UNIVERSITY of PENNSYLVANIA

Department of Criminology

Working Paper No. 2015-8.0

Using Regression Kernels to Forecast a Failure to Appear in Court

Richard Berk

Justin Bleich

Adam Kapelner

Jaime Henderson

Geoffrey Barnes

Ellen Kurtz

Using Regression Kernels to Forecast A Failure to Appear in Court

Richard Berk*[1,2], Justin Bleich[2], Adam Kapelner[2], Jaime Henderson[3], Geoffrey Barnes[3], and Ellen Kurtz[3]

[1]Department of Criminology, University of Pennsylvania
[2]Department of Statistics, University of Pennsylvania
[3]Philadelphia Adult Department of Probation and Parole

January 9, 2015

Abstract

Forecasts of prospective criminal behavior have long been an important feature of many criminal justice decisions. In this paper, we apply a form of kernel logistic regression to forecast at an arraignment whether an individual charged with drug possession will return to court when ordered to do so. The practical goal is to help inform a magistrate's release decision. We focus on individuals with drug possession charges because they have atypically high rates of failures to appear (FTAs). We apply a form of kernel logistic regression because recent work has shown that conventional logistic regression typically will not forecast as accurately as machine learning procedures. Our approach to kernel logistic regression, which can be seen as a hybrid of conventional logistic regression and machine learning, clearly dominates conventional logistic regression as a forecasting tool, and in some settings can be a legitimate competitor to machine learning procedures such as support vector machines, stochastic gradient boosting, and random forests. The methods applied are implemented in the `R` package `kernReg` currently available on `CRAN`.

*Electronic address: `berkr@sas.upenn.edu`; Corresponding author

1 Introduction

According to the Bail Reform Act of 1984, "risk of flight" should be a key determinant of release decisions made at arraignment (Adair, 2006). A defendant's "failure to appear" (FTA) at a subsequent court proceeding, after being ordered to do, can be thumb in the eye of judicial authority and may mean that the defendant's criminal charges are never adjudicated. When large numbers of defendants fail to appear at mandatory court proceedings, the legitimacy of the criminal justice system is challenged.

All arraignment release decisions informed by the risk of flight necessarily require forecasts. Magistrates or judges are tasked with looking into the future and determining that chances that a given defendant will return to court on a specified date. Clearly, a lot is at stake. Defendants may be held when there is no need or defendants may be released when some form of incapacitation is required. Forecasting accuracy really matters.

In this paper, we focus on defendants charged with drug possession in part because they often have atypically high rates of FTAs. They are also at the center of efforts to roll back the "war on drugs" that many believe "is doing more harm than good" (Drug Policy Alliance, 2014). In November of 2014, for example, New Jersey voters passed a ballot measure amending the state constitution, which when coupled with recent legislation, authorizes a range of pre-trial reforms that would benefit non-violent offenders, including the substantial fraction charged to drug possession (Amick, 2014). Better risk prediction is a key feature of these reforms (VanNostrand, 2013). Our intent here is to show how arraignment release decisions for defendants charged with drug possession can be better informed by sufficiently accurate forecasts of FTAs. We apply a form of kernel logistic regression, which has some attributes as machine learning and can perform far better than conventional logistic regression.

There is ample precedent for our interest in arraignments. At least since the Manhattan Bail Project in 1961, there have been serious efforts to reform the ways bail decisions are made (McElroy, 2011). Among the most important changes have been to take quantitative risk assessments far more seriously. Numerical risk scales are used to inform bail decisions and procedural reforms (Clarke et al., 1976; Goldkamp and White, 2006; VanNostrand and Keebler, 2009; Bornstein et al., 2012; Arnold Foundation, 2013).

Section 2 provides a brief summary of criminal justice risk assessment and some recent advances. The field is undergoing important changes. Section 3 addresses kernel logistic regression in general and variants on it. We

summarize the underlying statistical theory, a rationale for kernel dimension reduction, asymmetric costs from forecasting errors, the role of tuning, and a proper context for statistical inference. We also briefly discuss the implementation of our procedure, which is found in the `R` package `kernReg` currently available on `CRAN`. Section 4 is devoted to forecasting whether after an arraignment defendants charged with drug possession later return to court when required to appear. The forecasting task is very challenging because of important omitted variables and little convincing theory to guide model specification. Results from conventional stepwise logistic regression and our proposed "regularized" kernel logistic regression are compared. The legitimacy of such comparisons is discussed as well. In Section 5, we conclude and provide recommendations to the practitioner who wishes to employ our methods.

2 Some Background on Criminal Justice Risk Assessment

Conventional criminal justice risk assessments have their roots in parole decision-making dating back to the 1920s (Borden, 1928). The roots run deep. Even very recent methods typically rely on scaling approaches developed by Earnest Burgess shortly after World War I (Burgess, 1928). Compared to clinical judgment or craft lore, they have served devision-makers well (Dawes et al., 1989; Grove, and Meehl, 1996).

Criminal justice risk assessment technology typically begins with one or more behavioral outcomes to be forecasted. Felony arrests are an example. A search for "risk factors" associated with the outcomes usually follows. Prior record, age and gender are among the risk factors usually found. The risk factors are ordinarily combined in a linear fashion to produce a numerical score. The higher the score, the greater the presumptive risk. Subsequently, when measures of the risk factors exist, but a behavioral outcome is not yet known, a risk score can be computed and used to make a forecast. The forecast either can be a score that attempts to capture the degree of risk or can be translated into a binary outcome class such as "fail" or "not fail" using a threshold on that score. Scores above the threshold forecast one class. Scores at or below the threshold forecast the other class. Although it is often difficult to determine how accurate such forecasts are (Reiss, 1951; Farrington and Tarling, 2003; Gottfredson and Moriarty, 2006; Ridgeway, 2013b), they are now routinely used to inform a variety of criminal justice decisions (Berk and Bleich, 2013).

Because the operational outcomes are typically binary, most of the risk assessment instruments developed over the past several decades have used logistic regression to determine the relevant risk factors. There is now strong evidence that more accurate forecasts can be obtained directly from machine learning procedures (Berk, 2012; Berk and Bleich, 2013; Ridgeway, 2013; Bushway, 2013) such as random forests (Breiman, 2001), stochastic gradient boosting (Friedman, 2002), Bayesian additive regression trees (Chipman et al., 2010) and support vector machines (Vapnick, 1998). These are "black box" algorithms able to construct complicated "profiles" that sort individuals into discrete outcome classes such as an arrest for one of several different kinds of crime. There is no need for explicit identification of risk factors if accurate forecasts are the primary goal.

Recent extensions of logistic regression that work well in samples of modest size can be seen as another alternative to conventional logistic regression. Regressors are "kernel" transformations of the original predictors that will often include *a priori* the same kinds of subtle profiles inductively discovered by machine learning. In addition, a variety of methods can be applied that down-weight features of the kernel having weak associations with the response variable. The combination of kernel transformations and down-weighting allows for the number of predictors to be as large or larger than the number of observations. It costs only additional computer time to introduce far more potential predictors than conventional regression can possibly manage.

3 Kernel Principal Components Logistic Regression (KPCLR)

The kernel transformations, a defining feature of all kernel regression methods, are just the beginning. They set in motion a set of cascading consequences that change conventional logistic regression from a model-based account of how the data were generated to an algorithmic procedure bent on maximizing the goodness-of-fit (Breiman, 2001b). We turn now to some background material needed to motivate the algorithmic procedure we later apply to forecasts of FTAs. Some of what we consider is novel, especially the manner in which we incorporate the asymmetric of forecasting errors.

3.1 The Data Generation Mechanism

Conventional regression treats all predictors as fixed, and all results are then conditional on the predictor values in the data. Often this is not responsive to how the data were generated or to the empirical questions being asked. An alternative treats the predictors as random variables, and regression procedures are altered accordingly (White, 1980; 1982; Buja et al., 2014; Berk et al., 2014).

We will be using kernel regression methods as forecasting tools. A defining feature of all forecasting is that the predictors are not fixed. New observations for which forecasts are needed materialize, typically on some regular basis. We must proceed, therefore, within a framework that treats all predictors as random variables. Several key points follow. Details can be found in the paper by Buja et al.(2014).

The data on hand are viewed as a collection of random realizations from a joint probability distribution. The joint probability distribution has mathematically defined expectations, variances, and covariances. As a result, one can treat the joint probability distribution as a population. Alternatively, one heuristically can think of all possible realizations of the random variables as the population.

In any given dataset, the realized random variables can be (and usually are) a subset of the random variables that constitute the joint probability distribution. There is the prospect of "omitted variables." Moreover, the variables chosen to be predictors (i.e., $\mathbf{X}$) and the variable chosen to be the response (i.e., Y), are determined by a researcher's interests, subject-matter expertise and availabe data. There is no innate property of the joint distribution itself that determines which random variable should be the response variable and which random variables should be the predictors nor what the functional forms connecting the predictors to the response should be. In short, it is very difficult to make a convincing case that any statistical formulation derived from those variables is specified correctly. It follows that the proper estimation target usually cannot be the "true" response surface, but only an approximation of that true response surface, whose fidelity with respect to the "truth" is unknown. This presents no special problems for forecasting because the goal is to arrive at the most accurate forecasts possible with the data on hand. We will see shortly that the forecasts can have good statistical properties.

3.2 The ANOVA Kernel

In a regression setting, one begins with the usual predictor matrix $\mathbf{X}$ that has N rows and p columns. Suppose $\mathbf{X}$ is subjected to a set of linear basis expansions represented by $\Phi(\mathbf{X})$. A simple example of a linear basis expansion is a polynomial function of each column in $\mathbf{X}$. If cubic, there are now $3p$ columns. Thus, $\Phi(\mathbf{X})$ has N rows and q columns, and q can be as large or larger than N, and in principle, even infinite. $\Phi(\mathbf{X})$ can be, in turn, transformed into an $N \times N$ matrix $\mathbf{K}$ called a kernel. That is $\mathbf{X} \rightarrow \Phi(\mathbf{X}) \rightarrow \mathbf{K}$. For the moment, we focus on $\mathbf{K}$.

Because there are many kinds of linear basis expansions, there are many kinds of kernels. At this point, there exists little formal justification for applying one kernel rather than another (Gross et al., 2012; Devenaud et al., 2013). But experience suggests that in regression applications, the ANOVA kernel will often perform well. (We provide a brief tutorial on kernels in Appendix B.)

Beginning with $\mathbf{X}$, the ANOVA kernel is defined by

$$k(\boldsymbol{x}, \boldsymbol{x}') = \left(\sum_{j=1}^{p} \exp\left(-\gamma(x_j - x_j')^2\right) \right)^d , \tag{1}$$

where x_j and x_j' are two different observations' values for predictor j in $\mathbf{X}$. These calculations do not produce $\Phi(\mathbf{X})$. We have gone directly from $\mathbf{X}$ to $\mathbf{K}$. The connection to $\Phi(\mathbf{X})$ will be addressed shortly.[1]

Because the computations begin with differences, which after being transformed are added together, the calculations are linear when $d = 1$, and one has a linear (additive) representation. When $d = 2$, one is working with products that can be seen as two-way interactions and a squared representation. By the same reasoning, when $d = 3$, one has three-way interactions and a cubic representation.[2] One might think that $\mathbf{K}$ could simply replace

[1] To some readers, the notation used in Equation 1 may be unfamilar. Here are the steps to compute the kernel matrix $\mathbf{K}$: (1) for observations i and j do an element by element subtraction over each of the predictors; (2) square each of the differences; (3) multiply each of these squared differences by minus γ; (4) exponentiate each of these products; (5) sum the exponentiated products; (6) raise the sum to the power of d; and (7) Repeat steps 1-6 for all pairs of observations i, j to compute all $N \times N$ entries in $\boldsymbol{K}$. A kernel matrix can be seen as a similarity matrix. Smaller cell entries off the main diagonal imply that the given pair of observations is more alike in their predictor values.

[2] To take a simple example, suppose there are three predictors. For the pair of observations from the first and second row of $\mathbf{X}$ with $\gamma = 1$ and $d = 1$, the sum of differences is $\exp\left(-(x_{11} - x_{12})^2\right) + \exp\left(-(x_{12} - x_{22})^2\right) + \exp\left(-(x_{13} - x_{23})^2\right)$. This is

X in logistic regression. This thinking is a step in the right direction, but a much larger step is needed.

3.3 Reducing the Number of Expanded Predictor Terms

We have denoted the realized predictors by **X** and their linear basis expansions by $\Phi(\mathbf{X})$. Suppose, for the moment that we know how to compute the $N \times q$ $\Phi(\mathbf{X})$. When $q \geq N$, $\Phi(\mathbf{X})$ is not a viable regressor matrix as is. The full set of regression coefficients cannot be uniquely determined.

