From Fixed-X to Random-X Regression: Bias-Variance Decompositions, Covariance Penalties, and Prediction Error Estimation

Saharon Rosset Ryan J. Tibshirani *

Abstract

In statistical prediction, classical approaches for model selection and model evaluation based on covariance penalties are still widely used. Most of the literature on this topic is based on what we call the "Fixed-X" assumption, where covariate values are assumed to be nonrandom. By contrast, it is often more reasonable to take a "Random-X" view, where the covariate values are independently drawn for both training and prediction. To study the applicability of covariance penalties in this setting, we propose a decomposition of Random-X prediction error in which the randomness in the covariates contributes to both the bias and variance components. This decomposition is general, but we concentrate on the fundamental case of least squares regression. We prove that in this setting the move from Fixed-X to Random-X prediction results in an increase in both bias and variance. When the covariates are normally distributed and the linear model is unbiased, all terms in this decomposition are explicitly computable, which yields an extension of Mallows' Cp that we call RCp. RCp also holds asymptotically for certain classes of nonnormal covariates. When the noise variance is unknown, plugging in the usual unbiased estimate leads to an approach that we call $\widehat{\text{RCp}}$, which is closely related to Sp (Tukey 1967), and GCV (Craven and Wahba 1978). For excess bias, we propose an estimate based on the "shortcut-formula" for ordinary cross-validation (OCV), resulting in an approach we call RCp^+. Theoretical arguments and numerical simulations suggest that RCp^+ is typically superior to OCV, though the difference is small. We further examine the Random-X error of other popular estimators. The surprising result we get for ridge regression is that, in the heavily-regularized regime, Random-X variance is smaller than Fixed-X variance, which can lead to smaller overall Random-X error.

1 Introduction

A statistical regression model seeks to describe the relationship between a response $y \in \mathbb{R}$ and a covariate vector $x \in \mathbb{R}^p$, based on training data comprised of paired observations $(x_1, y_1), \ldots, (x_n, y_n)$. Many modern regression models are ultimately aimed at prediction: given a new covariate value x_0, we apply the model to predict the corresponding response value y_0. Inference on the prediction error of regression models is a central part of model evaluation and model selection in statistical learning (e.g., Hastie et al. 2009). A common assumption that is used in the estimation of prediction error is what we call a "Fixed-X" assumption, where the training covariate values $x_1, \ldots, x_n$ are treated as fixed, i.e., nonrandom, as are the covariate values at which predictions are to be made, $x_{01}, \ldots, x_{0n}$, which are also assumed to equal the training values. In the Fixed-X setting, the celebrated notions of optimism and degrees of freedom lead to covariance penalty approaches to estimate the prediction performance of a model (Efron, 1986, 2004; Hastie et al., 2009), extending and generalizing classical approaches like Mallows' Cp (Mallows, 1973) and AIC (Akaike, 1973).

*The authors thank Edgar Dobriban for help in formulating and proving Theorem 3, and Felix Abramovich, Trevor Hastie, Amit Moscovich, Moni Shahar, Rob Tibshirani and Stefan Wager for useful comments.

The Fixed-X setting is one of the most common views on regression (arguably the predominant view), and it can be found at all points on the spectrum from cutting-edge research to introductory teaching in statistics. This setting combines the following two assumptions about the problem.

(i) The covariate values $x_1, \ldots, x_n$ used in training are not random (e.g., designed), and the only randomness in training is due to the responses $y_1, \ldots, y_n$.

(ii) The covariates $x_{01}, \ldots, x_{0n}$ used for prediction exactly match $x_1, \ldots, x_n$, respectively, and the corresponding responses $y_{01}, \ldots, y_{0n}$ are independent copies of $y_1, \ldots, y_n$, respectively.

Relaxing assumption (i), i.e., acknowledging randomness in the training covariates $x_1, \ldots, x_n$, and taking this randomness into account when performing inference on estimated parameters and fitted models, has received a good deal of attention in the literature. But, as we see it, assumption (ii) is the critical one that needs to be relaxed in most realistic prediction setups. To emphasize this, we define two settings beyond the Fixed-X one, that we call the "Same-X" and "Random-X" settings. The Same-X setting drops assumption (i), but does not account for new covariate values at prediction time. The Random-X setting drops both assumptions, and deals with predictions at new covariates values. These will be defined more precisely in the next subsection.

