lceil m\rceil}\frac{\nu^T C\nu}{\nu^T\nu} \quad (13)$$

We call a feature "active" if its coefficient is non-zero. Now, each site may have different active features: let $s_h \leq kp$ be the sum of the number of active features over all sites. Similarly, s_p is the cardinality of the union of features that are active in at least one site ($s_h \leq ks_p, s_p \leq p$). Recall that when $\alpha \neq 0$, we add the Lasso penalty to the multi-site Lasso penalty. When the sparsity patterns are assumed to be similar across all sites, α is small. In contrast, to encourage site-specific sparsity patterns, we may set α to be large. We now analyze these cases independently.

Theorem 3.2 *Let $0 \leq \alpha \leq 0.4$. Assume there exist constants $0 \leq \rho_{\min} \leq \rho_{\max} \leq \infty$ such that*

$$\liminf_{n\to\infty}\phi_{\min}\left(s_p\left(1+\frac{2\alpha}{1-2\alpha}\right)^2\right) \geq \rho_{\min}$$

$$\limsup_{n\to\infty}\phi_{\max}(s_p + \min\{\sum_{i=1}^{k}n_i, kp\}) \leq \rho_{\max}. \quad (14)$$

Then, for $\lambda \propto \sigma\sqrt{\bar{n}\log(kp)}$, there exists a constant $\omega > 0$ such that, with probability converging to 1 for $n \to \infty$,

$$\frac{1}{k}\|\hat{B}^\lambda - B^*\|_F^2 \leq \omega\sigma^2\frac{\bar{s}\log(kp)}{\bar{n}}, \quad (15)$$

where $\bar{s} = \{(1-\alpha)\sqrt{s_p} + \alpha\sqrt{s_h/k}\}^2$, σ is the noise level.

Remarks: The above result agrees with known results for multi-task Lasso (Liu et al., 2009; Liu & Zhang, 2009b) when the sparsity patterns are the same across sites. The simplest way to interpret Theorem 3.2 is via the ratio $r = \frac{s_h}{s_p}$. Here, $r = k$ when the sparsity patterns are the same across sites. As r decreases, the sparsity patterns across sites start to differ, in turn, the sparse multi-site Lasso from (11) will provide stronger consistency compared to the multi-site Lasso (which corresponds to $\alpha = 0$). In other words, whenever we expect *site*-specific active features, the ℓ_2 consistency of (11) will improve as one includes an additional ℓ_1-penalty together with multi-site Lasso.

Observe that for the non-sparse β^j, we can verify that $\|\beta^j\|_1$ and $\sqrt{k}\|\beta^j\|_2$ have the same scale. On the other hand, for sparse β^j, $\|\beta^j\|_1$ has the same scale as $\|\beta^j\|_2$, i.e., with no $\sqrt{k}$ penalization (see appendix). Unlike Theorem 3.2 where the sparsity patterns across sites are similar, due to this scaling issue, the parameters α and λ need to be 'corrected' for the setting where sparsity patterns have little overlap. We denote these corrected versions by $\tilde{\alpha} = \frac{\alpha}{(1-\alpha)\sqrt{k}+\alpha}$ and $\tilde{\lambda} = ((1-\alpha)\sqrt{k} + \alpha)\lambda$.

Theorem 3.3 *Let* $0.4 \leq \tilde{\alpha} \leq 1$. *Assume there exist constants* $0 \leq \rho_{\min} \leq \rho_{\max} \leq \infty$ *such that*

$$\liminf_{n\to\infty} \phi_{\min}\left(s_h\left(1 + \frac{(1-\tilde{\alpha})}{\tilde{\alpha}}\right)^2\right) \geq \rho_{\min}$$

$$\limsup_{n\to\infty} \phi_{\max}(s_h + \min\{\sum_{i=1}^{k} n_i, kp\}) \leq \rho_{\max}. \tag{16}$$

Then, for $\tilde{\lambda} \propto \sigma\sqrt{\bar{n}\log(kp)}$, *there exists* $\omega > 0$ *such that, with probability converging to* 1 *for* $n \to \infty$, *we have* (15) *with* $\tilde{s} = \{(1-\tilde{\alpha})\sqrt{s_p/k} + \tilde{\alpha}\sqrt{s_h/k}\}^2$ *instead of* $\bar{s}$.

