Journal of Machine Learning Research 16 (2015) 3183-3217 Submitted 10/14; Revised 5/15; Published 12/15

# On Semi-Supervised Linear Regression in Covariate Shift Problems

**Kenneth Joseph Ryan** KJRYAN@MAIL.WVU.EDU
**Mark Vere Culp** MVCULP@MAIL.WVU.EDU
*Department of Statistics*
*West Virginia University*
*Morgantown, WV 26506, USA*

**Editor:** Xiaotong Shen

## Abstract

Semi-supervised learning approaches are trained using the full training (labeled) data and available testing (unlabeled) data. Demonstrations of the value of training with unlabeled data typically depend on a smoothness assumption relating the conditional expectation to high density regions of the marginal distribution and an inherent missing completely at random assumption for the labeling. So-called covariate shift poses a challenge for many existing semi-supervised or supervised learning techniques. Covariate shift models allow the marginal distributions of the labeled and unlabeled feature data to differ, but the conditional distribution of the response given the feature data is the same. An example of this occurs when a complete labeled data sample and then an unlabeled sample are obtained sequentially, as it would likely follow that the distributions of the feature data are quite different between samples. The value of using unlabeled data during training for the elastic net is justified geometrically in such practical covariate shift problems. The approach works by obtaining adjusted coefficients for unlabeled prediction which recalibrate the supervised elastic net to compromise: (i) maintaining elastic net predictions on the labeled data with (ii) shrinking unlabeled predictions to zero. Our approach is shown to dominate linear supervised alternatives on unlabeled response predictions when the unlabeled feature data are concentrated on a low dimensional manifold away from the labeled data and the true coefficient vector emphasizes directions away from this manifold. Large variance of the supervised predictions on the unlabeled set is reduced more than the increase in squared bias when the unlabeled responses are expected to be small, so an improved compromise within the bias-variance tradeoff is the rationale for this performance improvement. Performance is validated on simulated and real data.

**Keywords:** joint optimization, semi-supervised regression, usefulness of unlabeled data

## 1. Introduction

Semi-supervised learning is an active research area (Chapelle et al., 2006b; Zhu and Goldberg, 2009). Existing theoretical and empirical work typically invokes the missing completely at random (MCAR) assumption where the inclusion of a label is independent of the feature data and label. Under MCAR, there is theoretical work, mostly in classification, on finding borders that pass between dense regions of the data with particular emphasis on the cluster assumption (Chapelle et al., 2006b), semi-supervised smoothness assumptions (Lafferty and Wasserman, 2007; Azizyan et al., 2013), and manifold assumptions (Hein et al.,

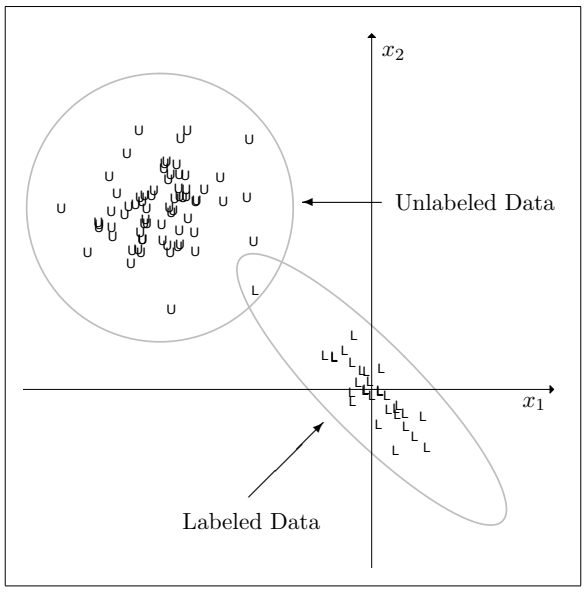


Figure 1: These feature data with $p = 2$ are referred to as the "block extrapolation" example because the unlabeled data "block" the 1st principal component of the labeled data. It is informative to think about how ridge regression would predict the unlabeled cases in this example. Favoring shrinking along the 2nd component will lead to high prediction variability. These block data are the primary working example throughout Sections 2-5, and it will be demonstrated that our semi-supervised approach has a clear advantage.

2005; Aswani et al., 2010). Many techniques including manifold regularization (Belkin et al., 2006) and graph cutting approaches (Wang et al., 2013) were developed to capitalize on unlabeled information during training, but beneath the surface of nearly all this work is the implicit or explicit use of MCAR (Lafferty and Wasserman, 2007).

Covariate shift is a different paradigm for semi-supervised learning (Moreno-Torres et al., 2008). It stipulates that the conditional distribution of the label given the feature data does not depend on the missingness of a label, but that the feature data distribution may depend on the missingness of a label. As a consequence, feature distributions can differ between labeled and unlabeled sets. Attempting to characterize smoothness assumptions between the regression function and the marginal of $\boldsymbol{X}$ (Azizyan et al., 2013) may not realize the value of unlabeled data if an implicit MCAR assumption breaks down. Instead, its value is in shrinking regression coefficients in an ideal direction to optimize the bias-variance tradeoff on unlabeled predictions. This is a novelty of our research direction.

The proposed approach is ideally suited for applications where the sequential generation of the labeled and unlabeled data causes covariate shift. Due to either matters of practicality or convenience the marginal distribution of the labeled feature data is likely to be profoundly different than that of the unlabeled feature data. Consider applications in drug discovery where the feature information consists of measurements on compounds and the responses are compound attributes, e.g., side effects of the drug, overall effect of the drug, or ability

to permeate the drug (Mente and Lombardo, 2005). Attributes can take years to obtain, while the feature information can be obtained much faster. As a result, the labeled data are often measurements on drugs with known attributes while the unlabeled data are usually compounds with unknown attributes that may potentially become new drugs (marketed to the public). Other applications mostly in classification include covariate shift problems (Yamazaki et al., 2007), reject inference problems from credit scoring (Moreno-Torres et al., 2008), spam filtering and brain computer interfacing (Sugiyama et al., 2007), and gene expression profiling of microarray data (Gretton et al., 2009). Gretton et al. (2009) further note that covariate shift occurs often in practice, but is under reported in the machine learning literature.

Many of the hypothetical examples to come do not conform to MCAR. The Figure 1 feature data are used to illustrate key concepts as they are developed in this work. Its labeled and unlabeled partitioning is unlikely if responses are MCAR. The vector of supervised ridge regression coefficients is proportionally shrunk more along the lower order principal component directions (Hastie et al., 2009). Such shrinking is toward a multiple of the unlabeled data centroid in the hypothetical Figure 1 scenario, so ridge regression may not deflate the variance of the unlabeled predictions enough. Standard methods for tuning parameter estimation via cross-validation do not account for the distribution of the unlabeled data either. Thus, supervised ridge regression is at a distinct disadvantage by not accounting for the unlabeled data during optimization. In general, the practical shortcoming of supervised regression (e.g., ridge, lasso, or elastic net) is to define regression coefficients that predict well for any unlabeled configuration. Our main contribution to come is a mathematical framework for adapting a supervised estimate to the unlabeled data configuration at hand for improved performance. It also provides interpretable "extrapolation" adjustments to the directions of shrinking as a byproduct.

Culp (2013) proposed a joint trained elastic net for semi-supervised regression under MCAR. The main idea was to use the joint training problem that encompasses the $S^3VM$ (Chapelle et al., 2006a) and $\psi$-learning (Wang et al., 2009) to perform semi-supervised elastic net regression. The concept was that the unlabeled data should help with decorrelation and variable selection, two known hallmarks of the supervised elastic net extended to semi-supervised learning (Zou and Hastie, 2005). Culp (2013), however, did not contain a complete explanation of how exactly the approach used unlabeled data and under what set of mathematical assumptions it is expected to be useful.

The joint trained elastic net framework is strengthened in this paper to handle covariate shift. Rigorous geometrical and theoretical arguments are given for when it is expected to work. Circumstances where the feature data distribution changes by label status is the primary setting. One could view the unlabeled data as providing a group of extrapolations (or a separate manifold) from the labeled data. Even if responses are MCAR, the curse of dimensionality stipulates that nearly all predictions from a supervised learner are extrapolations in higher dimensions (Hastie et al., 2009), so the utility of the proposed semi-supervised approach is likely to increase with $p$.

Presentation of major concepts often begins with hypothetical, graphical examples in $p = 2$, but is followed by general mathematical treatments of $p \geq 2$. The work is written carefully so that themes extracted from $p = 2$ generalize. Section 2 provides a conceptual overview of the general approach with emphasis on the value of unlabeled data in covariate

shift before diving into the more rigorous mathematics in later sections. The problem is set-up formally in Section 3. The nature of regularization approaches (e.g., ridge, lasso, and elastic net) is studied with emphasis on a geometric perspective in Section 4. The geometry helps articulate realistic assumptions for the theoretical risk results in Section 5, and the theoretical risk results help define informative simulations and real data tests in Section 6. In addition, the simulations and real data applications validate the theoretical risk results. The combined effect is a characterization of when the approach is expected to outperform supervised alternatives in prediction. Follow-up discussion is in Section 7, and a proof for each proposition and theorem is in Appendix A.

## 2. The Value of Unlabeled Data due to Covariate Shift

The purpose of this section is to motivate the proposed approach for covariate shift data problems. The data are partitioned into the set of the labeled $L$ and unlabeled $U$ observations with $n = |L| + |U|$, and a response variable is recorded only for labeled observations. Let $\boldsymbol{Y}_L$ denote the observed $|L| \times 1$ vector of mean centered, labeled responses and $\boldsymbol{Y}_U$ the $|U| \times 1$ missing, unlabeled responses. If data are sorted by label status, the complete response vector and $n \times p$ model matrix partition to

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_L \\ \boldsymbol{Y}_U \end{pmatrix} \qquad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_L \\ \boldsymbol{X}_U \end{pmatrix}.$$

The $\boldsymbol{X}_L$ data are mean centered and standardized so that $\boldsymbol{X}_L^T \boldsymbol{X}_L$ is a correlation matrix, and $\boldsymbol{X}_U$ is also scaled using the means and variances of the labeled data. A supervised linear regression coefficient vector $\widehat{\boldsymbol{\beta}}^{(\text{SUP})}$ is trained using only the labeled data: $\boldsymbol{X}_L$ and $\boldsymbol{Y}_L$. Our semi-supervised $\widehat{\boldsymbol{\beta}}$ is trained with data $\boldsymbol{X}$ and $\boldsymbol{Y}_L$ by trading off: (i) supervised predictions $\boldsymbol{X}_L\widehat{\boldsymbol{\beta}} = \boldsymbol{X}_L\widehat{\boldsymbol{\beta}}^{(\text{SUP})}$ on $L$ with (ii) shrinking $\boldsymbol{X}_U\widehat{\boldsymbol{\beta}}$ towards $\vec{0}$ on $U$, and the geometric value of this type of usage of the unlabeled data is presented in Section 2.1. A deeper presentation of this Section 2.1 concept is given by Sections 3 and 4. This work also demonstrates its theoretical performance under the standard linear model. In particular, the true coefficient vector must encourage shrinking as a good strategy in order for the unlabeled data to be useful in the proposed fashion. The introduction of this concept here in Section 2.2 precedes the corresponding mathematical presentation of performance bounds in Section 5.

### 2.1 Geometric Contribution of Unlabeled Data

The main strategy is to find a linear compromise between: (i) fully supervised prediction on the labeled data and (ii) predicting close to zero on the unlabeled data. Two examples of this are given below. In the "collinearity" example, it is possible to achieve both (i) and (ii). Thus, there is no need for a compromise. In the block extrapolation example, (i) and (ii) cannot be achieved simultaneously. The compromise is obtained by organizing the coefficient vector in terms of directions *orthogonal* to feature data extrapolation directions, so the predictions corresponding to more extreme unlabeled extrapolations are shrunk more.

**Collinearity Example:** Suppose $p = 2$, the two columns of labeled feature data are collinear with $\boldsymbol{X}_{L1} = \boldsymbol{X}_{L2}$, and the unlabeled data are also collinear and orthogonal to the labeled data with $\boldsymbol{X}_{U1} = -\boldsymbol{X}_{U2}$. The ordinary least squares estimator $\widehat{\boldsymbol{\beta}}^{(\text{OLS})}$ (i.e., a

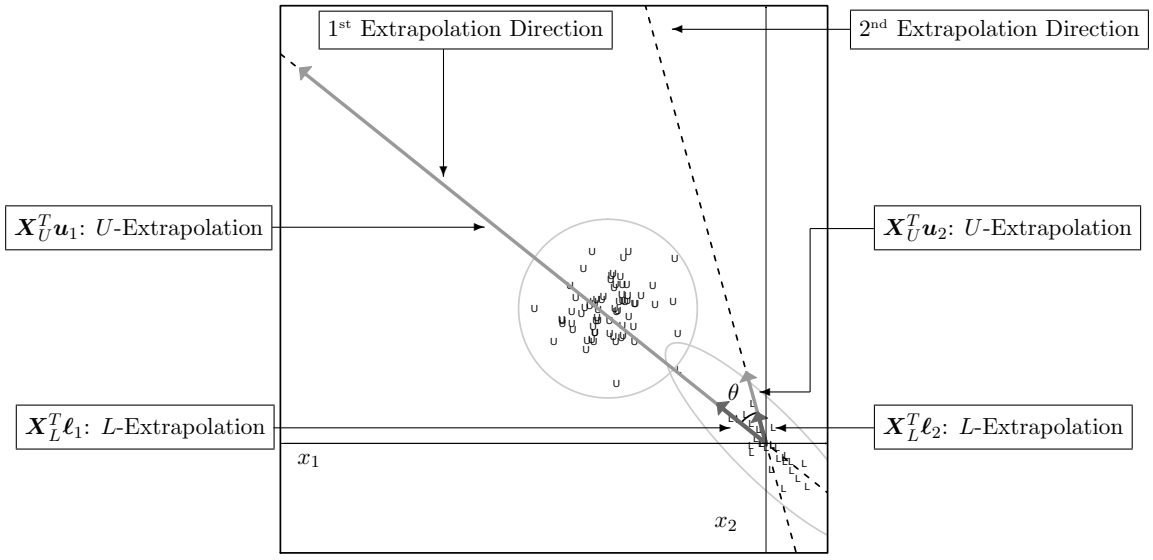


Figure 2: The dashed lines are the 1st and 2nd extrapolation directions for the block extrapolation example from Figure 1. The extent of $U$-extrapolation vector is a larger multiple of the extent of $L$-extrapolation vector in the 1st versus the 2nd extrapolation direction, so predictions corresponding to feature vectors on the 1st extrapolation direction are shrunk more than those on the 2nd extrapolation direction under the proposed method.

supervised linear regression estimator) is not unique since $\text{rank}(\boldsymbol{X}_L) = 1$, but the semi-supervised estimator $\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}_{L1}^T\boldsymbol{Y}_L/2\right)\vec{1}$ is the unique solution to

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{Y}_L - \boldsymbol{X}_L\boldsymbol{\beta}\|_2^2 + \|\boldsymbol{X}_U\boldsymbol{\beta}\|_2^2 . \tag{1}$$

This $\widehat{\boldsymbol{\beta}}$ is the ordinary least squares estimator with equal components, so it achieves objectives (i) $\boldsymbol{X}_L\widehat{\boldsymbol{\beta}} = \boldsymbol{X}_L\widehat{\boldsymbol{\beta}}^{(\text{OLS})}$ and (ii) $\boldsymbol{X}_U\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}_{L1}^T\boldsymbol{Y}_L/2\right)\boldsymbol{X}_U\vec{1} = \vec{0}$. Optimization Problem (1) is a special case of the joint training framework to come in Section 3, and our general semi-supervised approach is based on this type of estimator.

**Block Extrapolation Example:** These data in Figure 2 include two lines marked as 1st and 2nd extrapolation directions, and each direction has extent vectors of largest $U$- and $L$-extrapolations ($\boldsymbol{X}_L^T\boldsymbol{\ell}_1$, $\boldsymbol{X}_U^T\boldsymbol{u}_1$ and $\boldsymbol{X}_L^T\boldsymbol{\ell}_2$, $\boldsymbol{X}_U^T\boldsymbol{u}_2$). Each $L$-based extent vector in Figure 2 is the longest possible of the form $\boldsymbol{X}_L^T\boldsymbol{\ell}$ in a given direction for $\boldsymbol{\ell} \in \mathbb{R}^{|L|}$ such that $\|\boldsymbol{\ell}\|_2^2 = 1$. Similarly, the $U$-based extent vectors are the longest possible in a given direction based on a unit length linear combination of the rows of $\boldsymbol{X}_U$. While precise mathematics on determining the two extrapolation directions is deferred until Section 4, it also turns out that the ratio of $U$- to $L$-extent vector lengths in the 2nd direction is never bigger than that in the 1st direction, i.e.,

$$\frac{\left\|\boldsymbol{X}_U^T\boldsymbol{u}_2\right\|_2}{\left\|\boldsymbol{X}_L^T\boldsymbol{\ell}_2\right\|_2} \leq \frac{\left\|\boldsymbol{X}_U^T\boldsymbol{u}_1\right\|_2}{\left\|\boldsymbol{X}_L^T\boldsymbol{\ell}_1\right\|_2}. \tag{2}$$

The sought after compromise is struck with semi-supervised estimator $\widehat{\boldsymbol{\beta}}$ by shrinking a supervised estimator $\widehat{\boldsymbol{\beta}}^{(\text{SUP})}$ with respect to a basis of directions orthogonal to the extrapolation directions. With this in mind, define the decomposition of a supervised estimate

$$\begin{aligned} &\widehat{\boldsymbol{\beta}}^{(\text{SUP})} = \tilde{\boldsymbol{\nu}}_1 + \tilde{\boldsymbol{\nu}}_2, \text{ where} \\ &\tilde{\boldsymbol{\nu}}_1 \text{ is orthogonal to the } 1^{\text{st}} \text{ extrapolation direction} \\ &\tilde{\boldsymbol{\nu}}_2 \text{ is orthogonal to the } 2^{\text{nd}} \text{ extrapolation direction,} \end{aligned} \tag{3}$$

and consider a semi-supervised estimate of the form

$$\widehat{\boldsymbol{\beta}} = p_1\tilde{\boldsymbol{\nu}}_1 + p_2\tilde{\boldsymbol{\nu}}_2, \text{ where } p_1 = \frac{\left\|\boldsymbol{X}_L^T\boldsymbol{\ell}_1\right\|_2}{\left\|\boldsymbol{X}_L^T\boldsymbol{\ell}_1\right\|_2 + \left\|\boldsymbol{X}_U^T\boldsymbol{u}_1\right\|_2} \text{ and } p_2 = \frac{\left\|\boldsymbol{X}_L^T\boldsymbol{\ell}_2\right\|_2}{\left\|\boldsymbol{X}_L^T\boldsymbol{\ell}_2\right\|_2 + \left\|\boldsymbol{X}_U^T\boldsymbol{u}_2\right\|_2}. \tag{4}$$

Coefficient shrinking is more focused on the vector orthogonal to the $1^{\text{st}}$ extrapolation direction because $0 \leq p_1 \leq p_2 \leq 1$ by Inequality (2).

A semi-supervised $\widehat{\boldsymbol{\beta}}$ from Display (4) was decomposed with regard to a basis orthogonal to directions of extrapolations from Display (3) so that linear predictions $\boldsymbol{x}_0^T\widehat{\boldsymbol{\beta}}$ at an arbitrary feature vector $\boldsymbol{x}_0 \in I\!R^2$ are shrunk more heavily when $\boldsymbol{x}_0$ is in directions with larger extrapolations. To demonstrate this, define a closely related decomposition of a feature vector

$$\begin{aligned} &\boldsymbol{x}_0 = \boldsymbol{\nu}_1 + \boldsymbol{\nu}_2, \text{ where} \\ &\boldsymbol{\nu}_1 \text{ is on the } 1^{\text{st}} \text{ extrapolation direction} \\ &\boldsymbol{\nu}_2 \text{ is on the } 2^{\text{nd}} \text{ extrapolation direction.} \end{aligned} \tag{5}$$

Together, Decompositions (4) and (5) result in the semi-supervised prediction

$$\boldsymbol{x}_0^T\widehat{\boldsymbol{\beta}} = p_1\boldsymbol{\nu}_1^T\tilde{\boldsymbol{\nu}}_2 + p_2\boldsymbol{\nu}_2^T\tilde{\boldsymbol{\nu}}_1$$

because $\boldsymbol{\nu}_1^T\tilde{\boldsymbol{\nu}}_1 = \boldsymbol{\nu}_2^T\tilde{\boldsymbol{\nu}}_2 = 0$ by construction. Thus, with fixed length feature vectors $\boldsymbol{x}_0 = \boldsymbol{\nu}_i$ on the $1^{\text{st}}$ and $2^{\text{nd}}$ extrapolation directions, the $1^{\text{st}}$ direction corresponds to a semi-supervised prediction $\boldsymbol{x}_0^T\widehat{\boldsymbol{\beta}}$ that is a more heavily shrunken version of its supervised prediction $\boldsymbol{x}_0^T\widehat{\boldsymbol{\beta}}^{(\text{SUP})}$ whenever $p_1 < p_2$.