There are several popular ways to reduce the number of expansion terms in $\Phi(\mathbf{X})$ or to down-weight them accomplishing much the same thing. One option is to apply a conventional principle components analysis (PCA) to $\Phi(\mathbf{X})$. The resulting N principle components (PCs) would be linear combinations of the expansion terms in $\Phi(\mathbf{X})$ constructed to be uncorrelated with one another and collectively to incorporate all of the predictive information in $\Phi(\mathbf{X})$, or more technically, in its covariance matrix. How this is accomplished is addressed in considerable mathematical detail in Appendix C.

PCs can be ordered from high to low by their contribution to the variance of the set of expanded predictors. Typically, only a leading subset of the N principle components is used in a regression analysis because the leading PCs capture most of the variance for the expanded set of predictors. As a result, the number of PCs included can be less than N, and the $q \geq N$ problem is circumvented. When principle components of $\Phi(\mathbf{X})$ are used as a regressor matrix, we call the resulting regression procedure Kernel Principle Components Regression (KPCR) or, for a binary outcome, Kernel Principle Components Logistic Regression (KPCLR).

Another recent regularization development is "penalized regression." The basic idea is to include a penalty for model complexity as part of the fitting function being minimized. The more complex the estimated model, the larger the penalty can be. This makes the estimated regression coefficients smaller in absolute value and the fitted values more smooth. An important and somewhat counterintuitive benefit can be fitted values that have better *out-of-sample* performance, which can be especially important in forecasting applications. However, unlike principle components regression, the number

linear and additive. For $d = 2$, the result is $[\exp\left(-(x_{11} - x_{12})^2\right) + \exp\left(-(x_{12} - x_{22})^2\right) + \exp\left(-(x_{13} - x_{23})^2\right)]^2$. All of the terms are now products of two variables, which are two-way interaction effects. For $d = 3$, the result is $[\exp\left(-(x_{11} - x_{12})^2\right) + \exp\left(-(x_{12} - x_{22})^2\right) + \exp\left(-(x_{13} - x_{23})^2\right)]^3$. All of the terms are now products of three variables, which are three-way interaction effects.

of regressors does not need to be less than N.

We have found our KPCLR approach to be more robust then penalized regression for the kinds of problems often endemic in criminal justice data: highly unbalanced outcomes, long tailed predictor distributions, very different costs for false positives compared to false negatives, and a large number of highly correlated predictors. We will employ the KPCLR approach in our application.

3.4 The Kernel Trick

The idea of applying PCA to $\Phi(\mathbf{X})$ might seem like a routine instance of conventional multivariate statistics. However, in order to do so, the transformations in $\Phi(\mathbf{X})$ must be known. In empirical settings, they rarely are. One then has no way to determine the relationship between q and N and no way to apply PCA to $\Phi(\mathbf{X})$.[3] We appear to be at a dead end. But let's look a little deeper.

Equation 2 shows the conditional expectation of the response as an additive function of a set of linear basis expansions of the predictors.[4] Y can be quantitative or binary, the columns of $\mathbf{X}$ contain $1, 2, \ldots, p$ predictors, and there are q linear basis expansion terms constructed from the full set of original predictors. The q terms of $\phi_m(\mathbf{X})$ we have shown earlier to be the columns of $\Phi(\mathbf{X})$.

In principle, one can obtain estimates of β_0 and each β_m that have the usual desirable properties. The same holds for $\hat{Y}$. Were Y quantitative, one might apply ordinary least squares. Were Y binary, one might maximize a binomial regression likelihood function.

$$\mathbb{E}\left[Y \mid \mathbf{X}\right] = \sum_{m=1}^{q} \beta_m \phi_m(\mathbf{X}), \tag{2}$$

However, one must specify each of the basis expansion terms for each predictor. This is a daunting task unless there was credible subject-matter theory specifying the expansion for each predictor and data available to implement those expansions. To take what may seem like a simple example, what is the correct set of expansion terms for the relationship between the

[3]Recall that the use of linear basis expansions means that $q \geq p$, usually substantially larger, and q can easily be larger than N. There is no way to know for sure, but prudence dictates that one assume the worst. Some kind of dimension reduction procedure should be applied.

[4]Because the data on hand are composed of random variables, the estimation target is a conditional expectation of the response, not a conditional mean.

age of an offender and recidivism? The relationship is well known to be negative and nonlinear (Berk, 2012), but there is a very large number of potential nonlinear functions. Moreover, the particular expansion terms used will depend on the kind of recidivism (e.g., all new arrests versus all new felony arrests) and features of the offender (e.g., gender).

Kernel methods respond in a remarkable way. As we address in substantial technical detail in the Appendix C, there is a formal mapping from the original predictors to a set of linear basis expansions to a particular kernel. Recall that when the ANOVA kernel was introduced, the mapping went directly from $\boldsymbol{X}$ to $\boldsymbol{K}$. The intervening $\Phi(\mathbf{X})$ was bypassed.

The kernel trick (Hastie et al., 2009: 660), justifies proceeding directly from the original predictors contained in $\mathbf{X}$ to $\mathbf{K}$ because $\boldsymbol{K} = \Phi\left(\boldsymbol{X}\right)\Phi\left(\boldsymbol{X}\right)^{\top}$. Thus, there is no need to ever compute the linear basis expansions. The mathematics of PCA dictate that information they contain is incorporated in $\mathbf{K}$ and is sufficient undertaking a principle components analysis of $\Phi(\mathbf{X})$. The corresponding PCs can then be incorporated into standard regression methods. It follows that within the kernel framework, the basis functions themselves are unknown and unrecoverable. They are locked up in the black box. But then, how can one determine if the linear basis expansions are any good?

Kernels are designed to incorporate *a priori* a very rich menu of expansion terms, often hand-tailored for particular applications. Popular kernels typically are battle-tested in real scientific and policy settings. The ANOVA kernel we favor for regression is one example. Moreover, if the goal is forecasting, a kernel is judged by its forecasting accuracy. Is the accuracy good enough to usefully inform the decisions to be made and more accurate than competing forecasting procedures that rely on functional forms specified ~~using~~ **via** subject-matter knowledge? We consider these issues in the application presented later.

One might still worry that without $\Phi(\mathbf{X})$, important information is lost. If the primary goal is to understand better *why* the predictors are related to the response, the loss is important. Explanation is severely compromised. If the primary goal is forecasting accuracy, the loss may well be irrelevant. The kernel matrix $\mathbf{K}$ incorporates the *predictive* information contained in $\Phi(\mathbf{X})$.[5] To summarize, one would like to reduce the number of columns in

[5]At the same time, the values in the kernel matrix can be instructive. Consider the first column of an ANOVA kernel as an illustration. The values in that column are the set of similarities the first observation has with all other observations. To what degree do other observations have a profile like the first observation? Then one can use regression to determine if defendants who are more similar to the first defendant more likely to no-

$\Phi(\mathbf{X})$. Principle components analysis is one good option. However, $\Phi(\mathbf{X})$ is not available. One only has $\mathbf{K}$. Thanks to the kernel trick, one can apply principle components analysis to $\mathbf{K}$ to arrive at the desired result.

3.5 Introducing the Relative Costs of False Positives and False Negatives

In criminal justice policy settings, the costs of false negatives and false positives will generally be different. For example, when a release decision needs to be made at arraignment, there are two kinds of mistakes that can be made. An individual is released and then fails to appear at a subsequent court hearing or an individual is not released and would have appeared. The consequences and costs of these mistakes are rather different and should be built into the forecasts of failure to appear; they should affect the forecasts themselves (Berk, 2012).

Forecasting procedures can differ dramatically in the mechanisms by which the different relative costs of forecasting errors are be introduced. For conventional logistic regression, a popular option is to impose threshold on the fitted values (Seed, 2010). Values larger than some threshold imply that the associated observations belong in one outcome class, and fitted values equal to or smaller than that threshold imply that the associated observations belong in the other outcome class. A widespread "default" is to ignore asymmetric costs and employ a threshold of .50.

If the logistic regression model meets all of the conventional assumptions, the fitted values can be interpreted as asymptotically unbiased probability estimates. Then, if estimated probabilities are greater than .50 one outcome class is forecasted, and if estimated probabilities are equal to or smaller than .50 the other class is forecasted. Implicit is that the costs of false positives and false negatives are exactly the same.

A variety of relative costs can be easily introduced using different thresholds. Suppose there are two outcome classes following an arraignment, arrested for a felony (coded "1") or not arrested for a felony (coded "0"). One might call the arrested class a positive and the non-arrested class a negative. Here, the 1/0 values and the terms "positive" and "negative" are

show in court. In effect, one shared profile potentially associated with an FTA is being identified. For the second column, there is similar reasoning: are defendants who are more similar to the second defendant more likely to be arrested? All other columns can be interpreted in the same fashion. The kernel trick can help provide answers to such questions, but does *not* reveal what those profiles actually are. This is a consequence of the black box.

assigned arbitrarily. A *false positive* would be incorrectly forecasting an arrest. A *false negative* would be incorrectly forecasting the absence of an arrest. Then, suppose stakeholders determine that false negatives are twice as costly as false positives. Imposing a threshold of .33 on the fitted values forces the false negative to false positive cost ratio of 2 to 1 on the forecasted class (i.e., .67/.33). That is, for a case to be forecasted as an arrest, its fitted value must be in excess of only .33. If not, the forecasted class is the absence of an arrest. It is "easier" to forecast an arrest, which is consistent with the stated cost ratio. In contrast, a threshold of .75 implies a cost ratio of 1 to 3 (i.e., .25/.75). False positives are now three times more costly. It is "harder" to forecast an arrest. We will see later that this simple approach for introducing asymmetric costs does not live up to its promise.

For KPCLR, the introduction of relative costs is done somewhat differently. When a logistic regression is run, the data are weighted so that the marginal distribution of the response is altered, and the logistic regression is fit with these weights. Suppose a positive is an FTA, a negative is the absence of an FTA, and false positives are taken to be twice as costly as false negatives. Case weights are given to the logistic regression so that *actual* positives have twice the weight of *actual* negatives.[6] Not surprisingly, this does not immediately produce results in which *false* positives are twice as costly as *false* negatives. As we describe in more detail shortly, the KPCLR fitting algorithm we employ gradually increases the complexity of the fitted values. When the increasingly complex fitted values arrive at the specified cost ratio, one can have the requisite cost-ratio result.

3.5.1 Tuning and Statistical Inference

Researchers are gradually coming to realize that model selection is not without its inferential perils. Conventional approaches in which model selection and statistical inference are undertaken with the same data risk serious bias in parameter estimates and highly misleading confidence intervals and statistical tests (Leeb and Pötscher, 2005; 2006; Berk et al., 2010). Unfortunately, the importance of tuning parameters puts us squarely in the middle of these problems. We have found no practical solutions besides using split samples. The recent literature is quite rich, but key problems are not yet solved (Berk et al., 2013; Lockhard et al., 2013; Voorman et al. 2014).

Split sample approaches properly implemented promise valid statistical inference at the price of reduced statistical precision (Berk et al., 2010) and

[6]The weights are introduced as a vector of length N with values that leave the effective sample size unchanged (i.e., they have a mean of 1.0).

more complicated data management. Often, this is a tradeoff well worth making (Faraway, 2014). The basic idea is to use different random subsets of the data for different data analysis tasks. Model selection, parameter estimation, and model forecasting performance are not undertaken with the same data.

Our approach has some novel features. We will proceed sequentially using the following operations.

1. The data are randomly split into three disjoint subsets we will call *training data*, *validation data* and *test data*.

2. Stakeholders supply relative costs of false positives and false negatives, and these two costs are used to construct case weights for the *training* data.

3. A set of promising ANOVA kernels is specified with different tuning parameters (γ and d) from which KPCLR models will be built. For reasons discussed earlier, we favor ANOVA kernels for regression applications.

4. For each kernel, a set of proportions of variance explained is defined (i.e., values for ρ) that will be used to determine the number of principle components provisionally included (e.g., $\rho = 35\%, \rho = 40\%, \ldots, \rho = 95\%$).

5. For each kernel in step 3 and each value of ρ in step 4, a KPCLR model is built with the *training data*.

6. Using performance with the *validation* data as a guide, one preferred KPCLR model is chosen by applying procedures to be explained shortly. We have not seen this approach in the existing literature on dimension reduction (e.g., Bühlmann and van de Geer, 2011).

7. Forecasting accuracy is determined for the model selected in Step 5 by predicting into *test* data using the preferred regression model built from the *training* data.

These steps come bundled with many demanding particulars. To begin, there is currently no clear statistical guidance on the relative sizes of split samples, even when there are only two (Faraway, 2014). A lot depends on features of the data and the models being used. Samples of equal size are often reasonable; this we recommend as a default.