Remarks: This result agrees with known results for Lasso (Meinshausen & Yu, 2009) when the sparsity patterns are completely different across sites. In this case (i.e., α is large), the sparse multi-site Lasso has stronger consistency compared to Lasso ($\alpha = 1$). The sparse multi-site Lasso is preferable as $r = \frac{s_h}{s_p}$ increases. Note that although $\tilde{\alpha}$ and $\tilde{\lambda}$ are used for the results instead of α and λ, in practice, one can simply scale the chosen αs appropriately, e.g., with $k = 100$, we see that $\alpha \approx 0.99$ corresponds to $\tilde{\alpha} = 0.95$.

Performing hypothesis tests: Theorems 3.2 and 3.3 show consistency of sparse multi-site Lasso estimation. Hence, if the hyper-parameters α and λ are known, we can estimate the coefficients B^*. This variable selection phase can be followed by a hypothesis test, similar to Theorem 2.3 from Section 2. The only remaining issue is the choice of α. The existing methods show that joint cross-validation for α and λ performs unsatisfactorily and instead use a heuristic: set it to 0.05 when it is known that sparsity

patterns are similar across sites and 0.95 otherwise (Simon et al., 2013). Below, instead of a fixed α, we provide a data-driven alternative that works well in practice.

Choosing α using simultaneous inference: Our results in Thm. 3.2 (and Thm. 3.3 resp.) seem to suggest that increasing (and decreasing resp.) α will always improve consistency; however, this ends up requiring stronger m-sparsity conditions. We now describe a procedure to choose α. First, recall that an active feature corresponds to a variable with non-zero coefficient. We call a feature "site-active" if it is active at a site, an "always-active" feature is active at all k sites. The proposed solution involves three steps. **(1)** First, we apply simultaneous inference (like multi sample-splitting or de-biased Lasso) using all features at each of the k sites with FWER control. This step yields "site-active" features for each site, and therefore, gives the set of always-active features and the sparsity patterns. **(2)** Then, each site runs a Lasso and chooses a λ_i based on cross-validation. We then set $\lambda_{\mathrm{multi-site}}$ to be the minimum among the best λ's from each site. Using $\lambda_{\mathrm{multi-site}}$, we can vary α to fit various sparse multi-site Lasso models – each run will select some number of always-active features. We plot α versus the number of always-active features. **(3)** Finally, based on the sparsity patterns from the *site-active* set, we estimate whether the sparsity patterns across sites are similar or different (i.e., share few active features). Then, based on the plot from step (2), if the sparsity patterns from the site-active sets are different (similar) across sites, then the smallest (largest) value of α that selects the minimum (maximum) number of always-active features is chosen. The appendix includes details.

4. Experiments

Our experiments are two-fold. First we perform simulations evaluating the hypothesis test from §2 and sparse multi-site Lasso from §3. We then evaluate pooling two Alzheimer's disease (AD) datasets from different studies to evaluate improvements in power, and checking whether the proposed tests provide insights into the regimes when pooling is beneficial for regression, and will yield tangible statistical benefits in investigating scientific hypotheses.