The supervised estimate $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{(\text{SUP})}$ results whenever $p_1 = p_2 = 1$, by Displays (3) and (4). Thus, supervised predictions are favored when $L$-based extrapolations $\left\|\boldsymbol{X}_L^T\boldsymbol{\ell}_i\right\|_2$ dominate $U$-based extrapolations $\left\|\boldsymbol{X}_U^T\boldsymbol{u}_i\right\|_2$ because $p_i \approx 1$ follows from Display (4). On the other hand, predictions near zero are favored when $U$-based extrapolations dominate $L$-based extrapolations ($p_i \approx 0$). In both cases, the $p_i$ regulate the compromise (i) with (ii) for $\widehat{\boldsymbol{\beta}}$ term-by-term in each extrapolation direction. A significant contribution of this work is to provide a rigorous mathematical framework to study semi-supervised linear predictions for unlabeled extrapolations. In Section 4, directions of extrapolation and relative degrees of shrinking $p_i$ are shown to follow from the joint trained optimization framework.

### 2.2 Model-based Contributions of Unlabeled Data

Under the linear model ($\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{X}\boldsymbol{\beta}$ and $\mathbb{V}\text{ar}(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}$), the coefficient parameter space partitions into lucky $(\boldsymbol{\beta}, \sigma^2)$ and unlucky $(\boldsymbol{\beta}, \sigma^2)$ subsets. Lucky versus unlucky $\boldsymbol{\beta}$ directions are not equally likely but depend greatly on the range and shape of the unlabeled

data manifold and on the model parameter $\sigma^2$. The general theme is that lucky (unlucky) $\boldsymbol{\beta}$'s are in directions orthogonal (parallel) to the unlabeled feature data manifold, so lower variability within this manifold implies more lucky $\boldsymbol{\beta}$ directions where our approach improves performance. A general bound is presented in Section 5 to help understand when our semi-supervised linear adjustment is guaranteed to outperform its supervised baseline on unlabeled predictions. Next, the collinearity and block extrapolation examples from Section 2.1 are revisited to illustrate lucky versus unlucky (or favorable versus unfavorable) prediction scenarios.

**Collinearity Example:** This example had $p = 2$, $\boldsymbol{X}_{L1} = \boldsymbol{X}_{L2}$, and $\boldsymbol{X}_{U1} = -\boldsymbol{X}_{U2}$. A lucky $\boldsymbol{\beta}$ follows with $\boldsymbol{\beta} = (b, b)^T$ for some arbitrary $b \in I\!R$, since $\boldsymbol{X}_U\boldsymbol{\beta} = \vec{0}$ is clearly ideal for the semi-supervised approach. On the other hand, suppose the true $\boldsymbol{\beta} = (b, -b)^T$ for some scalar $b$ of large magnitude, and the components of $\boldsymbol{X}_{U1}$ are all of large magnitude with the same sign. This is an example of an unlucky $\boldsymbol{\beta}$ since the truth $\boldsymbol{X}_U\boldsymbol{\beta} = 2b\boldsymbol{X}_{U1}$ is far from the origin $\vec{0}$ with components of the same sign, so setting $\boldsymbol{X}_U\widehat{\boldsymbol{\beta}} = \vec{0}$ is less than ideal. Since $\boldsymbol{X}_L\boldsymbol{\beta} = \vec{0}$, the typical supervised linear regression estimators (e.g., ridge, lasso, and ENET) would predict the $\boldsymbol{X}_U$ cases close to $\vec{0}$ not $2b\boldsymbol{X}_{U1}$ and does not fair much better as a result. The bottom-line is that this unlucky $\boldsymbol{\beta}$ situation is not handled well by the conventional wisdom in machine learning of shrinking to optimize the bias-variance tradeoff (Hastie et al., 2009).

**Block Extrapolation Example:** This example was the block extrapolation from Figures 1 and 2. As it turns out, the ridge regression version of the Section 5 bound simplifies to a function of just $\boldsymbol{\beta}$ (call it $\sigma^2_{\text{LB}}(\boldsymbol{\beta})$) such that the semi-supervised approach is guaranteed to outperform the supervised approach whenever $\sigma^2 - \sigma^2_{\text{LB}}(\boldsymbol{\beta}) > 0$ at a given $\sigma^2$. Next, this bound is used to give a snapshot of parameter space $(\boldsymbol{\beta}, \sigma^2)$ in the context of the block extrapolation example, where lucky $\boldsymbol{\beta}$ correspond to $\sigma^2 - \sigma^2_{\text{LB}}(\boldsymbol{\beta}) > 0$ while unlucky $\boldsymbol{\beta}$ correspond to $\sigma^2 - \sigma^2_{\text{LB}}(\boldsymbol{\beta}) \leq 0$.

In order to investigate this, take all $\sigma^2 \in [0, 1]$ with all possible coefficient vectors

$$\boldsymbol{\beta}(\vartheta) = \begin{pmatrix} \sin(\vartheta) \\ \cos(\vartheta) \end{pmatrix} \text{ for } \vartheta \in [0, \pi]$$

on the right half of the unit circle. These parameters capture performance trends of an arbitrary fixed length $\boldsymbol{\beta}$ in all possible directions by the technical details in Section 5. Curves in Figure 3(a) are the bound $\sigma^2 - \sigma^2_{\text{LB}}(\boldsymbol{\beta}(\vartheta))$ as a function of $\vartheta$ at a given $\sigma^2$. Lighter (darker) curves correspond to smaller (larger) values $\sigma^2$ over an equally spaced grid on the interval $[0, 1]$, and the corresponding differences between unlabeled root mean-squared errors at the best supervised ($\text{RMSE}_U^{(\text{SUP})}$) and semi-supervised ($\text{RMSE}_U^{(\text{SEMI})}$) tuning parameter settings are provided in Figure 3(b). If $\vartheta$ is uniformly distributed on $[0, \pi]$, a lucky $\boldsymbol{\beta}$ is more likely than an unlucky $\boldsymbol{\beta}$, especially as $\sigma^2$ increases. The center for potentially large improvements in Figure 3(a) is roughly $\boldsymbol{\beta}(\pi/4) \approx (1, 1)^T/\sqrt{2}$. In addition, the unlabeled feature data centroid $\boldsymbol{X}_U^T\vec{1}/|U|$ in Figure 1 is roughly a multiple of $(-1, 1)^T$. Thus, $\vec{1}^T\boldsymbol{X}_U\boldsymbol{\beta}(\pi/4) \approx 0$. In other words, lucky $\boldsymbol{\beta}$ directions encourage shrinking predictions on $U$. On the other hand, unlucky $\boldsymbol{\beta}$ directions encourage large predictions. Take the center for little to no theoretically guaranteed improvement in Figure 3(a), i.e., $\boldsymbol{\beta}(3\pi/4) \approx (1, -1)^T/\sqrt{2}$. In this case, the true expected response at the unlabeled feature data centroid $\vec{1}^T\boldsymbol{X}_U\boldsymbol{\beta}(3\pi/4)/|U|$ is large because $\boldsymbol{\beta}(3\pi/4)$ is roughly a multiple of $\boldsymbol{X}_U^T\vec{1}/|U|$.

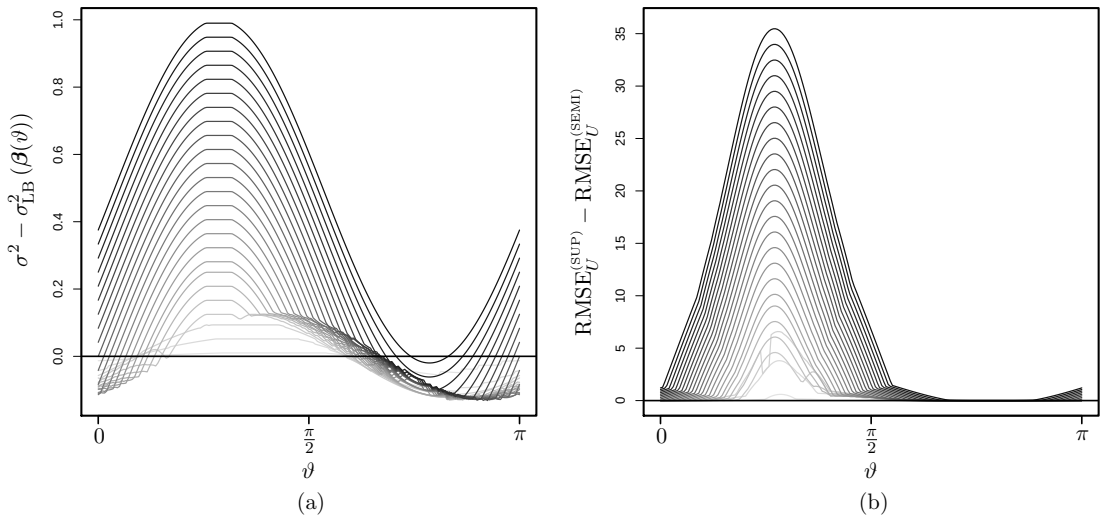


Figure 3: (a) The theoretical bound $\sigma^2 - \sigma^2_{\text{LB}}(\boldsymbol{\beta}(\vartheta))$ is plotted against $\vartheta$ for the block extrapolation example from Figures 1 and 2. Darker curves correspond to larger $\sigma^2$. Interest was in identifying $\vartheta$ such that $\sigma^2 - \sigma^2_{\text{LB}}(\boldsymbol{\beta}(\vartheta)) > 0$, since values greater than zero highlight the lucky unit length directions $\boldsymbol{\beta}(\vartheta)$ at a given $\sigma^2$ where our semi-supervised adjustment helps. (b) The corresponding differences between supervised and semi-supervised root mean squared errors (RMSEs) on the unlabeled set are displayed.

In general, the proposed approach is well suited for lucky $\boldsymbol{\beta}$ prediction problems, which include the following generalization of the Figure 1 block extrapolation example. The distance between feature data centroids (i.e., between the origin $\boldsymbol{X}_L^T\vec{1}/|L| = \vec{0}$ due to mean centering and $\boldsymbol{X}_U^T\vec{1}/|U|$) is increased relative to the variation about each centroid and the true coefficient vector $\boldsymbol{\beta}$ is not roughly a multiple of $\boldsymbol{X}_U^T\vec{1}/|U|$. One might conjecture lucky $\boldsymbol{\beta}$ to occur more often in practice during high-dimensional applications with large $p$ by a sparsity of effects assumption (i.e., the true $\boldsymbol{\beta}$ has few non-zero components). For example, if the unlabeled feature data are concentrated on a low dimensional manifold away from the labeled data, there are more lucky directions for the true coefficient vector to emphasize directions away from the unlabeled feature data manifold. Also note that the supervised RMSEs are no better than semi-supervised in the block example, i.e., no negative differences in Figure 3(b). In theory, our technique handles unlucky $\boldsymbol{\beta}$ by defaulting to supervised predictions; see Remark 1 for how unlucky scenarios are handled empirically in practice.

**Remark 1** *Nearly all supervised techniques would be challenged by an unlucky $\boldsymbol{\beta}$ direction since approaches typically improve predictive performance by shrinking (Hastie et al., 2009) and thus predicting large responses accurately on a covariate shifted data set is not what these techniques are designed to do. Supervised learning has a possible advantage over the proposed semi-supervised method in such situations by simply not shrinking extrapolation*

*directions in the unlabeled data, but there is no guarantee here either (i.e., the supervised technique may still perform much worse). In this work, we do not assume that the response is generated under a lucky $\boldsymbol{\beta}$ linear model. Instead, a tuning parameter is used to move the semi-supervised estimator closer to supervised in such cases to mitigate the losses relative to supervised for an unlucky $\boldsymbol{\beta}$. Cross-validation is used to estimate this parameter in the results Section 6.*

## 3. A Linear Joint Training Framework

The focus of this paper is the *joint trained elastic net*

$$\left(\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\gamma},\boldsymbol{\lambda}}, \widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma},\boldsymbol{\lambda}}\right) = \underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\arg\min} \|\boldsymbol{Y}_L - \boldsymbol{X}_L\boldsymbol{\beta}\|_2^2 + \gamma_1\|\boldsymbol{X}_U(\boldsymbol{\alpha}-\boldsymbol{\beta})\|_2^2 + \gamma_1\gamma_2\|\boldsymbol{\alpha}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1^1 + \lambda_2\|\boldsymbol{\beta}\|_2^2, \tag{6}$$

where $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma},\boldsymbol{\lambda}}$ is appropriately scaled and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2) \in [0,\infty]^2$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2) \in [0,\infty]^2$ are tuning parameter vectors. The joint trained elastic net is an example of a joint training optimization framework used in semi-supervised learning (Chapelle et al., 2006b). Comparisons will be made to the *supervised optimization*

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}^{(\text{ENET})} = \underset{\boldsymbol{\beta}}{\arg\min} \|\boldsymbol{Y}_L - \boldsymbol{X}_L\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1^1 + \lambda_2\|\boldsymbol{\beta}\|_2^2, \tag{7}$$

which is a partial solution to Joint Optimization (6) whenever $\gamma_1 = 0$ or $\gamma_2 = 0$.

Let $\boldsymbol{X}_U\boldsymbol{X}_U^T = \mathcal{O}_U\mathcal{D}_U\mathcal{O}_U^T$ be the eigendecomposition of this outer product and define

$$\boldsymbol{X}^{(\gamma_2)} \;=\; \begin{pmatrix} \boldsymbol{X}_L \\ \boldsymbol{X}_U^{(\gamma_2)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_L \\ \sqrt{\gamma_2}\left(\mathcal{D}_U + \gamma_2\boldsymbol{I}\right)^{-\frac{1}{2}}\mathcal{O}_U^T\boldsymbol{X}_U \end{pmatrix} \tag{8}$$

for $\gamma_2 > 0$. Proposition 2 establishes that the reduced problem

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma},\boldsymbol{\lambda}} = \underset{\boldsymbol{\beta}}{\arg\min} \|\boldsymbol{Y}_L - \boldsymbol{X}_L\boldsymbol{\beta}\|_2^2 + \gamma_1\left\|\boldsymbol{X}_U^{(\gamma_2)}\boldsymbol{\beta}\right\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1^1 + \lambda_2\|\boldsymbol{\beta}\|_2^2 \tag{9}$$

is an alternative to Joint Optimization (6) over $(\boldsymbol{\alpha},\boldsymbol{\beta}) \in I\!R^p \times I\!R^p$.

**Proposition 2** *If $\gamma_2 > 0$, then rank$(\boldsymbol{X}_U)$ = rank$\left(\boldsymbol{X}_U^{(\gamma_2)}\right)$ and a solution $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma},\boldsymbol{\lambda}}$ to Optimization Problem (9) is a partial solution to Optimization Problem (6).*

By Proposition 2, the semi-supervised estimate $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma},\boldsymbol{\lambda}}$ can be computed by an elastic net subroutine through data augmentation if the user simply inputs the supervised tuning parameters $\boldsymbol{\lambda}$ with model matrix $\left(\boldsymbol{X}_L^T,\ \sqrt{\gamma_1}\boldsymbol{X}_U^{(\gamma_2)^T}\right)^T$ and response vector $\left(\boldsymbol{Y}_L^T,\ \vec{0}^T\right)^T$ (i.e., impute $\boldsymbol{Y}_U = \vec{0}$). The Elastic Net Optimization Problem (7) is convex and can be solved quickly by the `glmnet` package in R (Friedman et al., 2010; R Core Team, 2015), so this helps make our semi-supervised adjustment computationally viable.

Matrix $\boldsymbol{X}_U^{(\gamma_2)^T}\boldsymbol{X}_U^{(\gamma_2)}$ from Optimization Problem (9) has the same eigenvectors as $\boldsymbol{X}_U^T\boldsymbol{X}_U$, but its eigenvalues homogenize to unity as $\gamma_2 \to 0$. As $\gamma_2 \to \infty$, $\boldsymbol{X}_U^{(\gamma_2)^T}\boldsymbol{X}_U^{(\gamma_2)} \to \boldsymbol{X}_U^T\boldsymbol{X}_U$, and Optimization Problem (9) goes to the *semi-supervised extreme*

$$\widehat{\boldsymbol{\beta}}_{(\gamma_1,\infty),\boldsymbol{\lambda}} = \underset{\boldsymbol{\beta}}{\arg\min} \|\boldsymbol{Y}_L - \boldsymbol{X}_L\boldsymbol{\beta}\|_2^2 + \gamma_1\|\boldsymbol{X}_U\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1^1 + \lambda_2\|\boldsymbol{\beta}\|_2^2. \tag{10}$$

Semi-Supervised Extreme (10) with $\boldsymbol{\lambda} = \vec{0}$ and $\gamma_1 = 1$ was seen earlier in Problem (1) during the conceptual overview. Finite $\gamma_2 > 0$ will later be seen to produce intermediate compromises between Supervised (7) and Semi-Supervised Extreme (10).

## 4. Geometry of Semi-Supervised Linear Regression

A geometrical understanding of the Joint Trained Elastic Net (6) is developed through the following logical progression: Section 4.1 joint trained least squares $\boldsymbol{\lambda} = \vec{0}$, Section 4.2 joint trained ridge $\boldsymbol{\lambda} = (0, \lambda_2)$, Section 4.3 joint trained lasso $\boldsymbol{\lambda} = (\lambda_1, 0)$, and then Section 4.4 joint trained elastic net regression $\boldsymbol{\lambda}$. Last, Section 4.5 provides a gallery of geometrical examples. The conceptual overview from Section 2.1 lines-up closely with the mathematics of Section 4.1 and is back-referenced extensively to help the reader make connections. The ridge, lasso, and elastic net semi-supervised geometries do, to some degree, simply follow from their well-known supervised properties when combined with the geometrical properties of joint trained (semi-supervised) least squares. However, an important subtlety is worth mentioning. This geometry section, especially Sections 4.3 and 4.4, establishes properties of the Joint Trained Elastic Net (6), and these properties are stated as the assumptions of Section 5 in order to derive general performance bounds that necessarily apply to the joint trained elastic net.

### 4.1 Joint Trained Least Squares

Optimization Problem (9) with $\boldsymbol{\lambda} = \vec{0}$ reduces to *joint trained least squares*

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{Y}_L - \boldsymbol{X}_L\boldsymbol{\beta}\|_2^2 + \gamma_1 \left\| \boldsymbol{X}_U^{(\gamma_2)}\boldsymbol{\beta} \right\|_2^2. \tag{11}$$

Briefly recall the collinearity example from Section 2.1, i.e., $p = 2$, $\boldsymbol{X}_{L1} = \boldsymbol{X}_{L2}$, $\boldsymbol{X}_{U1} = -\boldsymbol{X}_{U2}$, and $\boldsymbol{\gamma} = (1, \infty)$. A supervised $\widehat{\boldsymbol{\beta}}^{(\text{OLS})}$ was not unique, but the $\widehat{\boldsymbol{\beta}}^{(\text{OLS})}$ with equal components was the unique semi-supervised Estimator (11). In general, Estimator (11) is unique whenever $\boldsymbol{\gamma} > \vec{0}$ and $\text{rank}(\boldsymbol{X}) = p$. Henceforth, assume $\text{rank}(\boldsymbol{X}_L) = p$, so $\widehat{\boldsymbol{\beta}}^{(\text{OLS})} = \left(\boldsymbol{X}_L^T\boldsymbol{X}_L\right)^{-1}\boldsymbol{X}_L^T\boldsymbol{Y}_L$ is unique during this discussion of joint trained least squares. Section 4.2 on joint trained ridge regression is tailored for $\text{rank}(\boldsymbol{X}_L) < p$.

Figure 4(a) displays the semi-supervised extreme $\widehat{\boldsymbol{\beta}}_{\gamma_1,\infty}$ from the block extrapolation example for a particular $\gamma_1 > 0$ based on the calculus of Lagrangian multipliers. For general $p \geq 2$ with $\gamma_2 \geq 0$, there exists unique scalars $a_{\gamma_2}, b_{\gamma_2}$ such that the ellipsoids

$$\boldsymbol{\beta}^T {\boldsymbol{X}_U^{(\gamma_2)}}^T \boldsymbol{X}_U^{(\gamma_2)}\boldsymbol{\beta} \leq a_{\gamma_2} \tag{12}$$

$$\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{(\text{OLS})}\right)^T \boldsymbol{X}_L^T\boldsymbol{X}_L \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{(\text{OLS})}\right) \geq b_{\gamma_2} \tag{13}$$

have the same tangent slope at the point of intersection $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$. A novelty of the semi-supervised approach, that holds for general $p \geq 2$, is the use of origin-centered Ellipsoids (12) as opposed to the multidimensional spheres used in supervised ridge regression.