The weights used in the logistic regressions are determined *a priori* by the cost ratio of false positives and false negatives. Once a cost ratio is determined, the weights follow as a simple mathematical exercise. For our analysis of FTAs, false negatives are taken to be twice times the cost of false positives. Therefore, all actual non-FTA cases are given three times the weight of all actual FTA cases, scaled so that the effective sample size is unchanged. Perhaps counter-intuitively, this cost ratio implies that before a defendant is projected to a good risk for release, the statistical evidence must be quite strong.

Sets of values for the ANOVA kernel parameters γ and d must be specified. This will be easier if all predictors in $\mathbf{X}$ are standardized as z-scores.[7] For example, d might be 2 or 3, and initial values for γ could range from .01 to 100 as multiples of 10: .01, .10, 1.0, 10, 100. The search could then become more concentrated around the most promising initial values. Because the models are evaluated with out-sample-performance, the primary penalty of model searching is additional computation. We have found that a consideration of five to ten models can be sufficient and provide example in Section 4.

The number of principle components needed as regressors must be determined empirically by fitting models with different proportions of variance explained (e.g. $\rho = 30\%, 35\%, \ldots, 95\%$). For each combination of d, γ, and ρ, a KPCR model is built. This may seem quite daunting, but with our software, a rich set of models can be produced in a matter of minutes because many computationally demanding steps have been parallelized.

Model evaluation follows. For each model, we provide (1) the number of false negative errors and the number of false positive errors when predicting into the *validation* data, (2) the aggregate cost-weighted error when predicting into the *validation* data and (3) the proportion of PCs used. A first cut eliminates all models whose cost ratio of false negatives to false positives is not sufficiently close to the stakeholder-specified cost ratio. Such models are not responsive to stakeholder policy preferences. A second cut eliminates all models with an adequate cost ratio, but disappointing forecasting accuracy. Among the models that make the second cut, preference is given to models that use fewer principle components. There seems to be no point in wasting degrees of freedom. The reduced model we refer to as the "selected" model.

There are some tricky issues at this stage because of the temptation to treat each KPCLR as a usual linear model popular in the social sciences.

[7] If the predictors are not standardized, the values of the turning parameters can be dramatically affected by the units of measurement, which are here a distraction.

None are. Once the regressors are kernelized, logistic regression becomes a black box algorithm from which to construct useful forecasting procedures. There is no intent nor capability to reveal the subject-matter mechanisms by which the response is related to the original predictors. In addition, most of the usual regression diagnostics are not relevant and can even be misleading. One key reason is that we are interested in forecasting accuracy for each of the outcome classes (e.g., fail or not fail), not the logistic regression fitted values. Yet, most regression diagnostics build on the in-sample fitted values, often treated as probabilities.

For the selected model, the *test* data (i.e., the third random split) are then used to provide an honest assessment of forecasting accuracy. These are out-of-sample assessments because the third split had no role when constructing the model and no role when selecting the best model. Thus, they are not subject to overfitting. Nevertheless, there may be residual concerns about overfitting when earlier the "best" KPCLR model is selected. Just as in boosting of binary outcomes, the iterative process can produce dramatic overfitting. Yet in general, overfitting of the fitted value in *training data* can actually benefit *out-of-sample* forecasting accuracy (Mease et al., 2008).

This counter-intuitive result makes sense in the machine learning world of classification. With a rich menu of predictor transformations summarized by an increasing number of principle components, any training data fit will improve until it can improve no more. That fit will have two components. The first is systematic features of the training data that may normally be very hard to find, but can be captured by a sufficiently rich kernel expansion and a sufficient number of its principle components. The second is idiosyncratic patterns in the training data swept up in fitting process that are not features of the joint distribution from which the data came. The latter are often characterized as *chance variation*. When people speak of the dangers of overfitting, they are referring to the fitting of this chance variation only.

If one constructs a forecasting procedure and evaluates its performance using the same data, the two components cannot be disentangled. There is, then, a genuine reason for concern. But if there are data not used to build the forecasting procedure that can be used for performance assessments, one can obtain honest performance evaluations without the contaminating effects of overfitting. One is then left with the benefits of the hard-to-find, systematic features of the data.

We capitalize on just such thinking. KPCLR models are undertaken with the *training* data, in which overfitting can be a virtue. KPCLR model selection is done with the second split, the *validation* data to help counteract the misleading properties of in-sample assessments. Once a KPCLR model

is chosen, an honest evaluation of forecasting accuracy is provided by the third split, the *test* data. For the KPCLR model chosen, this evaluation is not a product of overfitting.

With this split sample approach, one has forecasting tools with good statistical properties. Perhaps most important, one has a forecasting procedure that provides asymptotically unbiased forecasts derived of the population response surface *approximation* (Buja et al., 2014). But, the price should now be clear. Each subsample will have many fewer observations than the original data set, and forecasting accuracy can be substantially reduced. In addition4, any estimator properties that depend on conventional asymptotics can be put in harm's way. In the example we turn to shortly, sample sizes will be relatively large to minimize the consequences of these difficulties.

4 An Application

For a large metropolitan area, we develop procedures to forecast FTAs among defendants charged at arraignment with drug possession. The long term goal is to develop forecasting procedures to help inform release decisions by magistrates. FTA is defined as a failure to appear in court after a preliminary arraignment in which formal charges are not dismissed.

As noted earlier, we focus on the subset of defendants charged with drug possession because they can be a substantial FTA risk, constitute a significant fraction of pre-trial defendants, and are at the center of many "bail reform" efforts. A common from of stepwise logistic regression and KPCLR are applied in an effort to obtain serviceable forecasts.[8] Their comparative forecasting performance is also a major interest. Both procedures will be challenged because, as we explain shortly, the forecasting task is very difficult.

We treat the forecasting exercise as illustrative. Should the results be sufficiently promising, a new analysis probably will be needed using the most recent data available at that time. It might also be possible to pool current data with new data, or even fold in older data, if the mix of defendants and circumstances surrounding FTAs has been relatively stable. Also, stakeholder preferences are still quite fluid. There are several other subsets of defendants that could be of interest or more likely, all defendants will be included in a single analysis. Likewise, the cost ratio of false negatives to

[8] A stepwise procedure is used because researchers typically select a relatively few "risk factors" from among a much larger set of predictors.

false positives is provisional.

4.1 The Data

We obtained data for defendants released during the most recent two weeks available that allowed for a two year follow-up period: the last 15 days of October, 2011, with the follow-up period ending on October 31st, 2013.[9] There are 596 observations. Each observation is a single bail decision for a single defendant referred to as a "case." Although in principle, a defendant can appear in the data more than once because of separate arrests for two or more crimes, the two-week interval effectively precludes that possibility. We treat the data as a set of realizations from the joint probability distribution applicable for those two weeks, although it probably applies far more generally.

From existing electronic criminal justice records, we have the usual background variables such as age, gender, and prior record. We have the types of charges heard at the arraignment. And, we have each defendant's prior record as a juvenile. For prior record as either an adult or a juvenile, we have the date on which each arrest occurred. Finally, we have any records of failures to appear in court for two years after their release.

Although an FTA is a violation, it often has a different etiology from conventional street crimes. Bornstein and his colleagues (2013) argue that FTAs often result from a faulty memory, an inability to get timely transportation, household responsibilities such as child care, or work-related obligations. Views about the fairness of the adjudication process can matter as well. They also show that written reminders can reduce FTA rates, especially if the reminder provides information about the negative consequences of a failure to appear.

Unfortunately, among our potential predictors we have virtually no measures of such factors. It follows that our ability to accurately forecast FTAs is seriously compromised from the start. We stress, however, that this application was not selected to make any particular statistical point. We had no idea how accurate our forecasts would be prior to our initial forecasting attempts. Nor did we know in advance how the different forecasting methods would perform. The forecasting problem was brought to us by real stakeholders who were seeking technical assistance. In retrospect, however, this forecasting exercise is well suited to our methodological purposes. When a

[9] Without a release, it would be impossible to learn how the individual performed "on the street."

forecasting task is easy, most any forecasting procedures will do well. Difficult forecasting tasks allow one to consider the relative performance of different methods.

4.2 Predictor Variables

There are 41 predictors. We have a few biographical variables such as age, and gender, but most come from adult and juvenile rap sheets and current charges.[10] Many of the predictors are correlated with one another, often strongly. Logistic regression can be sensitive to high multicollinearity, but with our main interest in forecasting accuracy, it does not cause serious problems. Fitted values are less affected by multicollinearity than estimates of the regression coefficients, and the dimension reduction tools we apply moderate much of the remaining instability.

As already noted, there is virtually no information on routine life circumstances that might influence a failure to appear. This makes the forecasting challenge substantial. Ideally, the variables we are able to include can serve at least in part as proxies for more important omitted predictors. The full set of predictors is listed in Appendix A.

4.3 Cost Ratios

For this exercise, a "positive" is an FTA. A "negative" is the absence of an FTA. A false positive is incorrectly predicting that an individual will not appear in court when ordered to do so. A false negative is incorrectly predicting that an individual will appear in court when the individual actually will not. For criminal justice stakeholders, both forecasting errors are undesirable and both have costs. When researchers accept the default procedures provided by the usual logistic regression software, they are accepting a default fitting function is which the two costs are treated the same: the costs are symmetric, and their cost ratio is 1 to 1. This is usually inconsistent with the preferences of stakeholders who will use and be affected by the forecasts.

[10]The rap sheet data only include arrests from within the state in which the metropolitan area is located. But most crime, like most politics, is local. Relatively few arrests are missed. Moreover, the impact of priors on forecasting accuracy is highly nonlinear. Although there are special concerns about "frequent flyers," variation of several prior arrests (e.g., 35 arrests versus 40 arrests) is for such offenders unrelated to forecasting accuracy. If out-of-state arrests matter for prediction, it is for offenders with very few in-state arrests. But then, there is some evidence that other predictors pick of the slack.

At the time the project began, the metropolitan area in which the forecasting was to be done had serious resource constraints. There was insufficient jail space should the forecasts produce a large number projected FTAs cases requiring incarceration. Many other jurisdictions face similar problems (VanNostrand, 2013). A key implication was that the false positives could be in the aggregate very costly. At the very least, there could be serious "overcrowding." Less costly options were being evaluated, but many had little demonstrable impact or were burdened by legal and political constraints. For example, methods that have been used to monitor individuals on probation may turn out to be illegal for defendants who have yet to be adjudicated.

There could also be collateral damage to the defendant if he or she were held unnecessarily. A job could be lost. Important contribution to a family, such as child care, could be disrupted. And incarceration in the county jail would at best be very unpleasant.

On the other side, defendants at arraignment have been charged with criminal offenses, often serious offenses that can be serial in nature. Should such defendants not return to court, they cannot be held accountable unless they are apprehended again. In this jurisdiction, FTAs often took a law enforcement back seat to other law enforcement priorities. Word on the street was that one might well be able to get away with ignoring court orders. There were also public relations issues should a defendant released awaiting trial be arrested for a violent crime.

Balancing all of these considerations is difficult and easily affected by changes in policy options. For example, if less costly alternatives to jail time could be found that effectively reduced FTAs, the preferred cost ratio of false negatives to false positives could change as well. Cost-effective diversion programs for drug users are an obvious example.

At this point, we provisionally set the cost ratio to favor accurate forecasts for those who are released. This means treating false negatives as more costly than false positives. Before an individual is released, the statistical evidence must be relatively strong. Relatively weak statistical evidence that someone will not return to court as required we will be enough to preclude a release. Because of the cost implications as currently understood, the cost ratio is set at 2 to 1. In many other criminal justice settings, the false negative to false positive cost ratio has been 10 to 1 or higher (Berk, 2012). The 2 to 1 cost ratio is an acknowledgement of the pressures false positives could place on available jail space.

4.4 Results for Stepwise Logistic Regression

The KPCLR analysis was designed around three even splits of the available data. How should the data be sampled for a stepwise logistic regression so that fair comparisons to the KPCLR results can be obtained? Because the KPCLR forecasting performance would be evaluated with the test sample of 199 cases, we decided that same should apply to the stepwise logistic regression. Therefore, two thirds of the data would constitute the training sample because there was no need for a validation sample. The remaining third would be used as the test sample. If anything, doubling the size of the training data would favor the stepwise logistic regression.

All of the available predictors were used in a stepwise selection (backward elimination using the AIC) applied to the training data. The result was a smaller model with 17 predictors and estimates for the regression parameters of that smaller model.[11] Fitted values were then obtained from the test data.

