Statistical learning theory for high dimensional prediction

Running Head: STATISTICAL LEARNING THEORY

Statistical Learning Theory for High Dimensional Prediction:

Application to Criterion-Keyed Scale Development

Abstract

Statistical learning theory (SLT) is the statistical formulation of machine learning theory, a body of analytic methods common in "big data" problems. Regression-based SLT algorithms seek to maximize predictive accuracy for some outcome, given a large pool of potential predictors, without overfitting the sample. Research goals in psychology may sometimes call for high dimensional regression. One example is criterion-keyed scale construction, where a scale with maximal predictive validity must be built from a large item pool. Using this as a working example, we first introduce a core principle of SLT methods: minimization of expected prediction error (EPE). Minimizing EPE is fundamentally different than maximizing the within-sample likelihood, and hinges on building a predictive model of sufficient complexity to predict the outcome well, without undue complexity leading to overfitting. We describe how such models are built and refined via cross-validation. We then illustrate how three common SLT algorithms--Supervised Principal Components, Regularization, and Boosting—can be used to construct a criterion-keyed scale predicting all-cause mortality, using a large personality item pool within a population cohort. Each algorithm illustrates a different approach to minimizing EPE. Finally, we consider broader applications of SLT predictive algorithms, both as supportive analytic tools for conventional methods, and as primary analytic tools in discovery phase research. We conclude that despite their differences from the classic null-hypothesis testing approach—or perhaps because of them--SLT methods may hold value as a statistically rigorous approach to exploratory regression.

Key words: Statistical Learning Theory; Exploratory Data Analysis; Prediction; Generalizability; Psychometrics; Criterion-Keyed Scales

Statistical Learning Theory for High Dimensional Prediction: Application to Criterion-Keyed Scale Development

Machine learning is an area of computer science focused on detecting patterns in data (Kodratoff, 2014). Machine learning has now become firmly ensconced within an underlying foundation of statistical principles, broadly referred to as *statistical learning theory* (SLT) (Hastie, Tibshirani, & Friedman, 2009). These principles give rise to a rich and extensive range of models and methods. SLT methods are gaining increasing popularity with the advent of "big data", which is now immigrating into psychology in several forms. These include, but are not limited to: internet-based data collection tools capable of rapidly generating large samples with many variables (Buhrmester, Kwang, & Gosling, 2011); large national studies assessing numerous psychological factors (Brim & Kessler, 2004); and the movement toward integrated data analysis—that is, either pooling or coordinating analysis across several data sets (Curran & Hussong, 2009; Hofer & Piccinin, 2009).

Laney (2001) suggested that three primary characteristics of "Big Data" are volume, variety, and velocity. Volume refers to the sheer amount of data--in a research setting, the number of participants and the number of variables collected for each participant. Variety reflects different types or kinds of variables, and velocity indicates the speed with which data can be collected (Laney, 2001). To some extent, what appears "big" to a given researcher is a matter of perspective. In comparison to a clinical or convenience sample of a few hundred, involving a few dozen variables, a multi-site or population study involving thousands of persons and a few hundred variables might seem to be "Big Data". Yet many (particularly those in computer science and genomics) are accustomed to far larger datasets, and increases in computing power

over time have continually demoted one period's "Big Data" to the next period's "unremarkably-sized" data. The methods we describe are useful in "Big Data" as well as data of a more standard size.

Big data carry the *curse of dimensionality*—the challenge of dealing with large numbers of variables in an analysis. In some instances, cases, the number of variables may even exceed the number of cases. Traditionally, the curse of dimensionality has been solved in through dimension reduction methods (i.e., factor or principal component analysis, multidimensional scaling). The result is a form of multivariate structure defined by a lower number of dimensions. Within SLT, such methods are called "unsupervised learning" techniques because there is no dependent variable—the goal is simply to ascertain patterns of aggregation within a set of variables (see Hastie et al., 2009, Chapter 14, for an overview). Any problem in which there is a dependent variable is called "supervised learning". We use the term "SLT" here to refer to supervised learning, with the caveat that it also encompasses unsupervised techniques. Supervised learning algorithms are essentially high-dimensional regression models designed to maximize out-of-sample predictive accuracy.

Such models are a marked departure from the use of regression in psychology. Most studies employ a regression model (usually linear) as a vehicle for testing a null hypothesis about a single parameter of interest. There may be covariates in the model, but they are limited in number and usually serve as "controls". Assuming reasonable power and model specification, this approach is a fruitful way to test statistical null hypotheses about a focal predictor variable. Of course this is not the only approach to scientific inquiry, and not all scientific hypotheses translate into statistical null hypotheses about a particular parameter. Situations often require exploration of a large pool of potential predictors, the development of a model with generalizable

predictive power, or both. At least three distinct varieties of research problems pose such challenges.

In the first case, the research question cannot be addressed without a statistical "fix" to methodological problems like selection effects, missing data, or sample weighting. The "fix" involves potentially high dimensional regression to correct or ameliorate the problem with the data, so that analysis of the substantive question can proceed. A poor model will yield a poor correction, jeopardizing the validity of the subsequent analysis. Second, the substantive research question may directly call for high dimensional regression. For instance, a researcher may be interested in anything related to a single outcome variable, rather than any specific predictor. The outcome is of such importance that a thorough understanding of everything predicting it is needed. A third case is when a "general" hypothesis exists about a broad kind or type of predictor. The class of predictors, however, involves numerous specific variables, and the state of knowledge is such that specific hypotheses cannot be credibly forwarded. We return to these kinds of research problems later, and now turn to a specific example combining elements of the latter two cases: criterion-keyed scale development.

Criterion-keyed scales are meant to predict a particular outcome, or criterion (Anastasi & Urbina, 1997). Developing such a scale presupposes a particular outcome of importance, and a "general hypothesis" about the kind or type of items to which it might be linked. Such scales are fundamentally different than those intended to measure a latent trait. Examples include the well-known Minnesota Multiphasic Personality Inventory basic scales, developed to predict psychiatric diagnosis (Butcher, Dahlstrom, Grahamn, Tellegen, & Kraemmer, 1989); adaptations of "Big Five" personality scales to predict job performance (Ones, Chockalingham, & Dilchert,

2005) and scales to predict health outcomes like cardiovascular disease, such as the Cook-Medley Hostility Scale short form (Barefoot, Dodge, Peterson, Dahlstrom, & Williams, 1989).

This paper is divided into three sections. First, we provide an overview of a core statistical principle underlying most SLT methods for high dimensional regression: the minimization of expected prediction error (EPE). Second, we discuss three common SLT algorithms: Supervised Principal Components (SPCA), regression regularization, and regression boosting. Each pursues the minimization of EPE in a different way. We illustrate each method by constructing a scale from the Eysenck Personality Inventory (EPI; Eysenck & Eysenck, 1964) item pool to predict 25-year all-cause mortality. Third, we consider the general use of SLT algorithms in psychology, including their contrasts with conventional methods and their potential applications.

The Foundational Principle of Predictive Generalizability

Mean Square Error and the Bias-Variance Trade-Off

SLT regression methods differ from classical statistical approaches in a number of ways, summarized in Table 1 (Breiman, 2001). Most, if not all of these differences arise from the SLT priority of maximizing out-of-sample predictive accuracy. Optimizing out-of-sample prediction is often neither the focus nor outcome of traditional null-hypothesis testing methods. To motivate discussion of SLT methods, consider the following general modeling framework. Let m be an arbitrary statistical model in a sample of n observations, relating a group of p predictors ($X_1, X_2, \ldots X_p$) in an $n \times p$ design matrix $\mathbf{X}$ to an $n \times 1$ outcome vector $\mathbf{y}$.[1] The model m maps the sample values in $\mathbf{X}$ to $\mathbf{y}$, through some function $f_m(\mathbf{X})$. Usually, the function involves a $p \times 1$ vector of parameters $\boldsymbol{\beta}$ which are estimated, yielding a corresponding vector of parameter

estimates $\widehat{\boldsymbol{\beta}}$. Often, $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$ are (p+1) x 1, since there is an additional parameter for the intercept in most models. Each person i's predicted value is computed by summing i's values on the p predictor variables, multiplied by the corresponding parameter estimates in $\widehat{\boldsymbol{\beta}}$**.** The resulting quantity is simply the weighted linear combination one would find in any regression context. In the generalized linear model literature this weighted linear combination is called the *linear predictor* for i (Hardin, Hilbe, & Hilbe, 2007), and denoted $\eta_i = \mathbf{x}_i^{\mathbf{T}}\widehat{\boldsymbol{\beta}}$. Note that an individual i's observed scores on some scale can be written as $\mathbf{x}_i^{\mathbf{T}}\widehat{\boldsymbol{\beta}}$, where $\mathbf{x_i}$ is the vector of item scores and $\widehat{\boldsymbol{\beta}}$ a vector of 1s.

In a likelihood framework, a probability distribution is chosen for the dependent variable, and (usually) its mean is parameterized as conditional on the predictors via $\boldsymbol{\beta}$**.** A set of estimating equations is then used to estimate the parameter values that maximize the log likelihood of the sample data. These Maximum Likelihood Estimates (MLEs) are often (but not always) unbiased (Cassella & Berger, 2002). While there are a variety of criteria for evaluating estimators, perhaps the most commonly utilized one is *accuracy.* An estimator's accuracy is defined as the inverse of its Mean Square Error (MSE). The MSE is the squared expectation of the difference between the estimator and the population parameter it seeks to estimate (Cassella & Berger, 2002). Formally, consider an estimator $\hat{\theta}$ of a population parameter θ (θ could be any kind of parameter, but in the present context consider a regression parameter β). The MSE of $\hat{\theta}$ has the following canonical decomposition (Cassella & Berger, 2002):

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}) = \mathrm{E}(\hat{\theta} - \theta)^2 = \mathrm{E}(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2 = \mathrm{E}\left(\hat{\theta} - \mathrm{E}(\hat{\theta})\right)^2 + (\mathrm{E}(\hat{\theta}) - \theta)^2 \\ = \mathrm{Var}(\hat{\theta}) + (\mathrm{Bias}(\hat{\theta}))^2 \qquad (1)\end{aligned}$$

The importance of (1) lies in the fact that an estimator's bias and variance are independent contributors to its overall accuracy. Unbiasedness simply refers to whether the

estimator produces an estimate equal to the parameter in probability theory expectation, or on average. The estimator may or may not show great variability around this average. Thus, unbiased estimators are not always the most accurate, because they may have large variance (see Cassella & Berger 2002, Chapter 7). As a result, accuracy in an estimator can be improved even when bias increases, if declines in variance occur at a rate greater than increases in the square of its bias.

Figure 1 depicts this idea graphically, displaying the sampling distributions of two estimators for a given parameter. The true value of the parameter is 3, represented by a vertical line. The sampling distribution of estimator A at the top has an expectation of 3, or takes that value on average—so it is unbiased. However, its sampling distribution has a standard deviation of 7, so its MSE is (0 bias + 49 variance) = 49. Estimator B on the bottom trades a small upward bias for reduced variance: its expectation is 5, but also it has a standard deviation of 5. Hence, its MSE is $(3\text{-}5)^2 + 25 = 29$. Thus, estimator B is more accurate despite its bias.

The independence of an estimator's bias and variance has an important implication when one strives to maximize a model's predictive power (Cassella & Berger, 2002). Maximizing prediction is equivalent to increasing variance explained. If one naively wants to increase R^2 or similar likelihood-based measures of model fit, one can do so by adding more and more predictors. The only technical restriction on the number of p (independent) parameters usually estimable is $p < n\text{-}1$ (or $p < n$ for models like Cox regression which have no intercept). In this context, interactions, polynomial terms, and other transformations of **X** variables count as separate predictors, so a relatively small number of conceptually distinct variables can quickly grow into a high dimensional regression scenario when one moves beyond linear main effects.

The reason that adding more and more predictors of even negligible importance can increase "variance explained" in the outcome lies in the fact that maximizing R^2–like measures is equivalent to minimizing $\text{Var}[Y|X_1, X_2, \ldots X_p]$. As long as X is sufficient statistic (a random variable is a sufficient statistic for itself), an important result known as the Rao-Blackwell theorem guarantees that $\text{Var}[Y|X] \leq \text{Var}[Y]$ (see Cassella & Berger, 2002, Chapter 7). This means that conditioning on more and more X's almost always drives down the conditional variance of the outcome, driving up variance explained by $f_m(\mathbf{X})$. With an unbiased estimator and the ability to so easily enhance the model's apparent predictive power, why not simply "force feed" more and more independent variables into a model?