When $\gamma_2 \approx 0$, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} \approx \widehat{\boldsymbol{\beta}}_{\gamma_1}^{(\text{RIDGE})} = \left(\boldsymbol{X}_L^T\boldsymbol{X}_L + \gamma_1\boldsymbol{I}\right)^{-1}\boldsymbol{X}_L^T\boldsymbol{Y}_L$ because Ellipsoids (12) are roughly spherical. When $\gamma_2$ is large, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ approximates a point on the semi-supervised

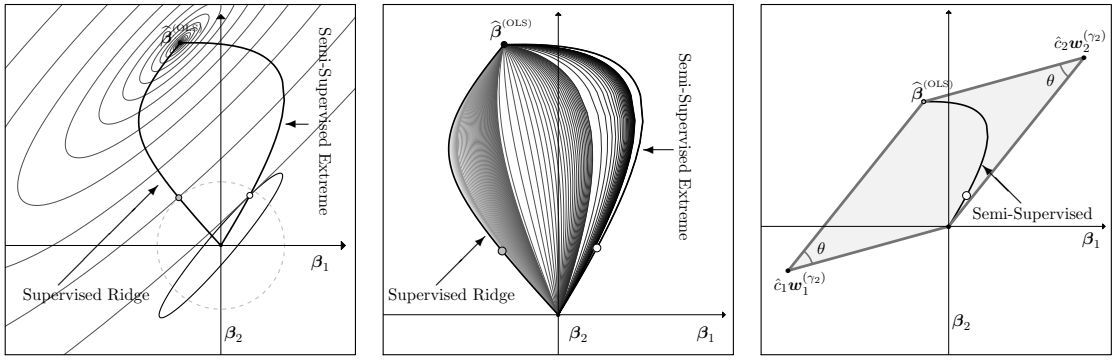


Figure 4: The Figure 1 block example is revisited. (a) A labeled response $\boldsymbol{Y}_L$ that resulted in the plotted estimate $\widehat{\boldsymbol{\beta}}^{(\mathrm{OLS})}$ is part of the assumed labeled data set. Each estimate on the semi-supervised extreme $\widehat{\boldsymbol{\beta}}_{\gamma_1,\infty}$, like the small white circle at $\gamma_1 = 0.18$, is the intersection of an origin-center Ellipse (12) and a $\widehat{\boldsymbol{\beta}}^{(\mathrm{OLS})}$-centered Ellipse (13) having the same tangent slope at this point of intersection. Similarly, each ridge estimate, like the small gray circle with $\lambda_2 = 5.9$, uses origin-centered, concentric circles instead of Ellipses (12). (b) Paths $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ varying $\gamma_1$ with darker curves for larger $\gamma_2$ fill-in all possible compromises between supervised ridge and the semi-supervised extreme. (c) The semi-supervised extreme $\widehat{\boldsymbol{\beta}}_{\gamma_1,\infty}$ is shrunk within its bounding parallelogram from supervised $\widehat{\boldsymbol{\beta}}^{(\mathrm{OLS})}$ toward the origin as $\gamma_1 \to \infty$.

extreme. For example, take the point along the supervised ridge (semi-supervised extreme) path indicated by the small gray (white) circle in Figure 4. Paths $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$, like those in Figure 4(b), start at $\widehat{\boldsymbol{\beta}}^{(\mathrm{OLS})}$ and converge to a point in the null space of $\boldsymbol{X}_U$ as $\gamma_1 \to \infty$.

The semi-supervised estimator for any $\boldsymbol{\gamma}$ is

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} &= \left(\boldsymbol{X}_L^T\boldsymbol{X}_L + \gamma_1 \boldsymbol{X}_U^{(\gamma_2)^T}\boldsymbol{X}_U^{(\gamma_2)}\right)^{-1}\boldsymbol{X}_L^T\boldsymbol{X}_L\widehat{\boldsymbol{\beta}}^{(\mathrm{OLS})} \\ &= \left(\boldsymbol{I} + \gamma_1 \boldsymbol{M}^{(\gamma_2)}\right)^{-1}\widehat{\boldsymbol{\beta}}^{(\mathrm{OLS})}, \text{ where } \boldsymbol{M}^{(\gamma_2)} = \left(\boldsymbol{X}_L^T\boldsymbol{X}_L\right)^{-1}\boldsymbol{X}_U^{(\gamma_2)^T}\boldsymbol{X}_U^{(\gamma_2)}. \end{aligned} \tag{14}$$

An eigenbasis $\left\{\left(\boldsymbol{w}_i^{(\gamma_2)}, \tau_i^{(\gamma_2)}\right)\right\}_{i=1}^p$ of $\boldsymbol{M}^{(\gamma_2)}$ such that $\left\|\boldsymbol{X}_L\boldsymbol{w}_i^{(\gamma_2)}\right\|_2^2 = 1$ will be used to help understand how joint trained least squares regression coefficients are shrunk. Proposition 3 establishes that this important eigenbasis is real whether or not matrix $\boldsymbol{M}^{(\gamma_2)}$ is symmetric.

**Proposition 3** *Any eigenbasis of the possibly non-symmetric matrix $\boldsymbol{M}^{(\gamma_2)}$ is real with eigenvalues $\tau_1^{(\gamma_2)} \geq \cdots \geq \tau_p^{(\gamma_2)} \geq 0$. Furthermore, $\tau_i^{(\gamma_2)} = 0$ iff $i > \mathrm{rank}(\boldsymbol{X}_U)$.*

While $\left\{\boldsymbol{w}_i^{(\gamma_2)}\right\}_{i=1}^p$ may be neither orthogonal nor unit length,

$$\widehat{\boldsymbol{\beta}}^{(\mathrm{OLS})} = \hat{c}_1^{(\gamma_2)}\boldsymbol{w}_1^{(\gamma_2)} + \cdots + \hat{c}_p^{(\gamma_2)}\boldsymbol{w}_p^{(\gamma_2)} \tag{15}$$

for some scalars $\hat{c}_i^{(\gamma_2)}$, and by Equations (14) and (15),

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = \left(\frac{1}{1+\gamma_1\tau_1^{(\gamma_2)}}\right)\hat{c}_1^{(\gamma_2)}\boldsymbol{w}_1^{(\gamma_2)} + \cdots + \left(\frac{1}{1+\gamma_1\tau_p^{(\gamma_2)}}\right)\hat{c}_p^{(\gamma_2)}\boldsymbol{w}_p^{(\gamma_2)}. \tag{16}$$

Equations (15) and (16) generalize Estimator (4) from Section 2.1 to $p \geq 2$. The terms on the right of Equation (15) were previously denoted by the $\tilde{\boldsymbol{\nu}}_i$ from Display (3), and these terms are weighted by proportions on the right of Equation (16) that were previously denoted by the $p_i$ from Display (4). Eigenvector $\hat{c}_1^{(\gamma_2)}\boldsymbol{w}_1^{(\gamma_2)}$ is proportionally shrunk the most at any fixed $\gamma_1 > 0$ because its proportion weight $1/\left(1+\gamma_1\tau_1^{(\gamma_2)}\right)$ is the smallest.

The bounding parallelogram in Figure 4(c) helps introduce another interpretation of Equation (16). This parallelogram has opposite corners at the origin and $\widehat{\boldsymbol{\beta}}^{(\text{OLS})}$ and sides parallel to the eigenvectors of $\boldsymbol{M}^{(\gamma_2)}$. The path $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ shrinks from $\widehat{\boldsymbol{\beta}}^{(\text{OLS})}$ to the origin along the sides with corner $\hat{c}_2^{(\gamma_2)}\boldsymbol{w}_2^{(\gamma_2)}$ as $\gamma_1 \in [0,\infty]$ increases and does so more closely when $\tau_1^{(\gamma_2)}$ and $\tau_2^{(\gamma_2)}$ differ in magnitude. Proposition 4 generalizes this concept to arbitrary $\gamma_2 \geq 0$ and $p \geq 2$.

**Proposition 4** *The path $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ as a function of $\gamma_1 \geq 0$ is bounded within a $p$-dimensional parallelotope with corners at each binary linear combination of $\left\{\hat{c}_1^{(\gamma_2)}\boldsymbol{w}_1^{(\gamma_2)}, \ldots, \hat{c}_p^{(\gamma_2)}\boldsymbol{w}_p^{(\gamma_2)}\right\}$. Furthermore, the terminal point as $\gamma_1 \to \infty$ is the corner $\sum_{i=1}^{p}\mathcal{I}_{\{i>rank(\boldsymbol{X}_U)\}}\hat{c}_i^{(\gamma_2)}\boldsymbol{w}_i^{(\gamma_2)}$ with indicator $\mathcal{I}_{\{\cdot\}}$.*

The conceptual overview in Section 2.1 made a careful distinction between shrinking regression coefficients $\widehat{\boldsymbol{\beta}}$ versus shrinking linear predictions $\boldsymbol{x}_0^T\widehat{\boldsymbol{\beta}}$. Vectors $\tilde{\boldsymbol{\nu}}_i$ from Display (3) were related to coefficient shrinking, whereas $\boldsymbol{\nu}_i$ from Display (5) were the feature vectors $\boldsymbol{x}_0$ related to prediction shrinking. Mathematically, eigenvectors $\boldsymbol{w}_i^{(\gamma_2)}$ determine directions of coefficient shrinking. Since $p = 2$, the Section 2.1 discussion in-fact concentrated on all feature vectors $\boldsymbol{w}_1^{(\gamma_2)^\perp}$ and $\boldsymbol{w}_2^{(\gamma_2)^\perp}$, and an eigenvector direction of maximum (minimum) coefficient shrinking was orthogonal to feature vectors of maximum (minimum) prediction shrinking. Generalizing this story to $p > 2$ also results in $p$ directions of coefficient shrinking and $p$ feature vector directions of interpretable prediction shrinking, but the mathematics has the following subtlety. When $p > 2$, a direction of coefficient shrinking $\boldsymbol{w}_i^{(\gamma_2)}$ is orthogonal to a $p-1$ dimensional vector space $\boldsymbol{w}_i^{(\gamma_2)^\perp}$ of feature vectors, so if $p-1 \geq 2$, vector space $\boldsymbol{w}_1^{(\gamma_2)^\perp}$ consists of an infinite number of directions. Proposition 5 below provides a convenient form for the line in common to all $\boldsymbol{w}_j^{(\gamma_2)^\perp}$ with $j \neq i$ for each $i \in \{1,\ldots,p\}$ by establishing a relationship between $\boldsymbol{w}_i^{(\gamma_2)}$, $\boldsymbol{w}_i^{(\gamma_2)^\perp}$, and $\boldsymbol{X}^{(\gamma_2)}$ from Equation (8). These $p$ lines of feature data vectors for arbitrary $p \geq 2$ will later be seen to have a clear interpretation when it comes to prediction shrinking, so we call them *extrapolation directions*.

**Proposition 5** *The* $\text{span}\left(\boldsymbol{X}^{(\gamma_2)^T}\boldsymbol{X}^{(\gamma_2)}\boldsymbol{w}_i^{(\gamma_2)}\right) = \bigcap_{j\in\{1,\ldots,p\}-\{i\}}\boldsymbol{w}_j^{(\gamma_2)^\perp}$ $\forall i \in \{1,\ldots,p\}$. *Henceforth, the line* $\text{span}\left(\boldsymbol{X}^{(\gamma_2)^T}\boldsymbol{X}^{(\gamma_2)}\boldsymbol{w}_i^{(\gamma_2)}\right)$ *is called the $i^{th}$ extrapolation direction $\forall i \in \{1,\ldots,p\}$.*

The $i^{\text{th}}$ extrapolation direction necessarily traces out a line because it's all scalar multiples of the nonzero vector $\boldsymbol{X}^{(\gamma_2)^T}\boldsymbol{X}^{(\gamma_2)}\boldsymbol{w}_i^{(\gamma_2)}$. Any feature vector on the $i^{th}$ *extrapolation direction*, i.e., $\boldsymbol{x}_0 \in \bigcap_{j\in\{1,\ldots,p\}-\{i\}} \boldsymbol{w}_j^{(\gamma_2)^\perp}$ from Proposition 5, is of special note. Their Equation (16) semi-supervised predictions simplify to $\boldsymbol{x}_0^T\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = \hat{c}_i^{(\gamma_2)}/\left(1+\gamma_1\tau_i^{(\gamma_2)}\right)\boldsymbol{x}_0^T\boldsymbol{w}_i^{(\gamma_2)}$ and are shrunk more (relative to the corresponding OLS supervised prediction $\boldsymbol{x}_0^T\widehat{\boldsymbol{\beta}}^{(\text{OLS})} = \hat{c}_i^{(\gamma_2)}\boldsymbol{x}_0^T\boldsymbol{w}_i^{(\gamma_2)}$) for smaller $i \in \{1,\ldots,p\}$ at any fixed $\gamma_1 > 0$ because $\tau_1^{(\gamma_2)} \geq \cdots \geq \tau_p^{(\gamma_2)}$.

Next, the $i^{\text{th}}$ extrapolation direction is shown to be one of more (or less) extreme unlabeled extrapolations. With this in mind, use the indicator function $\mathcal{I}_{\{\cdot\}}$ to define the positive number $\kappa_i^{(\gamma_2)} = \tau_i^{(\gamma_2)} + \mathcal{I}_{\{i>\text{rank}(\boldsymbol{X}_U)\}}$ and define the vectors

$$\boldsymbol{\ell}_i^{(\gamma_2)} = \boldsymbol{X}_L\boldsymbol{w}_i^{(\gamma_2)} \text{ and } \boldsymbol{u}_i^{(\gamma_2)} = \frac{\boldsymbol{X}_U^{(\gamma_2)}\boldsymbol{w}_i^{(\gamma_2)}}{\sqrt{\kappa_i^{(\gamma_2)}}}. \tag{17}$$

Vectors (17) in the semi-supervised extreme of $\gamma_2 = \infty$ were temporarily denoted by $\boldsymbol{\ell}_i$ and $\boldsymbol{u}_i$ during their more conceptual introduction within Section 2.1 (e.g., Figure 2). It was also stated previously during this overview that $\boldsymbol{\ell}_i$ and $\boldsymbol{u}_i$ were unit length. Proposition 6 is a generalization.

**Proposition 6** *If $\gamma_2 > 0$, vectors $\left\{\boldsymbol{\ell}_i^{(\gamma_2)}\right\}_{i=p}^{1}$ and $\left\{\boldsymbol{u}_i^{(\gamma_2)}\right\}_{i=1}^{rank(\boldsymbol{X}_U)}$ are orthonormal bases for the column spaces of $\boldsymbol{X}_L$ and $\boldsymbol{X}_U^{(\gamma_2)}$, and $\boldsymbol{u}_i^{(\gamma_2)} = \vec{0}$ if $i > rank(\boldsymbol{X}_U)$.*

Section 2.1 also introduced extents of $L$- and $U$-extrapolation. Vectors (17) are used to define these now for each $i \in \{1,\ldots,p\}$ as

$$\begin{aligned} \boldsymbol{X}_L^T\boldsymbol{\ell}_i^{(\gamma_2)} &\text{ Extent of } L\text{-Extrapolation (in the } i^{\text{th}} \text{ Direction)} \\ \boldsymbol{X}_U^{(\gamma_2)^T}\boldsymbol{u}_i^{(\gamma_2)} &\text{ Extent of } U\text{-Extrapolation (in the } i^{\text{th}} \text{ Direction), where} \\ \text{span}\left(\boldsymbol{X}^{(\gamma_2)^T}\boldsymbol{X}^{(\gamma_2)}\boldsymbol{w}_i^{(\gamma_2)}\right) &\text{ is the } i^{\text{th}} \text{ Direction of Extrapolation from Proposition 5.} \end{aligned} \tag{18}$$

Propositions 7 establishes that the $i^{\text{th}}$ extent vectors are in-fact on the $i^{\text{th}}$ extrapolation direction.

**Proposition 7** *For each $i \in \{1,\ldots,p\}$,*

$$\begin{aligned} \boldsymbol{X}_L^T\boldsymbol{\ell}_i^{(\gamma_2)} &= \frac{1}{1+\tau_i^{(\gamma_2)}}\boldsymbol{X}^{(\gamma_2)^T}\boldsymbol{X}^{(\gamma_2)}\boldsymbol{w}_i^{(\gamma_2)} \\ \boldsymbol{X}_U^{(\gamma_2)^T}\boldsymbol{u}_i^{(\gamma_2)} &= \frac{\tau_i^{(\gamma_2)}}{(1+\tau_i^{(\gamma_2)})\sqrt{\kappa_i^{(\gamma_2)}}}\boldsymbol{X}^{(\gamma_2)^T}\boldsymbol{X}^{(\gamma_2)}\boldsymbol{w}_i^{(\gamma_2)}, \end{aligned}$$

*so $\boldsymbol{X}^{(\gamma_2)^T}\boldsymbol{X}^{(\gamma_2)}\boldsymbol{w}_i^{(\gamma_2)}$, $\boldsymbol{X}_L^T\boldsymbol{\ell}_i^{(\gamma_2)}$, and $\boldsymbol{X}_U^{(\gamma_2)^T}\boldsymbol{u}_i^{(\gamma_2)}$ are parallel vectors in $\mathbb{R}^p$.*

Previously defined vectors are now verified to possess fundamental interpretations: (i) Extent Vectors (18) do indeed measure "extrapolation extents" in a sensible manner, (ii) Vectors (17) determine shrinking directions for joint trained least squares fits $\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$, and (iii) magnitudes of extent vectors regulate the shrinking of regression coefficients $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$. These three interpretations are gleaned by applying Propositions 6 and 7 in conjunction with well-known properties of orthogonal projection matrices and quadratic forms from linear algebra. The $n \times p$ matrix identity

$$\boldsymbol{X}^{(\gamma_2)} \begin{pmatrix} \boldsymbol{w}_1^{(\gamma_2)} & \cdots & \boldsymbol{w}_p^{(\gamma_2)} \end{pmatrix} = \left( \begin{pmatrix} \boldsymbol{\ell}_1^{(\gamma_2)} \\ \sqrt{\kappa_1^{(\gamma_2)}}\boldsymbol{u}_1^{(\gamma_2)} \end{pmatrix} \cdots \begin{pmatrix} \boldsymbol{\ell}_p^{(\gamma_2)} \\ \sqrt{\kappa_p^{(\gamma_2)}}\boldsymbol{u}_p^{(\gamma_2)} \end{pmatrix} \right) \tag{19}$$

follows from Definitions (17). The right of Equation (19) has orthogonal columns by Proposition 6, and the columns on the left of Equation (19) are eigenvectors with eigenvalue one of the orthogonal projection matrix $\boldsymbol{X}^{(\gamma_2)} \left( \boldsymbol{X}^{(\gamma_2)^T} \boldsymbol{X}^{(\gamma_2)} \right)^{-1} \boldsymbol{X}^{(\gamma_2)^T}$. Therefore, the columns of Matrix (19) are an orthogonal basis for the eigenspace of $\boldsymbol{X}^{(\gamma_2)} \left( \boldsymbol{X}^{(\gamma_2)^T} \boldsymbol{X}^{(\gamma_2)} \right)^{-1} \boldsymbol{X}^{(\gamma_2)^T}$ corresponding to eigenvalue one, because $\text{rank}\left( \boldsymbol{X}^{(\gamma_2)} \right) = p$ is a necessary condition for the joint trained least squares assumption that $\text{rank}(\boldsymbol{X}_L) = p$.

Projection matrix $\boldsymbol{X}^{(\gamma_2)} \left( \boldsymbol{X}^{(\gamma_2)^T} \boldsymbol{X}^{(\gamma_2)} \right)^{-1} \boldsymbol{X}^{(\gamma_2)^T}$ is nonnegative definite, so its main diagonal block sub matrices based on the $L$, $U$ data partition are also nonnegative definite. The nonnegative definite, rank-$p$, sub matrix $\boldsymbol{X}_L \left( \boldsymbol{X}^{(\gamma_2)^T} \boldsymbol{X}^{(\gamma_2)} \right)^{-1} \boldsymbol{X}_L^T$ has orthonormal eigenvectors $\left\{ \boldsymbol{\ell}_i^{(\gamma_2)} \right\}_{i=p}^{1}$ corresponding to its nonzero eigenvalues $1/(1+\tau_i^{(\gamma_2)})$ by Propositions 6 and 7. Similarly, nonnegative definite sub matrix $\boldsymbol{X}_U^{(\gamma_2)} \left( \boldsymbol{X}^{(\gamma_2)^T} \boldsymbol{X}^{(\gamma_2)} \right)^{-1} \boldsymbol{X}_U^{(\gamma_2)^T}$ has orthonormal eigenvectors $\left\{ \boldsymbol{u}_i^{(\gamma_2)} \right\}_{i=1}^{\text{rank}(\boldsymbol{X}_U)}$ corresponding to its nonzero eigenvalues $\tau_i^{(\gamma_2)}/(1+\tau_i^{(\gamma_2)})$. Well-known eigenvector solutions to constrained optimizations of quadratic forms imply

$$\begin{aligned} \boldsymbol{\ell}_i^{(\gamma_2)} &= \underset{\boldsymbol{v} \in I\!R^{|L|}: \boldsymbol{v}^T\boldsymbol{v}=1, \boldsymbol{v}^T\boldsymbol{\ell}_j^{(\gamma_2)}=0 \ \forall j>i}{\arg\max} \boldsymbol{v}^T \boldsymbol{X}_L \left( \boldsymbol{X}^{(\gamma_2)^T} \boldsymbol{X}^{(\gamma_2)} \right)^{-1} \boldsymbol{X}_L^T \boldsymbol{v} \\ \boldsymbol{u}_i^{(\gamma_2)} &= \underset{\boldsymbol{v} \in I\!R^{|U|}: \boldsymbol{v}^T\boldsymbol{v}=1, \boldsymbol{v}^T\boldsymbol{u}_j^{(\gamma_2)}=0 \ \forall j<i}{\arg\max} \boldsymbol{v}^T \boldsymbol{X}_U^{(\gamma_2)} \left( \boldsymbol{X}^{(\gamma_2)^T} \boldsymbol{X}^{(\gamma_2)} \right)^{-1} \boldsymbol{X}_U^{(\gamma_2)^T} \boldsymbol{v}. \end{aligned}$$

In other words, the unit length weight vectors on the rows of $\boldsymbol{X}_L$ (of $\boldsymbol{X}_U^{(\gamma_2)}$) that maximize a Mahalanobis distance measuring extent of extrapolation subject to orthogonality constraints are the eigenvectors $\left\{ \boldsymbol{\ell}_i^{(\gamma_2)} \right\}_{i=p}^{1}$ (eigenvectors $\left\{ \boldsymbol{u}_i^{(\gamma_2)} \right\}_{i=1}^{\text{rank}(\boldsymbol{X}_U)}$) sorted by descending positive eigenvalues. Proposition 7 also establishes that each eigenvalue

$$\tau_i^{(\gamma_2)} = \frac{\left\| \boldsymbol{X}_U^{(\gamma_2)^T} \boldsymbol{u}_i^{(\gamma_2)} \right\|_2}{\left\| \boldsymbol{X}_L^T \boldsymbol{\ell}_i^{(\gamma_2)} \right\|_2} \tag{20}$$

of the shrinking matrix $\boldsymbol{M}^{(\gamma_2)}$ from Display (14) is a ratio of parallel extent eigenvector lengths, so the extent of $U$-extrapolation is larger (smaller) than the corresponding $L$-extent in the $i^{\text{th}}$ direction of extrapolation if $\tau_i^{(\gamma_2)} > 1$ (if $\tau_i^{(\gamma_2)} < 1$).