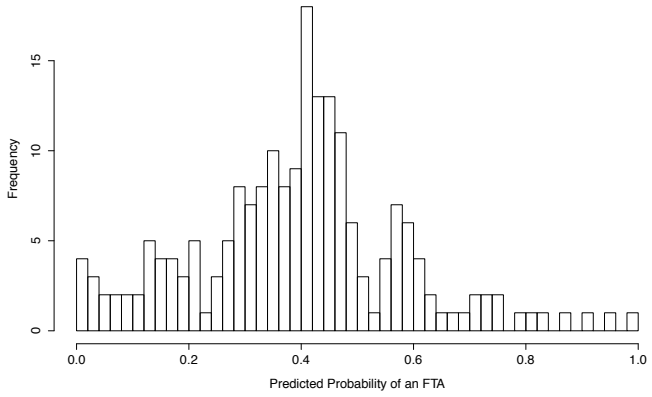

Figure 1: Histogram for Out-of-Sample Stepwise Logistic Regression Fitted Values (N=199)

Figure 1 shows that the fitted values from the test data are centered a little above .40. The median is .39, the first quartile is .30, and the third quartile is .48. But there are also a few cases at 0.0 and 1.0. It is clear that a substantial majority of the fitted values fall below .50. Some may choose to interpret these fitted values as probabilities, but because the

[11] We used the stepwise regression in `R` (stepAIC), which is part of the `MASS` library. Very similar results were obtained used forward selection.

logistic regression model is surely misspecified, it is not clear to what chance process these probabilities refer.

To arrive at forecasted classes, a threshold on the fitted values is required. Using the 2 to 1 cost ratio, that threshold is .33. (.67/.33 $\cong$ 2). Table 1 is the resulting "confusion table," a cross-tabulation of the actual outcome class by the forecasting outcome class. One would ordinarily examine the confusion table for insights about forecasting accuracy, but there is a major obstacle. The empirical cost ratio in the table should be about 2.0. But that ratio is actually 6.0 (84/14), triple the cost ratio that stakeholders specified. False negatives are being given far too much weight; they are being treated as far more important than stakeholder wish them to be. Any forecasts, therefore, are not responsive to the stated policy preferences. In particular, the limited available of jail space is being significantly discounted.

	Predict No FTA	Predict FTA	Model Error
Actual No FTA	50	84	0.63
Actual FTA	14	51	0.21
Forecasting Error	.21	.62	

Table 1: Failure to Appear (FTA) Stepwise Logistic Regression Confusion Table Constructed from the *Test* Data (N = 199)

But it's worse. The recommended logistic regression fix for settings in which the costs of false negatives and false positives are not the the same does not work as claimed (cf. Seed, 2010). Any confusion table depends on the distribution of the fitted values as well as the imposed cost ratio. Suppose an imposed cost ratio is .33. If all of the fitted values happen to be less than .33, no defendants would be forecasted to fail. If all of the fitted values happen to be greater than .33, all defendants would be forecasted to fail. In one case, the empirical cost ratio would be 0.0 and in the other case, the empirical cost ratio would be undefined. In short, there is no direct mapping from imposed cost ratio and the empirical cost ratio because the empirical cost ratio depends on the distribution of the fitted values. We address this problem explicitly in the KPCLR analysis.

4.5 Results for KPCLR

We proceeded with the very same three random splits of the data applying the procedures of Section 2.5, Steps 1-6, and searching over a parameter

grid with $d = 2$ or $d = 3$ and γ values of .01, 1, 3, 1, 3, 10, or 100. For each of the candidate kernels, the values of ρ were fixed at $30\%, 35\% \ldots, 95\%$. Each kernel's performance in a logistic regression was judged as principle components were added in 5% increments of the kernel variance accounted for. Regression parameters were estimated with the training data, and performance was evaluated using the validation data. Forecasting performance was assessed with the test data. Four promising ANOVA kernels were found: $(\gamma, d) = \{(3, 3),\ (10, 3),\ (100, 2),\ (100, 3)\}$

Figure 2 shows the diagnostic plots for the four kernels used to determine the preferred model (Step 6 in Section 2.5). Recall that for each logistic regression, the training data had been weighted by the imposed cost ratio of 2 to 1. These plots were constructed from the validation data, with the goal of arriving at an empirical cost ratio that was effectively the same as the imposed cost ratio combined with a small cost-weighted error and a small value for ρ.

On the horizontal axis are values of ρ, the fraction of variance of the transformed predictors accounted for by the principle components. With larger values of ρ, more PCs are used as predictors. The left vertical axis and black line show the ratio of the number of false negatives to the number of false positives. Because the target cost ratio of false negatives to false positives was 2 to 1, the goal was to arrive at empirical results in which there were two false positives for every false negative — two false positives are "worth" the same as one false negative. Therefore, there is a horizontal line at .5 representing the desired result (i.e., $1/2 = 0.5$). There are also horizontal lines as 0.0 and 1.0 defining a band in which the ratio of false negatives to false positives is reasonably close to .5. The right vertical axis and red line show the cost-weighted number of forecasting errors in the validation data.

The vertical blue line in the graph on the upper right, shows the value of ρ for the selected kernel. The target ratio of false negatives to false positives is achieved, the value of ρ is relatively small, and the cost-weighted error is as well. Note that each of the other kernels offer potential solutions that lead to similar results.[12]

Figure 3 shows the fitted values when the preferred regression is applied to the test data. The mean is .56, the first quartile .34 and the third quartile .77. Compared to the fitted values from the stepwise logistic regression, the

[12]With real data, there can be situations in which none of the proposed kernels are able to produce the desired cost ratio. As an empirical matter, the analysis fails. Here, there are several successful kernels. Because the kernels are not given substantive interpretations, it does not matter which one is chosen as long as forecasting accuracy is acceptable.

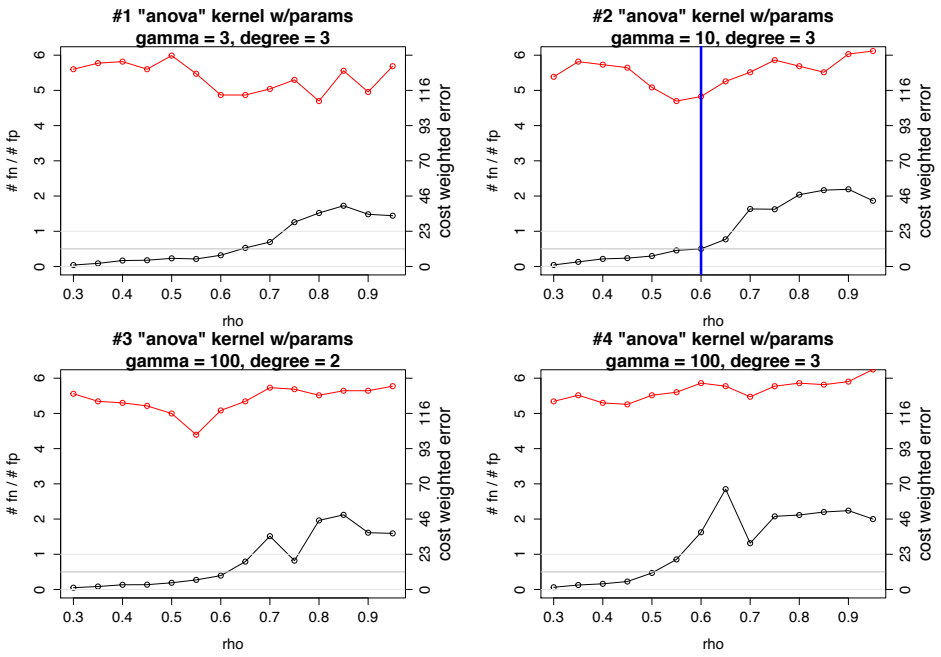

Figure 2: KPCLR Performance as a Function of ρ

fitted values are far more dispersed with much thicker tails, especially at the high end — the standard deviation of the fitted values increases from .19 to .26. Greater distinctions are being made between defendants; the fitted values discriminate better. Moreover, because the fitted values are asymptotically unbiased estimates of the fitted values in the population response surface *approximation*, they can be treated as estimates of the conditional expectations, which may be interpreted as conditional probabilities. None of the model misspecification issues that undermined the stepwise logistic regression are relevant because the estimation target is not the "true" model.

Table 2 shows the out-of-sample forecasting results from the test data. The cost ratio of false negatives to false positives approximates 2 to 1 quite well. (i.e., 64/26 = 2.5 which is within 25% of the target). In general, it is virtually impossible to hit the cost ratio exactly because, as required, the data were not tuned using the test data. Random sampling error complicates matters a bit, especially in small samples, but here, the difference between 2.0 and 2.5 makes no practical difference whereas the difference between 2 to 1 and 6 to 1 significantly consequential.

50% of the non-FTAs are incorrectly classified. 34% of the FTAs are

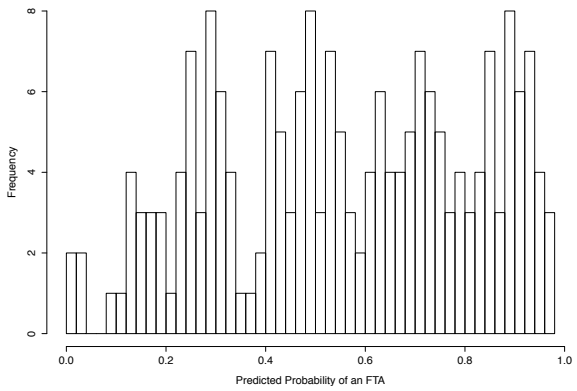

Figure 3: Histogram for Out-of-Sample Fitted Values from the KPCLR Procedure

incorrectly classified. By current standards (Berk, 2012), KPCLR performs reasonably well despite very weak predictors. The imbalance in model performance is an intended consequence of the 2 to 1 cost ratio.

	Predict No FTA	Predict FTA	Model Error
Actual No FTA	59	64	0.50
Actual FTA	26	49	0.34
Forecasting Error	.30	.56	

Table 2: Failure to Appear (FTA) Confusion Table Constructed from *Test* Data (N=199) for the KPCLR Procedure.

But, classification accuracy is a secondary consideration. In this research setting, what matters is forecasting accuracy. Within the data we have, nearly 36% of defendants arraigned for drug possession, who are not held in jail, subsequently fail by an FTA. Table 2 indicates that were the court to release individuals forecasted to return to court when ordered to do so, 30% would fail by an FTA (26/(26+59) = .30). Clearly, this is a modest improvement in percentage terms. But just as clearly, it is a policy-relevant improvement for a court system that arraigns about 15,000 drug possession offenders a year. If our results we used to determine whom to release, there could be approximately 800 fewer FTA incidents. 800 is a big number from a practical point of view. Modest relative improvements in the performance of large criminal justice systems can translate into dramatic absolute gains

that can make an important, real world difference.[13]

It is also possible to do much better. It cannot be overemphasized that the 30% failure rate among those predicted to succeed results substantially from the asymmetric 2 to 1 cost ratio. If as a policy matter, false positives were given less weight, more true negatives would be correctly identified, and the projected FTA failure rate would be reduced, perhaps dramatically.

One legitimately might wonder about the uncertainty in all of our KPCLR results. The data are treated as a set of random realizations from a joint probability distribution, and the analysis was conducted with three random splits of the data. Sampling variability is built in. However, we know of no formulaic way to represent the uncertainty in a credible manner. There may well be re-sampling strategies, but they come with important complications. At a practical level, each KPCLR would need to be hand tuned for every one of a large number of samples. At a theoretical level, uncertainty can be a particular problem in small samples because re-sampling procedures depend on asymptotics for their credibility (Efron and Tibshirani, 1993). When you may most need re-sampling, it is least well justified. And for these data, there is the additional challenge of highly non-normal distributions — more will be required of the asymptotics. Still, we intend to explore these issues further in the future. It has long been recognized that early bootstrap methods left lots of room for improvement, especially in small samples (Efron, 1987). Techniques like the bootstrap calibration (Loh, 1991) and the closely related double bootstrap (Nankervis, 2005) perhaps can help. In addition, we have some ideas about how the tuning can be automated.

We also applied random forests (Breiman, 2001a) to the data to see if KPCLR could be benchmarked against a machine learning procedure that has been successfully used for criminal justice forecasting. For comparability, random forests was trained on the 2/3rds sample and tested on the 1/3rd sample. The 2 to 1 cost ratio was imposed. Unfortunately the small size of the test sample meant that random sampling error was an important factor in all comparisons. Additional random sampling error was added because of the sampling built into the random forests algorithm. We made several comparisons, each with new random splits of the data. Both procedures were always able to approximate the 2 to 1 cost ratio reasonably well. Sometimes KPCLR forecasted more accurately. Sometimes random forests forecasted

[13]We have not considered here the possibility of additional benefits of requiring at least some individuals to post a bond before release. That is, these figures necessarily represent the policy status quo.

more accurately. Both always did better than the marginal FTA rate of nearly 36%. In short, KPCLR and random forests probably appear to be effective alternatives to logistic regression, but in small samples any further conclusions can be obscured by noise.[14]

5 Conclusions

Even with a very weak set of predictors, forecasts that a defendant will return to court as ordered can be serviceably accurate when constructed from a kernelized principle components regression (KPCLR). The forecasts can be made more accurate with cost ratios that increase the relative costs of false negatives. If one had information about defendants' life circumstances and views of the criminal justice system, there would be an additional way to improve such forecasts. Finally, were there effective interventions, such as mailed or texted reminders, the number of defendants who returned to court as required could be increased and made more predictable. Although at this point our findings only apply formally to a single jurisdiction, they support recent efforts elsewhere to reform pre-trial procedures.