Such a strategy risks *overfitting* $f_m(\mathbf{X})$ to the sample. Such a model has become erroneously complex, and will not predict the outcome very well in other samples. This is commonly understood at an intuitive level, and discouraged in practice by the use of adjusted R^2 or information criteria penalizing fit for model complexity (Vrieze, 2012). Overfitting happens because the more predictors that are added, the greater the chance that one or more parameter estimates may lie far away from the true value of the population parameter. Recall that unbiasedness means that the estimator yields estimates equal to the population parameter *on average*. Within any given sample, an unbiased estimator may produce an estimate that lies far from the true value of the parameter, or out at the tails of the sampling distributions in (1). The more predictors added to $f_m(\mathbf{X})$, the more estimates within its $\widehat{\boldsymbol{\beta}}$ are subject to this risk. And the more estimates in a model that deviate substantially from the corresponding population parameters, the more error is transmitted into the model's predicted values of the outcome. Developing an $f_m(\mathbf{X})$ that explains as much outcome variance as possible without overfitting the

sample is thus a difficult task. SLT approaches this problem by strategically trading bias for variance reductions, through estimation procedures that minimize Expected Prediction Error.

Expected Prediction Error

A model $f_m(\mathbf{X})$ acts as an estimator of the outcome Y. In a criterion-keyed scale context, $f_m(\mathbf{X})$ is a scale score based on items relevant to the outcome or criterion Y, arranged in a weighted linear combination. The weights are determined by regression, and depending on the type of model employed, $f_m(\mathbf{X})$ may involve a final transformation, such as exponentiation from a logit to an odds ratio if a logistic regression has been used. In any context, since $f_m(\mathbf{X})$ acts as an estimator of Y, its accuracy can be evaluated by expected prediction error (EPE; Hastie et al., 2009, Chapter 7). Denote the model estimate of Y at a particular set of inputs x_0 as $f_m(x_0)$. Expected Prediction Error (EPE) of the model at the particular set of values of X variables defined by x_0 is:

$$\mathrm{EPE}[f_m(x_0)] = \sigma_\varepsilon^2 + [\mathrm{E}[f_m(x_0)] - y_0]^2 + \mathrm{E}[\mathrm{E}[f_m(x_0)] - f_m(x_0)]^2 \qquad (2)$$

Where the first term on the right hand side of (2), σ_ε^2, is "irreducible error" that cannot be eliminated. Statistically, this reflects variation of the outcome about its average value that can never be explained. One might consider it an inherent property of the sample that cannot be eliminated by a model. The second term, $[\mathrm{E}[f_m(x_0)] - y_0]^2$, is the *squared bias* of the model-based estimate, or the extent to which the model systematically over- or under-estimates the outcome at the particular set of inputs x_0. The third term is the *variance* of the model-based estimate, or the degree of random variation about its average prediction at the set of inputs

E[$f_m(X_0)$]. These two sources of error are analogous to those of an estimator above, and in this case are tied to a particular model.

Note that in a scale-score context, $f_m(x_0)$ is merely a particular score on the scale corresponding to a pattern of responses on the scale's items. More than one individual observation may share a particular set of inputs, or have identical values of the X variables. Of course, the more numerous the X variables and/or smaller the sample, the less likely this will occur. Taken across the entire possible range of inputs on all X variables, EPE is conceptually akin to outcome variation *unexplained* by the model. In linear models, this is sometimes called the coefficient of alienation, 1 - R^2.

Minimization of EPE via Cross-Validation

EPE may not be minimized by the parameter estimates that maximize the likelihood of the sample data. This is most likely to be the case when the ratio of predictors *p* to observations *n* is large, although "rules of thumb" about *p*/*n* are at best tentative since every situation is different. At small *p*/*n* ratios, MLEs may perform reasonably well in out-of-sample predictions; we remark on this issue later. SLT methods represent a reaction to the "over-optimism" of MLEs, or their tendency to overfit samples in cases of larger *p*/*n*. Hastie and colleagues (2009; Chapter 7) provide a good overview of a variety of model selection strategies based on the minimization of EPE. A classic article-length introduction to cross-validation from a traditional statistical standpoint is (Harrell, Lee, & Mark, 1996), and a technical SLT-oriented overview can be found in (Arlot & Celisse, 2010). Here we focus on k-fold cross-validation, one of the most popular cross-validation strategies used to minimize EPE.

In k-fold cross-validation, a sample is split into *k* different segments, usually of equal size, with *k* commonly equal to five or ten. The parameters in $f_m(\mathbf{X})$ are estimated from the data

in *k*-1 folds, and then the model predictions are computed in the left-out fold and compared to the actual outcome value in that fold. The average residual in the left-out fold is the out-of-sample prediction error for that fold. This is repeated over all possible combinations of *k*-1 folds, and the out-of-sample prediction error from the left-out folds are averaged to yield an estimate of Generalized Cross Validation (GCV) error, or $\varepsilon_{\mathrm{GCV}}$ (discussed below). Table 2 schematizes this process for *k* = five.

In a large dataset, random allocation to folds will tend to produce similar folds, just as random assignment in experiments is presumed to create "equivalent" groups. If the split is repeated (without the same seed for the random number generator), the actual observations comprising each fold will be different, but randomization will still tend to create equivalent folds. GCV error is also an average across the folds, and will be more reliable from split to split than the error estimate of any particular fold. Thus, in a large sample, repeated k-fold cross validation will show relatively tight variation around a central tendency in the estimate of GCV error. Depending on the size of the sample, it may be difficult to have confidence that such a result obtains without some investigation. Since samples have finite size, there are a finite number of k-fold splits, and in theory each k-fold split could be examined and the exact distribution of the GCV error determined. At maximal *k*, where *k*=n, the well-known "jack knife" or leave-one-out cross-validation procedure results and the distribution of each observation's error (when left out of estimation) can be directly determined. With *k* in standard ranges (e.g., 3, 5, 10), obtaining the exact distribution of GCV error from every possibly split would be computationally prohibitive unless the sample is very small. Thus, an empirical sampling distribution of the GCV error could be estimated through repeating the procedure a large number of times. Analysis of this distribution reveals the extent to which the GCV error estimate

fluctuates strongly across splits, and may suggest focusing on the model producing the lowest mean GCV error across repeated splits, or reducing k—perhaps even to two—to obtain less variability from split to split. Arlot and Celisse (2010) consider these and other issues in detail.

While Hastie and colleagues tend to rely heavily on cross-validation in their work, others have been critical of it (Breiman, Friedman, Stone, & Olshen, 1984; Kuhn & Johnson, 2013). Breiman et al. (1984) has argued that GCV error may still yield models with less-than-optimal out-of sample performance, and has advocated a rule of thumb based on parsimony. His approach involves first identifying the key model parameters (called "tuning parameters", discussed below) that minimize GCV error. Then, the simplest model that lies within one standard error of the GCV error estimate is used.

When cross-validation methods are used in tandem with "training/test" sample splits, as is common in the SLT literature, confusion can easily arise over what part of the data is being used for what purpose. We thank an anonymous reviewer for suggesting the following description to clarify this issue. First, an overall data set is divided into two subsets of data - a training sample and test sample. The test sample is set aside to provide an independent estimate of model performance. We also refer to this as the hold-out sample, another term seen in the literature to emphasize the fact that this data is held apart from model development. The training sample is then used for model fitting, with the goal of producing the "best possible" model. Often, the training sample itself is partitioned into multiple parts or "folds" to conduct k-fold cross-validation estimation of GCV error. Bootstrap and other cross-validation methods are also conducted within the training sample. These strategies attempt to mimic model performance in independent data, in order to identify the values of tuning parameters most likely to lead to the best performance in actual independent data. Once cross-validation work has been conducted and

an optimal set of model parameters determined, assessment of the model's real performance in independent data is the final step. That assessment is conducted in the test sample.

We turn now to a more detailed discussion of GCV error.

Generalized Cross-Validation Error

EPE is a theoretical quantity that must be is estimated by generalized cross-validation (GCV) error (Hastie et al., 2009). GCV error can be defined generically as:

$$\varepsilon_{\text{GCV}} = g(\mathbf{y}_{cv}, f_{\text{fit}}(\mathbf{X_{cv}})) \tag{3}$$

where $\mathbf{y}_{cv}$ is an $n \times 1$ vector of values of the outcome in the test data, and $\mathbf{X_{cv}}$ is an $n \times p$ matrix of predictor variables in the test data. The function $f_{\text{fit}}(\cdot)$ is a model fit in the training data, and $g(\cdot)$ is a *loss function* reflecting misfit between model-based predictions and actual values of the outcome in the cross-validation data. The form of the loss function depends on the distribution of the outcome. Three frequently used loss functions are:

$$g_{ss}\big(\mathbf{y}, f_{\text{m}}(\mathbf{X})\big) = \sum_i^n (\text{y}_i - \mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta})^2 \tag{4}$$

$$g_{abs}\big(\mathbf{y}, f_{\text{m}}(\mathbf{X})\big) = \sum_i^n |\text{y}_i - \mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta}| \tag{5}$$

$$g_{gini}\big(\mathbf{y}, f_{\text{m}}(\mathbf{X})\big) = \hat{\text{p}}_{\text{v}}(1 - \hat{\text{p}}_{\text{v}}) \tag{6}$$

where (4) is sum-of-squares error, (5) is sum of absolute deviation errors, and (6) is the Gini Index for binary outcomes. These are general functions that can be used for any estimation problem, so they are not specifically subscripted CV. For (6), the fit model yields some probability of membership for individual *i* in a particular class *v*. Person *i* is assigned a predicted

class if the probability exceeds a certain threshold, and the proportion of correctly classified observations in class *v* is $\hat{p}_v$. Thus the Gini Index is simply the proportion of observations correctly classified, multiplied by the proportion of incorrectly classified. Other loss functions may be used in computing GCV error, such information criteria, the (log) Likelihood of the test data, or some derivation of it like the Deviance or an information criterion. As mentioned above in the context of k-fold cross-validation, loss functions are usually averaged across different folds to estimate average ε_{GCV}.

Once a training/testing strategy has been selected and a loss function determined to compute ε_{GCV}, an SLT algorithm can be used to estimate a model. Cross-validation strategies find values for model "tuning parameters" by minimizing ε_{GCV}, and these tuning parameters in turn structure other aspects of the model, such as the number of predictors selected or the extent to which their coefficients are altered. Tuning parameters are algorithm specific, and discussed below. The term "learning" in SLT comes from the "feedback" process about cross-validation prediction on which estimation is based.

Example SLT Algorithms

An extremely large number of SLT algorithms exist, with new variations continually emerging and journals such as the *Journal of Machine Learning Research* dedicated to the field. We illustrate three "basic" algorithms that are relatively common in the literature, each taking a different approach to minimizing EPE. Each also involves an element of potential familiarity to psychology: Supervised principal components, as its name suggests, incorporates principal components analysis (PCA); regularized regression has a connection to ridge regression; and boosted regression involves "residualization" of an outcome, a common technique used to

"regress out" or "partial" one variable from another (Cohen, Cohen, West, & Aiken, 2003). Most techniques extend from linear models to generalized linear models and other types, as shown in illustrations using the Cox proportional hazard model.

The example we present focuses on constructing a subscale embedded within the Eysenck Personality Inventory (EPI) (Eysenck & Eysenck, 1964). The EPI contains 57 items measuring the broad traits of Neuroticism and Extraversion, as well as a lie scale. The goal of our example is to create a scale from these items that predicts 25-year all-cause mortality, a pursuit of interest in health and personality psychology (Chapman, Roberts, & Duberstein, 2011). A general hypothesis motivating the effort is that some items on the EPI scales (even the lie scale) will predict mortality. However it is not clear which items will do so best, placing the project well within the realm of exploratory, rather than confirmatory, research. The data comes from the Health and Lifestyle Survey (HALS), a UK-wide study in which participants were interviewed and completed a number of questionnaire instruments in 1984. Survival status was assessed via the UK Registrar General's office in 2009 (Blaxter, 1987). The data set consists of HALS participants 40 and over, with EPI data: 3709 people (54% female, age M= 58 years, SD = 11.6, 49% deceased by 2009). We partitioned the dataset into a sample of 2472 that could be used for training and cross-validating the model. The remaining 1237 were put aside as a test or hold-out sample to provide an independent assessment of the final scale's predictive validity. The allocation represents a 2/3-1/3 split and is based on no hard and fast rules other than the fact that a larger training sample can accommodate more and/or larger cross-validation folds.