The joint trained least squares fits vector for all $n$ observations has the form

$$\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = \sum_{i=1}^{p} \hat{c}_i^{(\gamma_2)} \left( \frac{1}{1+\gamma_1\tau_i^{(\gamma_2)}} \right) \begin{pmatrix} \boldsymbol{\ell}_i^{(\gamma_2)} \\ \sqrt{\kappa_i^{(\gamma_2)}/\gamma_2}\mathcal{O}_U \left(\mathcal{D}_U + \gamma_2 \boldsymbol{I}\right)^{\frac{1}{2}} \boldsymbol{u}_i^{(\gamma_2)} \end{pmatrix}$$

by Equations (16) and (19) and the reverse of Transformation (8). Thus, eigenvectors $\boldsymbol{\ell}_i^{(\gamma_2)}$ and $\boldsymbol{u}_i^{(\gamma_2)}$ involved in constructing the $i^{\text{th}}$ extrapolation direction with smaller $i \in \{1, \ldots, p\}$ are used to shrink fits more as $\gamma_1$ is increased. By Equation (16) and Ratios (20), coefficient vector

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = \sum_{i=1}^{p} \left( \frac{\left\| \boldsymbol{X}_L^T \boldsymbol{\ell}_i^{(\gamma_2)} \right\|_2}{\left\| \boldsymbol{X}_L^T \boldsymbol{\ell}_i^{(\gamma_2)} \right\|_2 + \gamma_1 \left\| {\boldsymbol{X}_U^{(\gamma_2)}}^T \boldsymbol{u}_i^{(\gamma_2)} \right\|_2} \right) \hat{c}_i^{(\gamma_2)} \boldsymbol{w}_i^{(\gamma_2)}$$

is a generalization of Display (4) and balances the degree of coefficient shrinkage by the relative extents of $U$- versus $L$-extrapolations in the $i^{\text{th}}$ direction as tuning parameter $\gamma_1$ is increased.

The Figure 2 block extrapolation example is now revisited with the notation of Display (18) and other mathematical developments from this section in mind. Extrapolation directions can always be computed with Proposition 5. When $p = 2$, the $1^{\text{st}}$ extrapolation direction is comprised of all vectors orthogonal to $\boldsymbol{w}_2^{(\gamma_2)}$, and the $2^{\text{nd}}$ extrapolation direction is comprised of all vectors orthogonal to $\boldsymbol{w}_1^{(\gamma_2)}$. Directions and extents in Figure 2 were all based on the semi-supervised extreme setting $\gamma_2 = \infty$. In this example, the extent of $U$-extrapolation is a larger multiple of the $L$-extent in the $1^{\text{st}}$ direction, so $\tau_1^{(\gamma_2)} > \tau_2^{(\gamma_2)}$ is a strict inequality. In addition, $U$-extents have the larger magnitude, so $\tau_2^{(\gamma_2)} > 1$ is another artifact of this particular example. An example of $p > 2$ is deferred until discussion of Figure 6 in the examples Section 4.5.

### 4.2 Joint Trained Ridge Regression

Estimator (9) with $\boldsymbol{\lambda} = (0, \lambda_2)$ is motivated with augmented labeled data

$$\boldsymbol{X}_L^{(\lambda_2)} = \begin{pmatrix} \boldsymbol{X}_L \\ \sqrt{\lambda_2}\boldsymbol{I} \end{pmatrix} \text{ and } \boldsymbol{Y}_L^{\star} = \begin{pmatrix} \boldsymbol{Y}_L \\ \vec{0} \end{pmatrix} \tag{21}$$

having $p$ additional rows. The resulting *joint trained ridge estimator*

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma},(0,\lambda_2)} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{Y}_L - \boldsymbol{X}_L\boldsymbol{\beta}\|_2^2 + \gamma_1 \left\| \boldsymbol{X}_U^{(\gamma_2)}\boldsymbol{\beta} \right\|_2^2 + \lambda_2 \|\boldsymbol{\beta}\|_2^2$$

is equivalent to Joint Trained Least Squares (11) given Data (21). Hence,

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma},(0,\lambda_2)} = \left( {\boldsymbol{X}_L^{(\lambda_2)}}^T \boldsymbol{X}_L^{(\lambda_2)} + \gamma_1 {\boldsymbol{X}_U^{(\gamma_2)}}^T \boldsymbol{X}_U^{(\gamma_2)} \right)^{-1} \left( {\boldsymbol{X}_L^{(\lambda_2)}}^T \boldsymbol{X}_L^{(\lambda_2)} \right) \widehat{\boldsymbol{\beta}}_{\lambda_2}^{(\text{RIDGE})}, \tag{22}$$

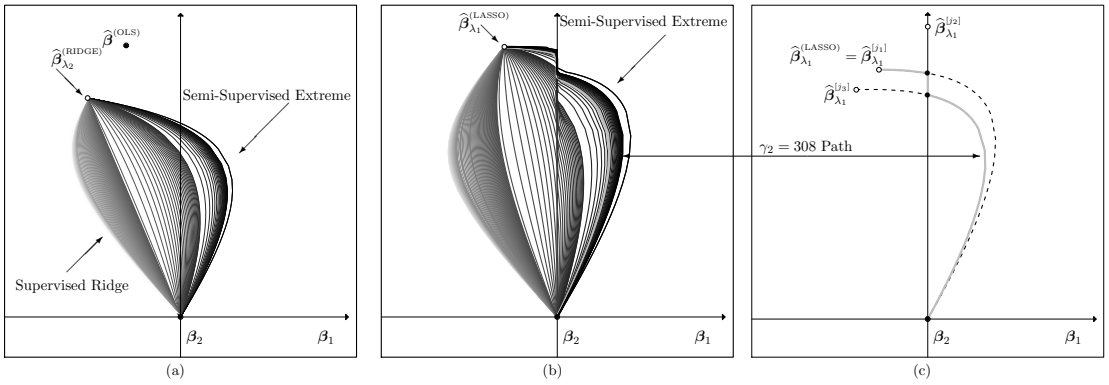


Figure 5: Paths of candidate $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma},\boldsymbol{\lambda}}$ for the Figure 1 block example varying $\gamma_1 > 0$ with darker curves for larger $\gamma_2 > 0$ are compared. (a) Joint trained ridge paths at a fixed $\boldsymbol{\lambda} = (0, 0.1)$ start at supervised ridge $\widehat{\boldsymbol{\beta}}^{(\text{RIDGE})}_{\lambda_2}$ instead of supervised OLS $\widehat{\boldsymbol{\beta}}^{(\text{OLS})}$. (b) Similarly, joint trained lasso paths at a fixed $\boldsymbol{\lambda} = (0.01, 0)$ start at supervised lasso $\widehat{\boldsymbol{\beta}}^{(\text{LASSO})}_{\lambda_1}$. However, these continuous paths are not differentiable at points where the active set changes. (c) The path from (b) with $\gamma_2 = 308$ is highlighted. Active set changes are marked by bullets $\bullet$, and the reference curves based on the right of Equation (23) are also displayed as dashed lines for $i = 1, 2, 3$. Each reference curve starts at a $\widehat{\boldsymbol{\beta}}^{[j_i]}_{\lambda_1}$ (marked by an open circle $\circ$) and terminates at the origin. The actual candidate path always equals one of the displayed reference curves. It starts at $\widehat{\boldsymbol{\beta}}^{(\text{LASSO})}_{\lambda_1} = \widehat{\boldsymbol{\beta}}^{[j_1]}_{\lambda_1}$ when $\gamma_1 = 0$ and switches reference curves whenever there is a change in the active set.

because $\widehat{\boldsymbol{\beta}}^{(\text{RIDGE})}_{\lambda_2} = \left( {\boldsymbol{X}_L^{(\lambda_2)}}^T \boldsymbol{X}_L^{(\lambda_2)} \right)^{-1} \boldsymbol{X}_L^T \boldsymbol{Y}_L$ is the OLS estimator given Data (21). Matrix ${\boldsymbol{X}_L^{(\lambda_2)}}^T \boldsymbol{X}_L^{(\lambda_2)} = \boldsymbol{X}_L^T \boldsymbol{X}_L + \lambda_2 \boldsymbol{I}$ with $\lambda_2 > 0$ is positive definite, so the inverse required to compute $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma},(0,\lambda_2)}$ exists. Estimates (22) for the block extrapolation example come out as expected in Figure 5(a). Paths start at $\widehat{\boldsymbol{\beta}}^{(\text{RIDGE})}_{\lambda_2}$ with $\lambda_2 = 0.1$ and converge to the origin.

### 4.3 Joint Trained Lasso Regression

Supervised Optimization (7) with $\lambda_2 = 0$ simplifies to $\widehat{\boldsymbol{\beta}}^{(\text{LASSO})}_{\lambda_1} = \widehat{\boldsymbol{\beta}}^{(\text{ENET})}_{\lambda_1,0}$, a well-understood technique for incorporating variable selection when $p$ is large and the columns of $\boldsymbol{X}_L$ are linearly independent (Friedman et al., 2010). The goal in this section is to use what is already known about $\widehat{\boldsymbol{\beta}}^{(\text{LASSO})}_{\lambda_1}$ to provide an understanding of the *joint trained lasso* $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma},(\lambda_1,0)}$ from Problem (9). Denote the *active set* of some estimate $\widehat{\boldsymbol{\beta}}$ by $\mathcal{A} \subset \{1, \ldots, p\}$, so $\left(\widehat{\boldsymbol{\beta}}\right)_{\mathcal{A}}$ is its $|\mathcal{A}| \times 1$ vector of nonzero components and $\left(\widehat{\boldsymbol{\beta}}\right)_{\bar{\mathcal{A}}} = \vec{0}$ is $(p - |\mathcal{A}|) \times 1$. Also denote its *sign vector* by $\boldsymbol{s} = \text{sign}\left(\left(\widehat{\boldsymbol{\beta}}\right)_{\mathcal{A}}\right)$ and the $|L| \times |\mathcal{A}|$ sub matrix of $\boldsymbol{X}$ with labeled rows and

active set columns by $\boldsymbol{X}_{L\mathcal{A}}$. The active set $\mathcal{A}^{(\text{SUP})}$ and sign vector $\boldsymbol{s}^{(\text{SUP})}$ of the supervised lasso at a given $\lambda_1$ satisfy the constraint

$$\boldsymbol{X}^T_{L\mathcal{A}^{(\text{SUP})}}\boldsymbol{X}_{L\mathcal{A}^{(\text{SUP})}}\left(\widehat{\boldsymbol{\beta}}^{(\text{LASSO})}_{\lambda_1}\right)_{\mathcal{A}^{(\text{SUP})}} = \boldsymbol{X}^T_{L\mathcal{A}^{(\text{SUP})}}\boldsymbol{Y}_L - \lambda_1\boldsymbol{s}^{(\text{SUP})}.$$

Estimates $\widehat{\boldsymbol{\beta}}^{(\text{LASSO})}_{\lambda_1}$ are a differentiable function in $\lambda_1$ with a finite number of exceptions. This function is continuous, but not differentiable when the active set changes.

The joint trained lasso $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma},(\lambda_1,0)}$ has properties similar to the supervised lasso by Optimization (9), because it's a lasso estimator with unlabeled imputations $\boldsymbol{Y}_U = \vec{0}$ and modified $\boldsymbol{X}$. Unlike joint trained ridge and joint trained least squares from Sections 4.1 and 4.2, the joint trained lasso coefficients are not always a linear combination of the supervised lasso, and this complicates its ensuing interpretation. There are $2^p + 2p + 1$ active-set/sign-vector combinations for any $p \geq 2$. For example, when $p = 2$, there are nine combinations, i.e., $2^2 = 4$ quadrants, $2 \times 2 = 4$ axial directions, and 1 origin. Each active-set/sign-vector combination has a set of reference coefficients $\left(\widehat{\boldsymbol{\beta}}^{[j]}_{\lambda_1}\right)_{\mathcal{A}_j} = \left(\boldsymbol{X}^T_{L\mathcal{A}_j}\boldsymbol{X}_{L\mathcal{A}_j}\right)^{-1}\left(\boldsymbol{X}^T_{L\mathcal{A}_j}\boldsymbol{Y}_L - \lambda_1\boldsymbol{s}_j\right)$ and $\left(\widehat{\boldsymbol{\beta}}^{[j]}_{\lambda_1}\right)_{\bar{\mathcal{A}}_j} = \vec{0}$ for $j = 1, \ldots, 2^p + 2p + 1$. These reference coefficients have important properties. First, $\widehat{\boldsymbol{\beta}}^{[j]}_{\lambda_1}$ are independent of $\boldsymbol{X}_U$. Second, there exists a $j \in \{1, \ldots, 2^p + 2p + 1\}$ such that $\widehat{\boldsymbol{\beta}}^{(\text{LASSO})}_{\lambda_1} = \widehat{\boldsymbol{\beta}}^{[j]}_{\lambda_1}$. Third, $\text{sign}\left(\left(\widehat{\boldsymbol{\beta}}^{[j]}_{\lambda_1}\right)_{\mathcal{A}_j}\right)$ does not necessarily equal $\boldsymbol{s}_j$. Next, the path of the joint trained lasso as a function of $\gamma_1$ at a given $\gamma_2$ is studied. Let the finite set $\{a_i\}_{i=1}^k$ be the finite values of $\gamma_1$ where the active set of the joint trained lasso changes and define $a_0 = 0$ and $a_{k+1} = \infty$. Also define the subsequence $j_1, \ldots, j_k$ such that $\mathcal{A}_{j_i}$ and $\boldsymbol{s}_{j_i}$ correspond to the joint trained lasso for any $\gamma_1 \in [a_{i-1}, a_i)$, so this subsequence tracks the evolution of the joint trained lasso's active set and sign vector. Thus, for all $\gamma_1 \in [a_{i-1}, a_i)$,

$$\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma},(\lambda_1,0)}\right)_{\mathcal{A}_{j_i}} = \left(\boldsymbol{X}^T_{L\mathcal{A}_{j_i}}\boldsymbol{X}_{L\mathcal{A}_{j_i}} + \gamma_1\boldsymbol{X}^{(\gamma_2)^T}_{U\mathcal{A}_{j_i}}\boldsymbol{X}^{(\gamma_2)}_{U\mathcal{A}_{j_i}}\right)^{-1}\boldsymbol{X}^T_{L\mathcal{A}_{j_i}}\boldsymbol{X}_{L\mathcal{A}_{j_i}}\left(\widehat{\boldsymbol{\beta}}^{[j_i]}_{\lambda_1}\right)_{\mathcal{A}_{j_i}}, \tag{23}$$

and shrinking of regression coefficients on the active set looks very much like Display (14).

| $i$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $\mathcal{A}_{j_i}$ | $\{1,2\}$ | $\{2\}$ | $\{1,2\}$ | $\emptyset$ |
| $\boldsymbol{s}^T_{j_i}$ | $(-1,1)$ | $(0,1)$ | $(1,1)$ | - |
| $\gamma_1$ | $[0, 0.004)$ | $[0.004, 0.008)$ | $[0.008, \infty)$ | $\infty$ |

Table 1: Block extrapolation active-set, sign-vector combinations are listed as a function of $\gamma_1$ for the joint trained lasso coefficients $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma},\boldsymbol{\lambda}}$ from Figure 5(c) with $\boldsymbol{\lambda} = (0.01, 0)$ and $\gamma_2 = 308$.

Figure 5(b) plots paths of vectors $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma},(\lambda_1,0)}$ by $\gamma_2$ as a function of $\gamma_1$ at $\lambda_1 = 0.01$ for the block extrapolation example. The semi-supervised path starts at the supervised estimate $\widehat{\boldsymbol{\beta}}^{(\text{LASSO})}_{\lambda_1}$ when $\gamma_1 = 0$. Equation (23) establishes a local property of the joint trained lasso. The approach has the same active set and sign vector as the supervised coefficient for a

small region $\gamma_1 \in [0, a_1)$, where $a_1 > 0$. This local property of the joint trained lasso, which was mathematically verified in this section, is stated as a key assumption while deriving the general performance bounds in Section 5. An example is the highlighted path with $\gamma_2 = 308$ from Figure 5(b) shown in Figure 5(c). This candidate path of semi-supervised regression coefficients visits four active-set, sign-vector combinations as a continuous function of $\gamma_1$ at given $\lambda_1$ and $\gamma_2$. These visited combinations are listed in Table 1 along with their corresponding values $\gamma_1$. Figure 5(c) also includes dashed reference curves based on the right of Equation (23) as a function of $\gamma_1$ for each non-empty active-set/sign-vector combination visited by the approach, i.e., $i = 1, 2, 3$. The candidate semi-supervised estimates follow along a reference path until the active set changes, and then the path switches to the reference path with the new active set and sign vector. This continues until the path terminates at the origin when $\gamma_1 = \infty$.

### 4.4 Joint Trained Elastic Net Regression

A general view of Problem (9) when all four tuning parameters are finite and positive comes from stringing concepts from Sections 4.2 and 4.3 together. In particular,

$$\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma},\boldsymbol{\lambda}}\right)_{\mathcal{A}_{j_i}} = \left(\boldsymbol{X}_{L\mathcal{A}_{j_i}}^{(\lambda_2)\ T}\boldsymbol{X}_{L\mathcal{A}_{j_i}}^{(\lambda_2)} + \gamma_1 \boldsymbol{X}_{U\mathcal{A}_{j_i}}^{(\gamma_2)\ T}\boldsymbol{X}_{U\mathcal{A}_{j_i}}^{(\gamma_2)}\right)^{-1}\left(\boldsymbol{X}_{L\mathcal{A}_{j_i}}^{(\lambda_2)\ T}\boldsymbol{X}_{L\mathcal{A}_{j_i}}^{(\lambda_2)}\right)\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}^{[j_i]}\right)_{\mathcal{A}_{j_i}},$$

where $\mathcal{A}_{j_i}$ and $\boldsymbol{s}_{j_i}$ depend on $(\boldsymbol{\gamma}, \boldsymbol{\lambda})$ and

$$\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}^{[j_i]}\right)_{\mathcal{A}_{j_i}} = (1+\lambda_2)\left(\boldsymbol{X}_{L\mathcal{A}_{j_i}}^{(\lambda_2)\ T}\boldsymbol{X}_{L\mathcal{A}_{j_i}}^{(\lambda_2)}\right)^{-1}\left(\boldsymbol{X}_{L\mathcal{A}_{j_i}}^{(\lambda_2)^T}\boldsymbol{Y}_L - \lambda_1 \boldsymbol{s}_{j_i}\right).$$

The order of operations are important: substitute $\boldsymbol{X}_{L\mathcal{A}_{j_i}}$ for $\boldsymbol{X}_L$ and then apply Equation (21) to get $\boldsymbol{X}_{L\mathcal{A}_{j_i}}^{(\lambda_2)}$, and similarly, $\boldsymbol{X}_{U\mathcal{A}_{j_i}}$ for $\boldsymbol{X}_U$ to then get $\boldsymbol{X}_{U\mathcal{A}_{j_i}}^{(\gamma_2)}$ from Equation (8). Increased $\gamma_1$ and $\gamma_2$ puts more emphasis on shrinking unlabeled fits. Increased $\lambda_2$ and/or decreased $\gamma_2$ results in the labeled and/or unlabeled directions being better approximated by an $|\mathcal{A}_{j_i}|$-sphere, and increased $\lambda_1$ for presumably more stringent variable selection. Cross-validation often selects the joint trained elastic net with strictly positive lasso $\lambda_1 > 0$ and ridge $\lambda_2 > 0$ tuning parameter values in practical applications, so the joint trained elastic net is showcased later through its performance on numerical examples (i.e., simulated and real data sets) in Section 6.