As a technical matter, KPCLR seems to be a useful alternative to conventional logistic regression when a researcher's primary interest is in classification and forecasting. By constructing in advance a rich menu of basis functions, complicated non-linear relationships and interaction effects often can be captured. Kernelized regression attempts to anticipate complexities that inductive methods like random forests discover on-the-fly (Brieman, 2001).

Can conventional logistic regression compete? When the predictor distinctions between outcome classes are uncomplicated, even very simple logistic regression models can perform well. Machine learning and kernel methods may then confer no special advantage. But as others have emphasized (Berk and Bleich, 2013, Ridgeway, 2013, Bushway 2013), one cannot know in advance how complicated the key relationships are. Prudence will often dictate assuming the worst.

When there is subject-matter knowledge indicating in a convincing manner how to correctly specify a regression model and data available to properly implement that correct specification, conventional logistic regression

[14] It would be possible to undertake a large number comparisons with random spits of the data and compare mean forecasting accuracy. Random sampling error would tend to cancel out. But whatever the conclusions, they would necessarily be data and cost ratio specific. Little of general interest would be learned.

may have no downside. One can have it all: excellent forecasts and genuine explanatory insight as well. But such requirements seem unrealistic for the processes by which some people return to court as ordered and some do not. There is probably not a jurisdiction in the United States with the requisite subject-matter knowledge and data. Prospects for both in the medium term are not encouraging. The use of Burgess-like scales constructed from risk factors determined by logistic regression have done yeoman service in the past. It is time to move on.

Acknowledgements

We would like to thank Larry Brown, Andreas Buja, Ed George, and Dean Foster for helpful discussions. Adam Kapelner acknowledges the National Science Foundation Graduate Research Fellowship for support.

Appendix A — Candidate Predictors

1. The age of the offender
2. The gender of the offender
3. The zipcode in which the offender lives if it is a high crime zip code
4. The length of the follow up period
5. The total number of prior charges as a juvenile
6. The number of "serious" prior charges as a juvenile
7. The number of "violent" prior charges as a juvenile
8. The number of sex crime prior charges as a juvenile
9. The number of firearm prior charges as a juvenile
10. The number of weapon prior charges as a juvenile
11. The number of drug prior charges as a juvenile
12. The number of property crime prior charges as a juvenile
13. Whether there was any prior charges as a juvenile
14. Whether there was any violent prior charges as a juvenile
15. The age of the first adult charge while a juvenile
16. The number of prior murder charges as an adult
17. The number of "serious" prior charges as an adult
18. The number of "violent" prior charges as an adult
19. The number of sex crime prior charges as an adult
20. The number of firearm prior charges as an adult
21. The number of weapon prior charges as an adult
22. The number of drug prior charges as an adult
23. The number of property crime prior charges as an adult

24. Whether there were any charges at the arraignment
25. The number of murder counts at the arraignment
26. The number of weapons counts at the arraignment
27. The number of property crime counts at the arraignment
28. The number of drug distribution counts at the arraignment
29. The number of domestic violence counts at the arraignment
30. The number of violent crime counts at the arraignment
31. The number of serious crime counts at the arraignment
32. The number of sex crime counts at the arraignment
33. The number of firearm crime counts at the arraignment
34. The number of drug possession crime counts at the arraignment
35. The number of sex crime counts at the arraignment
36. The number of prior FTAs
37. Whether the individual is currently on probation
38. The number of prior abscondings
39. The number of prior probation violations
40. The number of prior days in jail
41. The number of prior confinement days

Appendix B — A Brief Tutorial on Kernels and Regression

Kernel Regression Methods for Forecasting

In a conventional regression analysis, a primary goal is to represent how one or more predictors are related to a response. Often those relations are interpreted as causal. But there can also be interest in the fitted values. Sometimes the fitted values are plotted to provide information about the

possible nonlinear functional forms. There may be no regression coefficients to interpret, but the intent is still to characterize how the predictors are related to the response. Partial response plots used with generalized additive models are a good illustration (Hastie and Tibshirani, 1990).

Sometimes the fitted values by themselves are of interest. For example, when the response is categorical, the fitted values can be used for classification. The goal might be to determine whether particular transactions are fraudulent. Or the goal might be to provide a diagnosis for patients exhibiting certain symptoms. Such goals do not require that the relationships between the predictors and the response be captured in ways that are substantively interpretable. For instance, in principle components regression, the regressors are linear combinations from the full set of original predictors. Although post hoc interpretative overlays are sometimes employed, how the predictors are related to the response is typically obscured. The fitted values are the essential motivator.

Forecasting is another activity in which the role of predictors need not be a primary concern. An investor might be deciding which energy futures are a good bet based on their forecasted returns a year hence. Or a parole board may decide which inmates to release based on forecasts of whether a violent crime will be committed. One may achieve excellent forecasting accuracy with no real understanding about how the predictors are related to the response (Berk, 2012; Berk and Bleich, 2013; Ridgeway, 2013b, Bushway, 2013). Indeed, it is often productive to make forecasting and explanation separate data analysis objectives.

If the focus can be exclusively on forecasting, one has the opportunity to employ predictors in a manner that may dramatically improve forecasting accuracy even if explanation is severely compromised. Machine learning is a set procedures that commonly makes this tradeoff; kernel methods do likewise.

Here, we focus on kernel methods for regression applications in which the primary interest is in fitted values and subsequent forecasts. We will see that when complicated nonlinear and/or interaction effects may be needed, but the precise functions are unknown or the relevant variables are not in the data set, kernel methods can automatically assemble a very rich menu of functions that may serve as an effective alternative. Forecasts of useful accuracy can follow. Although we will later emphasize regression with binary outcomes, kernel methods can in principle be used in any form of regression when one is trying to characterize the distribution of a response variable conditional on a set of predictors.

Linear Basis Expansions

Linear basis expansions can be building blocks for a wide variety of statistical procedures and are a critical starting point for a discussion of how kernels can be employed in regression (Hastie et al., 2009: Section 5.1). For a single predictor X,

$$f(X) = \sum_{m=1}^{q} \beta_m \phi_m(X), \tag{3}$$

where the predictor X is replaced by a sum of q transformations of X, each transformation represented by $\phi_m(X)$. A cubic function in X is a simple illustration: $\phi_1(X) = X$, $\phi_2(X) = X^2$ and $\phi_3(X) = X^3$. Here, $q = 3$, and the three transformations of X are X, X^2, and X^3; $\phi_1(X) = X$ is a "trivial transformation." The corresponding weights β_1, β_2, and β_3 can be conventional regression coefficients. Where there was initially a single function of X, there are now three functions of X: hence the term "expansion." Other kinds of linear basis expansions include trigonometric functions, indicators variables, and various types of splines (Haste et al., 2009, Chapter 5). The formulation also is readily extended to more than one predictor. When linear basis expansions are used kernel applications, it is common to used the notation $\Phi(\mathbf{X})$ to represent the collection of linear basis expansions for the full set of predictors.

The benefits from linear basis expansions depend on two potential consequences of Equation 3. First, the expansion can directly alter how the relationships between the response and the predictors are characterized. In the example just given, a cubic function may fit the data better. Second, by transforming the *space* in which the observations are located, patterns may be found that are otherwise obscure. Precisely how this can be done is considered in later sections. For now, Figure 1 illustrates both possibilities.

The upper part of Figure 1 shows a scatter plot with two predictors, x_1 and x_2. For example, x_1 could be the number of prior arrests, and x_2 could be age.[15]

The open circles represent one of two response outcomes (e.g., rearrested while on parole). The solid circles represent the other response outcome (e.g., not rearrested while on parole). There are also two overlays representing two different decision boundaries. The solid line is a linear decision boundary. The dashed line is a nonlinear decision boundary.[16]

[15] For ease of exposition, we are playing a little fast and loose with notation at this point because formally both predictors are vectors. We will get more formal shortly.

[16] The term "decision boundary" is used because depending on which side of the decision

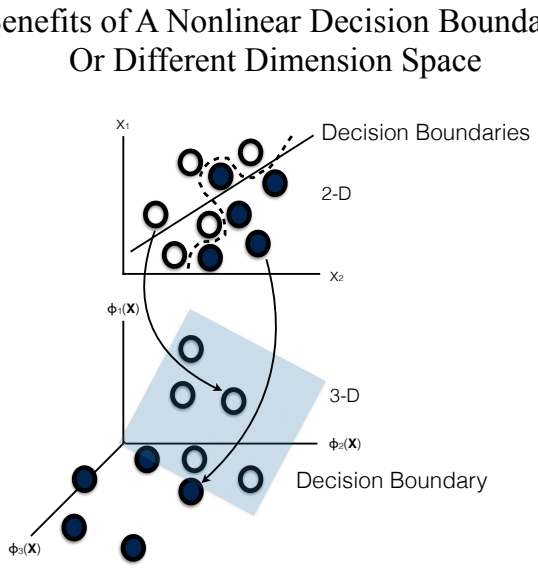

Figure 1: An Illustration of the Gains From a Nonlinear Fit (top figure) or a Transformed Predictor Space (bottom figure)

The data analyst's task is to partition the predictor space using a decision boundary so that all of the open circles fall in one partition and all of the solid circles fall in the other partition. If one were able to do so, the pair of predictors x_1 and x_2 would be able to perfectly distinguish between the two outcomes. The predictors would be able to classify these observations without error.

It is impossible here to find a linear decision boundary in the 2-D predictor space that perfectly distinguishes between the open and solid circles. Any linear attempt to classify cases in these two dimensions will result in two partitions of the space, with at least one having a mix of open and solid circles. More technically phrased, there can be no linear "separation" between the open and solid circles. In Figure 1, for instance, one solid circle

boundary an observation falls, different decisions about that observations can be justified. For example, one decision might be to release an inmate on parole and another decision might be to be keep the inmate incarcerated.

falls in the partition dominated by open circles.

However, Figure 1 shows a nonlinear decision boundary that can partition the 2-D predictor space discriminating the two response types perfectly. Nonlinear transformations of the predictors can in principle be helpful in precisely this way.[17]

The lower part of Figure 1 has three predictors, $\Phi_1(\mathbf{X})$, $\Phi_2(\mathbf{X})$, and $\Phi_3(\mathbf{X})$, where $\mathbf{X}$ is matrix notation representing both x_1 and x_2. Thus each predictor is a different basis expansion term using *both* x_1 and x_2. For example, the expansion might be a cubic polynomial of an element by element product $x_1 \times x_2$, where $q = 3$.[18]

Each dimension represents a term of the expansion that as a group defines a 3-D space in which the observations can be located. In this new space, one can see that the open circles are separated perfectly from the solid circles because the former are located toward the back of the figure, and the latter are located toward the front of the figure. Therefore, it is possible to construct a 2-D plane that can perfectly discriminate between the two response types.

Linear basis expansions are easily extended to higher dimensions. Figure 1 provides an initial sense of the benefits that we will addressed in more depth later. But in practice, perfect separation is still very difficult to achieve. Rather, we seek substantially improved separation.

Kernel Functions and Kernel Matrices

A powerful way to construct and deploy linear basis expansions is to apply "kernel transformations." Kernel transformations are defined by a "kernel functions." There are many such functions. Some are typically employed in highly specialized applications. Still, this coupling of kernel to application is usually justified by little more than hunch or craft lore (Gross et al., 2012; Duvenaud et al., 2013). We consider here two kernel functions commonly used in regression settings that seem to work well.

We begin with a toy predictor matrix $\mathbf{X}$:

[17] There are two ways to think about this. In the original units of the predictors, the decision boundary is nonlinear. Or in the units of the transformed predictors, the decision boundary is linear. We show the former in Figure 1.

[18] "Element by element" means $x_{11} \times x_{12},\ x_{21} \times x_{22},\ \ldots\ , x_{N1} \times x_{N2}$, where the first subscript denotes the observation number and the second subscript is for the predictor number. $\Phi_1(\mathbf{X})$ is then the element by element product, $\Phi_2(\mathbf{X})$ is the element by element product squared, and $\Phi_3(\mathbf{X})$ is the element by element product cubed. One might view the three terms as an interaction effect variable, an interaction effect variable squared, and an interaction effect variable cubed.

$$\mathbf{X} = \begin{bmatrix} 1 & 2 & 3 & 2 & 0 \\ 2 & 6 & 1 & 1 & 1 \\ 0 & 6 & 0 & 1 & 2 \end{bmatrix} \tag{4}$$

There are 3 rows representing 3 observations, where the number of observations is conventionally denoted by N. There are 5 columns representing 5 predictors, where the number of predictors is conventionally denoted by p. To illustrate some important features of kernels, there are more predictors than cases (i.e., $p > N$). This does not present an immediate problem but would if we considered off-the-shelf regression tools.