Algorithm 1: Supervised Principal Components

Background and motivation. Supervised Principal Components Analysis (SPCA) (Bair, Hastie, Debashis, & Tibshirani, 2006) is an adaptation of principal components regression. The

intuition is that one will achieve better prediction from a pool of predictors by first screening out those unrelated to the outcome, instead of simply creating principal components based on all possible predictors. The technique was developed for so-called $p > n$ problems, where the number of predictors exceeds the number of observations (Bair et al., 2006). However, it is useful whenever one is faced with a large number of potential predictors. For instance, Weiss and colleagues recently employed it to identify items of the Minnesota Multiphasic Personality Inventory (MMPI) that predict mortality (Weiss, Gale, Batty, & Deary, 2013).

Formally, the goal is to take the n (observations) × p (items) matrix $\mathbf{X}$ and partition it into an $n \times q$ submatrix $\mathbf{X}_{\boldsymbol{\tau}}$, with columns containing scores on the q items related to the outcome, and an $n \times r$ submatrix $\mathbf{X}_{\mathbf{D}}$ with columns containing scores on items unrelated to the outcome that will be discarded from further use. The p items are ranked based on the absolute value of their univariate regression coefficients predicting the outcome. Some threshold τ is set so that only the q items with coefficients > $|\tau|$ are selected. The threshold is the value that minimizes ε_{GCV} during k-fold cross-validation. After selecting the items, the $\mathbf{X}_{\boldsymbol{\tau}}$ matrix is subjected to the singular value decomposition (Bair et al., 2006):

$$\mathbf{X}_{\boldsymbol{\tau}} = \mathbf{U}_{\boldsymbol{\tau}}\mathbf{D}_{\boldsymbol{\tau}}\mathbf{V}_{\boldsymbol{\tau}}^{\mathbf{T}} \tag{7}$$

where $\mathbf{U}_{\boldsymbol{\tau}}$ in an $n \times r$ matrix with the columns consisting of the r principal components of $\mathbf{X}_{\boldsymbol{\tau}}$, $\mathbf{D}_{\boldsymbol{\tau}}$ is an $r \times r$ diagonal matrix of singular values in which $d_1 > d_2 > \ldots d_r > 0$, and $\mathbf{V}_{\boldsymbol{\tau}}$ is an $r \times q$ matrix.[2] Unless there is linear dependence between two items, there are as many principal components as items, so $r = q$ even though not all r will be used. For each of the $j = 1 \ldots r$ principal components (typically $r \leq 3$), the j^{th} component score for observation i is computed as

$u_{ij} = \sum_{k=1}^{q} \alpha_{jk} x_{ik}$ (Bair et al., 2006). This is simply a familiar regression-weighted component score with α_{jk} as the component *j* scoring coefficient for item *k*. In the case of unrotated components, α_{jk} is a loading in the pattern matrix (see Grice & Harris, 1998). Subject *i*'s scores on the first *j* components denoted by the 1 × *j* vector $\mathbf{u_{ij}}$, are then used to predict his or her outcome value y_i (Bair et al., 2006):

$$\hat{y}_i = f(\hat{\beta}_0 + \mathbf{u_{ij}^T}\hat{\boldsymbol{\beta}}) \tag{8}$$

where *f* is the model regression function (often a generalized linear model involving a link function), $\hat{\beta}_0$ the intercept (if the model has one), and $\hat{\boldsymbol{\beta}}$ is a *j* × 1 vector of parameter estimates. In the context of a criterion-keyed scale, $\hat{y}_i$ is persons *i's* scale score. It is a model-based estimate of the outcome, and computed from the *q* items, their loadings on the *r* retained components, and the parameter estimates for those coefficients—that is, substituting the expression for component scores into (8):

$$\hat{y}_i = f(\hat{\beta}_0 + \sum_{j=1}^{r} \hat{\beta}_j (\sum_{k=1}^{q} \alpha_{jk} x_{ik})) \tag{9}$$

The threshold parameter used to select items is one key tuning parameter in SPCA. The other is the number of components to extract from $\mathbf{X_\tau}$.[3] These tuning parameters are determined by fitting several SPCA models with a series of threshold values and one, two, or three components. The threshold value and number of components that jointly minimize GCV error are then chosen. Typically, one, two, or three components are extracted, based on whether one, two, or three minimize ε_{GCV} (Bair et al., 2006). Rotation of components would destroy the

orthogonality intended by the procedure, so is not used. Instead, unrotated loadings from the original, independent components are used. Rotation would redistribute variance, possibly evening out predictive power of components, but they are all added together ultimately in (9) to produce a single predicted value, or scale score. The advantage of extracting only one component is that the first component often dominates, and limiting the number of components minimizes complexity.

Also note that the linear combination of predictors within $f(\cdot)$ in (9) is simply the linear predictor η_i in a generalized linear model. In the context of a criterion-keyed scale, this quantity represents an untransformed scale score based on the q items selected for the scale. When filtered through the link function of $f(\cdot)$, the linear predictor is transformed to $\hat{y}_i$, which is on the metric of the actual outcome. In a scale score context, this can be thought of as a transformed scale score expressed in the units of the criterion. For instance, if Y is income in dollars, with f a log link function, $\beta_0 + \sum_{\mathbf{j=1}}^{\mathbf{r}} \beta_j (\sum_{k=1}^{q} \alpha_{ij} x_{ij})$ in (9) is an untransformed scale score reflecting the natural logarithm of a dollar amount. In contrast, $\exp(\beta_0 + \sum_{\mathbf{j=1}}^{\mathbf{r}} \beta_j (\sum_{k=1}^{q} \alpha_{ij} x_{ij}))$ is a transformed scale score in the metric of dollars.

Illustration. We applied SPCA to the item pool of the EPI to construct a scale criterion-keyed predicting all-cause mortality. Analyses were performed using the R package **superpc** (Bair & Tibshirani, 2010). Sample code for this and other examples appears in Appendix A. In the first step, all 57 EPI items were screened for their association with mortality using Cox proportional hazard models. After this step, items above a log hazard threshold determined by 10-fold cross-validation were subjected to PCA. Final scale scores were computed from component scoring coefficients and regression coefficients as in (9) (although note in Cox

models there is no intercept). The metric of this scale score is a log hazard rate which, when exponentiated, produces a hazard ratio for all-cause mortality over the follow-up period.

The location of the optimal threshold for the log-hazards of the 57 items is shown in the right portion of Figure 2. The test-data likelihood ratio (vs. a null model) was used as the measure of GCV error, and was maximized by a threshold of a log-hazard of |1.93|. Sixteen items lay beyond this threshold. In the right hand portion of Figure 2, the test-data likelihood ratio corresponding to one, two, and three principal components across various thresholds for item inclusion is shown. As can be seen GCV error is minimized by two or three components at the selected threshold of |1.93|. A small difference appears in the gains from a third component. In general, some judgment is required around whether added complexity in any model is justified by any prediction gains. In this particular context, an additional linear combination (component score) would add substantially to the complication of scale scoring. Therefore, we opted not to include it. Table 3 shows component scoring coefficients, corresponding to pattern matrix loadings for the two retained components, in the columns “SPCA 1” and “SPCA 2”. The log hazard rates for each component with respect to 25-year all-cause mortality are at the bottom of the SPCA 1 and 2 columns. All analyses up to this point were conducted in the training sample. Once this final model had been developed, we used it to generate scale scores and examined their accuracy in the test sample.

Table 4 shows various measures of predictive accuracy for the SPCA scale scores in the first column, in both the training/test and hold out samples. The peak accuracy is the total percentage of correct positive and negative classifications, using the scale cut-point that maximizes sensitivity and specificity. The point-biserial correlation is a familiar validity coefficient metric (McGrath & Meyer, 2006), with Spearman’s rank-order correlation coefficient

included for comparison. Also included is the hazard rate for death associated with a 1 SD difference in scale scores (i.e., a one unit increase in a scale *z*-score). When this value is much smaller in new data, one or more regression coefficients have been over-estimated in the fitting sample (Harrell et al., 1996). A relatively similar HR (or log HR) in independent data suggests good generalization. The pseudo R^2 is the Cox-Snell version, $R^2 = 1 - \exp[\frac{-2*(LL_{null} - LL_{model})}{N}]$ where *LL* subscripted "null" and "model" indicate the log likelihood of a null and the fit model, respectively. The *p*-values for these associations are very low and omitted because they are uninformative. The area under the curve (AUC) represents the percentage of true positives (i.e., predicted decedents who actually were deceased) out of all predicted positives, averaged across all possible cut-points for a positive prediction.

The key point in Table 4 is that the SPCA-constructed scale shows comparable predictive validity in both the training and test samples. SPCA thus effectively avoided overfitting by effective minimization of an estimate of EPE, constructing a? model that produces a criterion-keyed scale with generalizable predictive validity.

Algorithm 2: Regularization

Background and Motivation. A second family of algorithms tries to minimize EPE by shrinking regression coefficients based on their instability, and is called "regularization". In a standard MLE context, instability is reflected in the size of a coefficient's standard error. In likelihood theory, imprecision is determined by the curvature of the (log) likelihood function at its maximum (Gould, 2006). The parameter estimates solve a set of estimating equations, written in vector form as:

$$\widehat{\boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial}{\partial \widehat{\boldsymbol{\beta}}} \log(L(\mathrm{y_i}, f_m(\mathrm{x}_i))) = 0 \qquad (10)$$

where $L(\mathrm{y_i}, f_m(\mathrm{x}_i))$ is the likelihood for an observation i , there are $i = 1 \ldots n$ independent observations and partial derivatives are with respect to the other parameters in $\widehat{\boldsymbol{\beta}}$. The Hessian, or matrix of second derivatives, describes the curvature of the log likelihood function at its optimum:

$$\mathbf{H}(\widehat{\boldsymbol{\beta}}) = \frac{\partial}{\partial \widehat{\boldsymbol{\beta}} \partial \widehat{\boldsymbol{\beta}}^{\mathbf{T}}} [\sum_{i=1}^{n} \frac{\partial}{\partial \widehat{\boldsymbol{\beta}}} \log(L(\mathrm{y_i}, f_m(\mathrm{x}_i)))] \tag{11}$$

The negative inverse of the Hessian, the Fisher information matrix $\mathbf{I}(\widehat{\boldsymbol{\beta}}) = -\mathbf{H}(\widehat{\boldsymbol{\beta}})^{-1}$ yields (squared) standard errors for the model coefficients. When the likelihood function is rounded or nearly flat in a particular dimension corresponding to some particular $\hat{\beta}$, a relatively wide range of estimates for that parameter describes the data almost equally well. The small second derivative for $\hat{\beta}$ produces a large Information Matrix standard error, reflecting uncertainty about the precise value of $\hat{\beta}$. Since model predictions are then based on $\hat{\beta}$, its imprecision erodes out-of-sample prediction accuracy.

By shrinking coefficients proportional to a model's GCV error, regularization methods "damp down" the deleterious effects of predictors with large standard errors. For generalized linear models, shrunken parameters are achieved by optimizing a penalized log likelihood of the form (Park & Hastie, 2007):

$$\widehat{\boldsymbol{\beta}}_{\delta} = arg \min_{\widehat{\beta}} \{-\log[\mathrm{L}(\mathbf{y}; \widehat{\boldsymbol{\beta}}) + \delta]\} \tag{12}$$

where log $L(\mathbf{y}; \widehat{\boldsymbol{\beta}})$ is the log likelihood of **y**, the outcome vector in the sample, given the vector of parameter estimates $\widehat{\boldsymbol{\beta}}$, and δ is a penalty term. More generally, δ keeps the sum of all $j = 1 \ldots p$ regression coefficients in $\widehat{\boldsymbol{\beta}}$ from exceeding a specified total. Two common forms for δ under this general scenario are the so-called L1 and L2 penalties (Park & Hastie, 2007):

$$\delta_{L1} = \lambda \sum_{j=1}^{p} |\hat{\beta}_j| \quad (13)$$

$$\delta_{L2} = \lambda \sum_{j=1}^{p} \hat{\beta}_j^2 \quad (14)$$

where λ is between 0 and 1. In practice, λ is found by trying a series of values between 0 and 1, and selecting the value producing the lowest GCV error. Penalization with the L1 term of (13), referred to as "Lasso" regression (Tibshirani, 1996), subtracts from the likelihood some portion of the summed absolute values of the regression coefficients. The L2 penalty, which corresponds to ridge regression (Hoerl & Kennard, 1988), subtracts a portion of the sum of squared coefficients. If λ is 0, δ disappears from (12) and the parameters are standard MLEs. As λ approaches 1, the penalty term reflects the full sum (L1) or sum of squares (L2) of coefficients.