### 4.5 Geometric Extrapolation Examples

The purpose of this section is learn more about the properties of our semi-supervised adjustment through additional geometrical examples of joint trained least squares from Section 4.1. Recall the joint trained least squares example in Figures 1, 2, and 4 for the heavily studied block extrapolation example. The first row of Figure 6 motivates additional discussion by simply changing the unlabeled feature data as follows.

- "Pure" – Extrapolations of larger magnitude are roughly in-line with the 2$^{\text{nd}}$ principal component, so supervised and semi-supervised shrinking are in similar directions at varying degrees.

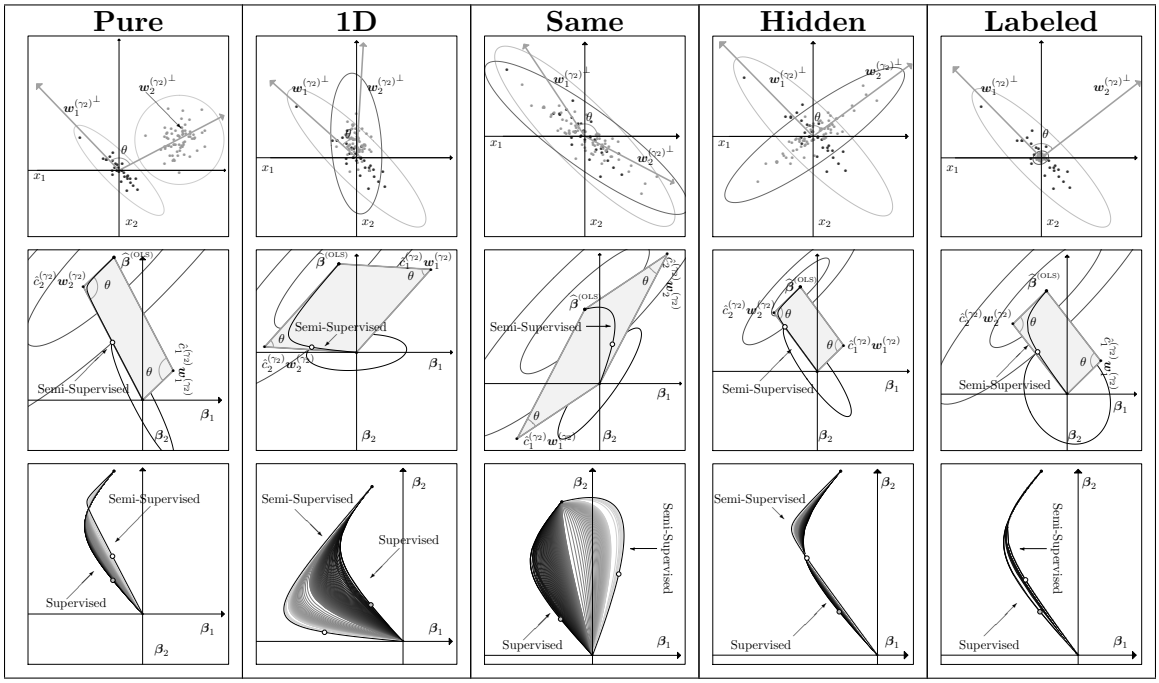


Figure 6: An additional geometrical example of joint trained least squares is displayed in each column. Row 1: Only the unlabeled feature data $\boldsymbol{X}_U$ from the "working" block extrapolation example from Figures 1, 2, and 4 were changed. Row 2: Ellipses (12) and (13) intersect at a point on the semi-supervised extreme. Row 3: Paths $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ are plotted by $\gamma_2$ varying $\gamma_1$. The gray circle is the supervised ridge solution from Figure 4(a).

- "1D" – The unlabeled marginal distribution is more volatile in one dimension $x_2$.
- "Same" – Minor discrepancies arise naturally in empirical distributions when taking independent samples from the same distribution.
- "Hidden" – Components $x_1$ and $x_2$ have roughly the same marginal distributions in both sets, but unlabeled extrapolations are hidden in the bivariate distribution of $(x_1, x_2)$.
- "Labeled" – Only the labeled feature data deviate substantially from the origin.

Broader sets of candidate $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ are entertained in the block, 1D, and same extrapolation examples. On the other hand, directions of extrapolations are roughly the principal components in the pure, hidden, and labeled extrapolation examples, and these examples have smaller candidates sets $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ as a result. In general, such smaller candidates sets are expected whenever the semi-supervised eigenvector directions of shrinking based on $\left(\boldsymbol{X}_L^T \boldsymbol{X}_L\right)^{-1} {\boldsymbol{X}_U^{(\gamma_2)}}^T \boldsymbol{X}_U^{(\gamma_2)}$ are approximately those in supervised ridge regression based on

| $\tau_i^{(\gamma_2)}$ | Block | Pure | 1D | Same | Hidden | Labeled |
|---|---|---|---|---|---|---|
| $\tau_1^{(\gamma_2)}$ | 95.1 | 662.5 | 11.2 | 4.2 | 38.0 | 0.30 |
| $\tau_2^{(\gamma_2)}$ | 22.6 | 1.3 | 0.3 | 1.2 | 0.1 | 0.01 |

Table 2: The eigenvalues of $\boldsymbol{M}^{(\gamma_2)}$ with $\gamma_2 = \infty$ are listed.

$\boldsymbol{X}_L^T \boldsymbol{X}_L$, but this does not imply that supervised and semi-supervised ridge techniques are approximately the same (see Remark 8).

The block and pure examples emphasize profoundly different directions of extrapolation, but have eigenvalues of large magnitude in Table 2. Extrapolations are on separate manifolds, and the approach shrinks predictions much more in these two examples at a given $\gamma_1 > 0$, by Equation (16). The semi-supervised extreme path closely maps the sides of its bounding parallelogram from Proposition 4 in the pure and hidden examples because their $\tau_i^{(\gamma_2)}$ in Table 2 are of different orders of magnitude. This phenomena is not present in the block and same examples when eigenvalues are of the same order of magnitude. The semi-supervised extreme in the 1D example is of special note. Its labeled feature data are negatively correlated, so the extreme emphasizes $x_1$ to shrink the influence of the component $x_2$ which is volatile in the unlabeled data.

Figure 7 is a 3D example. In the semi-supervised extreme, the shrinking matrix $\boldsymbol{M}^{(\gamma_2)}$ has eigenvalues $\tau_i^{(\gamma_2)} = 2090, 21.3, 1.08$, so shrinking of regression coefficients is much more heavily focused in direction $\boldsymbol{w}_1^{(\gamma_2)}$ because these eigenvalues differ in magnitude. The 1st direction of extrapolation is based on the other $p - 1 = 2$ directions of coefficient shrinking $\boldsymbol{w}_2^{(\gamma_2)}$ and $\boldsymbol{w}_3^{(\gamma_2)}$ and is defined as the set of all feature vectors that are orthogonal to both of these directions. The desired effect of using the unlabeled data to shrink unlabeled extrapolations more is achieved through Equation (16) at any $\gamma_1 > 0$. Semi-supervised predictions are $\boldsymbol{x}_0^T \widehat{\boldsymbol{\beta}}^{(\text{OLS})}/(1 + \gamma_1 2090)$ if $\boldsymbol{x}_0$ is a feature vector on the 1st direction of extrapolation; $\boldsymbol{x}_0^T \widehat{\boldsymbol{\beta}}^{(\text{OLS})}/(1 + \gamma_1 21.3)$ if $\boldsymbol{x}_0$ is on the 2nd direction; and $\boldsymbol{x}_0^T \widehat{\boldsymbol{\beta}}^{(\text{OLS})}/(1 + \gamma_1 1.08)$ if $\boldsymbol{x}_0$ is on the 3rd direction. Candidate vectors $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ in Figure 7(b) form a curved surface between supervised and semi-supervised extreme.

**Remark 8** *Even if supervised and semi-supervised candidate sets $\widehat{\boldsymbol{\beta}}$ are approximately equal, semi-supervised training with the unlabeled feature data $\boldsymbol{X}_U$ may pick a very different (and hopefully more advantageous) estimate $\widehat{\boldsymbol{\beta}}$ within the candidate set during cross-validation. In general, whether or not such apparent "parameter redundancies" exist, we always advocate the use of supervised regularization ($\boldsymbol{\lambda} \neq \vec{0}$) together with semi-supervised regularization ($\boldsymbol{\gamma} \neq \vec{0}$), especially when p is large. Many parameter redundancies noted in the $p = 2$ examples are not present in large p applications. If one briefly backs up to the case of $p = 1$, all candidate paths from Section 4.1 essentially start on the number line at the OLS estimate and then shrink to zero. When $p = 3$, one could overlay $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma},(0,\lambda_2)}$ for all $\boldsymbol{\gamma} \in [0, \infty]^2$ at fixed $\lambda_2 > 0$, and this in-fact adds a distinct layer to the 3D surface in Figure 7(b). The key point is to broaden the choices in an intelligent manner as needed so that a most desirable $\widehat{\boldsymbol{\beta}}$ can be selected for the purpose of unlabeled prediction.*

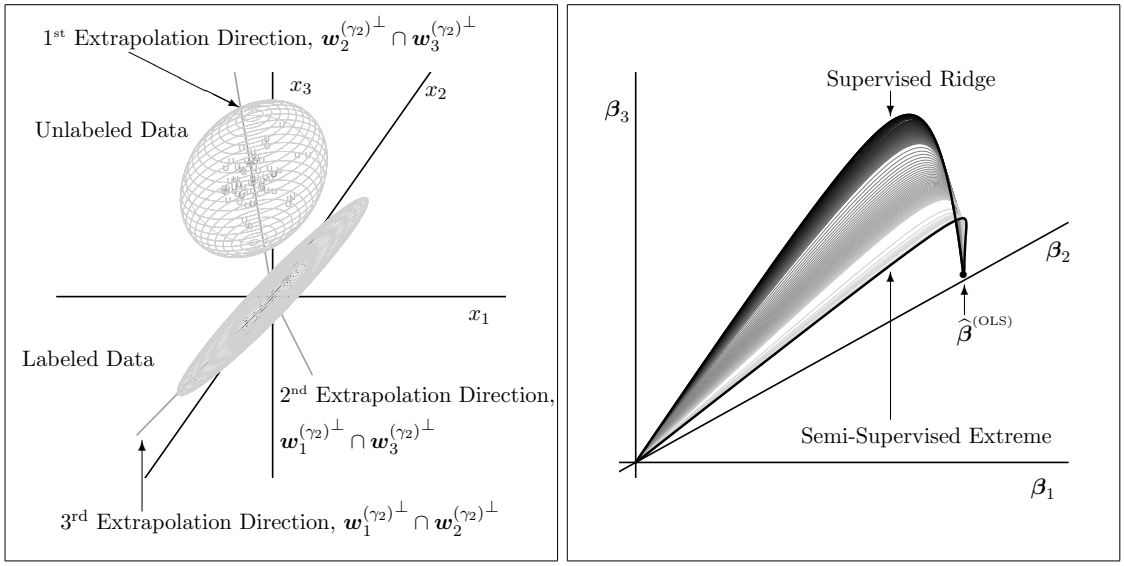


Figure 7: A $p = 3$ extrapolation data set is displayed. (a) The feature data along with the three extrapolation directions in the semi-supervised extreme of $\gamma_2 = \infty$ are plotted. Each direction of extrapolation is a line that equals the intersection of two planes by Proposition 5. (b) Candidate paths $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ by $\gamma_2$ varying $\gamma_1$ have a nonlinear compromise between supervised ridge and the semi-supervised extreme.

## 5. Performance Bounds

A general sufficient condition is given in this section for when a semi-supervised adjustment improves expected unlabeled prediction performance for a large class of linear supervised approaches. Assumption 1 on the class of supervised approaches is a necessary but not a sufficient condition for the elastic net; this generality was intentional. Assumption 2 characterizes a local property of our semi-supervised adjustment that follows from its Section 4 geometry.

*Assumption 1:* *The supervised estimate $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}^{(SUP)}$ is unique for data $(\boldsymbol{X}_L, \boldsymbol{Y}_L)$ and some $\boldsymbol{\lambda}$. Let $\boldsymbol{\phi} = \{\boldsymbol{\lambda}, \mathcal{A}, \boldsymbol{s}\}$ and $q = |\mathcal{A}|$ denote its fixed properties.*

*Assumption 2:* *$\exists\ \delta > 0$ such that $\forall\ \gamma_1 \in [0, \delta)$ semi-supervised estimates $\widehat{\boldsymbol{\beta}}_{\gamma_1}^{(\boldsymbol{\phi})}$ have the supervised active set $\mathcal{A}$ and sign vector $\boldsymbol{s}$, and*

$$\begin{aligned} \left(\widehat{\boldsymbol{\beta}}_{\gamma_1}^{(\boldsymbol{\phi})}\right)_{\mathcal{A}} &= \left(\boldsymbol{I} + \gamma_1 \boldsymbol{M}_{\mathcal{A}}^{(\lambda_2,\infty)}\right)^{-1} \left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}^{(SUP)}\right)_{\mathcal{A}}, \text{ where} \\ \boldsymbol{M}_{\mathcal{A}}^{(\lambda_2,\gamma_2)} &= \left(\boldsymbol{X}_{L\mathcal{A}}^{(\lambda_2)^T} \boldsymbol{X}_{L\mathcal{A}}^{(\lambda_2)}\right)^{-1} \boldsymbol{X}_{U\mathcal{A}}^{(\gamma_2)^T} \boldsymbol{X}_{U\mathcal{A}}^{(\gamma_2)}. \end{aligned}$$

Assumptions 1 and 2 always hold for the Joint Trained Optimization Problem (6) when $\lambda_2 > 0$ or $\mathrm{rank}(\boldsymbol{X}_L) = p$. For example, consider the joint trained lasso example from Figure 5(b) and Table 1. Assumption 1 holds with $\boldsymbol{\phi} = \{(0.01, 0), \{1, 2\}, (-1, 1)\}$, and Assumption 2 holds with $\delta = 0.004$ from Table 1.

Results to come focus on the impact of semi-supervised learning with $\gamma_2 = \infty$, so Propositions 3-6 are applied to $\widehat{\boldsymbol{\beta}}_{\gamma_1}^{(\boldsymbol{\phi})}$ and $\boldsymbol{M}_{\mathcal{A}}^{(\lambda_2,\infty)}$. Let $\left\{\left(\boldsymbol{w}_i^{(\boldsymbol{\phi})}, \tau_i^{(\boldsymbol{\phi})}\right)\right\}_{i=1}^{q}$ be an eigenbasis of $\boldsymbol{M}_{\mathcal{A}}^{(\lambda_2,\infty)}$ such that $\left\|\boldsymbol{X}_{L\mathcal{A}}^{(\lambda_2)}\boldsymbol{w}_i^{(\boldsymbol{\phi})}\right\|_2^2 = 1$ and $\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}^{(\text{SUP})}\right)_{\mathcal{A}} = \sum_{i=1}^{q}\hat{c}_i^{(\boldsymbol{\phi})}\boldsymbol{w}_i^{(\boldsymbol{\phi})}$ generalize Equation (15). Assumption 2 implies that Equation (16) generalizes to

$$\left(\widehat{\boldsymbol{\beta}}_{\gamma_1}^{(\boldsymbol{\phi})}\right)_{\mathcal{A}} = \left(\tfrac{1}{1+\gamma_1\tau_1^{(\boldsymbol{\phi})}}\right)\hat{c}_1^{(\boldsymbol{\phi})}\boldsymbol{w}_1^{(\boldsymbol{\phi})} + \cdots + \left(\tfrac{1}{1+\gamma_1\tau_q^{(\boldsymbol{\phi})}}\right)\hat{c}_q^{(\boldsymbol{\phi})}\boldsymbol{w}_q^{(\boldsymbol{\phi})}. \tag{24}$$

Assume the linear model with $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{X}\boldsymbol{\beta}$ and $\mathbb{V}\text{ar}(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}$ and project

$$\boldsymbol{\beta}_{\mathcal{A}} = c_1^{(\boldsymbol{\phi})}\boldsymbol{w}_1^{(\boldsymbol{\phi})} + \cdots + c_q^{(\boldsymbol{\phi})}\boldsymbol{w}_q^{(\boldsymbol{\phi})}. \tag{25}$$

If small $c_i^{(\boldsymbol{\phi})}$ correspond to large $\tau_i^{(\boldsymbol{\phi})}$, a performance improvement on the evaluation function $\left\|\boldsymbol{X}_{U\mathcal{A}}\left(\boldsymbol{\beta}_{\mathcal{A}} - \left(\widehat{\boldsymbol{\beta}}_{\gamma_1}^{(\boldsymbol{\phi})}\right)_{\mathcal{A}}\right)\right\|_2^2$ appears likely by Equations (24) and (25). If $\tau_i^{(\boldsymbol{\phi})}$ is large for a small subset $i \in \Omega \subset \{1,\ldots,p\}$ and small otherwise, then semi-supervised performance is expected to be better over a larger percentage of the possible directions for the true $\boldsymbol{\beta}$. Such high performance circumstances occur when a low dimensional manifold of $\boldsymbol{X}_{U\mathcal{A}}$ concentrates away from that of $\boldsymbol{X}_{L\mathcal{A}}$ and the true coefficient vector $\boldsymbol{\beta}$ emphasizes directions dominated by labeled extrapolations. Assumption 3 helps establish a general transductive bound for when semi-supervised learning is better than supervised on evaluation function $\mathbb{E}\left[\left\|\boldsymbol{X}_U\left(\widehat{\boldsymbol{\beta}}_{\gamma_1}^{(\boldsymbol{\phi})} - \boldsymbol{\beta}\right)\right\|_2^2\middle|\boldsymbol{\phi}\right]$.

*Assumption 3:* $\mathbb{E}\left[\hat{c}_i^{(\boldsymbol{\phi})}\middle|\boldsymbol{\phi}\right] = \mu_i < \infty$ *and* $\mathbb{V}\text{ar}\left[\hat{c}_i^{(\boldsymbol{\phi})}\middle|\boldsymbol{\phi}\right] = \sigma_i^2 < \infty\ \forall i \in \{1,\ldots,q\}$.

Let $\bar{\mathcal{A}} = \{1,\ldots,p\} - \mathcal{A}$ be the supervised non-active set and define $\boldsymbol{X}_{U\emptyset}\boldsymbol{\beta}_{\emptyset} = \vec{0}$. Theorem 9 provides a sufficient condition on parameters $\left(\boldsymbol{\beta}, \sigma^2\right)$ for when semi-supervised outperforms supervised given the feature data and $\boldsymbol{\phi}$.

**Theorem 9** *Let Assumptions 1-3 hold. Also, let* $q \geq 1$, $\tau_1^{(\boldsymbol{\phi})} > 0$, *and* $p_i\left(\boldsymbol{\tau}^{(\boldsymbol{\phi})}\right) = \frac{\tau_i^{(\boldsymbol{\phi})^2}\sigma_i^2}{\sum_{j=1}^{q}\tau_j^{(\boldsymbol{\phi})^2}\sigma_j^2}$. *If* $\sum_{i=1}^{q} p_i\left(\boldsymbol{\tau}^{(\boldsymbol{\phi})}\right)\left(\frac{\mu_i\left(c_i^{(\boldsymbol{\phi})} + \boldsymbol{u}_i^{(\boldsymbol{\phi})^T}\boldsymbol{X}_{U\bar{\mathcal{A}}}\boldsymbol{\beta}_{\bar{\mathcal{A}}}/\sqrt{\kappa_i^{(\gamma_2)}}\right) - \mu_i^2}{\sigma_i^2}\right) < 1$, *then*

$$\mathbb{E}\left[\left\|\boldsymbol{X}_U\left(\widehat{\boldsymbol{\beta}}_{\gamma_1}^{(\boldsymbol{\phi})} - \boldsymbol{\beta}\right)\right\|_2^2\middle|\boldsymbol{\phi}\right] < \mathbb{E}\left[\left\|\boldsymbol{X}_U\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}^{(SUP)} - \boldsymbol{\beta}\right)\right\|_2^2\middle|\boldsymbol{\phi}\right].$$

As stated earlier, Assumptions 1 and 2 hold for the general $\boldsymbol{\lambda}$ joint trained elastic net regression of Section 4.4. In the case of $\boldsymbol{\lambda} = \vec{0}$ least squares, it is also easily verified that $\mu_i = c_i^{(\boldsymbol{\phi})}$ and $\sigma_i^2 = \sigma^2$ for Assumption 3. The mathematical form of the extreme version of joint trained least squares in Equation (14) is equivalent to that for generalized ridge regression. Corollary 10 in conjunction with Casella (1980) shows that joint trained least squares is

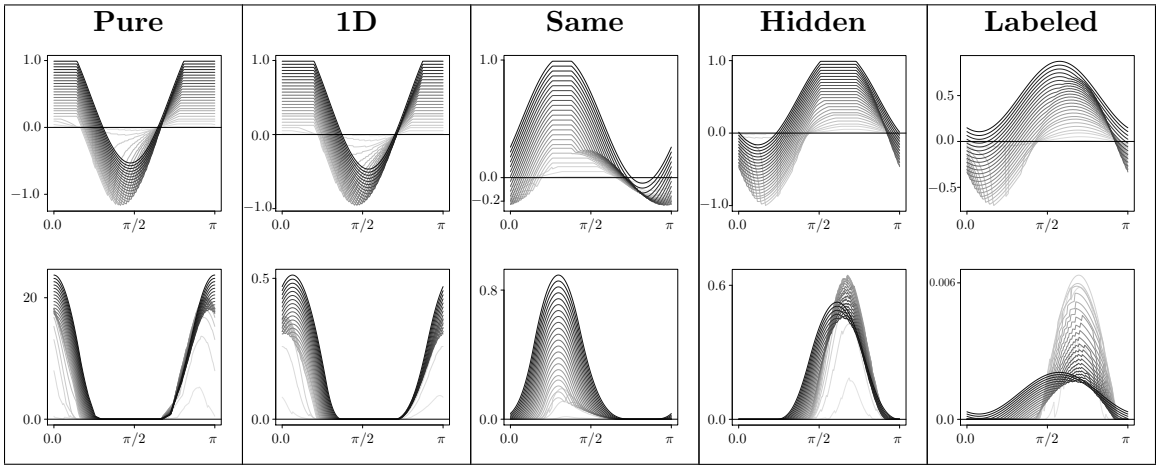


Figure 8: The five examples with $p = 2$ from Figure 6 are revisited. Row 1: Theoretical bound $\sigma^2 - \sigma^2_{\text{LB}}(\boldsymbol{\beta}(\vartheta))$ is plotted against $\vartheta$. Darker curves correspond to larger $\sigma^2 \in [0, 1]$. Row 2: The corresponding differences $\text{RMSE}_U^{(\text{SUP})} - \text{RMSE}_U^{(\text{SEMI})}$ are plotted against $\vartheta$.

asymptotically minimax with respect to loss function $\mathbb{E}\left[\left\|\boldsymbol{X}_U\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right\|_2^2\right]$ as $|L| \to \infty$. In the case of ridge regression, $\mu_i = c_i^{(\boldsymbol{\phi})} - \lambda_2 \boldsymbol{w}_i^{(\boldsymbol{\phi})^T}\boldsymbol{\beta}$ and $\sigma_i^2 = \boldsymbol{w}_i^{(\boldsymbol{\phi})^T}\boldsymbol{X}_L^T\boldsymbol{X}_L\boldsymbol{w}_i^{(\boldsymbol{\phi})}\sigma^2$ for any $i \in \{1, \ldots, p\}$ are also straightforward to derive, so Theorem 9 reduces to Corollary 11.