Power and Type I Error of Theorem 2.3: The first set of simulations evaluate the setting from Section 2.1 where the coefficients are same across two different sites. The inputs for the two sites are set as X_1, $X_2(\in \mathbb{R}^{n\times 3}) \sim \mathcal{N}(0, \Sigma)$ with $\Sigma = 0.5(I + E)$ (where I is identity and E is a 3×3 matrix of 1s). The true coefficients are given by $\beta_1 \sim U(0, 4I)$ and $\beta_2 = \beta_1 + 0.1$ (where $U(\cdot)$ is multivariate uniform), and the noise corresponds to $\epsilon_1 \sim \mathcal{N}(0, 3I)$ and $\epsilon_2 \sim \mathcal{N}(0, 0.5I)$ for the two sites respectively. With this design, the responses are set as $y_1 = X_1\beta_1 + \epsilon_1$ and $y_2 = X_2\beta_2 + \epsilon_2$. Using $\{X_1, y_1\}$ and $\{X_2, y_2\}$, the *shared*

$\hat{\beta}$ are estimated. The simulation is repeated 100 times with 9 different sample sizes ($n = 2^b$ with $b = 4, \ldots, 12$) for each repetition. Fig. 3(a) shows the MSE of two-site (blue bars) and a baseline single-site (red bars) model computed using the corresponding $\hat{\beta}$s on site 1. Although both MSEs decrease as n increases, the two-sites model consistently produces smaller MSE – with large gains for small sample sizes (left-end of Fig. 3(a)). Fig. 3(d) shows the acceptance rates of our proposed hypothesis test (from (6) and (8)) with 0.05 significance level. The purple solid line is the sufficient condition from Theorem 2.3, while the dotted line is where the MSE of the baseline single-site model starts to decrease below that of the two-site model. The trend in Fig. 3(d) implies that as n increases, the test tends to reject pooling the multi-site data with power $\rightarrow 1$. Further, the type I error is well-controlled to the left of the solid line, and is low between the two lines. See appendix for additional details about Figs. 3(a,d).

Power and Type I Error of Theorem 2.5: The second set of simulations evaluate the confounding variables setup from Section 2.2. Similar to Section 4, here we have (X_1, Z_1), $(X_2, Z_2) \sim \mathcal{N}(0, \Sigma)$ with $\Sigma = \begin{pmatrix} 0.5I_{3\times3} + 0.5E_{3\times3}, & 0.2E_{3\times5} \\ 0.2E_{5\times3}, & 0.8I_{5\times5} + 0.2E_{5\times5} \end{pmatrix}$. β_1 and β_2 are the same as before. $\gamma_1 = (1, 1, 2, 2, 2)^T$ and $\gamma_2 = (2, 2, 2, 1, 1)^T$ are the coefficients for Z_1 and Z_2 respectively. The new responses y_1 and y_2 will have the extra terms $Z_1\gamma_1$ and $Z_2\gamma_2$ respectively. Fig. 3(b,e) shows the results. All the observations from Fig. 3(a,d) hold here as well. For small n, MSE of two-site model is much smaller than baseline, and as sample size increases this difference reduces. The test accepts with high probability for small n, and as sample size increases it rejects with high power. The regimes of low type I error and high power in Fig. 3(e) are similar to those from Fig. 3(d).

4.1. Sparse multi-sites Lasso ℓ_2-consistency

We now use 4 sites with $n = 150$ samples each and $p = 400$ features to test the sparse multi-site model from §3. We set the design matrices X_i ($i = 1, \ldots, 4) \sim \mathcal{N}(0, \Sigma)$ with $\Sigma = 0.8I_{p\times p} + 0.2E_{p\times p}$. We consider the two cases (sparsity patterns shared/not shared) separately.

Few sparsity patterns shared: 6 shared features and 14 site-specific features (out of the 400) are set to be active in 4 sites. Each shared feature is sampled from $U(0, 4)$ for the first two sites and $U(0, 0.5)$ for the rest. All the site-specific features are $\sim U(0, 4)$. The noise $\epsilon_i \sim \mathcal{N}(0, 1)$, and the responses are $y_i = X_i\beta_i + \epsilon_i$. Fig. 3(c) shows the 10-fold cross validation error as λ changes (i.e., solution path) for different α settings, including the value from our proposed selection procedure (from Section 3.1), Lasso ($\alpha = 1$), group Lasso ($\alpha = 0$) and arbitrary values $\alpha = 0.05, 0.95$ (as suggested by (Simon et al., 2013)).