How does this penalty term reflect the imprecision of predictors? If $\hat{\beta}$s are large but also have large standard errors, they will lead to high ε_{GCV} in test data. As a result, a higher value of λ will be chosen. Since this creates a larger penalty term for the likelihood, smaller values of $\hat{\beta}$ will be needed to maximize the penalized likelihood. Because the set of coefficients as a whole cannot exceed a certain size, predictors with large standard errors will be "crowded out" by predictors with smaller standard errors. With an L1 penalty, the worst predictors may have their coefficients set to 0 in order to improve the penalized likelihood, ejecting them from the model completely. The Lasso model (L1 penalty) thus performs both variable selection and shrinkage,

while the ridge model (L2 penalty) performs only shrinkage. However, a drawback is that the L1 penalty tends to arbitrarily select only one of several correlated predictors (Park & Hastie, 2007). The L2 penalty does not eliminate correlated predictors, but simply shrinks their coefficients. To combine the desirable features of both penalties, a penalty based on both L1 and L2 terms was developed, called the "elastic net" (Zou & Hastie, 2005), taking the form :

$$\delta_{\mathrm{L1-L2}} = \lambda[\sum_{j=1}^{p} \alpha|\hat{\beta}_j| + (1-\alpha)\hat{\beta}_j^2] \quad (15)$$

The first term inside the brackets in (15) is merely the L1 penalty of (13) with a new scaling factor (α, a value ranging from zero to one), while the second term is merely the L2 penalty of (14) multiplied by 1- α. α has the effect of balancing the overall penalty between the behavior of a pure L1 (α = 1) and a pure L2 (α = 0) penalty (Zou & Hastie, 2005). The elastic net penalty of (15) can thus be used to move flexibly between a pure Lasso model, shrinking some predictors to zero (L1), or a pure ridge model (L2), incorporating all predictors with shrinkage. Values between 1 and 0 allow some correlated predictors to enter the model, while still removing some. A slight variation on (15) appears in (Friedman, Hastie, & Tibshirani, 2010), in which (1- α) is replaced by (1- α) / 2, and in Park and Hastie (2007) the scaling factor on the L2 portion of the term is independent of the scaling factor on the L1 portion of the term.

Park and Hastie (2007) suggested fixing the L2 scaling factor at some small positive value. Friedman et al. similarly suggest fixing α at a value close to 1 so that a small coefficient is attached to the L2 term, and Hastie et al. (2009; p. 663) suggest that α be pre-selected or estimated via cross-validation, necessitating a search over a tuning parameter space jointly defined by α and λ. A common practice is to graph the size of predictors' coefficients across

different values of λ assessed in test data. This is called the "solution path" of the model, since it charts the path of a predictor's coefficient shrinkage as λ moves from 0 toward 1.

Illustration. Using the R package glmnet (Friedman, Hastie, & Tibshirani, 2011), we fit a Cox regression model with elastic net penalization to the training sample, beginning with all 57 items in the pool. We utilized 10-fold cross-validation to select the optimal value of λ, at each of eleven different values of α in (15). In other words, we conducted a two-dimensional search across independent tuning parameters. Values of α ranged from 0 (ridge penalty) to 1 (Lasso penalty), in increments of .1. Very little difference in the GCV error estimate was noted for the best λ at varying α. The 10-fold cross-validated Deviance ranged from 16.59-16.61, indicating that whatever the value of α, the algorithm could find an optimum λ leading to comparable overall GCV error. The number of items selected varied considerably however, as expected. At α = 0, a pure ridge penalty, all 57 items were included, while at α = 1, a pure Lasso penalty, 38 items were selected. Intermediate values of α resulted in 46-48 items being selected, consistent with suggestions in the literature that when α is neither zero nor one, its exact value may sometimes make little difference (Friedman et al., 2010; Hastie et al., 2009; Park & Hastie, 2007).

In this particular instance, we chose to place a high premium on parsimony, and even the pure Lasso penalty selected a relatively large number of predictors. Thus, to settle on a more economical model, we applied Breiman's (1984) rule and selected the simplest Lasso model within one standard error of the cross-validated Deviance. This model corresponded to a λ value of .038, and contained 21 items. Two of the items had coefficients close to zero (as shown in Table 3) and, in effect, the resulting scale would have 19 items.

The entire solution path for this final model, across different values of (log) λ is shown in Figure 3. At the far left of the path, λ is near 0, virtually no shrinkage is applied, and most items are in the model with coefficients close their MLEs. As one moves right, values of λ increase, shrinking coefficients toward zero and eliminating many items from the model. The top x-axis shows the number of items in the model at each step. The final items and their shrunken coefficients are shown in Table 3. Table 4 shows that scale scores from the elastic net model perform virtually as well in the test sample as within the training sample. Regularization in this case effectively minimized EPE, yielding a scale with generalizable predictive validity.

Algorithm 3: Boosting

Background and Motivation. Boosting partitions the outcome variance into small slices and builds a model by fitting the best variables from a predictor pool to each successive slice. For this reason, some simple applications of boosting have been called "additive stagewise" modeling, because they predict portions of an outcome's variance in stages that are sequentially additive (see Hastie et al. 2009, Chapter 10). A large number of boosting algorithms exist, differing (among other things) in the "base learner" or model that is boosted (i.e., different kinds of regression models, regression trees); the loss function that is optimized; the number of predictors that can enter at each iteration; the use of subsampling or bootstrap aggregating ("bagging"); the type of cross-validation used to decide tuning parameters; and the reweighting of observations at each iteration based on prediction error. However, all algorithms involve two fundamental tuning parameters: the learning rate or "step length", and the number of stages boosted. We discuss a simple algorithm involving these two tuning parameters (analogous to the "generic" or L2 algorithm in Buhlmann and Hothorn (2007), and forward stagewise additive modeling algorithm in Hastie et al., 2009, Chapter 10).

A general overview of a basic boosting process is as follows: first, a regression model is fit with one predictor, its coefficient shrunk, and then used to predict the outcome. The residual is then computed. Second, from the candidate pool of predictors, the one most associated with the outcome residuals is selected, added to the model, and its coefficient shrunk. The model is again used to predict the residuals. The residuals from that model are then computed. The process next repeats for a third step, and so forth. A "cumulative" model is kept, consisting of the predictors used at each stage and their shrunken regression coefficients. A variable used at a previous step can re-enter the model at a subsequent step, if it is the one most associated with the outcome residuals at that point. Since the re-entering predictor was initially used in the model with a shrunken coefficient, only part of the relationship between that predictor and outcome was removed or "partialed out". Thus, some association may remain. When a predictor enters the model repeatedly, its coefficient in the "cumulative" model is the sum of its shrunken coefficients at each step. Table 5 summarizes the steps of this basic algorithm, referencing a shrinkage factor discussed next.

A simple example can illustrate boosting. Table 6 shows a sample of $n = 10$ individuals, each with a standardized normal (z-scored) outcome Y regressed on a single standardized normal predictor, X. Ordinary least squares (OLS) regression is boosted over $s = 1 \ldots t$ steps/stages/iterations. At the first step, $s = 0$, the outcome value of each observation, y_i is subtracted from the mean of Y. This forms a residual for each observation, $y_{i0} = (y_i - \bar{y})$, where the subscript i denotes the individual and the second subscript the boosting stage $s = 0$. Beginning at stage $s = 1$, a standard linear regression is fit, predicting the stage 0 residuals, y_{i0}, with a least-squares parameter estimate $\hat{\beta}$. At this point, boosting procedures introduce a penalty or shrinkage parameter (sometimes called a "step length"), λ, similar to regularization. The

shrinkage parameter is often fixed at a small value such as .1 or less, but in general $0 < \lambda < 1$. The OLS coefficient $\hat{\beta}$ is subscripted to reflect stage *s* =1, and called $\hat{\beta}_1$. $\hat{\beta}_1$ is multiplied by λ (.1 in this example), and the residuals for each of the *i* observations in stage one are then computed as $y_{i1} = (y_{i0} - \lambda\hat{\beta}_1 x_i)$. Note that the predictor has not been fully "regressed out" of the outcome because its shrunken regression coefficient, $\lambda\hat{\beta}_1$, leaves $\hat{\beta}_1 - \lambda\hat{\beta}_1 = (1-\lambda)\hat{\beta}_1$ of the regression relationship in the stage one residuals. Now, updating to stage *s* = 2, a linear regression model is again fit predicting the residuals of stage one, y_{i1}, with X. Normally there would be other predictors that might be more related to the stage one residuals, but in this simple example there is only a single predictor. This results in a stage two OLS estimate $\hat{\beta}_2$ that is again shrunk by a factor of λ to produce the second-stage residuals $y_{i2} = (y_{i1} - \lambda\hat{\beta}_2 x_i)$. This process is repeated for *t* stages. In the Table 6 example, *t* = 3. The "cumulative model" is additive over these stages. At the end of stage 1, the coefficient is $\hat{\beta}_{\text{cum }1} = \lambda\hat{\beta}_1 x_i$, and at the end of stage two, this is updated to $\hat{\beta}_{\text{cum }2} = \lambda\hat{\beta}_1 + \lambda\hat{\beta}_2$, and at the end of stage three it becomes $\hat{\beta}_{\text{cum }3} = \lambda\hat{\beta}_1 + \lambda\hat{\beta}_2 + \lambda\hat{\beta}_3$. The prediction of *i* at the third stage is, correspondingly, $\hat{y}_{i,\text{cum }3} = \lambda\hat{\beta}_1 x_i + \lambda\hat{\beta}_2 x_i + \lambda\hat{\beta}_3 x_i$. With a single variable, one can determine the exact number of boosting iterations needed to produces its standard OLS coefficient, $\hat{\beta}$; $\frac{1}{\lambda}$ iterations are needed. As the procedure approaches this value, $\lim_{s \to \frac{1}{\lambda}} \hat{\beta}_{cum\,s} = \hat{\beta}$ for $\lambda \in (0,1]$ and the cumulative model coefficient grows arbitrarily close to the standard OLS coefficient (more generally, to the MLE).

In reality, *p* predictors rather than a single one are used. At each stage, the algorithm selects the single predictor most related to the current residuals. Each additional variable selected thus has incremental validity at its stage of selection. The shrinkage and residualization yield a more conservative procedure than a standard forward step-wise procedure because each step is

relatively small, and only "part" of a predictor enters (Hastie et al., 2009; Chapter 10). Damage from poor predictors is minimized in this way, and truly good predictors will tend to be drawn into the algorithm repeatedly to dominate the cumulative model.

There is another mechanism in boosting to guard against overfitting, however. In the *stochastic gradient boosting* algorithm, a subset of the sample is drawn without replacement at the beginning of each step *s* (Friedman, 2002). The model is fit and its predictions are computed for that subsample. If a sampling fraction of 50% is used for instance, only a random 50% of the observations are residualized at a given stage. Thus, estimation at each stage involves small subsections of data, in addition to estimating "small pieces" of predictors' coefficients. Reducing the residualization of the data at each iteration also prevents sporadic mistakes—for instance, allowing a poor predictor to residualize too much of the outcome (Hastie et al., 2009; Chapter 10). k-fold cross-validation procedures or bagging can introduce similar random variation to guard against overfitting. Other modifications such as re-weighting observations according to how hard they are to predict, or varying the size λ at each stage are also sometimes used (Binder & Schumacher, 2009; Buhlmann & Hothorn, 2007)

Even at a slow pace, boosting a model through too many stages will bring the model coefficients arbitrarily close to their MLEs, as the simple example above shows, potentially overfitting the data (Buhlmann & Hothorn, 2007). Therefore, the critical tuning parameter in boosting is at what stage $s = 1 \ldots t$ one should stop. Unlike the single variable case above where the precise number of iterations needed to reach the predictor's MLE can be ascertained based on a value of λ, such an exact determination cannot be made with a large number of possible predictors and the randomness in bootstrap draws. Thus, one "rule of thumb" is that $\frac{\lambda}{t} \in [100, 10{,}000]$ (Schonlau, 2005). This "rule of thumb" is actually a sizable interval, and of

general use only in identifying potentially under-iterated (i.e., $t < 100$) or over-iterated ($t >$ 10,000) models.