**Corollary 10** *Joint trained least squares with $\gamma_2 = \infty$ dominates supervised least squares in prediction on $\boldsymbol{X}_U$ if $q \geq 1$ and $\tau_1^{(\boldsymbol{\phi})} > 0$.*

**Corollary 11** *The extreme version of joint trained ridge regression dominates supervised ridge regression in prediction on $\boldsymbol{X}_U$ if $q \geq 1$, $\tau_1^{(\boldsymbol{\phi})} > 0$, and*

$$\sigma^2_{LB}(\boldsymbol{\beta}) = \left(\sum_{i=1}^{p} p_i\left(\boldsymbol{\tau}^{(\boldsymbol{\phi})}\right)\left(\frac{\left(c_i^{(\boldsymbol{\phi})} - \lambda_2 \boldsymbol{w}_i^{(\boldsymbol{\phi})^T}\boldsymbol{\beta}\right)\left(\lambda_2 \boldsymbol{w}_i^{(\boldsymbol{\phi})^T}\boldsymbol{\beta}\right)}{\boldsymbol{w}_i^{(\boldsymbol{\phi})^T}\boldsymbol{X}_L^T\boldsymbol{X}_L\boldsymbol{w}_i^{(\boldsymbol{\phi})}}\right)\right)_+ < \sigma^2.$$

The block feature data from Figure 1 were used to construct Figure 3 and introduce the reader to the semi-supervised ridge bound $\sigma^2_{\text{LB}}(\boldsymbol{\beta})$ earlier in Section 2.2. The analog of that figure for the five examples from Figure 6 is given in this section by Figure 8. A technical explanation of how these figures were constructed precedes the qualitative discussion of their interpretations in the next paragraph. First, note that $\sigma^2_{\text{LB}}(\boldsymbol{\beta}(\vartheta))$ from Corollary 11 is independent of $\sigma^2$. It was computed for all $\boldsymbol{\beta}(\vartheta) = (\sin(\vartheta), \cos(\vartheta))^T$ over a fine grid of $\vartheta \in [0, \pi]$, and the $\sigma^2_{\text{LB}}(\boldsymbol{\beta}(\vartheta))$ were compared to a fine, equally spaced grid of $\sigma^2 \in [0, 1]$. Only the right half of the unit circle was considered for $\boldsymbol{\beta}$ because $\sigma^2_{\text{LB}}(\boldsymbol{\beta}(\vartheta)) = \sigma^2_{\text{LB}}(\boldsymbol{\beta}(\vartheta + \pi))$. Also, $\sigma^2_{\text{LB}}(r\boldsymbol{\beta}(\vartheta)) = r^2\sigma^2_{\text{LB}}(\boldsymbol{\beta}(\vartheta))$, so the same trend results from the scaled parameters $r\boldsymbol{\beta}(\vartheta)$ with $\sigma^2 \in [0, r^2]$. The ridge parameter was set to the "best" supervised attempt

of $\lambda_2^{(\text{opt})}$ minimizing $\mathbb{E}\left[\left\|\boldsymbol{X}_L\left(\boldsymbol{\beta}(\vartheta)-\widehat{\boldsymbol{\beta}}_{\lambda_2}^{(\text{RIDGE})}\right)\right\|_2^2\right]$. Interest was in identifying $\vartheta$'s when a semi-supervised adjustment helps, i.e., when $\sigma^2-\sigma^2_{\text{LB}}\left(\boldsymbol{\beta}(\vartheta)\right)>0$.

Angles $\vartheta$ corresponding to lucky $\boldsymbol{\beta}(\vartheta)$ and to reductions in RMSE due to semi-supervised learning line-up vertically across the rows of Figure 8 (i.e., $\vartheta$ with a positive vertical coordinate in row 1 also have a positive coordinate in row 2 and vice versa). Row 2 is the magnitude of the improvements, and the examples with the largest magnitude (i.e., the pure example and the block example from Figure 3(b)) are those with the largest eigenvalues in Table 2 as expected. The labeled example with the smallest improvements also has the smallest eigenvalues. Direction $\boldsymbol{w}_2^{(\boldsymbol{\phi})}$ (eyeballed from the row 1 of Figure 6) should be compared to row 1 of Figure 8. In each example, the center for potentially large improvements is roughly $\boldsymbol{\beta}(\vartheta)\propto\boldsymbol{w}_2^{(\boldsymbol{\phi})}$, and the center for little to no potential improvement is roughly $\boldsymbol{\beta}(\vartheta)\propto\boldsymbol{w}_2^{(\boldsymbol{\phi})^\perp}$. The generalization to $p\geq 2$ in Proposition 12 below extends this interpretation to that given back in Section 2.2. That is, if $\boldsymbol{\beta}$ is orthogonal to an unlabeled manifold, then $\widehat{\boldsymbol{\beta}}_{\gamma_1}^{(\boldsymbol{\phi})}$ has an unlabeled prediction advantage over $\widehat{\boldsymbol{\beta}}_{\lambda_2}^{(\text{RIDGE})}$, whereas $\boldsymbol{\beta}$ parallel to the unlabeled manifold yields no theoretical advantage.

**Proposition 12** *If $\tau_1^{(\boldsymbol{\phi})}>0$ and $\boldsymbol{\beta}_i\in\bigcap_{j\in\{1,\ldots,p\}-\{i\}}\boldsymbol{w}_j^{(\boldsymbol{\phi})^\perp}$ is unit length, then the joint trained ridge performance bound from Corollary 11 satisfies $\sigma^2_{LB}\left(\boldsymbol{\beta}_i\right)\geq\lambda_2 p_i\left(\boldsymbol{\tau}^{(\boldsymbol{\phi})}\right)$ for $i\in\{1,\ldots,p\}$ and $\sigma^2_{LB}\left(\boldsymbol{\beta}_i\right)\geq\sigma^2_{LB}\left(\boldsymbol{w}_j^{(\boldsymbol{\phi})}/\left\|\boldsymbol{w}_j^{(\boldsymbol{\phi})}\right\|_2\right)$ if $j\geq i$.*

Given a lasso estimate $\widehat{\boldsymbol{\beta}}_{\lambda_1}^{(\text{LASSO})}$, response $\boldsymbol{Y}_L\in\mathcal{Y}_L(\boldsymbol{\phi})=\left\{\boldsymbol{y}\in I\!R^{|L|}:\widehat{\boldsymbol{\beta}}_{\lambda_1}^{(\text{LASSO})}\text{ has }\boldsymbol{\phi}\right\}$, and the sets $\mathcal{Y}_L(\boldsymbol{\phi})$ partition $I\!R^{|L|}$ at fixed $\boldsymbol{\lambda}=(\lambda_1,0)$. If we additionally assume a normal theory linear model, $\boldsymbol{Y}_L|\boldsymbol{\phi}$ has a truncated normal distribution on $\mathcal{Y}_L(\boldsymbol{\phi})$, so means $\mu_i$ and variances $\sigma_i^2$ also depend on $\left(\boldsymbol{\beta},\sigma^2\right)$. Although the extreme versions of the lasso and elastic net are intractable, the interpretation of Theorem 9 still applies.

## 6. Numerical Examples

In this Section, both simulated and real data scenarios are presented for the Joint Trained Elastic Net (JT-ENET). The simulation is run with both lucky and unlucky $\boldsymbol{\beta}$ examples. For the ridge regression version of our estimator, the theoretical bound from Proposition 12 implies that a lucky $\boldsymbol{\beta}$ is perpendicular to the unlabeled centroid and a unlucky $\boldsymbol{\beta}$ is parallel to the unlabeled centroid. The result in Theorem 9 presumably extends the generality of this concept. The simulation was designed in part to assess whether the notion of lucky versus unlucky $\boldsymbol{\beta}$ extends to the JT-ENET. The real data sets provide covariate shift applications, so the JT-ENET should have some advantage over supervised learning in terms of a prediction focused objective function on the unlabeled set. It is important to note that only $\boldsymbol{X}_L$, $\boldsymbol{X}_U$, and $\boldsymbol{Y}_L$ were used during training throughout this section.

In all cases, comparisons were made to the supervised elastic net using the R package `glmnet` (Friedman et al., 2010; R Core Team, 2015). This particular implementation is optimized for estimating $\lambda_1+2\lambda_2$ with 10-fold cross validation given $\lambda_1/(\lambda_1+2\lambda_2)$. First, the supervised elastic net was implemented by varying $\lambda_1/(\lambda_1+2\lambda_2)\in[0,1]$ over an equally spaced grid of length 57 to optimize parameters $\boldsymbol{\lambda}$.

Second, the semi-supervised JT-ENET was implemented by estimating its parameters $(\boldsymbol{\lambda}, \boldsymbol{\gamma})$ simultaneously. Calls to the `glmnet` with data augmentations from Proposition 2 were used for all low-level fittings. Parameter $\lambda_1 + 2\lambda_2$ was estimated using 10-fold cross-validation given $\lambda_1/(\lambda_1 + 2\lambda_2)$, $\gamma_1$, and $\gamma_2$. Parameter $\lambda_1/(\lambda_1 + 2\lambda_2)$ was optimized over the grid $\{0, 0.25, 0.5, 0.75, 1, \hat{a}\}$, where $\hat{a}$ was the optimal supervised setting for this parameter. Fixed grids $\gamma_1 \in \nu^{-1}$ and $\gamma_2 \in \nu$ were used for the other parameters, where $\nu = \{0.1, 0.5, 1, 10, 100, 1000, 10000, \infty\}$ and $\nu^{-1} = \{1/r : r \in \nu\}$. For $K$-fold cross-validation in the semi-supervised setting, the $L$ cases were partitioned into $K$ folds, $\{L_k\}_{k=1}^K$. Let $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma},\boldsymbol{\lambda}}^{(-k)}$ be the estimate from labeled data $L - L_k$ and unlabeled data $U \cup L_k$, and let the $K$-fold cross-validated variance be $\widehat{\sigma}_K^2 = \sum_{k=1}^K \left\| \boldsymbol{Y}_{L_k} - \boldsymbol{X}_{L_k} \widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma},\boldsymbol{\lambda}}^{(-k)} \right\|_2^2 / |L|$. The JT-ENET estimate $\widehat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\lambda}}}$ minimized $\widehat{\sigma}_3^2$ over the grid for $\lambda_1/(\lambda_1 + 2\lambda_2)$, $\gamma_1$, and $\gamma_2$.

Our objective function was the RMSE on the unlabeled set. The RMSE of $\boldsymbol{X}_U \widehat{\boldsymbol{\beta}}$ from $\boldsymbol{X}_U \boldsymbol{\beta}$ was computed within simulations, but was computed from the withheld responses $\boldsymbol{Y}_U$ in the real data examples. Let ENET and JT-ENET represent this unlabeled set RMSE for the supervised elastic net and our proposed method using the true $\boldsymbol{\beta}$ for the simulations and their empirical versions in real data examples. Percent improvement $\%\text{JT-ENET} = \frac{\text{ENET} - \text{JT-ENET}}{\text{ENET}} \times 100\%$ was used to assess semi-supervised performance. A baseline comparison to the theoretical best parameter settings for the semi-supervised technique was also computed in the simulations, and its percent improvement is denoted by %BEST. Two regression based covariate shift competitors were also applied to the real data examples: adaptive importance-weighted kernel regularized least-squares (AIWKRLS) (Sugiyama et al., 2007) and plain kernel regularized least-squares (PKRLS) (Kananmori et al., 2009). The `caret` package in R (Kuhn, 2008) was also used to fit the SVM with a polynomial kernel on the real data examples.

### 6.1 Simulations

Same and extrapolated feature data distributions were constructed to study three, high-dimensional scenarios. Each scenario had $|L| = |U| = 100$, $p = 1,000$, true active set $\mathcal{T} = \{1, \ldots, 10\}$, $(\boldsymbol{X}_L)_{ij} \overset{\text{i.i.d.}}{\sim} \text{N}(0, 0.4)$, and $\boldsymbol{Y}_L = \boldsymbol{X}_L \boldsymbol{\beta} + \epsilon$ with $\epsilon \sim \text{N}\left(\vec{0}, \sigma^2 \boldsymbol{I}\right)$. Define indicator vector $\boldsymbol{\mu}(\mathcal{A}) \in I\!R^p$ with entries $\boldsymbol{\mu}_j(\mathcal{A}) = \mathcal{I}_{\{j \in \mathcal{A}\}}$ for some active set $\mathcal{A}$, $\boldsymbol{\beta}^{(\text{unlucky})} = 5\boldsymbol{\mu}(\mathcal{T})/\sqrt{10}$, and $\boldsymbol{\beta}^{(\text{lucky})} = 5\left(\boldsymbol{\mu}(\mathcal{T}_1) - \boldsymbol{\mu}(\mathcal{T}_2)\right)/\sqrt{10}$ with $\mathcal{T}_1 = \{1, \ldots, 5\}$ and $\mathcal{T}_2 = \{6, \ldots, 10\}$. The three scenarios were

1. **Same Distribution:** $(\boldsymbol{X}_U)_{ij} \overset{\text{i.i.d.}}{\sim} \text{N}(0, 0.4)$ and $\boldsymbol{\beta} = \boldsymbol{\beta}^{(\text{lucky})}$
2. **Extrapolation (Lucky $\boldsymbol{\beta}$):** $(\boldsymbol{X}_U)_{ij} \overset{\text{ind}}{\sim} \text{N}\left(10\boldsymbol{\mu}_j(\mathcal{T}), 0.4\right)$ and $\boldsymbol{\beta} = \boldsymbol{\beta}^{(\text{lucky})}$
3. **Extrapolation (Unlucky $\boldsymbol{\beta}$):** $(\boldsymbol{X}_U)_{ij} \overset{\text{ind}}{\sim} \text{N}\left(10\boldsymbol{\mu}_j(\mathcal{T}), 0.4\right)$ and $\boldsymbol{\beta} = \boldsymbol{\beta}^{(\text{unlucky})}$.

If the truth $\boldsymbol{X}_U \boldsymbol{\beta}$ is large, any type of shrinking may be detrimental, so shrinking methods (supervised or semi-supervised) should struggle in the extrapolation scenario with unlucky $\boldsymbol{\beta}$ because $\boldsymbol{\beta} = \boldsymbol{\beta}^{(\text{unlucky})}$ is parallel to the unlabeled data centroid $\boldsymbol{\mu}(\mathcal{T})$. On the other hand, $\boldsymbol{\beta}^{(\text{lucky})} \perp \boldsymbol{\mu}(\mathcal{T})$, so shrinking directions of extrapolation is more desirable. There is an unlucky $\boldsymbol{\beta}$ direction, but a $|\mathcal{T}| - 1$ or 9-dimensional vector space of lucky $\boldsymbol{\beta}$ directions. Setting $\boldsymbol{\beta} = \boldsymbol{\beta}^{(\text{lucky})}$ versus $\boldsymbol{\beta} = \boldsymbol{\beta}^{(\text{unlucky})}$ is not critical in the same distribution scenario.

| | Same Distribution | | | Extrapolation (Lucky $\boldsymbol{\beta}$) | | | Extrapolation (Unlucky $\boldsymbol{\beta}$) | | |
|---|---|---|---|---|---|---|---|---|---|
| $\sigma^2$ | ENET | %JT-ENET | %BEST | ENET | %JT-ENET | %BEST | ENET | %JT-ENET | %BEST |
| 2.5 | 0.43 | 19.27 | 30.67 | 0.91 | 18.08 | 27.58 | 15.05 | −2.55 | 2.06 |
| | 0.03 | 3.41 | 3.63 | 0.08 | 3.50 | 3.65 | 0.11 | 0.83 | 0.65 |
| 5.0 | 0.69 | 31.88 | 46.67 | 1.25 | 26.10 | 38.73 | 15.19 | −1.74 | 2.90 |
| | 0.06 | 4.88 | 4.43 | 0.12 | 4.32 | 4.43 | 0.14 | 0.96 | 0.86 |
| 7.5 | 0.93 | 35.36 | 54.63 | 1.61 | 33.49 | 46.98 | 15.30 | −1.47 | 4.75 |
| | 0.09 | 5.68 | 4.58 | 0.17 | 4.53 | 4.50 | 0.21 | 1.44 | 1.25 |

Table 3: Unlabeled root mean squared error performance is summarized on high-dimensional ($p = 1,000$), simulated data sets: supervised elastic net (ENET), percent improvement over ENET with the joint trained elastic net (%JT-ENET), and the hypothetical maximum of %JT-ENET based on "cheating" with the "answers" $\boldsymbol{X}_U\boldsymbol{\beta}$ while picking the point $(\boldsymbol{\lambda}, \boldsymbol{\gamma})$ in the cross-validation grid (%BEST). Fifty data sets were generated per level combination of scenario (i.e., same, lucky, and unlucky) and model error variance $\sigma^2 = 2.5, 5.0, 7.5$. Cell entries are the sample mean (top) and standard error (bottom).

These probability models were used to conduct simulations studies in the following manner.

Model matrix $\boldsymbol{X}$ was generated once and fixed by scenario, and 50 independent response vectors $\boldsymbol{Y}_L$ were generated from the assumed linear model for each level combination of scenario $= 1, 2, 3$ and $\sigma^2 = 2.5, 5.0, 7.5$. Cross-validation took an average of 3.5 minutes per data set on a 2.6 GHz Intel Core i7 Power Mac. The supervised ENET is best suited for the same distribution prediction task, and its RMSEs are smallest in this scenario. The significant performance advantage due to our semi-supervised adjustment in the same distribution scenario relates to the curse of dimensionality, because extrapolations are likely in a high-dimensional empirical distribution. There was also substantial improvement in the extrapolation with lucky $\boldsymbol{\beta}$ scenario, while both approaches struggled at extrapolation with unlucky $\boldsymbol{\beta}$.

The %BEST values reported in Table 3 correspond to the best possible points $(\boldsymbol{\lambda}, \boldsymbol{\gamma})$ in the cross validation grid and provide at least two points of useful discussion. First, values %BEST increased with $\sigma^2$, and this is consistent with what one might expect given the factorization of the bound in Corollary 11. Its left hand side is a nonnegative number that is independent of $\sigma^2$, and a semi-supervised improvement is possible when $\sigma^2$ exceeds this nonnegative number. The values %BEST in Table 3 supports that a similar concept holds with the bound in Theorem 9 that applies to the JT-ENET. Second, most points in the cross validation grid corresponded to negative percent improvements, and some of these are the largest in magnitude. Thus, while the method of cross validation is not getting the very best point in the grid, its performance is competitive.