Our chosen $\alpha = 0.97$ (the blue curve in Fig. 3(c)) has the smallest error, across all λs, thereby implying a better ℓ_2 consistency. Table 1 in the appendix includes more details, including $\alpha = 0.97$ discovers more always-active features, while preserving the ratio of correctly discovered active features to all the discovered ones.

Most sparsity patterns shared: Unlike the earlier case, here we set 16 shared and 4 site-specific features (both $\sim U(0, 4)$) to be active among all 400 features. The result, shown in Fig. 3(f), is similar to Fig. 3(c). The proposed choice of $\alpha = 0.25$ competes favorably with alternate choices while preserving the correctly discovered number of always-active features. Unlike the previous case, the ratio of correctly discovered active features to all discovered ones increases here (see appendix).

4.2. Combining AD datasets from multiple sites

We now evaluate whether two AD datasets acquired at different sites – an Alzheimer's Disease Neuroimage Initiative (ADNI) dataset and a local dataset from Wisconsin ADRC (ADlocal) can be combined (appendix has dataset details). The sample sizes are 318 and 156 respectively. Cerebrospinal fluid (CSF) protein levels are the inputs, and the response is hippocampus volume. Using 81 age-matched samples from each dataset, we first perform domain adaptation (using a maximum mean discrepancy objective as a measure of distance between the two marginals), and then transform CSF proteins from ADlocal to match with ADNI. The transformed data is then used to evaluate whether adding ADlocal data to ADNI will improve the regression performed on the ADNI data. This is done by training a regression model on the 'transformed' ADlocal and a subset of ADNI data, and then testing the resulting model on the remaining ADNI samples. We use two baseline models each of which are trained using – ADNI data *alone*; and *non-transformed* ADlocal (with ADNI subset).

Fig. 4(a,b) show the resulting mean prediction error (MPE) scaled by the estimated noise level in ADNI responses, and the corresponding acceptance rate (with significance level 0.05) respectively. The x-axis in Fig. 4(a,b) represents the size of ADNI subset used for training. As expected, the MPE reduces as this subset size increases. Most importantly, pooling after transformation (green bars) seems to be the most beneficial in terms of MPE reduction. As shown in Fig. 4(a), to the left of purple line where the subset size is smaller than ADlocal datasize, pooling the datasets improves estimation. This is the small sample size regime which necessitates pooling efforts in general. As the dataset size increases (to the right of x-axis in Fig. 4(a)) the resulting MPE for the pooled model is close to what we will achieve using the ADNI data by itself.