Illustration. We boosted a Cox model in the training sample of HALS using the R package mboost (Hothorn, Buehlmann, Kneib, Schmid, & Hofner, 2011), setting λ to .1, and choosing the stopping iteration *t* based on test data likelihood in 10-fold cross-validation. The initial stopping stage suggested by this procedure was 2495 iterations. However, the cross-validation data log-likelihood at this advanced point was only fractionally better than at earlier iterations. Since the model had been boosted through so many stages, it included a host of items with extremely small coefficients. Few applied researchers or clinicians would want a scale laden with minimally relevant items, so we examined earlier stopping points. The cross-validation data likelihood appeared to plateau at a much earlier iteration (950), which yielded a more parsimonious model. Table 3 shows the items selected and their coefficients from this model in the "boosting" column. Table 4 shows the performance of scores from a scale based on the boosting model—that is, the linear predictor from the cumulative model coefficients at their final stage. Again, predictive validity in the test sample was virtually as good as in the training sample, suggesting that boosting yielded a scale with generalizable criterion validity.

Selecting Algorithms

Often in SLT prediction applications, multiple modeling algorithms are deployed and compared. Each empirical context is unique, and different algorithms may exhibit different strengths or weaknesses under different conditions. Few, if any studies imply global superiority for any algorithm or class of algorithms (see Hastie et al., 2009, p. 350-352, for some considerations). Thus, both applied and simulation studies pitting one SLT predictive model against another are an active area of research.

In our examples, the difference in predictive performance between scales produced by the three algorithms was trivial. When differences do emerge, they may point toward features of the data that bear consideration. In the analyses presented here, substantial superiority of regularization or boosting over SPCA would reinforce the need for shrinkage of item coefficients, and/or the inadequacy of the SPCA item selection method. In that case, an SPCA approach with shrinkage might be considered (Bair et al., 2006). Substantial differences between regularization and boosting would suggest that one method of shrinkage employed here was better than the other. Boosting begins with no variables in the model and builds up predictors' coefficients from zero, while regularization begins with all variables in the model and reduces coefficients from their full MLE (Binder & Schumacher, 2009; Buhlmann & Hothorn, 2007). In cases of similar performance, selecting the "best" algorithm may be dictated by practical considerations. In measurement contexts like ours, for instance, shorter scales might be preferable and lead to the use of the 16-item SPCA scale.

Another issue in model comparison concerns whether a "standard" model should be included in the set that is evaluated. The final column of Table 4 shows the performance of a "naïve" Cox model with standard MLEs (i.e., no variable selection, no shrinkage). In our example, the MLEs show reasonable out-of-sample performance, although the overall model is obviously less parsimonious. There would appear to be at least two reasons for this: first, the sample is quite large, relative to the number of predictors (i.e., $n >> p$ rather than $p > n$). Therefore, copious information is available to support the estimation of a large number of parameters, good variable selection is less important, and the naïve model is not as overfit as it would be in a smaller sample. Second, the test data is actually from the same overall sample as

the training data in which the naïve MLEs were obtained which, along with a large sample size, likely reduces over-fitting to some extent.

To illustrate the impact of sample size, we refit all the models in only a random 10% of the training and test samples. As can be seen in Table 7, the naïve MLEs show a large drop from an "over-optimistic" fit in the training sample. By contrast, both SPCA and regularization models show very good generalization to the test sample. The boosting model falls somewhere in between, evidencing some over-fitting. These observations suggest that in large samples, and/or scenarios in which the resulting model is intended for use in specific samples likely to vary little from one occasion to the next (e.g, patients from the same clinic), standard methods ought to be compared to SLT-based models. Depending on the degree of difference in performance and parsimony afforded by each approach, SLT-based models may or may not be a useful improvement over standard models.

Regarding the issue of similarity between future samples and those used to develop the model, a key point is that cross-validation techniques inherently presume that generalization across random splits of the training data mimics generalization to new, future data. When convenience or other idiosyncratic samples are used for model development, the resulting model is not guaranteed to perform well in future samples, which are likely to come from substantively different populations. SLT methods provide neither unqualified promises of generalizability nor a substitute for good sampling. In psychometrics the generalizability and stability of measurement properties are key issues for any scale. Before moving on to more general matters, we briefly consider the concept of "reliability" for SLT-based scales.

Predictive Validity vs. Internal Consistency in SLT Scales

Criterion-keyed scales, whether built by SLT methods or not, are fundamentally different than scales intended to capture an underlying latent factor. Lest confusion arise, it is helpful to explicitly note why. When a scale is not built to capture a latent factor, internal consistency is not a measure of reliability (Streiner, 2003). Indeed criterion-keyed scales with high predictive validity often might have poor internal consistency. This may seem contrary to the Classical Test Theory point of view that "reliability caps validity." But several independent predictors will yield a better predictive linear combination because a) each captures a unique or non-redundant portion of outcome variances and b) the precision of the estimates is maximized when collinearity is 0, leading to more precise predicted values of the outcome (Seber & Lee, 2012). Indeed, this result has been known for well over half a century, and has been called the "attenuation paradox" (Loevinger, 1954). Consider the following correlation matrix from a joint multivariate normal distribution of six variables, with the first being Y, the outcome, and the remaining five items X_1-X_5:

$$\Sigma = \begin{bmatrix} 1 & .4 & .4 & .4 & .4 \\ .4 & 1 & .05 & .05 & .05 \\ .4 & .05 & 1 & .05 & .05 \\ .4 & .05 & .05 & 1 & .05 \\ .4 & .05 & .05 & .05 & 1 \end{bmatrix}$$

A simulation drawing 100 samples of varying size from this distribution leads to a scale with an average Cronbach's alpha of .17. Yet when the items are formed into a predictive scale using the weighted linear combination from a linear regression, the average validity coefficient between the scale and the outcome is .75. Although this is a trivial example, it illustrates the point that scales built for latent trait measurement vs. criterion prediction differ and thus must be evaluated with different standards.

Consider a criterion-keyed scale administered repeatedly within a sample, each time with a corresponding criterion measured at some fixed interval in the future. If the scale's correlation with the criteria—its validity coefficient—is relatively similar across these occasions, one might argue that this is a kind of reliability, in that it reflects a regularity or consistency in validity. This is not, however, a conventional notion of reliability, and the concept of validity generalization is perhaps more apt. Whether one looks for consistent criterion validity within a sample over time or across samples, regularity in the scale's predictive capacity, rather than its internal consistency, is the key standard by which it should be judged. SLT builds consistency or generalizability of prediction into scales. Further exploration of the distinction between predictive and latent-variable appears in the literature on "clinimetrics" (Fayers & Hand, 2002).

General Conclusions

We have thus far reviewed the core objective of SLT regression methods--minimization of prediction error--and illustrated three common algorithms in a research application requiring high dimensional exploratory regression. As we noted in the introduction, there are many types of psychology research problems for which SLT algorithms may hold utility. These scenarios may occur in both "Big Data" and standard-sized data, and are characterized by a high number of potential predictors. We deem the first set of circumstances "supportive applications," by which we mean methodological problems that must be solved for an analysis to proceed to primary objectives. The solution to these problems involves some sort of model used to reduce biases in the data, so that analysis of the a priori question can proceed with greater rigor. We caution against viewing "supportive" applications as somehow less important than the primary, hypothesis-testing analysis in the study. The quality of the latter depends in large part on that of

the former. The second potential use of SLT methods is as a primary mode of analysis itself, in inductive or discovery phases of research and in theory refinement.

Supportive Applications of SLT

Often, methodological roadblocks arise due to limitations in the study design, and must be statistically ameliorated. Perhaps the most ubiquitous example is a causal research question posed in the context of non-experimental design where it is simply not possibly to randomize the putatively causal factor. Propensity score analysis has been on the rise in psychology as a way to approximate random assignment, and thus strengthen confidence in the possibility that observed associations are actually causal (Shadish, 2010). Propensity score methods necessitate prediction of membership in a treatment or control group, which is then used in analysis of the main research question for matching, weighting, or occasionally as a global covariate (Haviland, Nagin, & Rosenbaum, 2007).

Regardless of whether one considers such analyses truly causal, the degree to which this method approximates randomization is only as good as the model used to produce propensity scores. A weak model with poor prediction of group membership does not "match" or equate groups on very many factors effectively. In most observational studies, a host of variables distinguish groups or treatment levels, meaning that the propensity score model is likely to be high dimensional. The propensity score model is also hard to specify a priori, since factors that researchers cannot anticipate may be strong predictors of group membership. However, one does not simply want to fit a model with every variable in the data set, unless the sample size is large enough to justify such an approach. If it is not, the model will then be severely overfit, and its predicted values—the propensity scores—will have such large standard errors that they will be virtually useless. Thus, whether or not to move to an SLT-based model involves some judgment

about the number of predictors one wishes or needs to include for maximal accuracy, relative to the amount of information available in the data. This consideration is relevant to all applications of SLT subsequently discussed. SLT methods—particularly those that perform variable selection and shrinkage—are ideal tools with which to balance the predictive accuracy and the complexity of the propensity score model. Boosting, for instance, has been used at least once for this purpose (McCaffrey, Ridgeway, & Morral, 2004).

A second area involves selection models (DeMaris, 2014), of which the Heckman selection model (Heckman, 1979) is most prominent. Selection models are needed when researchers wish to draw conclusions about an entire sample, but some natural process leads to the availability of data only in a subsample. A classic example is the analysis of links between education and wages, which can only be conducted in those within the sample who actually have jobs. In the Heckman model, selection into the workforce is first predicted by a probit model. The probit model is used to adjust the model of education-wage relations, through the correlation of selection model and wage equation residuals. As with propensity scores models, a potentially large number of predictors may be involved in the selection process and are not known a-priori. A large predictor pool may therefore need to be scanned for the probit model. Misspecification of the selection model—for instance, carelessly entering large numbers of variables--will yield incorrect adjustment to the primary model. Again, some balance between optimal prediction and model complexity is needed, and might be achieved by SLT methods performing variable selection.

A third possible support area for SLT involves regression imputation for missing data. The specification of the imputation model is often a critical issue in this area, whether one is performing single or multiple imputations (Schafer & Graham, 2002). In a single regression

imputation, a model that predicts the missing variable(s) poorly will yield imputed values that are quite variable and inaccurate. It is therefore useful to harness any other information available in the data set that may help predict the missing values. Often, variables within a subset are imputed based only on other variables in the subset. Yet these may or may not provide the best imputation model for any given variable, so a good imputation model may necessitate a search over a large number of auxiliary variables. A trustworthy method of predictor selection is therefore important. The effect of shrinkage methods like boosting or regularization would be conservative predictions of missing values--meaning imputations less driven by sampling idiosyncrasies. Since multiple imputation approaches strive to capture the uncertainty of imputation itself, the conservative nature of SLT predictions and safeguards against over-fitting would seem useful.

A fourth "supportive application" in which SLT models might be considered involves sample weighting. There are many different types of weighting, and we refer here to inverse probability weights, or IPWs (Woolridge, 2002). IPWs are a general way to render some subsample representative of a broader reference sample, and constructed similar to propensity scores through a logistic model; however, they are, as their name suggests, the inverse of predicted probabilities, rather than the probability itself represented by the propensity score. IPW estimators are used in a variety of scenarios like attrition in longitudinal studies (Woolridge, 2007). Usually, a large number of factors predict sample attrition (or similar selection phenomenon), and reweighting to match the original sample requires a weight model capturing all these factors. Again however, overfitting will lead to weights that are themselves highly imprecise, so modeling the missing mechanism must be done with care. SLT methods are ideal tools for striking the right balance between comprehensiveness and complexity in this regard.

Double-robustness methods combine a single regression imputation model with an IPW model for missing data (Funk et al., 2011), and might also profit from the judicious application of SLT models.

To our knowledge, SLT methods have not been yet been considered in selection models, imputation, or IPW estimators, and are only beginning to be studied in the context of propensity score estimators. These four areas are by no means an exhaustive list of the statistical problems for which SLT may be well suited. "Supportive" applications of SLT would appear ripe for further research.

Primary Applications of SLT

The SLT emphasis on model generalizability dovetails with growing interest in psychology in the replicability of research findings. While the so-called "replicability crisis" has largely played out within experimental psychology (Francis, 2012), it is also a concern in observational research (Asendorpf et al., 2013). In non-experimental work, the specification and estimation of regression models is often crucial: small differences in predictor sets and coefficient estimates can have large implications for whether a result is perceived as "strong" or "weak", "statistically significant" or "non-significant". SLT methods may be useful tools of primary analysis in two general scenarios where scientific questions require exploratory regression analysis. In both cases, the state of knowledge is in a *discovery phase*—that is, deductive hypotheses cannot be usefully formulated because not enough is known about the phenomenon. The principle scientific task is thus an inductive one involving observation and description of the phenomenon, from which a theory might be built and working hypotheses generated (Rozeboom, 1997).