Because of the nature of the kernel functions to be applied, all of the elements in $\mathbf{X}$ must be numerical. This includes categorical variables. If C is the number of categories, it is conventional to use $C - 1$ indicator variables, all coded numerically in the same way (e.g., 0 or 1). For example, if there are four different kinds of employment (including not being employed), there would be three indicator variables, where for each, "1" represents the presence of that form of employment and "0" represents the absence of that form of employment. This is consistent with common practice in many different kinds regression applications.

The Radial Basis Kernel

It is nearly universal to denote a kernel function by $k(\boldsymbol{x}, \boldsymbol{x}')$ where $\boldsymbol{x}$ and $\boldsymbol{x}'$ are two different row vectors in $\mathbf{X}$.[19] The *radial basis kernel* is defined as

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\gamma \|\boldsymbol{x} - \boldsymbol{x}'\|^2\right), \tag{5}$$

with $\|.\|$ denoting the squared Euclidian distance (i.e. the "sum of squared differences" also, known as the "norm"), and γ denoting a scale parameter greater than 0.

The kernel transformation is created by applying the kernel function to the data $\boldsymbol{X}$ producing the *kernel matrix,* a matrix which, as we will see shortly, contains the predictive information for a proper regression analysis. To arrive at the *kernel matrix* $\mathbf{K}$, one computes k for each combination of rows i, j and inserts the kernel value in the i, j location of $\mathbf{K}$. Since there are N observations, there are N comparisons for each observation yielding

[19] In standard notation, the two row vectors are in $\mathbb{R}^p$, which is a p-dimensional Euclidian space. Here, $p = 5$ and a row vector is for a given observation its value for each predictor. For example, age might be 24, years since the last arrest might be 2.5, the number of prior prison terms might be 2, gender might be male (i.e. "1"), and the number of prior convictions might be 3.

an $N \times N$ matrix. As an example, consider the second and third row of our toy $\mathbf{X}$. One has for the sum of squared differences: $(2-0)^2+(6-6)^2+(1-0)^2+(1-1)^2+(1-2)^2 = 6$. The sum of squared differences is multiplied by scale parameter γ, negated, and then exponentiated. If the scale parameter were 0.01, one can perform all $3 \times 3 = 9$ calculations to compute

$$\mathbf{K} = \begin{bmatrix} 1.0 & .79 & .73 \\ .79 & 1.0 & .94 \\ .73 & .94 & 1.0 \end{bmatrix}. \tag{6}$$

All kernels matrices are symmetric. Element i, j is the same as element j, i.

The diagonal entries of the radial basis kernel are always 1 (because the squared distance between any $\mathbf{x}$ and itself is 0 and $\exp(0) = 1$), and the off-diagonal entries are between 0 and 1 (because squared distances are positive and $\exp\left(-\Delta d^2\right)$ is bounded between 0 and 1 for positive Δd^2).[20] Radial kernels and others that build on Euclidian distances yield $\mathbf{K}$'s that can be viewed as similarity matrices. Because of the $-\gamma$ in Equation 5, larger off-diagonal values imply less distance between a given pair of observations, which means that they have more similar profiles over variables — they are more similar. Radial basis kernels have proved to be useful in a wide variety of applications but for regression, there can be a better choice.

The ANOVA Radial Basis Kernel

The ANOVA radial basis kernel is closely related to the radial basis kernel. Using common notation for the ANOVA kernel,

$$k(\boldsymbol{x}, \boldsymbol{x}') = \left(\sum_{j=1}^{p} \exp\left(-\gamma(x_j - x_j')^2\right) \right)^d, \tag{7}$$

where x_j and x_j' are two different observations' values for predictor j, and p is the number of predictors in $\mathbf{X}$.[21] Because the computations begin with differences, which after being transformed are added together, the calculations are linear when $d = 1$, and one has a linear (additive) similarity

[20] Thus there are only $\binom{N}{2} - N$ computations to construct $\boldsymbol{K}$.

[21] The computational translation is a little tricky. Here are the steps to compute $\mathbf{K}$: (1) for observations i and j do an element by element subtraction over each of the predictors; (2) square each of the differences; (3) multiply each of these squared differences by minus γ; (4) exponentiate each of these products; (5) sum the exponentiated products; (6) raise the sum to the power of d; and (7) Repeat steps 1-6 for all pairs of observations i, j to compute all $N \times N$ entries in $\boldsymbol{K}$.

formulation. When $d = 2$, one has a formulation with products that can be seen as two-way interactions and a squared similarity formulation. By the same reasoning, when $d = 3$, one has three-way interactions and a cubic similarity formulation.[22] The result here for $\gamma = .01$, and $d = 2$ is

$$\mathbf{K} = \begin{bmatrix} 25.00 & 22.88 & 12.16 \\ 22.88 & 25.00 & 24.41 \\ 22.16 & 24.41 & 25.00 \end{bmatrix}. \tag{8}$$

The value of d is commonly set at 1, 2, or 3. In our experience, using 2 or 3 seems to work well in practice. The values for γ are generally far more important and much more difficult to determine. With larger values of γ, the off-diagonal values of $\mathbf{K}$ become smaller. Their *different* Euclidian distances are reduced. In a regression setting, this will make the support from transformed predictor matrix more localized so that more complex relationships with the response variable can be captured. In the language of smoothers, a smaller window (or band width) is being used.

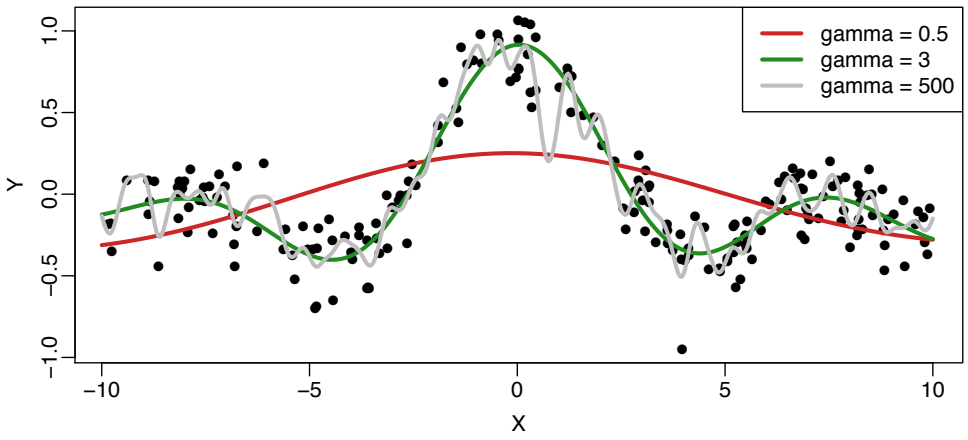

Figure 2: A Simulation of How Fitted Values Depend on the Value of γ

Figure 2 illustrates these points. The figure is a conventional scatterplot

[22] To take a simple example, suppose there are three predictors. For the pair of observations from the first and second row of $\mathbf{X}$ with $\gamma = 1$ and $d = 1$, the sum of differences is $\exp\left(-(x_{11} - x_{12})^2\right) + \exp\left(-(x_{12} - x_{22})^2\right) + \exp\left(-(x_{13} - x_{23})^2\right)$. This is linear and additive. For $d = 2$, the result is $[\exp\left(-(x_{11} - x_{12})^2\right) + \exp\left(-(x_{12} - x_{22})^2\right) + \exp\left(-(x_{13} - x_{23})^2\right)]^2$. All of the terms are now products of two variables, which are two-way interaction effects. For $d = 3$, the result is $[\exp\left(-(x_{11} - x_{12})^2\right) + \exp\left(-(x_{12} - x_{22})^2\right) + \exp\left(-(x_{13} - x_{23})^2\right)]^3$. All of the terms are now products of three variables, which are three-way interaction effects.

showing the results of a simulation in which the fitted values from a simple kernel regression depend on the value of γ. To help make the plot more visually instructive, both the response and single predictor are quantitative. For the same reason, we use a rectangular predictor distribution so there is no significant data sparsity, and the response is highly nonlinear function of the predictor.

In this simulation, $d = 1$ because there is only one regressor (no interactions are possible), and the fit is quite good when $\gamma = 3$. For a γ of 0.5, the fitted values are far too smooth. Important patterns are not captured, although there is much less variance to contend with. For a γ of 500, the fitted values are much too rough. Patterns are captured that are dominated by noise.

There is a lot going on beneath the surface. The $\mathbf{K}$ constructed for the simulation is not a conventional covariance matrix nor a conventional smoother matrix, and nowhere are the predictors in $\mathbf{X}$ or the linear basis expansions $\Phi(\mathbf{X})$ explicitly represented. We will see that this all makes sense because of the "kernel trick" in which $\boldsymbol{K} = \Phi(\boldsymbol{X})\,\Phi(\boldsymbol{X})^{\top}$. In the next several pages, we summarize the reasoning.

How the Kernel Works for Regression

Broadly stated, the operational procedures for the kernel-based forecasting procedures we apply are relatively straightforward. The set of predictors is transformed using an ANOVA kernel. Principle components analysis is applied to the kernel. Then, logistic regression is implemented using a subset of the principle components as regressors. The process of kernel construction, principle component analysis and logistic regression is repeated a number of times with different tuning parameters for the kernel and different subsets of principle components. Using out-of-sample performance, a "best" logistic regression forecasting model is selected.

Beneath these operational steps, however, there are many details and a substantial statistical foundation that provides a rigorous rationale. A technical treatment available in Appendix C. The first subsection addresses the data generation mechanism, which is "assumption lean" compared to conventional regression, which is "assumption laden" (Buja et al., 2014). This background is necessary to understand what a forecast is estimating. The second subsection considers regularization that is needed to reduce the number of columns of $\Phi(\mathbf{X})$. The problem for regression applications is that the number of expansion terms can be equal to or larger than the number of observations. Principle components analysis provides a solution. The

third subsection introduces the "kernel trick" that allows the regularization to proceed even though $\Phi(\mathbf{X})$ is not known. As already noted, the trick depends on $\boldsymbol{K} = \Phi(\boldsymbol{X})\Phi(\boldsymbol{X})^{\top}$, and $\Phi(\boldsymbol{X})\Phi(\boldsymbol{X})^{\top}$ differs fundamentally from $\Phi(\boldsymbol{X})^{\top}\Phi(\boldsymbol{X})$. The fourth discusses how the relative costs of forecasting errors are properly introduced in classification exercises. The fifth examines the role of tuning parameters and how to obtain valid statistical inference along with honest measures of regression performance.

Appendix C — A More Formal Treatment of Kernel Principle Components Regression

Principle Components and the Kernel Trick

Using subset of principle components (PCs) as regressors is an old regression story (see Hastie et al., 2009: Section 3.5.1), commonly motivated by unacceptably high multicollinearity among the predictors. An $N \times p$ matrix of predictors is transformed into an $N \times p$ matrix of PCs that are orthogonal by construction and thereby uncorrelated with one another. The complete set of p PCs account for all of the variance in the matrix of predictors. The PCs employed in the regression as predictors are then a subset of the p PCs. Often they are a small subset because a small fraction of the PCs can account for most of the variance in the original predictor matrix.

Using a subset of PCs is a form of *regularization* because the fitted values will be less variable than had a larger number of PCs been used. One hopes to introduce only a small amount of bias into estimates of the fitted values for a large reduction in the fitted values' variances. We build on these ideas for kernel principle components regression (KPCR). There is a lot of detail, but we provide a summary near the end.

There are N data vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^p$, each with dimension $1 \times p$. For our application in Section 3, these vectors correspond to the characteristics of the individual offenders, such as age and number of prior jail terms. As a simple running example to illustrate the conceptual material, suppose that each offender has three covariates: age, number of prior jail terms, and number of drug prior charges ($p = 3$).

Imagine there is a function $\Phi : \mathbb{R}^p \rightarrow \mathbb{R}^q$ that transforms $\boldsymbol{x}$ into $\Phi(\boldsymbol{x})$, a set of q basis terms with $q > p$. Often, q is larger than N, and possibly even be infinite. As we explain shortly, this function Φ need not be explicitly specified or even known, but we have already used for simple illustrative purposes computing polynomial powers of the predictors. If, for instance,

Φ expanded the original covariate vectors to include all polynomial terms up to cubic terms for each of the $p = 3$ predictors, then q would be 9. The vector $\Phi(\boldsymbol{x})$ would contain terms such as age, age^2, and age^3. Therefore, Φ has expanded the number of predictors available for a regression problem and introduced the flexibility to fit nonlinearities.