Since pooling after transformation is at least as good as us-
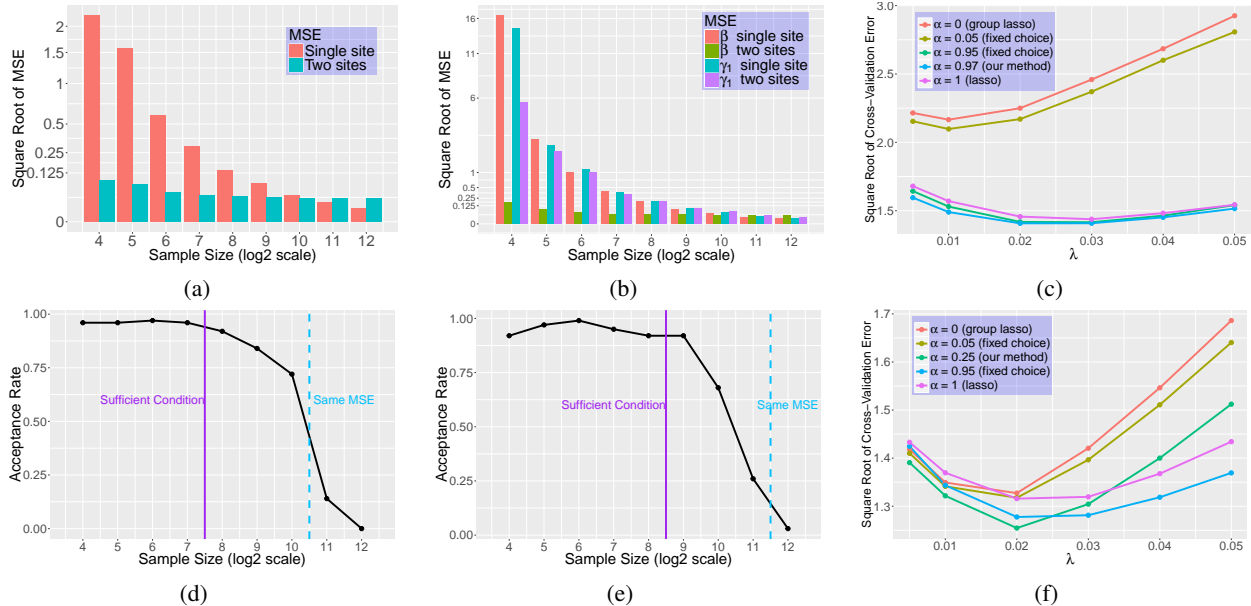
Figure 3. (a,d) $\hat{\beta}$'s MSE and the acceptance rate (Sec 2.1), (b,e) MSE of $\hat{\beta}$ and $\hat{\gamma}_1$, and the acceptance rate (Sec 2.2) using 100 bootstrap repetitions. Solid line in (d,e) is when the condition from Theorem 2.3 is 1. Dotted line is when MSE of single-site and multi-site models are the same. (c) λ error path when sparsity patterns are dissimilar across sites, (f) The regime where sparsity patters are similar.
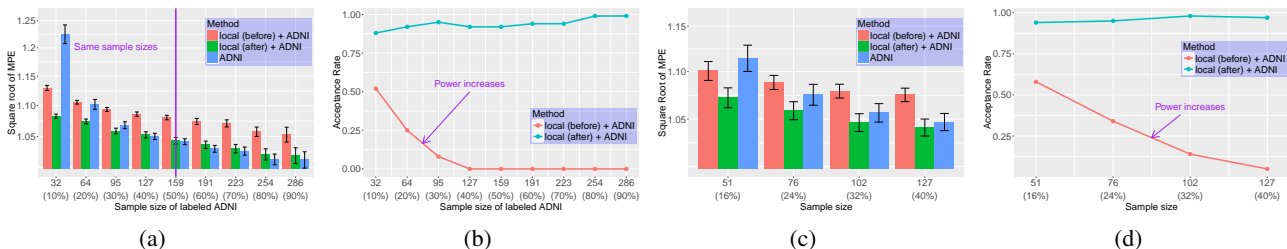

Figure 4. (a,c) MPE for the pooled regression model after/before transformations (green/red) compared to baseline (blue) plotted against training subset size of ADNI. x-axis is number/fraction of ADNI labeled samples used in training (apart from ADlocal). (b,d) show the acceptance rates for (a,c). Unlike in (a), (c) restricts same training data size for ADNI and ADlocal.

ing ADNI data alone, our hypothesis test accepts the combination with high rate ($\approx 95\%$), see Fig. 4(b). The test rejects the pooling strategy with high power for combining before domain adaptation (see Fig. 4(b)), as one would expect. This rejection power increases rapidly as sample size increases, see red curve in Fig. 4(b). The results in Fig. 4(c,d) show the setting where one *cannot* change the dataset sizes at the sites i.e., the training set uses an equal number of labeled samples from both the ADNI and ADlocal (x-axis in Fig. 4(c)), and the testing set always corresponds to 20% of ADNI data. This is a more interesting scenario for a practitioner compared to Fig. 4(a,b), because in Fig. 4(c,d) we use same sample sizes for both datasets. The trends in Fig. 4(c,d) are the same as Fig. 4(a,b).