In the first case, the construct or phenomenon of interest is the outcome, rather than predictor, in an analysis. This usually occurs when the construct is of such importance that the scientific priority is to determine how to prevent or promote it. Whether the phenomenon precedes, causes, or in other ways acts as "predictor" of other things is of secondary interest. Examples might be severe mental illness, IQ, racial discrimination, and suicide. There are social and scientific consequences to artificially narrowing the pool of predictors when potentially important but unexplored determinants of the phenomenon go undetected. In the absence of knowledge about specific processes and developmental antecedents, a wide range of candidates factors are probably best considered in order to capture the etiology or mechanisms giving rise to the phenomenon.

Researchers are often hesitant to "admit" to such exploratory work, and may pursue it (with or without guilt) by manually fitting scores of regressions in search of $p < .05$. Issues of how or whether to adjust for multiple comparisons arise and generate great consternation. Whatever binary decision rule is selected, a list of "Yes/no" findings is generated, highly dependent on the power of the study and subject to sampling error. SLT methods would be particularly useful, since they provide a *degree* of relatedness for the predictor set in the form of regression coefficients rather than an accept/reject decision rule, with decisions based on cross-validation, and designed to maximize generalizability rather than exaggerate findings in the training data. Predictors identified in this way can be used to generate hypotheses, leading to a deductive step of a priori hypothesis testing with new data.

A second case in which SLT regression methods might serve as the principle tool of analysis involves a "general hypothesis" that entails many specific predictors. For example, one might forward an omnibus conjecture that specific facets of personality predict net worth. The

motivation for the study is not to determine every possible predictor of net worth in the dataset, but is instead a "broad-bandwidth" question. In these cases, there is enough a priori theory to suggest that a class or type of phenomenon may predict the outcome, but the class entails a potentially large number of predictors, creating a high-dimensional exploratory regression problem for which SLT methods would be well suited.

We suspect that these two situations are far more common than one might guess, but are often disguised as classical hypothetico-inductive investigations and pursued with the statistical apparatus of null-hypothesis testing. In other words, a fundamentally exploratory research problem is presented as a carefully derived and tested a priori hypothesis (Kerr, 1998). Some have argued that this occurs because of the domination of the hypothetico-deductive framework in psychology, which discourages inductive work (Haig, 2005; Nickerson, 2000; Rozeboom, 1997). Exploratory work must be disguised or risk widespread distaste[4]. Yet it is not the pursuit of exploratory work itself that warrants scientific policing, but its improper conduct and misrepresentation. A common "Big Data" turn-of-phrase is that researchers will "interrogate the data". Doing so in a naïve search for low *p*-values is at best "enhanced interrogation" however, and more often degenerates into "torturing the data". SLT methods would seem to be a step forward from this state of affairs[5].

SLT and machine learning theory are fundamentally different approaches to, analysis and often do not even produce *p*-values (see Breiman, 2001, for a review of other tensions with classical approaches). In a field dominated by null hypothesis testing, methods that downplay statistical significance may be met with perplexity or skepticism (Haig, 2005; Nickerson, 2000; Rozeboom, 1997). Resistance to new approaches in workaday psychological research (Sharpe, 2013) and aversion to serious modeling (Borsboom, 2006) are oft lamented, and SLT methods

are likely to be no exception. Nevertheless effective inductive approaches, theory-building, and exploration are mainstay components of psychological science, and statistical adaptation to their demands would appear wise. It is within those areas that there may be a potential home for SLT methods within psychology.

References

Anastasi, A., & Urbina, S. (1997). *Psychological Testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys, 4*, 40-79.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., . . . Nosek, B. A. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*(2), 108-119.

Bair, E., Hastie, T., Debashis, P., & Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association, 101*, 119-137.

Bair, E., & Tibshirani, R. (2010). *Superpc: supervised principal components. R package version 1.07.* Retrieved from http://www-stat.standord.edu/~tibs/superpc

Barefoot, J. C., Dodge, K. A., Peterson, B. L., Dahlstrom, W. G., & Williams, R. B., Jr. (1989). The Cook-Medley hostility scale: item content and ability to predict survival. *Psychosomatic Medicine, 51*, 46-57.

Benjamini, Y. (2010). Simultaneous and selective inference: Current successes and future challenges. *Biometrical journal. Biometrische Zeitschrift, 52*(6), 708-721. doi:10.1002/bimj.200900299

Binder, H., & Schumacher, M. (2009). Incorporating pathway information into boosting estimation of high-dimensional risk prediction models. *BMC Bioinformatics, 10*(1), 18. Retrieved from http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-18

Blaxter, M. (1987). Sample and data collection, the Health and Lifestyle Survey: A preliminary report. In B. M. Cox B.D., Buckle A.L.J., Fenner, N.P., Golding, J.F., Gore, M., Huppert, F.A., Nickson, J., Roth, M., Stark, J., Wadsworth, M.E.J., & Whichelow, M. (Ed.). London: Health Promotion Trust.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*(3), 425-440.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16(3)*, 199-231.

Breiman, L., Friedman, J., Stone, C., & Olshen, R. (1984). *Classification and Regression Trees*. Boca Raton, FL: CRC press.

Brim, O. G. , Ryff, C.D., & Kessler, R. C. (2004). The MIDUS National Survey: An Overview. In O. G. Brim, C.D. Ryff, & R. C. Kessler (Eds.), *How Healthy? Are We? A National Study of Well Being at Midlife* (pp. 1-34). Chicago, IL: University of Chicago Press.

Buhlmann, P., & Hothorn, T. (2007). Boosting algorithms: regularization, prediction, and model fitting. *Statistical Science, 22*(4), 477-505.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3-5.

Butcher, J. N., Dahlstrom, W. G., Grahamn, J. R., Tellegen, A., & Kraemmer, B. (1989). *Minnesota Multiphasic Personality Inventory-2 (MMPI-2): Manual for administration and scoring*. Minneapolis, MN: University of Minnesota Press.

Chapman, B. P., Roberts, B., & Duberstein, P. (2011). Personality and longevity: knowns, unknowns, and implications for public health and personalized medicine. *Journal of Aging Research, 2011*, 24. doi:10.4061/2011/759170

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/ Correlation Analysis for the Behavioral Sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods, 14*(2), 81-100.

DeMaris, A. (2014). Combating unmeasured confounding in cross-sectional studies: Evaluating instrumental-variable and Heckman selection models. *Psychological Methods, 19*(3), 380-397.

Efron, B. (2010). *Large-scale inference: Empirical bayes methods for estimation, testing, and prediction*. New York: Cambridge University Press.

Eysenck, S. B. G., & Eysenck, H. J. (1964). *Manual of the Eysenck Personality Inventory*. London: University of London Press.

Fayers, P. M., & Hand, D. J. (2002). Causal variables, indicator variables and measurement scales: an example from quality of life. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 165*(2), 233-253.

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review, 19*(6), 975-991.

Friedman, J. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis, 38*(4), 367-378.

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

Friedman, J., Hastie, T., & Tibshirani, R. (2011). Glmnet: Lasso and elastic-net regularized generalized linear models. R package 1.6. Retrieved from http://www.jstatsoft.org/v33/i01/.

Funk, M. J., Westreich, D., Wiesen, C., Sturmer, T., Brookhart, M. A., & Davidian, M. (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology, 173*(7), 761-767. doi:kwq439 [pii]10.1093/aje/kwq439

Gould, W., Pitblado, J., & Sribney, W. (2006). *Maximum Likelihood Estimation With Stata* (3rd ed.). College Station, Texas: Stata Press.

Grice, J. W., & Harris, R. J. (1998). A comparison of regression and loading weights for the computation of factor scores. *Multivariate Behavioral Research, 33*, 221-247.

Haig, B. D. (2005). An Abductive Theory of Scientific Method. *Psychological Methods, 10*(4), 371-388.

Hardin, J. W., Hilbe, J. M., & Hilbe, J. (2007). *Generalized linear models and extensions*. College Station, TX: Stata Press.

Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine, 15*, 361-387.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*: Springer Science & Business Media.

Haviland, A., Nagin, D. S., & Rosenbaum, P. R. (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods, 12*(3), 247.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica, 47(1)*, 153-161.

Hoerl, A., & Kennard, R. (1988). Ridge regression *Encyclopedia of Statistical Sciences* (Vol. 8, pp. 129-136). New York, NY: Wiley.

Hofer, S. M., & Piccinin, A. M. (2009). Integrative data analysis through coordination of measurement and analysis protocol across independent longitudinal studies. *Psychological Methods, 14*(2), 150-164.

Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M., & Hofner, B. (2011). Mboost: Model-based boosting. R package 2.0-11. Retrieved from http://cran.rproject.org/web/packages/mboost/index.html

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196-217.

Kodratoff, Y. (2014). *Introduction to machine learning*. San Mateo, CA: Morgan Kaufmann.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York, NY: Springer.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note, 6*, 70.

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin, 51*(5), 493-504.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods, 9*(4), 403-425. doi:2004-21445-001 [pii]10.1037/1082-989X.9.4.403

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of r and d. *Psychological Methods, 11*(4), 386-401.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 5*(2), 241-301.

Ones, D. S., Chockalingham, V., & Dilchert, S. (2005). Personality at work: Raising awareness and correcting misconceptions. *Human Performance, 18*(4), 389-404.

Park, M. Y., & Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69*, 659-677.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds): *What if there were no significance tests*, 335-392. Lawrence Erlbaum: Mahwah NJ

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177.

Schonlau, M. (2005). Boosted regression (boosting): An inroductory tutorial and a Stata plugin. *The Stata Journal, 5*(3), 330-354.

Seber, G. A., & Lee, A. J. (2012). *Linear regression analysis*. Hoboken, NJ: John Wiley & Sons.

Shadish, W. R. (2010). Campbell and Rubin: A primer and comparison of their approaches to causal inference in field settings. *Psychological Methods, 15*(1), 3-17.

Sharpe, D. (2013). Why the resistance to statistical innovations? Bridging the communication gap. *Psychological Methods, 18*(4), 572-582.

Streiner, D. L. (2003). Being inconsistent about consistency: When coefficient alpha does and doesn't matter. *Journal of PersonalityAssessment, 80*(3), 217-222.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods, 17*(2), 228-243.

Weiss, A., Gale, C. R., Batty, G. D., & Deary, I. J. (2013). A questionnaire-wide association study of personality and mortality: The Vietnam Experience Study. *Journal of Psychosomatic Research, 74*(6), 523-529.

Woolridge, J. M. (2002). Inverse probability weighted M-estimators for sample selection, attrition, and stratification. *Portugese Economic Journal, 1(2)*, 117-139.

Woolridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics, 141(2)*, 1281-1301.

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

Table 1: Classical Statistical Models Versus Statistical Learning Theory Models

	Classical Approach	Statistical Learning Theory
Statistical Emphasis	Test a statistical null hypothesis about an association of interest	Maximize generalized predictive accuracy for an outcome
Estimation strategy	Maximize sample likelihood	Minimize cross-validation error
Number of predictors	Usually one or a small number of focal predictors, with additional control variables.	Small to very large predictor pool
Role of sample size	Dictate size of standard errors and power for hypothesis testing	Dictate thoroughness of cross-validation efforts and
General Scientific Application	Testing a theory	Building or refining a theory
Psychometric Purpose	Applying developed scales to test hypothesized associations	Building criterion-keyed scales from a large item pool

Table 2: Schematic of Five-Fold Cross Validation

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

Note. Five-fold cross-validation splits a sample into five equally sized folds. A model is fit in all but one fold of the data (the shaded portions), then model prediction error is evaluated in the left-out (unshaded) fold. This is repeated for all possible combinations of four fitting and one left-out fold, and the prediction error across the left-out folds is averaged to provide a cross-validated error rate for the model.