1.1 Notation and assumptions

We assume that the training data $(x_1, y_1), \ldots, (x_n, y_n)$ are i.i.d. according to some joint distribution P. This is an innocuous assumption, and it means that we can posit a relationship for the training data,

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n \tag{1}$$

where $f(x) = \mathbb{E}(y|x)$, and the expectation here is taken with respect to a draw $(x, y) \sim P$. We also assume that for $(x, y) \sim P$,

$$\epsilon = y - f(x) \text{ is independent of } x, \tag{2}$$

which is less innocuous, and precludes, e.g., heteroskedasticity in the data. We let $\sigma^2 = \mathrm{Var}(y|x)$ denote the constant conditional variance. It is worth pointing out that some results in this paper can be adjusted or modified to hold when (2) is not assumed; but since other results hinge critically on (2), we find it is more convenient to assume (2) up front.

For brevity, we write $Y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ for the vector of training responses, and $X \in \mathbb{R}^{n \times p}$ for the matrix of training covariates with ith row x_i, $i = 1, \ldots, n$. We also write Q for the marginal distribution of x when $(x, y) \sim P$, and $Q^n = Q \times \cdots \times Q$ (n times) for the distribution of X when its n rows are drawn i.i.d. from Q. We denote by $\tilde{y}_i$ an independent copy of y_i, i.e., an independent draw from the conditional law of $y_i|x_i$, for $i = 1, \ldots, n$, and we abbreviate $\tilde{Y} = (\tilde{y}_1, \ldots, \tilde{y}_n) \in \mathbb{R}^n$. These are the responses considered in the Same-X setting, defined below. We denote by (x_0, y_0) an independent draw from P. This the covariate-response pair evaluated in the Random-X setting, also defined below.

Now consider a model building procedure that uses the training data (X, Y) to build a prediction function $\hat{f}_n : \mathbb{R}^p \to \mathbb{R}$. We can associate to this procedure two notions of prediction error:

$$\mathrm{ErrS} = \mathbb{E}_{X,Y,\tilde{Y}}\bigg[\frac{1}{n}\sum_{i=1}^{n}\big(\tilde{y}_i - \hat{f}_n(x_i)\big)^2\bigg] \quad \text{and} \quad \mathrm{ErrR} = \mathbb{E}_{X,Y,x_0,y_0}\big(y_0 - \hat{f}_n(x_0)\big)^2,$$

where the subscripts on the expectations highlight the random variables over which expectations are taken. (We omit subscripts when the scope of the expectation is clearly understood by the context.) The *Same-X* and *Random-X* settings differ only in the quantity we use to measure prediction error: in Same-X, we use ErrS, and in Random-X, we use ErrR. We call ErrS the Same-X prediction error

and ErrR the Random-X prediction error, though we note these are also commonly called in-sample and out-of-sample prediction error, respectively. We also note that by exchangeability,

$$\mathrm{ErrS} = \mathbb{E}_{X,Y,\tilde{y}_1}\big(\tilde{y}_1 - \hat{f}_n(x_1)\big)^2.$$

Lastly, the *Fixed-X* setting is defined by the same model assumptions as above, but with $x_1, \ldots, x_n$ viewed as nonrandom, i.e., we assume the responses are drawn from (1), with the errors being i.i.d. We can equivalently view this as the Same-X setting, but where we condition on $x_1, \ldots, x_n$. In the Fixed-X setting, prediction error is defined by

$$\mathrm{ErrF} = \mathbb{E}_{Y,\tilde{Y}}\bigg[\frac{1}{n}\sum_{i=1}^{n}\big(\tilde{y}_i - \hat{f}_n(x_i)\big)^2\bigg].$$

(Without $x_1, \ldots, x_n$ being random, the terms in the sum above are no longer exchangeable, and so ErrF does not simplify as ErrS did.)

1.2 Related work

From our perpsective, much of the work encountered in statistical modeling takes a Fixed-X view, or when treating the covariates as random, a Same-X view. Indeed, when concerned with parameter estimates and parameter inferences in regression models, the randomness of new prediction points plays no role, and so the Same-X view seems entirely appropriate. But, when focused on prediction, the Random-X view seems more realistic as a study ground for what happens in most applications.