### 6.2 Real Data Examples

The 10 tests listed in Table 4 were constructed using 8 publicly available data sets and a simulated toy extrapolation data set. Each is expected to have a covariate shifted empirical feature data distribution either because the characteristic used to define the labeled set is

| Data Set $(n, p)$ | Labeled Set $L$ | Response $y$ | Data Set Source |
|---|---|---|---|
| Toy Cov. Shift $(1200, 1)$ | Training Set | $\text{sinc}(x) + \epsilon$ | Sugiyama et al. (2007) |
| Auto-MPG $(398, 8)$ | P1: Domestics | Fuel (mpg) | Lichman (2013) |
| Auto-MPG $(398, 8)$ | P2: $\leq 4$ Cyl. | Fuel (mpg) | Lichman (2013) |
| Heart $(462, 8)$ | No History | $\sqrt{}$Sys. BP | Hastie et al. (2009) |
| U.S. News $(1004, 19)$ | Private Schools | SAT.ACT | ASA Data Expo '95 |
| Auto-Import $(205, 24)$ | Low Risk Cars | Price | Lichman (2013) |
| Blood Brain $(208, 135)$ | Cmpds. 1-52 | log(BBB) | Kuhn (2008) |
| Eye $(120, 200)$ | Rats 1-30 | $\sqrt{}$Express | Scheetz et al. (2006) |
| Cookie $(72, 700)$ | Training Set | Water | Osborne et al. (1984) |
| Ethanol $(589, 1037)$ | Sols. 1-294 | Ethanol | Shen et al. (2013) |

Table 4: These ten covariate shift tests are used to establish benchmarks in Table 5.

| Data Set | $p$ | $\lvert L\rvert$ | $\lvert U\rvert$ | ENET | SVM | AIWKRLS | PKRLS | JT-ENET | %JT-ENET |
|---|---|---|---|---|---|---|---|---|---|
| Toy Cov. Shift | 1 | 200 | 1000 | 0.527 | 0.186 | **0.103** | 0.129 | 0.169 | **67.83** |
| Auto-MPG (P1) | 8 | 149 | 249 | 5.361 | 5.272 | 5.974 | 8.459 | **4.341** | **19.02** |
| Auto-MPG (P2) | 8 | 208 | 190 | 8.296 | 13.478 | 15.374 | 39.570 | **6.723** | **18.96** |
| Heart | 8 | 192 | 270 | 0.789 | 0.790 | 0.795 | 0.802 | **0.788** | **0.13** |
| U.S. News | 19 | 640 | 364 | 1.738 | 1.724 | 1.928 | 1.918 | **1.684** | **3.11** |
| Auto-Import | 24 | 113 | 92 | 4995 | 4223 | 6292 | 6376 | **4201** | **15.89** |
| Blood Brain | 135 | 52 | 156 | 1.684 | 6.424 | 0.797 | 0.815 | **0.649** | **61.46** |
| Eye | 200 | 30 | 90 | 0.019 | 0.425 | 0.027 | 0.027 | **0.016** | **15.79** |
| Cookie | 700 | 40 | 32 | 0.388 | 0.580 | 1.466 | 1.309 | **0.342** | **11.86** |
| Ethanol | 1037 | 294 | 295 | 1.461 | 1.422 | 2.626 | 2.625 | **1.391** | **4.79** |

Table 5: Empirical unlabeled root mean squared errors are listed for the ten covariate shift tests defined by Table 4 and a field of five competitors: the supervised elastic net (ENET), a support vector machine (SVM), adaptive importance-weighted kernel regularized least-squares (AIWKRLS), plain kernel regularized least-squares (PKRLS), and joint trained ENET (JT-ENET). The top performer is in bold. The final column is percent improvement of JT-ENET over its supervised ENET alternative with positive values in bold.

associated with other variables in the model matrix, because of the curse of dimensionality, or because the simulated toy data were generated from a model with covariate shift. Since covariate shift is our focus, randomized subsetting of the data (i.e., MCAR) was not performed. When $p$ is larger in the blood brain, eye, cookie, and ethanol applications, the unlabeled set is likely to contain extrapolations. In all cases, the bounds from Section 5 together with the Section 4 geometry of the JT-ENET are at play here behind the scenes. The *U.S. News & World Report* data required preprocessing. SAT scores were transformed to their ACT equivalent, and the new variable with either transformed SAT, ACT, or their average was used instead. Median imputation was used for all other missing values across the board. In the Toy Covariate Shift example, we forced $\boldsymbol{\lambda} = \vec{0}$ for both the ENET and JT-ENET to make comparisons consistent with Sugiyama et al. (2007).

RMSEs for the various approaches and the empirical percent improvement for JT-ENET are reported in Table 5. The JT-ENET appears to have worked in the ideal manner independent of what caused the empirical covariate shift. In their toy covariate shift example, competitors AIWKRLS and PKRLS performed strongly, but their edge went away with increased $p$. AIWKRLS and PKRLS are principled on estimating empirical density ratios, and this can be a challenging task in practical applications with large $p$. The SVM and ENET are very close competitors for most of the examples. The results provide further evidence that the JT-ENET is achieving the goal of out-performing the ENET in covariate shift problems.

The JT-ENET fit fairly quickly on a 2.6 GHz Intel Core i7 Power Mac. Thus, if the range of possible improvements is from roughly none to substantial in any given prediction focused application, the associated computational overhead of the JT-ENET appears worthwhile. In addition, it is embarrassingly parallel. Just consider the fixed $6 \times 8 \times 8$ grid search over $(\lambda_1/(\lambda_1 + 2\lambda_2)) \times \gamma_1 \times \gamma_2$ in our implementation. Effective times can essentially be divided by 6 if one sends $1 \times 8 \times 8$ grid searches to each of 6 computers or divided by 48 with grids of $1 \times 1 \times 8$ to 48 computers.

## 7. Discussion

This work provided a clear and succinct mathematical framework for semi-supervised linear predictions of the unlabeled data. Our joint trained elastic net has two pairs of tuning parameters: supervised $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ and semi-supervised $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$. Adjusting the semi-supervised parameters has an interpretable, geometrical effect on the unlabeled predictions. Furthermore, we provided theoretical bounds for when this interpretable adjustment guarantees a performance improvement under the standard linear model, and this main theme of these theoretical results was validated with simulated data. This practical approach was also competitive with existing approaches throughout a set of challenging, high-dimensional, real data applications, where the unlabeled data contained extrapolations. Extrapolations in the unlabeled set are expected to occur often in practice, due to the curse of dimensionality with large $p$ or practical constraints that result in covariate shift applications, and our method is unique among existing approaches in its direct and effective accounting for these circumstances. Simultaneous estimation of the supervised and semi-supervised tuning parameters was feasible in the high-dimensional examples we tested.

## Acknowledgments

The authors thank the AE and three anonymous referees. Their comments and suggestions led to substantial improvements in the presentation of this work. The authors also thank Professor Stephen B. Vardeman. Extensive in-person discussions between the first author and Professor Vardeman led to an understanding of the Joint Training Optimization Problem (6) that ultimately helped both authors articulate its applications. These in-person conversations were made possible through the visiting faculty program within the Statistical Sciences Group at Los Alamos National Laboratory, and the first author is also thankful to be a part of that research program. The work of Mark Vere Culp was supported in part

by the NSF CAREER/DMS-1255045 grant. The opinions and views expressed in this paper are those of the authors and do not reflect the opinions or views at the NSF.

## Appendix A. Proofs

Proofs of Propositions and Theorems follow.

### A.1 Joint Training Framework

**Proposition 2** *If $\gamma_2 > 0$, then rank$(\boldsymbol{X}_U)$ = rank$\left(\boldsymbol{X}_U^{(\gamma_2)}\right)$ and a solution $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma},\boldsymbol{\lambda}}$ to Optimization Problem (9) is a partial solution to Optimization Problem (6).*

**Proof** Clearly, $\text{rank}\left(\boldsymbol{X}_U\right) = \text{rank}\left(\boldsymbol{X}_U^{(\gamma_2)}\right)$ whenever $\gamma_2 > 0$ by Equation (8). Based on Objective (6), the optimal $\boldsymbol{\alpha}$ at any $\boldsymbol{\beta}$ is $\boldsymbol{\alpha} = \left(\boldsymbol{X}_U^T\boldsymbol{X}_U + \gamma_2\boldsymbol{I}\right)^{-1}\boldsymbol{X}_U^T\boldsymbol{X}_U\boldsymbol{\beta}$ and does not depend on $\gamma_1 > 0$. The derivative with respect to $\boldsymbol{\beta}$ of the objective is proportional to $-\gamma_1\boldsymbol{X}_U^T\boldsymbol{X}_U(\boldsymbol{\alpha}-\boldsymbol{\beta})$ as a function of the unlabeled data, and after plugging-in the optimal $\boldsymbol{\alpha}$ it simplifies to

$$\begin{aligned} -\gamma_1\boldsymbol{X}_U^T\boldsymbol{X}_U(\boldsymbol{\alpha}-\boldsymbol{\beta}) &= -\gamma_1\boldsymbol{X}_U^T\boldsymbol{X}_U\left\{\left(\boldsymbol{X}_U^T\boldsymbol{X}_U+\gamma_2\boldsymbol{I}\right)^{-1}\boldsymbol{X}_U^T\boldsymbol{X}_U - \boldsymbol{I}\right\}\boldsymbol{\beta} \\ &= \gamma_1{\boldsymbol{X}_U^{(\gamma_2)}}^T\boldsymbol{X}_U^{(\gamma_2)}\boldsymbol{\beta}, \end{aligned} \tag{26}$$

where ${\boldsymbol{X}_U^{(\gamma_2)}}^T\boldsymbol{X}_U^{(\gamma_2)} = \gamma_2\boldsymbol{X}_U^T\boldsymbol{X}_U\left(\boldsymbol{X}_U^T\boldsymbol{X}_U+\gamma_2\boldsymbol{I}\right)^{-1}$ used in Equality (26) holds because

$$\begin{aligned} \gamma_2\boldsymbol{X}_U^T\boldsymbol{X}_U &= \gamma_2\boldsymbol{X}_U^T\left(\boldsymbol{X}_U\boldsymbol{X}_U^T+\gamma_2\boldsymbol{I}\right)^{-1}\left(\boldsymbol{X}_U\boldsymbol{X}_U^T+\gamma_2\boldsymbol{I}\right)\boldsymbol{X}_U \\ &= {\boldsymbol{X}_U^{(\gamma_2)}}^T\boldsymbol{X}_U^{(\gamma_2)}\left(\boldsymbol{X}_U^T\boldsymbol{X}_U+\gamma_2\boldsymbol{I}\right). \end{aligned}$$

Thus, the optimal $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma},\boldsymbol{\lambda}}$ from Problem (6) must also solve Problem (9) by Identity (26). ■

### A.2 Geometry Results

**Proposition 3** *Any eigenbasis of the possibly non-symmetric matrix $\boldsymbol{M}^{(\gamma_2)}$ is real with eigenvalues $\tau_1^{(\gamma_2)} \geq \cdots \geq \tau_p^{(\gamma_2)} \geq 0$. Furthermore, $\tau_i^{(\gamma_2)} = 0$ iff $i >$ rank$(\boldsymbol{X}_U)$.*

**Proof** Let $\boldsymbol{X}_L^T\boldsymbol{X}_L = \boldsymbol{O}_L\boldsymbol{D}_L\boldsymbol{O}_L^T$ be the eigendecomposition, assume $\text{rank}(\boldsymbol{X}_L) = p$, and define the linear transformation

$$\tilde{\boldsymbol{w}} = \boldsymbol{D}_L^{1/2}\boldsymbol{O}_L^T\boldsymbol{w} \tag{27}$$

that changes the coordinate basis to $\boldsymbol{O}_L$ and then rescales by $\boldsymbol{D}_L^{1/2}$. The symmetric matrix

$$\widetilde{\boldsymbol{M}}^{(\gamma_2)} = \boldsymbol{D}_L^{-1/2}\boldsymbol{O}_L^T{\boldsymbol{X}_U^{(\gamma_2)}}^T\boldsymbol{X}_U^{(\gamma_2)}\boldsymbol{O}_L\boldsymbol{D}_L^{-1/2} \tag{28}$$

has an orthonormal eigenvector decomposition $\left\{\left(\tilde{\boldsymbol{w}}_i^{(\gamma_2)}, \tau_i^{(\gamma_2)}\right)\right\}_{i=1}^p$, so $\boldsymbol{M}^{(\gamma_2)}$ has the real eigendecomposition $\left\{\left(\boldsymbol{w}_i^{(\gamma_2)}, \tau_i^{(\gamma_2)}\right)\right\}_{i=1}^p$ by the reverse of Transformation (27) because

$$\tau_i^{(\gamma_2)}\tilde{\boldsymbol{w}}_i^{(\gamma_2)} = \widetilde{\boldsymbol{M}}^{(\gamma_2)}\tilde{\boldsymbol{w}}_i^{(\gamma_2)} \iff \tau_i^{(\gamma_2)}\boldsymbol{w}_i^{(\gamma_2)} = \boldsymbol{M}^{(\gamma_2)}\boldsymbol{w}_i^{(\gamma_2)}.$$

Furthermore, $\tau_i^{(\gamma_2)} = \tilde{\boldsymbol{w}}_i^{(\gamma_2)^T} \widetilde{\boldsymbol{M}}^{(\gamma_2)} \tilde{\boldsymbol{w}}_i^{(\gamma_2)} = \boldsymbol{w}_i^{(\gamma_2)^T} \boldsymbol{X}_U^{(\gamma_2)^T} \boldsymbol{X}_U^{(\gamma_2)} \boldsymbol{w}_i^{(\gamma_2)} = 0$ iff $\tau_i^{(\gamma_2)} = 0$. ■

**Proposition 4** *The path $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ as a function of $\gamma_1 \geq 0$ is bounded within a $p$-dimensional parallelotope with corners at each binary linear combination of $\left\{\hat{c}_1^{(\gamma_2)} \boldsymbol{w}_1^{(\gamma_2)}, \ldots, \hat{c}_p^{(\gamma_2)} \boldsymbol{w}_p^{(\gamma_2)}\right\}$. Furthermore, the terminal point as $\gamma_1 \to \infty$ is the corner $\sum_{i=1}^p \mathcal{I}_{\{i > rank(\boldsymbol{X}_U)\}} \hat{c}_i^{(\gamma_2)} \boldsymbol{w}_i^{(\gamma_2)}$ with indicator $\mathcal{I}_{\{\cdot\}}$.*

**Proof** Decomposing $\widehat{\boldsymbol{\beta}}^{(\text{OLS})}$ in Equation (15) onto the real eigenbasis $\left\{\boldsymbol{w}_i^{(\gamma_2)}\right\}_{i=1}^p$ from Proposition 2 and then applying Equation (14) to establish Equation (16) are the main steps. Path $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ goes to the terminal point as $\gamma_1 \to \infty$ because the probability weights $1/\left(1 + \gamma_1 \tau_i^{(\gamma_2)}\right)$ in Equation (16) have limits of 0 or 1 when $\tau_i^{(\gamma_2)} > 0$ or $\tau_i^{(\gamma_2)} = 0$. Next, consider the set of all vectors within the $p$-dimensional parallelotope defined by each binary linear combination of $\left\{\hat{c}_i^{(\gamma_2)} \boldsymbol{w}_i^{(\gamma_2)}\right\}_{i=1}^p$ and those for the $p$-dimensional rectangle defined by each binary linear combination of $\left\{\hat{c}_i^{(\gamma_2)} \tilde{\boldsymbol{w}}_i^{(\gamma_2)}\right\}_{i=1}^p$, where $\left\{\tilde{\boldsymbol{w}}_i^{(\gamma_2)}\right\}_{i=1}^p$ are orthonormal eigenvectors of Matrix (28). Transformation (27) is a bijection from the parallelotope to the rectangle. This bijective mapping replaces the $\boldsymbol{w}_i^{(\gamma_2)}$ on the right of Equation (16) with $\tilde{\boldsymbol{w}}_i^{(\gamma_2)}$, and so $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} \mapsto \boldsymbol{D}_L^{1/2} \boldsymbol{O}_L^T \widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ is clearly within the rectangle. ■

**Proposition 5** *The* $\text{span}\left(\boldsymbol{X}^{(\gamma_2)^T} \boldsymbol{X}^{(\gamma_2)} \boldsymbol{w}_i^{(\gamma_2)}\right) = \bigcap_{j \in \{1,\ldots,p\} - \{i\}} \boldsymbol{w}_j^{(\gamma_2)^\perp}$ $\forall i \in \{1, \ldots, p\}$. *Henceforth, the line* $\text{span}\left(\boldsymbol{X}^{(\gamma_2)^T} \boldsymbol{X}^{(\gamma_2)} \boldsymbol{w}_i^{(\gamma_2)}\right)$ *is called the $i^{th}$ extrapolation direction $\forall i \in \{1, \ldots, p\}$.*

**Proof** If $\left\{\tilde{\boldsymbol{w}}_i^{(\gamma_2)}\right\}_{i=1}^p$ are orthonormal eigenvectors of the Symmetric Matrix (28),

$$\boldsymbol{w}_i^{(\gamma_2)^T} \boldsymbol{X}_L^T \boldsymbol{X}_L \boldsymbol{w}_j^{(\gamma_2)} = \mathcal{I}_{\{i=j\}} \tag{29}$$

$$\boldsymbol{w}_i^{(\gamma_2)^T} \boldsymbol{X}_U^{(\gamma_2)^T} \boldsymbol{X}_U^{(\gamma_2)} \boldsymbol{w}_j^{(\gamma_2)} = \mathcal{I}_{\{i=j\}} \tau_i^{(\gamma_2)} \tag{30}$$

by Transformation (27). Let $\boldsymbol{\nu} \in \text{span}\left(\boldsymbol{X}^{(\gamma_2)^T} \boldsymbol{X}^{(\gamma_2)} \boldsymbol{w}_i^{(\gamma_2)}\right)$. Summing Equations (29) and (30) implies that $\boldsymbol{\nu}^T \boldsymbol{w}_j^{(\gamma_2)} = 0$ and hence $\boldsymbol{\nu} \in \boldsymbol{w}_j^{(\gamma_2)^\perp}$ for each $j \neq i$. Now, let $\boldsymbol{\nu} \in \bigcap_{j \neq i} \boldsymbol{w}_j^{(\gamma_2)^\perp} \subseteq \mathbb{R}^p$, so $\boldsymbol{\nu}^T \boldsymbol{w}_j^{(\gamma_2)} = 0$ for each $j \neq i$. There exists a unique sequence $\{a_k\}_{k=1}^p$ such that $\boldsymbol{\nu} = \sum_{k=1}^p a_k \boldsymbol{X}^{(\gamma_2)^T} \boldsymbol{X}^{(\gamma_2)} \boldsymbol{w}_k^{(\gamma_2)}$ by the assumption $\text{rank}(\boldsymbol{X}_L) = p$, so $\boldsymbol{\nu}^T \boldsymbol{w}_j^{(\gamma_2)} = a_j \left(1 + \tau_j^{(\gamma_2)}\right)$ by Equations (29) and (30). Thus, $a_j = 0$ for each $j \neq i$ and $\boldsymbol{\nu} \in \text{span}\left(\boldsymbol{X}^{(\gamma_2)^T} \boldsymbol{X}^{(\gamma_2)} \boldsymbol{w}_i^{(\gamma_2)}\right)$. ■

**Proposition 6** *If $\gamma_2 > 0$, vectors $\left\{\boldsymbol{\ell}_i^{(\gamma_2)}\right\}_{i=p}^1$ and $\left\{\boldsymbol{u}_i^{(\gamma_2)}\right\}_{i=1}^{rank(\boldsymbol{X}_U)}$ are orthonormal bases for the column spaces of $\boldsymbol{X}_L$ and $\boldsymbol{X}_U^{(\gamma_2)}$, and $\boldsymbol{u}_i^{(\gamma_2)} = \vec{0}$ if $i > rank(\boldsymbol{X}_U)$.*

**Proof** The orthonormality holds by Definitions (17) and Identities (29) and (30). Note $\boldsymbol{u}_i^{(\gamma_2)} = \vec{0}$ if $i > \text{rank}(\boldsymbol{X}_U)$ by Identity (30). The column space result follows from Equation (19) and the joint trained least squares assumption of $\text{rank}(\boldsymbol{X}_L) = p$. ■

**Proposition 7** *For each $i \in \{1,\ldots,p\}$,*

$$\begin{aligned}\boldsymbol{X}_L^T\boldsymbol{\ell}_i^{(\gamma_2)} &= \frac{1}{1+\tau_i^{(\gamma_2)}}\boldsymbol{X}^{(\gamma_2)^T}\boldsymbol{X}^{(\gamma_2)}\boldsymbol{w}_i^{(\gamma_2)}\\ \boldsymbol{X}_U^{(\gamma_2)^T}\boldsymbol{u}_i^{(\gamma_2)} &= \frac{\tau_i^{(\gamma_2)}}{(1+\tau_i^{(\gamma_2)})\sqrt{\kappa_i^{(\gamma_2)}}}\boldsymbol{X}^{(\gamma_2)^T}\boldsymbol{X}^{(\gamma_2)}\boldsymbol{w}_i^{(\gamma_2)},\end{aligned}$$

*so $\boldsymbol{X}^{(\gamma_2)^T}\boldsymbol{X}^{(\gamma_2)}\boldsymbol{w}_i^{(\gamma_2)}$, $\boldsymbol{X}_L^T\boldsymbol{\ell}_i^{(\gamma_2)}$, and $\boldsymbol{X}_U^{(\gamma_2)^T}\boldsymbol{u}_i^{(\gamma_2)}$ are parallel vectors in $\mathbb{R}^p$.*