Table 3: Item Selection and Weighting for Three Criterion-Keyed Scales Produced by Different SLT Algorithms

EPI Item	**EPI scale**	**SPCA 1**	**SPCA 2**	**Regularization**	**Boosting**
Often longs for excitement	Extraversion	-0.33	-0.07	0.04	0.1
Often needs understanding friends to cheer up	Neuroticism				
Usually carefree	Extraversion				-0.04
Finds it hard to take no for an answer	Neuroticism				
Does NOT stop and think things over before doing them	Extraversion				
If says will do something, always keeps promise even if inconvenient	Lie				
Mood often goes up and down	Neuroticism				
Generally does and says things quickly, without stopping to think	Extraversion				
Sometimes feels 'just miserable' for no good reason	Neuroticism				0.01
Would do almost anything for a dare	Extraversion				
Does NOT feel suddenly shy when wants to talk to attractive stranger	Extraversion	0.18	0.37	-0.11	-0.18
Does NOT once in a while lose temper and get angry	Lie	0.17	-0.05	-0.25	-0.21
Often does things on the spur of the moment	Extraversion				
Often worries about things should not have done or said	Neuroticism	-0.34	-0.25	0.01	-0.03
Does NOT prefer reading to meeting people	Extraversion			0.00*	0.1
Feelings are rather easily hurt	Neuroticism				0.05
Likes going out a lot	Extraversion				
Occasionally has thoughts/ideas does not want others to know about	Lie	0.39	-0.05	-0.01	-0.02
Sometimes bubbling over with energy and sometimes very sluggish	Neuroticism	-0.32	-0.17	0.06	0.15
Does NOT prefer to have a few, but special friends	Extraversion				
Daydreams a lot	Neuroticism				
When people shout at, shouts back	Extraversion				
Often troubled by feelings of guilt	Neuroticism				
All habits are good and desirable ones	Lie	0.32	-0.22	-0.14	-0.14
Lets self go and enjoys self a lot at a lively party	Extraversion				0.07
Calls self tense or 'highly-strung'	Neuroticism				
Others think of as being very lively	Extraversion				
After doing something important, feels could have done better	Neuroticism				
NOT mostly quiet when with other people	Extraversion	-0.08	0.43	0.02	0.1

Sometimes gossips	Lie	0.29	-0.18	-0.14	-0.16
Ideas run through head so that cannot sleep	Neuroticism				0.01
Would rather NOT look up something wants to know in book than talk to someone about it	Extraversion	-0.06	0.18	0.09	0.15
Gets palpitations or thumping in heart	Neuroticism				
Does NOT likes the kind of work needs to pay close attention to	Extraversion				
Gets attacks of shaking or trembling	Neuroticism			-0.07	-0.09
Always declares everything at customs, even if could not be found out	Lie	0.22	-0.28	-0.11	-0.17
Does NOT hate being with a crowd that plays jokes on one another	Extraversion				
Is an irritable person	Neuroticism				
Likes doing things in which has to act quickly	Extraversion				
Worries about awful things that might happen	Neuroticism				
Is NOT slow and unhurried in way of movement	Extraversion	-0.07	0.21	0.31	0.36
Has NOT ever been late for an appointment or work	Lie	0.24	-0.2	-0.08	-0.12
Has many nightmares	Neuroticism				
Likes talking so much never misses chance of talking to stranger	Extraversion			-0.07	-0.13
Troubled by aches and pains	Neuroticism	-0.19	-0.37	-0.22	-0.3
Would be very unhappy if could not see lots of people most of time	Extraversion				-0.04
Would call self a nervous person	Neuroticism				
Of all people known, NONE who definitely does not like	Lie				
Is fairly self-confident	Neuroticism			-0.07	-0.17
Is easily hurt when people find fault with	Neuroticism				
Does NOT find it hard to enjoy self at a lively party	Extraversion				
Is troubled with feelings of inferiority	Neuroticism				
Can easily get some life into a rather dull party	Extraversion			-0.00*	-0.12
Does NOT sometimes talk about things knows nothing about	Lie	0.26	-0.21	-0.05	-0.12
Worries about health	Neuroticism	-0.22	-0.34	-0.03	-0.13
Likes playing pranks on others	Extraversion				
Suffers from sleeplessness	Neuroticism				
Component weights		1.54	1.49		

Note. EPI = Eysenck Personality Inventory. Items are in order of appearance. SPCA = Supervised Principle Components, component 1 and 2. For SPCA, numbers are component loadings (and regression scoring coefficients for components), while “component

weights” in bottom row reflect raw regression coefficients (log hazards) for the association between each component and -cause mortality over a 25-year span. For boosting and elastic net models, coefficients are log-hazards for all-cause mortality over a 25-year span. Blanks indicate that an item was not selected by a particular model for the final scale. * = item included in final model with coefficient <.01.

Table 4: Predictive Validity of Scales in Training/Test and Hold-Out Samples

Performance Measure	**SPCA**	**Regularization**	**Boosting**	**Naïve MLE**
Scale Length (number of items)	16	21	27	57
Training/Test Sample N = 2472				
Peak Accuracy (% Correct Classification)	61%	64%	64%	65%
Point Biserial Correlation	0.26	0.32	0.35	.37
Spearman Correlation	0.26	0.33	0.35	.37
Pseudo R^2	0.08	0.12	0.15	0.16
HR for +1 *SD* Scale Score	1.51	1.6	1.78	1.87
AUC	.65	.65	.69	.71
Hold-Out Sample N = 1237				
Peak Accuracy (% Correct Classification)	61%	65%	65%	63%
Biserial Correlation	0.25	0.31	0.33	.33
Spearman Correlation	0.24	0.31	0.33	.33
Pseudo R^2	0.07	0.11	0.13	0.14
HR for +1 *SD* Scale Score	1.47	1.57	1.73	1.76
AUC	0.64	0.68	0.69	.69

Notes: AUC = Area Under the Receiver Operating Curve; HR = Hazard Rate; MLE = Maximum Likelihood Estimate; SPCA = Supervised Principal Components. The final column, "Naïve MLE", pertains to a Cox model with no shrinkage or variable selection (i.e., standard Cox model MLEs for all 57 items).

Table 5: General Algorithm for Boosting

Step Number	**Statistical Procedure**	**Conceptual Purpose**
1 [a]	Initialize predicted values of y to sample mean	Need an initial unconditional "best guess" for outcome
	*****ITERATE ******	
2	Draw a bootstrap sample from the data	Mimic sampling variation in the observations used to fit the model
3[b]	Compute current residuals: loss function of y vs. current prediction	Obtain part of outcome currently unaccounted for the model
4[c]	Fit a model to current residuals	Attempt to predict a portion of the unaccounted-for outcome
5[d]	Generate predictions from the current model	Predict an additional portion of unaccounted-for outcome
6[e]	Take only predictions from a randomly selected 50% of cases in the bootstrap sample	Introduce further random variation in an attempt to avoid overfitting
7[f]	Add scaled predictions of current model to the running prediction. Scaling involved multiplying by a learning rate, or a constant between 0 and 1.	Update the current prediction of the outcome with the results of current iteration, but scale the update so no particular iteration exerts an undue impact on the cumulative prediction
8[g]	Go back to step 2	Move to next iteration

Notes: The scaling factor in step 7 is λ, the learning rate. Smaller values correspond to greater shrinkage at each step and a slower fitting process that may require more iterations. The iteration procedure is stop when some form of GCV error, usually the error in the out of bag observations, is stopped.

Table 6: Example Simple Boosting of a Linear Regression Model with One Predictor for Ten Observations Over Three Iterations

		Stage 0	Stage 1			Stage 2			Stage 3		
Predictor and Outcome Variables		residual-ized outcome	predict-ed value	residual-ized outcome	cumu-lative model predict-ion	predict-ed value	residual-ized outcome	cumu-lative model predict-ion	predict-ed value	residual-ized outcome	cumu-lative model predict-ion
x	y	$y_{i0} = (y_i - \bar{y})$	$\hat{y}_{i1} = \lambda\hat{\beta}_1 x_i$	$y_{i1} = (y_{i0} - \lambda\hat{\beta}_1 x_i)$	$\hat{\beta}_{\text{cum }1} = \lambda\hat{\beta}_1 x_i$	$\hat{y}_{i2} = \lambda\hat{\beta}_2 x_i$	$y_{i2} = (y_{i1} - \hat{y}_{i2})$	$\hat{\beta}_{\text{cum }2} = \lambda\hat{\beta}_1 x_i + \lambda\hat{\beta}_2 x_i$	$\hat{y}_{i3} = \lambda\hat{\beta}_3 x_i$	$y_{i3} = (y_{i2} - \hat{y}_{i3})$	$\hat{\beta}_{\text{cum }3} = \lambda\hat{\beta}_1 x_i + \lambda\hat{\beta}_2 x_i + \lambda\hat{\beta}_3 x_i$
1.04	2.37	1.81	.12	1.69	.12	.11	1.58	.22	.10	1.49	.32
1.04	.38	-.18	.12	-.30	.12	.11	-.40	.22	.10	-.50	.32
-.51	-.85	-1.42	-.06	-1.36	-.06	-.05	-1.31	-.11	-.05	-1.26	-.16
-1.55	-.58	-1.15	-.18	-.98	-.18	-.16	-.82	-.33	-.14	-.68	-.47
.24	1.92	1.35	.03	1.32	.03	.02	1.30	.05	.02	1.28	.07
-2.16	-2.43	-2.99	-.24	-2.75	-.24	-.22	-2.53	-.46	-.20	-2.33	-.66
.79	2.55	1.98	.09	1.89	.09	.08	1.81	.17	.07	1.74	.24
.86	1.55	.98	.10	.89	.10	.09	.80	.18	.08	.72	.26
1.49	1.78	1.21	.17	1.04	.17	.15	.89	.32	.14	.76	.46
.31	-1.03	-1.59	.04	-1.63	.04	.03	-1.66	.07	.03	-1.69	.10
		Model Coefficients and Error Terms at Each Stage									
Estimate		Stage 0	Stage 1			Stage 2			Stage 3		
β_s		--	1.13			1.02			.92		
$\lambda\,\beta$		--	.11			.10			.09		
$\beta_{cum,s}$			.11			.21			.31		
ε_{RSS}		26.3	23.14			2.58			18.50		
R^2_{cum}		0	.12			.22			.31		

Note. Simulated sample of ten observations with standardized normal predictor and outcome.

Table 7: Model Performance at Sample Size 10% of Primary Analyses

Performance Measure	**SPCA**	**Regularization**	**Boosting**	**Naïve MLE**
Scale Length (number of items)	9	6	11	57
Training/Test Sample N = 247				57
Peak Accuracy (% Correct Classification)	65%	62%	69%	78%
Point Biserial Correlation	.28	.29	.42	.59
Spearman Correlation	.28	.28	.41	.60
Pseudo R2	.08	.10	.18	.39
HR for +1 SD Scale Score	1.53	1.56	1.80	3.8
AUC	.66	.60	.74	.85
Hold-Out Sample N = 124				
Peak Accuracy (% Correct Classification)	64%	63%	62%	63%
Biserial Correlation	.26	.26	.24	.30
Spearman Correlation	.27	.26	.25	.30
Pseudo R2	.07	.10	.04	0
HR for +1 SD Scale Score	1.46	1.58	1.29	1.51
AUC	.66	.65	.64	.68