Recall that the sample covariance matrix of the data matrix $\boldsymbol{X}$ can be defined as:[23]

$$\boldsymbol{C}' = \frac{1}{N}(\boldsymbol{X} - \bar{\boldsymbol{X}})^\top(\boldsymbol{X} - \bar{\boldsymbol{X}}) \tag{9}$$

where $\bar{\boldsymbol{X}}$ is the matrix of sample averages duplicated over columns so that $\boldsymbol{X} - \bar{\boldsymbol{X}}$ is "centered" i.e. each column's average is 0. The $p \times p$ matrix $\boldsymbol{C}'$ is commonly the input matrix for principle components analysis (PCA).

In kernel regression, the linear expanded basis $\Phi(\boldsymbol{X})$ is used as the predictor matrix. Its sample covariance matrix is:

$$\boldsymbol{C} = \frac{1}{N}(\Phi(\boldsymbol{X}) - \overline{\Phi(\boldsymbol{X})})^\top(\Phi(\boldsymbol{X}) - \overline{\Phi(\boldsymbol{X})}) = \frac{1}{N}\widetilde{\Phi(\boldsymbol{X})}^\top\widetilde{\Phi(\boldsymbol{X})}. \tag{10}$$

where $\widetilde{\Phi(\boldsymbol{X})}$ is the matrix of expanded bases and is centered analogously to $\boldsymbol{X} - \bar{\boldsymbol{X}}$. It is important to note the $\boldsymbol{C}$ *is not* the kernel matrix. It is the sample covariance matrix across the expanded set of predictors.

PCA as Eigendecomposition

In order to carry out PCA, one first computes the eigendecomposition of the matrix portion of $\boldsymbol{C}$ given as

$$\widetilde{\Phi(\boldsymbol{X})}^\top\widetilde{\Phi(\boldsymbol{X})} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top = [\boldsymbol{v}_1 \dots \boldsymbol{v}_q]\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_q \end{bmatrix}\begin{bmatrix} \boldsymbol{v}_1^\top \\ \vdots \\ \boldsymbol{v}_q^\top \end{bmatrix} \tag{11}$$

where the $\boldsymbol{v}$'s are the $q \times 1$ eigenvectors of $\widetilde{\Phi(\boldsymbol{X})}^\top\widetilde{\Phi(\boldsymbol{X})}$ joined column-wise into the matrix $\boldsymbol{V}$, and the λ's are the corresponding eigenvalues (which are non-negative) that form the diagonal of the matrix $\boldsymbol{\Lambda}$. By convention, the eigenvalues are sorted in decreasing order, and their eigenvectors follow suit

[23] This is the maximum likelihood estimate of the covariance matrix, not the usual unbiased estimate using $1/(N-1)$,

when packed into $\boldsymbol{V}$. Recall that as a consequence of eigendecomposition, each of the eigenvectors are mutually orthogonal.[24]

Without loss of generality, we consider the orthonormal set of eigenvectors $\boldsymbol{V}$, which are obtained by rescaling each eigenvector by the reciprocal of its norm $\boldsymbol{v}_k \leftarrow (\sum_{i=1}^{q} v_{k,i}^2)^{-1/2}\boldsymbol{v}_k)$.

Note that with N data points, the matrix $\widetilde{\Phi(\boldsymbol{X})}^\top \widetilde{\Phi(\boldsymbol{X})}$ can be at most rank N (it can be less than N, but for simplicity, we assume it is exactly rank N). Thus, when $q > N$, we only need consider the first N eigenvectors because the eigenvalues of the remaining $q - N$ eigenvalues will all be 0. It follows that Equation 11 can be rewritten as

$$\widetilde{\Phi(\boldsymbol{X})}^\top \widetilde{\Phi(\boldsymbol{X})} = [\boldsymbol{v}_1 \dots \boldsymbol{v}_N] \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1^\top \\ \vdots \\ \boldsymbol{v}_N^\top \end{bmatrix} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top. \quad (12)$$

In standard principle components regression, it is conventional to compute the eigenvalues normalized by their sum, $\lambda'_k = \lambda_k / \sum_{j=1}^{N} \lambda_j$. This results in the convenient interpretation that each of the λ'_k's represent the percentage of variation explained by the kth most important dimension. Because the λ's are sorted from high to low, the first eigenvector $\boldsymbol{v}_1$ represents the "most important" dimension, the second eigenvector $\boldsymbol{v}_2$ represent the second most important dimension, and so on. One can decide from the cumulative sum of the λ'_k's how many eigenvectors (and thereby PCs) should be used in the subsequent analysis. Earlier, we used ρ to denote this cumulative sum. A ρ of 90% means that the number of PCs retained for later use "accounted for" 90% of the variance of the covariance matrix.

However, in KPCR for logistic regression, instead of selecting a number of PCs directly using solely the value of ρ, one proceeds in a three step process that begins by trying to match the validation data cost ratio as closely as possible the desired cost ratio. (See Section 4.5.) In KPCR for a numerical response variable, there is no cost ratio to approximate and selection depends on the lowest out-of-sample squared error in the validation data (not shown in this paper).

After deciding to employ the first r eigenvectors, one must transform each observation in the expanded space $\Phi(\boldsymbol{x}_1), \dots, \Phi(\boldsymbol{x}_n)$. All were $1 \times q$ vectors

[24] $\boldsymbol{V}$ is $q \times N$, so that each row represents an expansion term and each column represents an observation. $\boldsymbol{\Lambda}$ is $N \times N$. The covariance matrix can be expressed as a simple function of its eigenvalues and eigenvectors.

in $\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_n$ after being expanded. With the application of PCA, the dimension can be reduced to $1 \times r$ vectors by a transformation that eliminates the minor $q - r$ dimensions. How does one accomplish this transformation? Some straightforward linear algebra shows that orthogonal projection onto a vector $\boldsymbol{v}_k$ is given by

$$\boldsymbol{X}'_k = \Phi(\boldsymbol{X})\,\boldsymbol{v}_k. \tag{13}$$

The notation $\boldsymbol{X}'_k$ (without the "$\Phi(\cdot)$") is used to indicate that this is the kth column of the new regressor matrix, or kth PC arrived at through projection of $\Phi(\boldsymbol{X})$ onto the subset of eigenvectors of the expanded basis. The new regressor matrix can be used in linear model in the same way that the original regressor matrix $\boldsymbol{X}$ was used.[25]

However, this procedure only works if one can calculate $\boldsymbol{V}$. That creates a serious problem because one can only calculate $\boldsymbol{V}$ if $\Phi(\boldsymbol{X})$ is known, and the transformation Φ is unknown. Fortunately, the "kernel trick" permits recovery of $\boldsymbol{X}'$ without knowledge of Φ.

The Singular Value Decomposition and the Kernel Trick

Consider the usual singular value decomposition (SVD) that is a tripartite decomposition valid for any matrix. Applying the SVD, our expanded bases matrix can be decomposed into $\widetilde{\Phi(\boldsymbol{X})} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{V}^\top$, where $\boldsymbol{V}$ is as above; it is the matrix of the column-wise eigenvectors of $\widetilde{\Phi(\boldsymbol{X})}^\top \widetilde{\Phi(\boldsymbol{X})}$ sorted in decreasing eigenvalue order and becomes size $q \times N$ after dropping the dimensions associated with an eigenvalue of zero. $\boldsymbol{\Lambda}$ also is as above, implying that the middle matrix in the SVD is diagonal and is composed of the square roots of the eigenvalues. $\boldsymbol{U}$ is the matrix of the eigenvectors of $\widetilde{\Phi(\boldsymbol{X})}\widetilde{\Phi(\boldsymbol{X})}^\top$ likewise sorted in decreasing eigenvalue order and is $N \times N$.

In Section 3.2 we called $\boldsymbol{K} = \Phi(\boldsymbol{X})\,\Phi(\boldsymbol{X})^\top$ the "kernel matrix." Here, we call $\widetilde{\boldsymbol{K}} = \widetilde{\Phi(\boldsymbol{X})}\widetilde{\Phi(\boldsymbol{X})}^\top$ the "centered kernel matrix."[26] Simple linear algebra shows that they are related via

[25] After taking the transposes, $\boldsymbol{v}_k^\top$ is $1 \times q$, and $\Phi(\boldsymbol{X})^\top$ is $q \times N$. The projected values for the kth column of the new regressor matrix are the N linear combinations of expansion terms of $\Phi(\mathbf{X})$, each weighted the kth eigenvector values. They have some of the look and feel of regression fitted values.

[26] This is achieved implicitly by centering the kernel matrix $\boldsymbol{K}$ to $\bar{\boldsymbol{K}} = \boldsymbol{K} - \mathbf{1}_N\boldsymbol{K} - \boldsymbol{K}\mathbf{1}_N + \mathbf{1}_N\boldsymbol{K}\mathbf{1}_n$ where $\mathbf{1}_N$ is an $N \times N$ matrix of elements $1/N$. the centering is need so that different means across predictors do not dominate the results.

$$\widetilde{\boldsymbol{K}} = \boldsymbol{K} - \frac{1}{N}\boldsymbol{J}_N\boldsymbol{K} - \frac{1}{N}\boldsymbol{K}\boldsymbol{J}_N + \frac{1}{N^2}\boldsymbol{J}_N\boldsymbol{K}\boldsymbol{J}_n, \tag{14}$$

where $\boldsymbol{J}_N$ represents a $N \times N$ matrix with all entries being 1. Another property of SVD is that $\widetilde{\boldsymbol{K}}$ has the same eigenvalues as $\boldsymbol{\Lambda}$.

Additionally, SVD makes $\boldsymbol{V}$ an orthonormal basis for the rowspace of $\widetilde{\Phi(\boldsymbol{X})}$ and $\boldsymbol{U}$ an orthonormal basis for the column space of $\widetilde{\Phi(\boldsymbol{X})}$. They are related via

$$\widetilde{\Phi(\boldsymbol{X})}\boldsymbol{v}_k = \sqrt{\lambda_k}\boldsymbol{u}_k \quad \text{and} \quad \widetilde{\Phi(\boldsymbol{X})}^\top \boldsymbol{u}_k = \sqrt{\lambda_k}\boldsymbol{v}_k. \tag{15}$$

Thus, an arbitrary eigenvector $\boldsymbol{v}_k$ can be written as

$$\boldsymbol{v}_k^\top = \frac{1}{\sqrt{\lambda_k}}\boldsymbol{u}_k^\top \widetilde{\Phi(\boldsymbol{X})}. \tag{16}$$

Finally, one has the solution for not having $\boldsymbol{V}$. One can project $\Phi(\boldsymbol{X})$ onto the s-dimensional subset of $\boldsymbol{V}$ *despite being unable to construct* $\boldsymbol{V}$ *explicitly*. This is the key step in kernel regression and is somewhat counterintuitive.

The projected $\boldsymbol{X}'_k$ of $\Phi(\boldsymbol{X})$ onto the kth eigenvector $\boldsymbol{v}_k$ is given by Equation 13. We now substitute Equation 16, which we learned from the SVD, into the projection formula to arrive at

$$\boldsymbol{X}_k'^\top = \boldsymbol{v}_k^\top \Phi(\boldsymbol{X})^\top = \left(\frac{1}{\sqrt{\lambda_k}}\boldsymbol{u}_k^\top \widetilde{\Phi(\boldsymbol{X})}\right) \widetilde{\Phi(\boldsymbol{X})}^\top = \frac{1}{\sqrt{\lambda_k}}\boldsymbol{u}_k^\top \widetilde{\boldsymbol{K}}. \tag{17}$$

Equation 17 employs the kernel trick (the equivalence of the outer product in the centered expanded bases with the centered kernel matrix).