**Proof** By Definitions (8), (14), and (17),

$$\begin{aligned}\tau_i^{(\gamma_2)}\boldsymbol{w}_i^{(\gamma_2)} &= \boldsymbol{M}^{(\gamma_2)}\boldsymbol{w}_i^{(\gamma_2)}\\ \tau_i^{(\gamma_2)}\boldsymbol{X}_L^T\boldsymbol{X}_L\boldsymbol{w}_i^{(\gamma_2)} &= \boldsymbol{X}_U^{(\gamma_2)^T}\boldsymbol{X}_U^{(\gamma_2)}\boldsymbol{w}_i^{(\gamma_2)} &(31)\\ \tau_i^{(\gamma_2)}\boldsymbol{X}_L^T\boldsymbol{\ell}_i^{(\gamma_2)} &= \sqrt{\kappa_i^{(\gamma_2)}}\boldsymbol{X}_U^{(\gamma_2)^T}\boldsymbol{u}_i^{(\gamma_2)} &(32)\\ \left(1+\tau_i^{(\gamma_2)}\right)\boldsymbol{X}_L^T\boldsymbol{\ell}_i^{(\gamma_2)} &= \boldsymbol{X}^{(\gamma_2)^T}\boldsymbol{X}^{(\gamma_2)}\boldsymbol{w}_i^{(\gamma_2)}. &(33)\end{aligned}$$

Hence, Vectors (31)-(33) are parallel, and the stated identities follow from Equations (32) and (33). ■

### A.3 Performance Bounds

**Theorem 9** *Let Assumptions 1-3 hold. Also, let $q \geq 1$, $\tau_1^{(\boldsymbol{\phi})} > 0$, and $p_i\left(\boldsymbol{\tau}^{(\boldsymbol{\phi})}\right) = \frac{\tau_i^{(\boldsymbol{\phi})^2}\sigma_i^2}{\sum_{j=1}^q \tau_j^{(\boldsymbol{\phi})^2}\sigma_j^2}$. If $\sum_{i=1}^q p_i\left(\boldsymbol{\tau}^{(\boldsymbol{\phi})}\right)\left(\frac{\mu_i\left(c_i^{(\boldsymbol{\phi})}+\boldsymbol{u}_i^{(\boldsymbol{\phi})^T}\boldsymbol{X}_{U\bar{\mathcal{A}}}\boldsymbol{\beta}_{\bar{\mathcal{A}}}/\sqrt{\kappa_i^{(\gamma_2)}}\right)-\mu_i^2}{\sigma_i^2}\right) < 1$, then*

$$\mathbb{E}\left[\left\|\boldsymbol{X}_U\left(\widehat{\boldsymbol{\beta}}_{\gamma_1}^{(\boldsymbol{\phi})}-\boldsymbol{\beta}\right)\right\|_2^2\middle|\boldsymbol{\phi}\right] < \mathbb{E}\left[\left\|\boldsymbol{X}_U\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}^{(SUP)}-\boldsymbol{\beta}\right)\right\|_2^2\middle|\boldsymbol{\phi}\right].$$

**Proof** Let $\gamma_1 \in [0,\delta)$ for $\delta > 0$ from Assumption 2, and define $\boldsymbol{u}_i^{(\boldsymbol{\phi})} = \boldsymbol{X}_U\boldsymbol{w}_i^{(\boldsymbol{\phi})}/\sqrt{\kappa_i^{(\boldsymbol{\phi})}}$, where $\kappa_i^{(\boldsymbol{\phi})} = \tau_i^{(\boldsymbol{\phi})} + \mathcal{I}_{\{i>\mathrm{rank}(\boldsymbol{X}_U)\}} > 0$ and hence $\kappa_i^{(\boldsymbol{\phi})}\tau_i^{(\boldsymbol{\phi})} = \tau_i^{(\boldsymbol{\phi})^2}$. Vectors $\left\{\boldsymbol{u}_i^{(\boldsymbol{\phi})}\right\}_{i=1}^q$ are an orthonormal basis for the column space of $\boldsymbol{X}_U$ by Proposition 6, and

$$\boldsymbol{X}_{U\mathcal{A}}\left(\left(\widehat{\boldsymbol{\beta}}_{\gamma_1}^{(\boldsymbol{\phi})}\right)_{\mathcal{A}}-\boldsymbol{\beta}_{\mathcal{A}}\right) = \sum_{i=1}^q\left(\frac{\hat{c}_i^{(\boldsymbol{\phi})}}{1+\gamma_1\tau_i^{(\boldsymbol{\phi})}}-c_i^{(\boldsymbol{\phi})}\right)\boldsymbol{u}_i^{(\boldsymbol{\phi})}\sqrt{\kappa_i^{(\boldsymbol{\phi})}} \quad (34)$$

by Equations (24) and (25). Next, define loss function

$$Q = \left\|\boldsymbol{X}_U\left(\widehat{\boldsymbol{\beta}}_{\gamma_1}^{(\boldsymbol{\phi})}-\boldsymbol{\beta}\right)\right\|_2^2 = Q_1+Q_2+Q_3, \quad (35)$$

where $Q_1 = \left\|\boldsymbol{X}_{U\mathcal{A}}\left(\left(\widehat{\boldsymbol{\beta}}_{\gamma_1}^{(\boldsymbol{\phi})}\right)_{\mathcal{A}}-\boldsymbol{\beta}_{\mathcal{A}}\right)\right\|_2^2$, $Q_2 = -2\left(\left(\widehat{\boldsymbol{\beta}}_{\gamma_1}^{(\boldsymbol{\phi})}\right)_{\mathcal{A}}-\boldsymbol{\beta}_{\mathcal{A}}\right)^T\boldsymbol{X}_{U\mathcal{A}}^T\boldsymbol{r}$, $Q_3 = \|\boldsymbol{r}\|_2^2$, and $\boldsymbol{r} = \boldsymbol{X}_{U\bar{\mathcal{A}}}\boldsymbol{\beta}_{\bar{\mathcal{A}}}$. If $\gamma_1 = 0$, the supervised estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}^{(\mathrm{SUP})}$ follows, so an improvement is guaranteed if the gradient of $\mathbb{E}\left[Q|\boldsymbol{\phi}\right]$ with respect to $\gamma_1$ evaluated at $\gamma_1 = 0$ is negative.

By Equation (34) and Assumption 3,

$$\begin{aligned}\mathbb{E}\left[Q_1|\boldsymbol{\phi}\right] &= \sum_{i=1}^{q}\mathbb{E}\left[\left(\frac{\hat{c}_i^{(\boldsymbol{\phi})}}{1+\gamma_1\tau_i^{(\boldsymbol{\phi})}}-c_i^{(\boldsymbol{\phi})}\right)^2\middle|\boldsymbol{\phi}\right]\kappa_i^{(\boldsymbol{\phi})} \\ &= \sum_{i=1}^{q}\left(\left(\frac{1}{1+\gamma_1\tau_i^{(\boldsymbol{\phi})}}\right)^2\left(\sigma_i^2+\mu_i^2\right)-2\frac{\mu_i}{1+\gamma_1\tau_i^{(\boldsymbol{\phi})}}c_i^{(\boldsymbol{\phi})}+{c_i^{(\boldsymbol{\phi})}}^2\right)\kappa_i^{(\boldsymbol{\phi})}, \qquad (36)\end{aligned}$$

and the gradient of Equation (36) is

$$\frac{\partial\mathbb{E}\left[Q_1|\phi\right]}{\partial\gamma_1} = -2\sum_{i=1}^{q}\frac{\tau_i^{(\boldsymbol{\phi})}\kappa_i^{(\boldsymbol{\phi})}}{\left(1+\gamma_1\tau_i^{(\boldsymbol{\phi})}\right)^3}\left(\sigma_i^2+\mu_i^2-c_i^{(\boldsymbol{\phi})}\mu_i-\gamma_1\mu_i c_i^{(\boldsymbol{\phi})}\tau_i^{(\boldsymbol{\phi})}\right). \qquad (37)$$

Similarly for the second term $Q_2$ on the right of Equation (35),

$$\begin{aligned}\mathbb{E}\left[Q_2|\boldsymbol{\phi}\right] &= -2\sum_{i=1}^{q}\left(\frac{\mu_i}{1+\gamma_1\tau_i^{(\boldsymbol{\phi})}}-c_i^{(\boldsymbol{\phi})}\right)\sqrt{\kappa_i^{(\boldsymbol{\phi})}}{\boldsymbol{u}_i^{(\boldsymbol{\phi})}}^T\boldsymbol{r} \\ \frac{\partial\mathbb{E}\left[Q_2|\phi\right]}{\partial\gamma_1} &= 2\sum_{i=1}^{q}\frac{\tau_i^{(\boldsymbol{\phi})}\sqrt{\kappa_i^{(\boldsymbol{\phi})}}\mu_i}{\left(1+\gamma_1\tau_i^{(\boldsymbol{\phi})}\right)^2}{\boldsymbol{u}_i^{(\boldsymbol{\phi})}}^T\boldsymbol{r}. \qquad (38)\end{aligned}$$

The third term $Q_3$ on the right of Equation (35) is constant with respect to $\gamma_1$ and thus ignored, and the sum of Scores (37) and (38) with $\gamma_1 = 0$ is negative whenever

$$\begin{aligned}-2\sum_{i=1}^{q}\tau_i^{(\boldsymbol{\phi})}\kappa_i^{(\boldsymbol{\phi})}\left(\sigma_i^2+\mu_i^2-c_i^{(\boldsymbol{\phi})}\mu_i-\mu_i{\boldsymbol{u}_i^{(\boldsymbol{\phi})}}^T\boldsymbol{r}/\sqrt{\kappa_i^{(\boldsymbol{\phi})}}\right) &< 0 \\ \sum_{i=1}^{q}{\tau_i^{(\boldsymbol{\phi})}}^2\left(\mu_i\left(c_i^{(\boldsymbol{\phi})}+{\boldsymbol{u}_i^{(\boldsymbol{\phi})}}^T\boldsymbol{r}/\sqrt{\kappa_i^{(\boldsymbol{\phi})}}\right)-\mu_i^2\right) &< \sum_{i=1}^{q}{\tau_i^{(\boldsymbol{\phi})}}^2\sigma_i^2 \\ \sum_{i=1}^{q}p_i\left(\boldsymbol{\tau}^{(\boldsymbol{\phi})}\right)\left(\mu_i\left(c_i^{(\boldsymbol{\phi})}+{\boldsymbol{u}_i^{(\boldsymbol{\phi})}}^T\boldsymbol{r}/\sqrt{\kappa_i^{(\boldsymbol{\phi})}}\right)-\mu_i^2\right)/\sigma_i^2 &< 1.\end{aligned}$$

■

**Proposition 12** *If $\tau_1^{(\boldsymbol{\phi})} > 0$ and $\boldsymbol{\beta}_i \in \bigcap_{j\in\{1,\ldots,p\}-\{i\}}{\boldsymbol{w}_j^{(\boldsymbol{\phi})}}^{\perp}$ is unit length, then the joint trained ridge performance bound from Corollary 11 satisfies $\sigma_{LB}^2\left(\boldsymbol{\beta}_i\right) \geq \lambda_2 p_i\left(\boldsymbol{\tau}^{(\boldsymbol{\phi})}\right)$ for $i \in \{1,\ldots,p\}$ and $\sigma_{LB}^2\left(\boldsymbol{\beta}_i\right) \geq \sigma_{LB}^2\left(\boldsymbol{w}_j^{(\boldsymbol{\phi})}/\left\|\boldsymbol{w}_j^{(\boldsymbol{\phi})}\right\|_2\right)$ if $j \geq i$.*

**Proof** The desired vectors are $\boldsymbol{\beta}_j = {\boldsymbol{X}_L^{(\lambda_2)}}^T\boldsymbol{\ell}_i^{(\boldsymbol{\phi})}/\left\|{\boldsymbol{X}_L^{(\lambda_2)}}^T\boldsymbol{\ell}_i^{(\boldsymbol{\phi})}\right\|_2$ by Proposition 5, so

$${\boldsymbol{w}_i^{(\boldsymbol{\phi})}}^T\boldsymbol{\beta}_j = \mathcal{I}_{\{i=j\}}/\left\|{\boldsymbol{X}_L^{(\lambda_2)}}^T\boldsymbol{\ell}_i^{(\boldsymbol{\phi})}\right\|_2 \qquad (39)$$

by Equation (29). Constraints (39) imply that only term $i = j$ of $\sigma_{\text{LB}}^2(\boldsymbol{\beta}_j)$ can be nonzero. For any $\boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{\beta} = \sum_{i=1}^{p} c_i^{(\boldsymbol{\phi})}\boldsymbol{w}_i^{(\boldsymbol{\phi})}$ with $c_i^{(\boldsymbol{\phi})} = {\boldsymbol{w}_i^{(\boldsymbol{\phi})}}^T{\boldsymbol{X}_L^{(\lambda_2)}}^T\boldsymbol{X}_L^{(\lambda_2)}\boldsymbol{\beta}$ by Equation (29), so

$c_j^{(\phi)} = \left\| \boldsymbol{X}_L^{(\lambda_2)^T} \boldsymbol{\ell}_j^{(\phi)} \right\|_2$ if $\boldsymbol{\beta} = \boldsymbol{\beta}_j$. These facts can help simplify the bound to

$$\sigma^2_{\mathrm{LB}}(\boldsymbol{\beta}_j) \;=\; \lambda_2 p_j \left(\boldsymbol{\tau}^{(\phi)}\right) \left( 1 + \lambda_2 \frac{\left( \boldsymbol{w}_j^{(\phi)^T} \boldsymbol{w}_j^{(\phi)} - 1/ \left\| \boldsymbol{X}_L^{(\lambda_2)^T} \boldsymbol{\ell}_j^{(\phi)} \right\|_2^2 \right)}{\boldsymbol{w}_j^{(\phi)^T} \boldsymbol{X}_L^T \boldsymbol{X}_L \boldsymbol{w}_j^{(\phi)}} \right)_+ . \tag{40}$$

Next, define $\boldsymbol{G} = \left[ \boldsymbol{w}_i^{(\phi)^T} \boldsymbol{w}_j^{(\phi)} \right]_{i,j=1}^p$ as the Gram matrix of vectors $\boldsymbol{w}_i^{(\phi)}$. Let $\boldsymbol{G}^{(-j)}$ be the $(p-1) \times (p-1)$ sub matrix of $\boldsymbol{G}$ obtained by deleting the $j^{\text{th}}$ row and column, and let $\boldsymbol{G}_j$ be the $1 \times (p-1)$ vector obtained by deleting the $j^{\text{th}}$ entry from the $j^{\text{th}}$ row of $\boldsymbol{G}$. Matrix $\boldsymbol{G}^{(-j)}$ is positive definite by Proposition 3, and it can be shown that $\left( \boldsymbol{w}_j^{(\phi)^T} \boldsymbol{w}_j^{(\phi)} - 1/ \left\| \boldsymbol{X}_L^{(\lambda_2)^T} \boldsymbol{\ell}_j^{(\phi)} \right\|_2^2 \right) = \boldsymbol{G}_j^T \left( \boldsymbol{G}^{(-j)} \right)^{-1} \boldsymbol{G}_j \geq 0$ by Constraints (39). Therefore, Bound (40) further reduces to

$$\sigma^2_{\mathrm{LB}}(\boldsymbol{\beta}_j) \;=\; \lambda_2 p_j \left(\boldsymbol{\tau}^{(\phi)}\right) \left( 1 + \lambda_2 \frac{\boldsymbol{G}_j^T \left( \boldsymbol{G}^{(-j)} \right)^{-1} \boldsymbol{G}_j}{\boldsymbol{w}_j^{(\phi)^T} \boldsymbol{X}_L^T \boldsymbol{X}_L \boldsymbol{w}_j^{(\phi)}} \right) \geq \lambda_2 p_j \left(\boldsymbol{\tau}^{(\phi)}\right). \tag{41}$$

For the second part, define $\nu_{ij} = \frac{\left( \boldsymbol{w}_j^{(\phi)^T} \boldsymbol{w}_i^{(\phi)} \right)^2}{\left\| \boldsymbol{w}_i^{(\phi)} \right\|_2^2 \boldsymbol{w}_i^{(\phi)^T} \boldsymbol{X}_L^T \boldsymbol{X}_L \boldsymbol{w}_i^{(\phi)}} \geq 0$, so

$$\sigma^2_{\mathrm{LB}} \left( \boldsymbol{w}_j^{(\phi)} / \left\| \boldsymbol{w}_j^{(\phi)} \right\|_2 \right) \;=\; \left( \lambda_2 p_j \left(\boldsymbol{\tau}^{(\phi)}\right) - \lambda_2^2 \sum_{i \neq j} p_j \left(\boldsymbol{\tau}^{(\phi)}\right) \nu_{ij} \right)_+ . \tag{42}$$

The result is trivial if Bound (42) is zero, and the difference of Bounds (41) and (42)

$$\sigma^2_{\mathrm{LB}}(\boldsymbol{\beta}_i) - \sigma^2_{\mathrm{LB}} \left( \boldsymbol{w}_j^{(\phi)} / \left\| \boldsymbol{w}_j^{(\phi)} \right\|_2 \right) \geq \lambda_2 \left( p_i \left(\boldsymbol{\tau}^{(\phi)}\right) - p_j \left(\boldsymbol{\tau}^{(\phi)}\right) \right) + \lambda_2^2 \sum_{i \neq j} p_j \left(\boldsymbol{\tau}^{(\phi)}\right) \nu_{ij}$$

is no less than the sum of two non-negative terms if Bound (42) is positive and $j \geq i$. ■

## References

A Aswani, P Bickel, and C Tomlin. Regression on manifolds: estimation of the exterior derivative. *The Annals of Statistics*, 39(1):48–81, 2010.

M Azizyan, A Singh, and L Wasserman. Density-sensitive semisupervised inference. *The Annals of Statistics*, 41(2):751–771, 2013.

M Belkin, P Niyogi, and V Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

G Casella. Minimax ridge regression estimation. *The Annals of Statistics*, 8:937–1178, 1980.

O Chapelle, M Chi, and A Zien. A continuation method for semi-supervised SVMs. In *International Conference on Machine Learning*, 2006a.

O Chapelle, B Schölkopf, and A Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006b. URL `http://www.kyb.tuebingen.mpg.de/ssl-book`.

M Culp. On the semi-supervised joint trained elastic net. *Journal of Computational Graphics and Statistics*, 22(2):300–318, 2013.

J Friedman, T Hastie, and R Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

A Gretton, A Smola, J Huang, M Schmittfull, K Borgwardt, and B Schölkopf. Covariate shift by kernel mean matching. In J Quiñonero-Candela, M Sugiyama, A Schwaighofer, and N Lawrence, editors, *Dataset Shift in Machine Learning*, pages 1–38. The MIT Press, 2009.

T Hastie, R Tibshirani, and J Friedman. *The Elements of Statistical Learning (Data Mining, Inference, and Prediction)*. Springer Verlag, 2009.

M Hein, J Audibert, and U von Luxburg. From graphs to manifolds–weak and strong pointwise consistency of graph Laplacians. In *Conference on Learning Theory*, pages 470–485, 2005.

T Kananmori, S Hido, and M Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, 2009.

M Kuhn. Building predictive models in R using the `caret` package. *Journal of Statistical Software*, 28(5):1–26, 2008.

J Lafferty and L Wasserman. Statistical analysis of semi-supervised regression. In *Advances in NIPS*, pages 801–808. MIT Press, 2007.

M Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

S Mente and F Lombardo. A recursive-partitioning model for blood-brain barrier permeation. *Journal of Computer-Aided Molecular Design*, 19:465–481, 2005.

J Moreno-Torres, T Raeder, R Alaiz-Rodrguez, N Chawla, and F Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 2008.

BG Osborne, T Fearn, AR Miller, and S Douglas. Application of near infrared reflectance spectroscopy to the compositional analysis of biscuits and biscuit doughs. *Journal of the Science of Food and Agriculture*, 35:99–105, 1984.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2015. URL `http://www.R-project.org/`.

T Scheetz, K Kim, R Swiderski, A Philp, T Braun, K Knudtson, A Dorrance, G DiBona, J Huang, T Casavant, V Sheffield, and E Stone. Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, 103(39):14429–14434, 2006.

X Shen, M Alam, F Fikse, and L Rönnegård. A novel generalized ridge regression method for quantitative genetics. *Genetics*, 193(4):1255–1268, 2013. URL `http://www.genetics.org/content/193/4/1255.full`.

M Sugiyama, M Krauledat, and K Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

J Wang, X Shen, and W Pan. On efficient large margin semisupervised learning: Method and theory. *Journal of Machine Learning Research*, 10:719–742, 2009.

J Wang, T Jebara, and S Chang. Semi-supervised learning using greedy max-cut. *Journal of Machine Learning Research*, 14:771–800, 2013.

K Yamazaki, M Kawanabe, S Watanabe, M Sugiyama, and K Müller. Asymptotic Bayesian generalization error when training and test distributions are different. In *Proceedings of the 24th International Conference on Machine Learning*, pages 1079–1086, 2007.

X Zhu and A Goldberg. *Introduction to Semi-Supervised Learning*. Morgan and Claypool Publishers, 2009.

H Zou and T Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 67(2):301–320, 2005.