5. Conclusions

We present a hypothesis test to answer whether pooling multiple datasets acquired from different sites is guaran-

teed to increase statistical power for regression models. For both standard and high dimensional linear regression, we identify regimes where such pooling is sensible, and show how such policy decisions can be made via simple checks executable on each site before any data transfer ever happens. We also show empirical results by combining two Alzheimer's disease datasets in the context of different regimes proposed by our analysis, and see that the regression fit improves as suggested by the theory. The code is available at https://github.com/hzhoustat/ICML2017.

References

Ando, Rie Kubota and Zhang, Tong. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(Nov):1817–1853, 2005.

Baktashmotlagh, Mahsa, Harandi, Mehrtash T, Lovell, Brian C, and Salzmann, Mathieu. Unsupervised domain adaptation by domain invariant projection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 769–776, 2013.

Ben-David, Shai and Schuller, Reba. Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines*, pp. 567–580. Springer, 2003.

Ben-David, Shai, Blitzer, John, Crammer, Koby, Kulesza, Alex, Pereira, Fernando, and Vaughan, Jennifer Wortman. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

Chatterjee, Soumyadeep, Steinhaeuser, Karsten, Banerjee, Arindam, Chatterjee, Snigdhansu, and Ganguly, Auroop. Sparse group lasso: Consistency and climate applications. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, pp. 47–58. SIAM, 2012.

Chen, Xi, He, Jinghui, Lawrence, Rick, and Carbonell, Jaime G. Adaptive multi-task sparse learning with an application to fmri study. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, pp. 212–223. SIAM, 2012.

Cortes, Corinna and Mohri, Mehryar. Domain adaptation in regression. In *International Conference on Algorithmic Learning Theory*, pp. 308–323. Springer, 2011.

Dezeure, Ruben, Bühlmann, Peter, Meier, Lukas, Meinshausen, Nicolai, et al. High-dimensional inference: Confidence intervals, *p*-values and r-software hdi. *Statistical Science*, 30(4):533–558, 2015.

Dezeure, Ruben, Bühlmann, Peter, and Zhang, Cun-Hui. High-dimensional simultaneous inference with the bootstrap. *arXiv preprint arXiv:1606.03940*, 2016.

ENCODE Project Consortium. The encode (encyclopedia of dna elements) project. *Science*, 306(5696):636–640, 2004.

Fortin, Jean-Michel and Currie, David J. Big science vs. little science: how scientific impact scales with funding. *PloS one*, 8(6):e65263, 2013.

Ganin, Yaroslav, Ustinova, Evgeniya, Ajakan, Hana, Germain, Pascal, Larochelle, Hugo, Laviolette, François, Marchand, Mario, and Lempitsky, Victor. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Gong, Boqing, Grauman, Kristen, and Sha, Fei. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *ICML (1)*, pp. 222–230, 2013.

Greco, T, Zangrillo, A, Biondi-Zoccai, G, and Landoni, G. Meta-analysis: pitfalls and hints. *Heart Lung Vessel*, 5 (4):219–25, 2013.

Group, Glioma Meta-analysis Trialists GMT. Chemotherapy in adult high-grade glioma: a systematic review and meta-analysis of individual patient data from 12 randomised trials. *The Lancet*, 359(9311):1011–1018, 2002.

Haase, Michael, Bellomo, Rinaldo, Devarajan, Prasad, Schlattmann, Peter, Haase-Fielitz, Anja, and Group, NGAL Meta-analysis Investigator. Accuracy of neutrophil gelatinase-associated lipocalin (ngal) in diagnosis and prognosis in acute kidney injury: a systematic review and meta-analysis. *American Journal of Kidney Diseases*, 54(6):1012–1024, 2009.

Hoerl, Arthur E and Kennard, Robert W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Huang, Jiayuan, Smola, Alexander J, Gretton, Arthur, Borgwardt, Karsten M, Schölkopf, Bernhard, et al. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19: 601, 2007.