A Summary

In summary, one considers the eigenvalues $\boldsymbol{\Lambda}$, which are calculated via the eigendecomposition of $\widetilde{\boldsymbol{K}}$ (because it shares the same eigenvalues as $\widetilde{\Phi(\boldsymbol{X})}^\top \widetilde{\Phi(\boldsymbol{X})}$). One then picks the first $r < N$ eigenvectors to form a subspace that explains a large enough percentage of the variance in $\boldsymbol{C}$. Then $\Phi(\boldsymbol{X})$ is rotated onto this lower dimensional space to obtain the new regressor matrix $\boldsymbol{X}'$. Taking the transpose of Equation 17 and absorbing the $1/\sqrt{\lambda_k}$ into $\boldsymbol{U}$[27], one arrives at the simple

[27] Scaling predictors will not affect fitted values in a linear model. Moreover, the columns of $\boldsymbol{X}'$ are generally uninterpretable and are not an inferential target.

$$\boldsymbol{X}' = \widetilde{\boldsymbol{K}}\boldsymbol{U}_{1...r}, \tag{18}$$

where the $\boldsymbol{U}_{1...r}$ denotes the first r columns of the full $\boldsymbol{U}$ matrix.[28]

Forecasting for New Cases

Fitted values then follow as usual via ordinary least squares or logistic regression, and with each new $\boldsymbol{x}^*$ for which one wishes to obtain a forecast. But one must first recapitulate the steps that that transform the original regressors into the principle components used in the logistic regression. That is, $\boldsymbol{x}^*$ must transformed into $\Phi(\boldsymbol{x}^*)$ then rotated onto the selected $\boldsymbol{v}_1 \ldots \boldsymbol{v}_r$ chosen during the modeling phase. Following Equation 13, one obtains $\boldsymbol{x}_k'^{*\top} = \boldsymbol{v}_k^\top \Phi(\boldsymbol{x}^*)^\top$. The kernel trick is then used to resolve $\boldsymbol{v}_k$ in the style of Equation 16. Again absorbing the eigenvalue constants into $\boldsymbol{U}$, transposing as in Equation 18 and generalizing for all r dimensions, the result is

$$\boldsymbol{x}'^* = \widetilde{\boldsymbol{K}}(\boldsymbol{x}^*, \boldsymbol{X})\boldsymbol{U}_{1...r}, \tag{19}$$

where the function $\widetilde{\boldsymbol{K}}(\boldsymbol{x}^*, \boldsymbol{X})$ is a function that returns the $1 \times N$ vector of the kernel evaluated between $\boldsymbol{x}^*$ and all $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ and then centered. Following Equation 14, it can be shown that

$$\widetilde{\boldsymbol{K}}(\boldsymbol{x}^*, \boldsymbol{X}) = \boldsymbol{K}(\boldsymbol{x}^*, \boldsymbol{X}) - \frac{1}{N}\mathbf{1}_N^\top \boldsymbol{K} - \frac{1}{N}\boldsymbol{K}(\boldsymbol{x}^*, \boldsymbol{X})\mathbf{1}_N\mathbf{1}_N^\top + \left(\frac{1}{N^2}\mathbf{1}_N^\top \boldsymbol{K}\mathbf{1}_N\right)\mathbf{1}_n^\top$$

where $\mathbf{1}_N$ is the $n \times 1$ vector of all 1's. Finally, to get a prediction for $\boldsymbol{x}^*$, we take the rotated $\boldsymbol{x}'^*$ and use the slope estimates from the generalized linear model in the usual fashion. Hence, computing a predicted value requires not only the estimated regression coefficients, but also the original data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ and the kernel function as well. One cannot simply drop an $\boldsymbol{x}^*$ into the estimated logistic regression equation.

[28] $\widetilde{\boldsymbol{K}}$ is $N \times N$ and $\boldsymbol{U}_{1...r}$ is $N \times r$.

References

Adair, D.N., (2006) *The Bail Reform Act of 1984.* Washington, DC: Federal Judicial Center.

Amick, G. (2014) "Bail Reform in New Jersey, for a Fairer, Safer State." Trenton, N.J.: Times of Trenton (11/1/2014).

Arnold Foundation (2013) "Developing a National Model for Pretrial Risk Assessment." Research Summary from the Laura and John Arnold Foundation, www.arnoldfoundation.org.

Berk, R.A., (2012) *Criminal Justice Forecasts of Risk: A Machine Learning Approach.* New York: Springer.

Berk, R.A., Brown, L., and L. Zhao (2010) "Statistical Inference After Model Selection." *Journal of Quantitative Criminology* 26(2): 217–236.

Berk, R.A., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013) "Valid Post-Selection Inference." *Annals of Statistics* 41(2): 401–1053.

Berk, R.A., Brown, L., Buja, A., George, E., Pitkin, E., Zhang, K., and Zhao, L. (2014) "Misspecified Mean Function Regression: Making Good Use of Regression Models That Are Wrong." *Sociological Methods and Research*, 43: 422 – 451.

Berk, R.A., and Bleich, J. (2013) "Statistical Procedures for Forecasting Criminal Behavior: A Comparative Assessment." *Journal of Criminology and Public Policy* 12(3): 513–544.

Bishop, C.M. (2006) *Pattern Recognition and Machine Learning.* New York: Springer.

Borden, H.G. (1928) "Factors Predicting Parole Success." *Journal of the American Institute of Criminal Law and Criminology* 19: 328–336.

Bornstein, B.H., Tomkins, A.J., Neeley, E.M., Herian, M.N., and Hamm, J.A. (2013) "Reducing Courts Failure-to-Appear Rate by Written Reminders." *Psychology, Public Policyand Law* 19 (1): 70–80.

Brieman, L. (2001a) "Random Forests." *Machine Learning*, 45: 5–32.

Breiman, L. (2001b) "Statistical Modeling: The Two Cultures." *Statistical Science* 16(3): 199–231.

Buja, A, Berk, R., Brown, L., George, E., Pitkin. E., Traskin, M., Zhang, K., and Zhao. L. (2104) "A Conspiracy of Random Predictors and Model Violations against Classical Inference in Regression." *stat.ME*, arXIV:1404.1578v1.

Bülmann, P., and van de Geer, S. (2011) *Statistics for High-Dimensional Data*. New York: Springer.

Burgess, E. M. (1928) "Factors Determining Success or Failure on Parole." In A. A. Bruce, A. J. Harno, E. .W Burgess, & E. W., Landesco (eds.) *The Working of the Indeterminate Sentence Law and the Parole System in Illinois* (pp. 205–249). Springfield, Illinois, State Board of Parole.

Bushway, S.D. (2013) "Is There Any Logic to Using Logit: Find the Right Tool for the Increasingly Important Job of Risk Prediction." *Criminology and Public Policy* 12(3): 563–567.

Chipman, H,A., George, E.I., and McCulloch, R.E. (2010) "BART: Bayesian Additive Regression Trees." *Annals of Applied Statistics* 4(1): 266–298.

Cule, E., and De Iorio, M. (2012) "A Semi-Automatic method to Guide the Choice of Ridge Parameter in Ridge Regression." *Annals of Applied Statistics*, working paper, Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London.

Cule, E., and De Iorio, M. (2013) "Ridge Regression in Prediction Problems: Automatic Choice of the Ridge Parameter." **Genetic Epidemiology** 37(7): 704–714.

Dawes, R. M., Faust, D., Meehl, P. E. (1989). "Clinical Versus Actuarial Judgment." *Science*, 243(4899): 1668-1674.

Drug Policy Alliance (2014) http://www.drugpolicy.org/about-drug-policy-alliance.

Duvenaud, D., Lloyd, J.R., Grosse, R., Tenenbaum, J.B., and Ghahramani, Z. (2013) "Structure Discovery in Nonparametric Regression through Compositional Kernel Search." *Proceedings of the 30th International Conference on Machine Learning*, Atlanta, Georgia, USA.

Efron, B. (1987) "Better Bootstrap Confidence Intervals" (with discussion). *Journal of the American Statistical Association* 82: 171-200.

Efron, B., and Tibshirani, R.J., (1993) *An Introduction to the Bootstrap.* NewYork: Chapman and Hall.

Faraway, J.J. (2014) "Does Data Splitting Improve Prediction?" arXiv:1301.2983v2.

Farrington, D. P. and Tarling, R. (2003) *Prediction in Criminology.* Albany: SUNY Press.

Friedman, J.H. (2002) "Stochastic Gradient Boosting." *Computational Statistics and Data Analysis* 38: 367–378.

Goldkamp, J. S., and White, M. D. (2006). "Restoring Accountability in Pretrial Release: The Philadelphia Pretrial Release Supervision Experiments." *Journal of Experimental Criminology* 2, 143–181.

Gottfredson, S. D., and Moriarty, L. J. (2006) "Statistical Risk Assessment: Old Problems and New Applications." *Crime & Delinquency*52(1): 178–200.

Grosse, R.B., Salakhutdinov, R., Freeman, W.T., and Tennebaum, J.B. (2012) "Exploiting Compositionality to Explore a Large Space of Model Structures." 30th Conference on Uncertainty in Artificial Intelligence, Quebec City, Quebec, Canada.

Grove, W. M., Meehl, P. E. (1996). "Comparative Efficiency of Informal and Formal Prediction Procedures: The ClinicalStatistical Controversy." *Psychology, Public Policy, and Law*, 2(2), 293 – 323.

Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models.* New York: Chapman & Hall/CRC.

Hastie, T., Tibshirani, R., and Friedman, J. (2009) *Elements of Statistical Learning: Data Mining. Inference, and Prediction*, second edition. New York: Springer.

Hoerl, A.E., Kennard, R.W., and Baldwin, KF. (1975) "Ridge Regression: Some Simulations. *Communications in Statistocs - Theory and Methods* 4:105 – 123.

Leeb, H. and B.M. Pötscher (2005) "Model Selection and Inference: Facts and Fiction." *Econometric Theory* 21: 21–59.

Leeb, H., B.M. Pötscher (2006) "Can One Estimate the Conditional Distribution of Post-Model-Selection Estimators?" *The Annals of Statistics* 34(5): 2554–2591.

Le Cressie, S., and Van Houwelingen (1992) "Ridge Estimators ro Logistic Regression." *Journal of Applied Statistics* 41 (1): 191–201.

Lockhard, R., Taylor, J., Tibshirani, R., and Tibshirani, R. (2014) "A Significance Test for the Lasso." (with discussion) *Annals of Statistics*, forthcoming.

Loh, W.-L. (1991) "Bootstrap Calibration for Confidence Interval Construction and Selection." *Statistica Sinica* 1: 477–491.

McElroy, J.E. (2011) "Introduction to the Manhattan Bail Project." *Federal Sentencing Reporter* 24(1): 8–9.

Mease, D., Wyner, A.J., and Buja, A. (2008) "Boosted Classification Trees and Class Probability/Quantile Estimation." *Journal of Machine Learning Research* 8: 409–439.

Nankervis, J. (2005) "Computational Algorithms for Double Bootstrap Confidence Intervals." *Computational Statistics & Data Analysis* 45: 461–475

Qin, Z., B. Huang, B., Chandramouli, S.S, He, J., and Kumar, S. (2011) "Large-scale Sparse Kernel Logistic Regression - with a Comparative Study on Optimization Algorithms. Spotlight Oral and Poster session at 6th Annual NYAS Machine Learning Symposium.

Reiss, A.J. (1951) "The Accuracy, Efficiency, and Validity of a Prediction Instrument." *American Journal of Sociology* 56: 552–561.

Ridgeway, G. (2013a) "The Pitfalls of Prediction." *NIJ Journal* 271.

Ridgeway, G. (2013b) "Linking Prediction to Prevention." *Criminology and Public Policy* 12(3): 545–562.

Searle, S.R. (1982) *Matrix Algebra Useful for Statistics* New York: Wiley.

Seed, P. (2010) "The Use of Cost Information When Defining Critical Values for Prediction of Rare Events by Using Logistic Regression and Similar Methods." *Journal of the Royal Statistical Society* 173(1): 255–256.

Schaefer, R., Roi, L., and Wolfe, R. (1984) A Ridge Logistic Estimator. *Communications in Statistics – Theory and Methods* 13: 99113.

Schölkopf, B., Herbrich,R.,and Smola, A. J. (2001). "A Generalized Representer Theorem." *Computational Learning Theory. Lecture Notes in Computer Science* 2111: 416 – 426.

VanNostrand, M., and Keebler, G. (2009). *Pretrial Risk Assessment in the FederalCoourt.* Washington, DC: Office of the Federal Detention Trustee, U.S. Department of Justice.

VanNostrand, M. (2013) "Identifying Opportunities to Safely and Responsibly Reduce the Jail Population." http://luminosity-solutions.com/site /wp-content/uploads/2014/02/New-Jersey-Jail-Population- Analysis-Identifying-Opportunities-to-Responsibly-Reduce-the-Jail-Population-4.pdf

Vapnick, V. (1998) *Statistical Learning Theory.* New York; Wiley.

Voorman, A., Shokaie, A., and Witten, D. (2014) "Inference in High Dimensions with the Penalized Score Test." posted on arXiv: 1401.2678v1

Wahba, G., Wang, Y., Gu, C., Klein, R., and Klein, B. (1994) "Smoothing Spline ANOVA for Exponential Families, with Application to the Wisconsin Epidemiological Study of Diabetic Retinopathy." *The Annals of Statistics* 33(6): 1865–1895.

White, H. (1980) "Using Least Squares to Approximate Unknown Regression Functions." *International Economic Review* 21(1): 149–170.

White, H. (1982) "Maximum Likelihood Estimation of Misspecified Models." *Econometrica* 50(1): 1–25.

Zhu, J., and Hastie, T. (2005) "Kernel Logistic Regression and The Import Vector Machine." *Journal of Computational and Graphical Statistics* 14: 185–205.