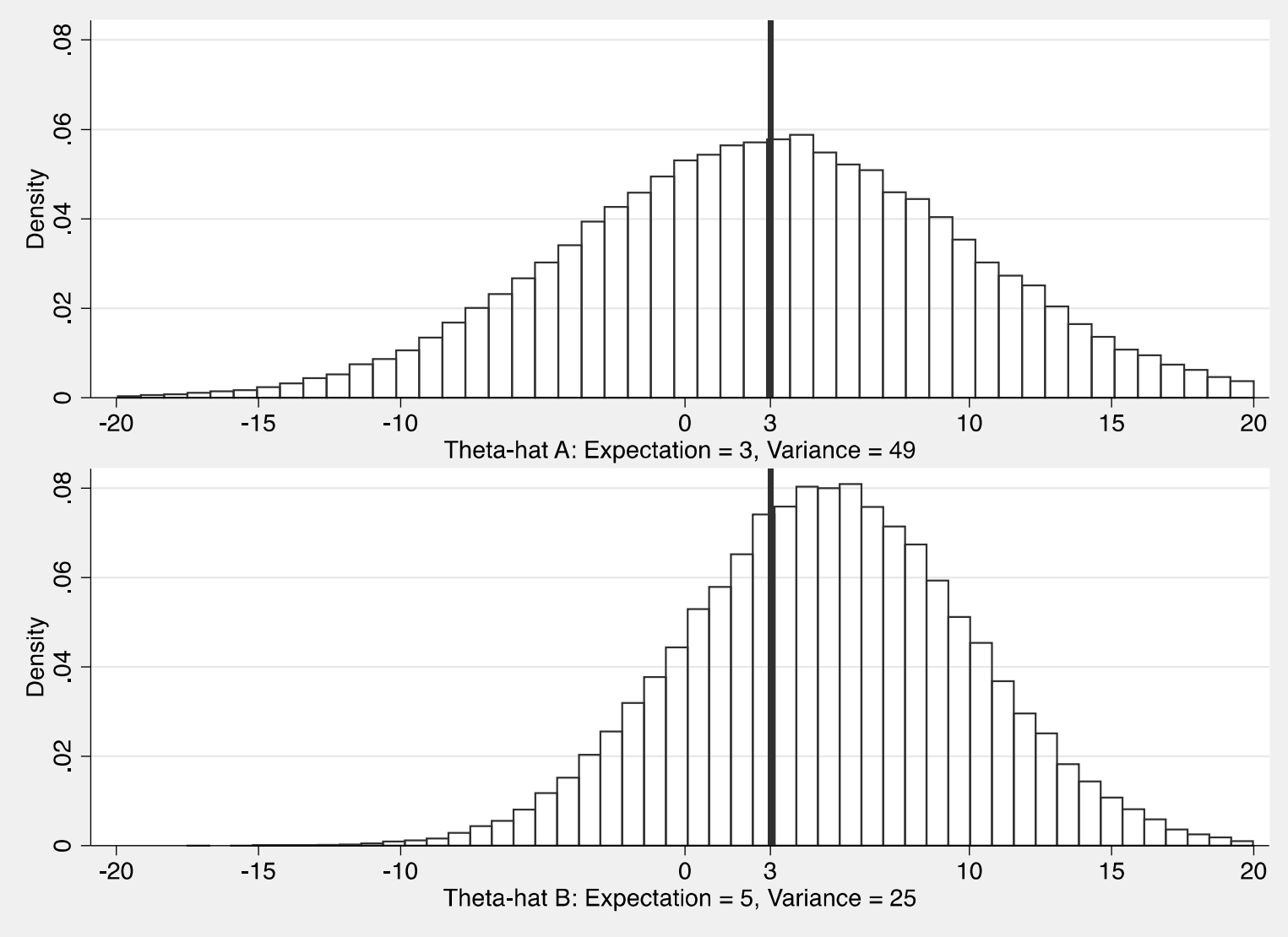

Figure 1. Sampling distribution of an estimator $\hat{\theta}$ for $\theta, \theta \in (-\infty, \infty)$. The top estimator is unbiased, with an expectation of 3—the parameter's true value. The bottom estimator is slightly upwardly biased, but has a smaller variance and smaller Mean Square Error.

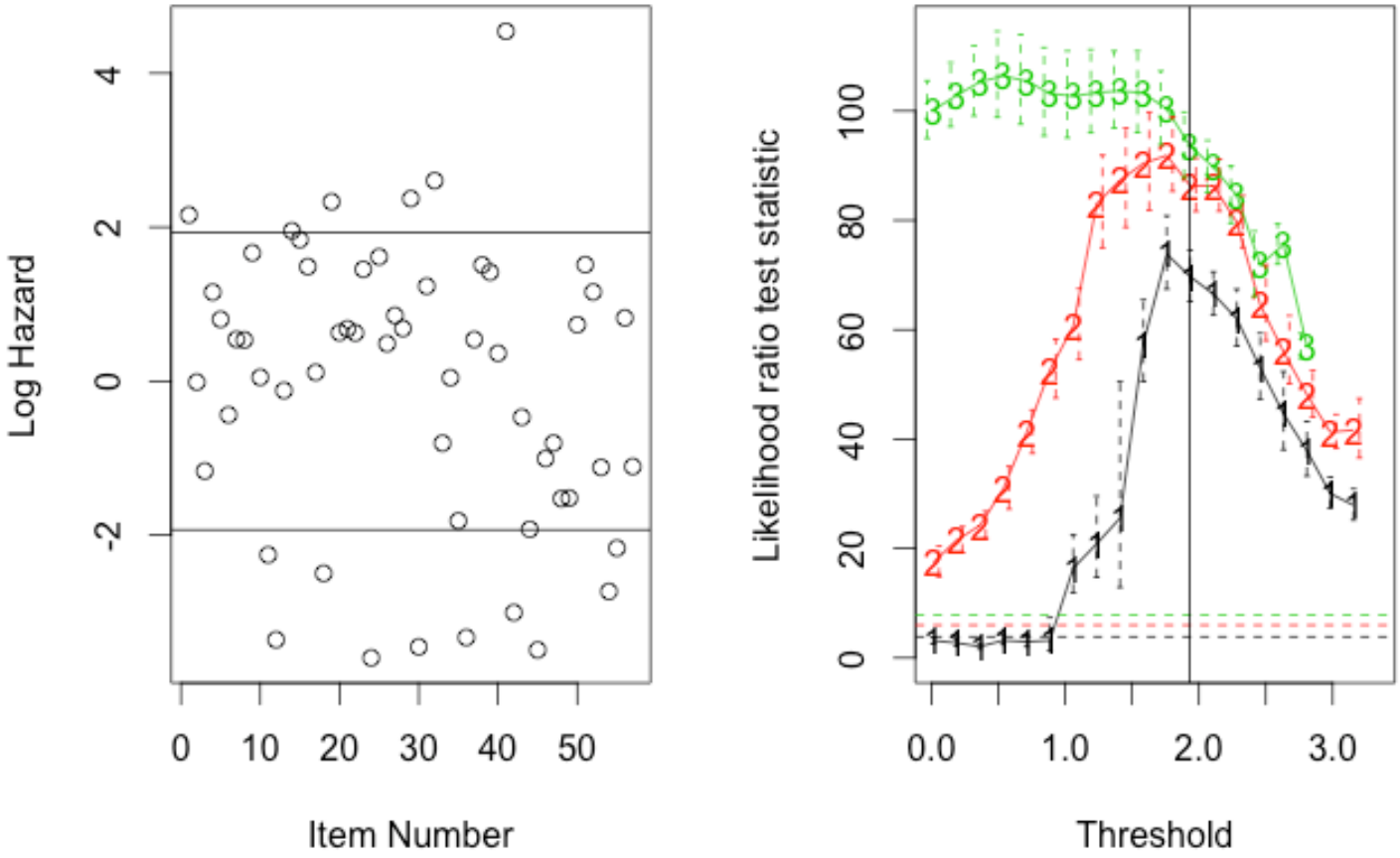

Figure 2. Left: univariate log hazards plotted against item number, with horizontal lines indicating the threshold value for item retention selected by 10-fold cross validation. Right: Likelihood ratio from 10-fold cross-validation for differing threshold values of item selection and differing number of principal components using selected items.

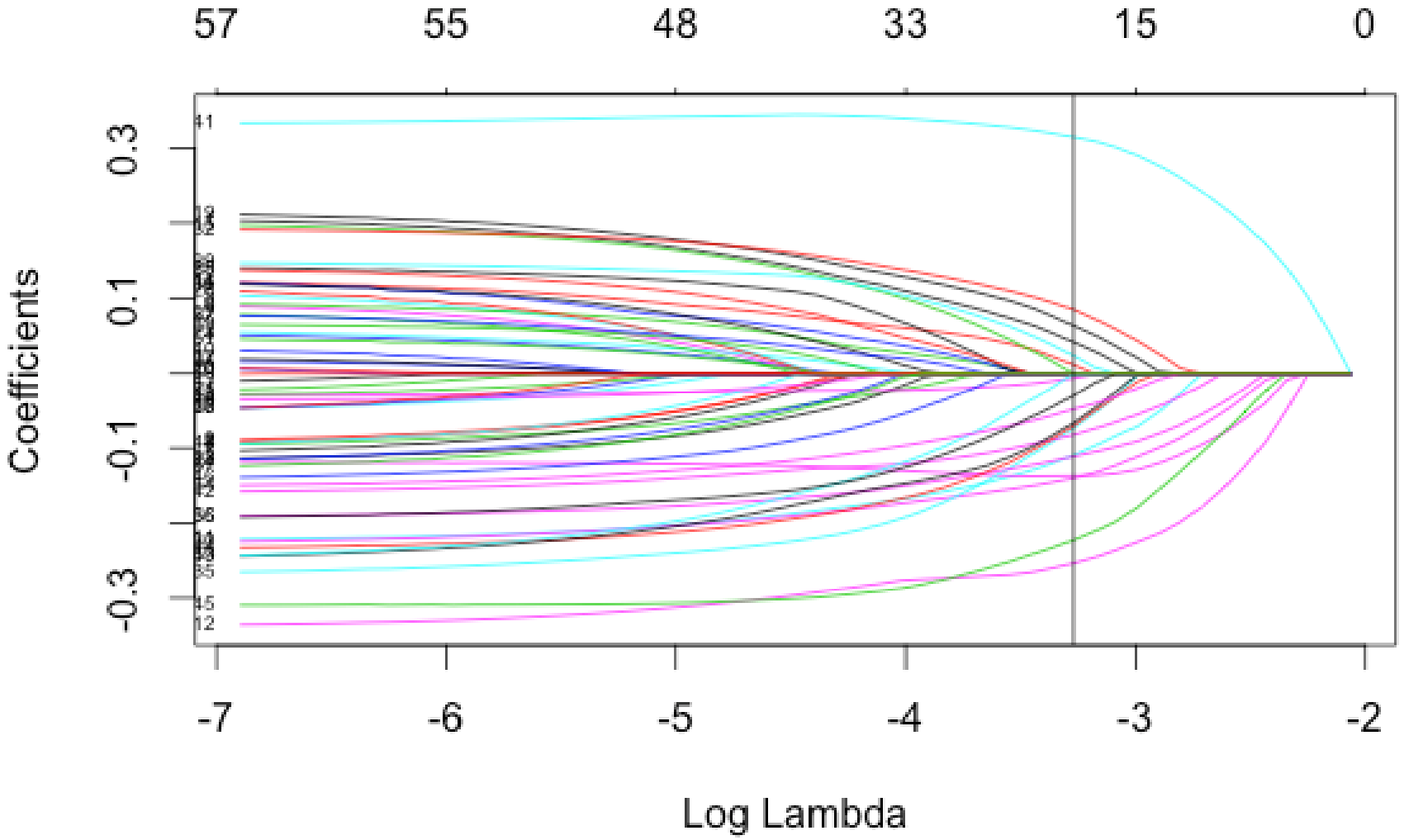
57
55
48
33
15
0
Coefficients
0.3
0.1
-0.1
-0.3
-7
-6
-5
-4
-3
-2
Log Lambda

Figure 3: Solution path for the L1/L2 regularization model. The bottom x-axis shows values for the log of λ_1 in (12), while the y-axis shows the log-hazard of items resulting from the model fit at each λ_1. Each line represents an item. Starting from the left, λ_1 is very small (virtually no penalization, coefficients near MLEs) and all items are included in the model. Moving from left to right, with increasingly more shrinkage, coefficients tend toward zero and items are eliminate when they hit the horizontal "0" line. The top x-axis lists the number of items included at each step along the path of λ_1 values. The vertical line represents the selected by 10-fold cross-validation, .038, corresponding to a model with 21 items.

Footnotes

[1] Our notation generally follows that of Hastie et al. (2009). Capital letters such as X or Y refer to random variables in a general sense. A specific value of a variable is denoted with a lower case letter. For instance, X is a variable with any possible value, while x is a specific value. A vector is denoted by a lower case bold letter such as **x** and is a column vector unless otherwise noted, and a matrix is an upper case bold letter such as **X**. Capital script letters refer to sets. Greek letters indicate parameters, and bolded Greek letters are parameter vectors.

[2] (7) is equivalent to the eigen-decomposition of the $q \times q$ inter-item covariance matrix $\mathbf{\Sigma}$ associated with PCA because $\mathbf{\Sigma} = \mathbf{X^T X} = (\mathbf{UDV^T})^\mathbf{T}(\mathbf{UDV^T}) = \mathbf{U^T UVD^2 V^T} = \mathbf{VD^2 V^T}$ where $\mathbf{D^2}$ is a diagonal q × q matrix of eigenvalues, the square root of which are singular values (Mardia, 1979).

[3] Considerations in psychology around the use of components vs. factors have not entered the SPCA literature.

[4] An interesting, and possibly instructive exception is the perennial popularity of exploratory factor analysis (EFA) and, more recently, mixture models (i.e., latent class and growth-mixture models), which are also fundamentally exploratory.

[5] Large scale null-hypothesis testing can be reasonably conducted, but requires approaches to multiple testing such as the False Discovery Rate (Benjamini, 2010). Even here, emphasis is sometimes shifted away from “statistical significance” in favor of “interesting” or “uninteresting” terminology and p-values are interpreted rather differently (Efron, 2010).