Jahanshad, Neda, Kochunov, Peter V, Sprooten, Emma, Mandl, René C, Nichols, Thomas E, Almasy, Laura, Blangero, John, Brouwer, Rachel M, Curran, Joanne E, de Zubicaray, Greig I, et al. Multi-site genetic analysis of diffusion images and voxelwise heritability analysis: A pilot project of the enigma–dti working group. *Neuroimage*, 81:455–469, 2013.

James, William and Stein, Charles. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pp. 361–379, 1961.

Kim, Seyoung and Xing, Eric P. Tree-guided group lasso for multi-task regression with structured sparsity. 2010.

Klunk, William E, Koeppe, Robert A, Price, Julie C, Benzinger, Tammie L, Devous, Michael D, Jagust, William J, Johnson, Keith A, Mathis, Chester A, Minhas, Davneet, Pontecorvo, Michael J, et al. The centiloid project: standardizing quantitative amyloid plaque estimation by pet. *Alzheimer's & Dementia*, 11(1):1–15, 2015.

Lauer, Mike. Fy2015 by the numbers, and a quick look at recent trends. 2016. URL https://nexus.od.nih.gov/all/2016/03/14/fy2015-by-the-numbers/.

Lee, Seunghak, Zhu, Jun, and Xing, Eric P. Adaptive multi-task lasso: with application to eqtl detection. In *Advances in neural information processing systems*, pp. 1306–1314, 2010.

Li, Qi. Literature survey: domain adaptation algorithms for natural language processing. *Department of Computer Science The Graduate Center, The City University of New York*, pp. 8–10, 2012.

Liu, Han and Zhang, Jian. Estimation consistency of the group lasso and its applications. In *AISTATS*, pp. 376–383, 2009a.

Liu, Han and Zhang, Jian. Estimation consistency of the group lasso and its applications. In *AISTATS*, pp. 376–383, 2009b.

Liu, Han, Palatucci, Mark, and Zhang, Jian. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 649–656. ACM, 2009.

Long, Mingsheng, Cao, Yue, Wang, Jianmin, and Jordan, Michael I. Learning transferable features with deep adaptation networks. In *ICML*, pp. 97–105, 2015.

Maurer, Andreas, Pontil, Massimiliano, and Romera-Paredes, Bernardino. Sparse coding for multitask and transfer learning. In *ICML (2)*, pp. 343–351, 2013.

Maurer, Andreas, Pontil, Massimiliano, and Romera-Paredes, Bernardino. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

Meinshausen, Nicolai and Yu, Bin. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, pp. 246–270, 2009.

Murdoch, Travis B and Detsky, Allan S. The inevitable application of big data to health care. *Jama*, 309(13): 1351–1352, 2013.

Pan, Sinno Jialin, Tsang, Ivor W, Kwok, James T, and Yang, Qiang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

Patel, Vishal M, Gopalan, Raghuraman, Li, Ruonan, and Chellappa, Rama. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015.

Simon, Noah, Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

Tarca, Adi L, Carey, Vincent J, Chen, Xue-wen, Romero, Roberto, and Drăghici, Sorin. Machine learning and its applications to biology. *PLoS Comput Biol*, 3(6):e116, 2007.

Weiner, Michael W, Veitch, Dallas P, Aisen, Paul S, Beckett, Laurel A, Cairns, Nigel J, Cedarbaum, Jesse, Donohue, Michael C, Green, Robert C, Harvey, Danielle, Jack, Clifford R, et al. Impact of the alzheimer's disease neuroimaging initiative, 2004 to 2014. *Alzheimer's & Dementia*, 11(7):865–884, 2015.

Zhou, Hao H, Ravi, Sathya N, Ithapu, Vamsi K, Johnson, Sterling C, Wahba, Grace, and Singh, Vikas. Hypothesis testing in unsupervised domain adaptation with applications in alzheimer's disease. In *Advances in Neural Information Processing Systems*, pp. 2496–2504, 2016.