On the other hand, while the Fixed-X view is common, the Same-X and Random-X views have not exactly been ignored, either, and several groups of researchers in statistics, but also in machine learning and econometrics, fully adopt and argue for such random covariate views. A scholarly and highly informative treatment of how randomness in the covariates affects parameter estimates and inferences in regression models is given in Buja et al. (2014, 2016). We also refer the reader to these papers for a nice review of the history of work in statistics and econometrics on random covariate models. It is also worth mentioning that in nonparametric regression theory, it is common to treat the covariates as random, e.g., the book by Gyorfi et al. (2002), and the random covariate view is the standard in what machine learning researchers call statistical learning theory, e.g., the book by Vapnik (1998). Further, a stream of recent papers in high-dimensional regression adopt a random covariate perspective, to give just a few examples: Greenshtein and Ritov (2004); Chatterjee (2013); Dicker (2013); Hsu et al. (2014); Dobriban and Wager (2015).

In discussing statistical models with random covariates, one should differentiate between what may be called the "i.i.d. pairs" model and "signal-plus-noise" model. The former assumes i.i.d. draws (x_i, y_i), $i = 1, \ldots, n$ from a common distribution P, or equivalently i.i.d. draws from the model (1); the latter assumes i.i.d. draws from (1), and additionally assumes (2). The additional assumption (2) is not a light one, and it does not allow for, e.g., heteroskedasticity. The books by Vapnik (1998); Gyorfi et al. (2002) assume the i.i.d. pairs model, and do not require (2) (though their results often require a bound on the maximum of $\mathrm{Var}(y|x)$ over all x.)

More specifically related to the focus of our paper is the seminal work of Breiman and Spector (1992), who considered Random-X prediction error mostly from an intuitive and empirical point of view. A major line of work on practical covariance penalties for Random-X prediction error in least squares regression begins with Stein (1960) and Tukey (1967), and continues onwards throughout the late 1970s and early 1980s with Hocking (1976); Thompson (1978a,b); Breiman and Freedman (1983). Some more recent contributions are found in Leeb (2008); Dicker (2013). A common theme to these works is the assumption that (x, y) is jointly normal. This is a strong assumption, and is one that we avoid in our paper (though for some results we assume x is marginally normal); we will discuss comparisons to these works later. Through personal communication, we are aware of work in progress by Larry Brown, Andreas Buja, and coauthors on a variant of Mallows' Cp for a setting

in which covariates are random. It is out understanding that they take somewhat of a broader view than we do in our proposals RCp, $\widehat{\text{RCp}}$, RCp$^+$, each designed for a more specific scenario, but resort to asymptotics in order to do so.

Finally, we must mention that an important alternative to covariance penalties for Random-X model evaluation and selection are resampling-based techniques, like cross-validation and bootstrap methods (e.g., Efron 2004; Hastie et al. 2009). In particular, ordinary leave-one-out cross-validation or OCV evaluates a model by actually building n separate prediction models, each one using $n-1$ observations for training, and one held-out observation for model evaluation. OCV naturally provides an almost-unbiased estimate of Random-X prediction error of a modeling approach ("almost", since training set sizes are $n-1$ instead of n), albeit, at a somewhat high price in terms of variance and inaccuracy (e.g., see Burman 1989; Hastie et al. 2009). Altogether, OCV is an important benchmark for comparing the results of any proposed Random-X model evaluation approach.

2 Decomposing and estimating prediction error

2.1 Bias-variance decompositions

Consider first the Fixed-X setting, where $x_1, \ldots, x_n$ are nonrandom. Recall the well-known decomposition of Fixed-X prediction error (e.g., Hastie et al. 2009):

$$\text{ErrF} = \sigma^2 + \frac{1}{n}\sum_{i=1}^n \big(\mathbb{E}\hat{f}_n(x_i) - f(x_i)\big)^2 + \frac{1}{n}\sum_{i=1}^n \text{Var}\big(\hat{f}_n(x_i)\big)$$

where the latter two terms on the right-hand side above are called the (squared) *bias* and *variance* of the estimator $\hat{f}_n$, respectively. In the Same-X setting, the same decomposition holds conditional on $x_1, \ldots, x_n$. Integrating out over $x_1, \ldots, x_n$, and using exchangeability, we conclude

$$\text{ErrS} = \sigma^2 + \underbrace{\mathbb{E}_X\Big(\mathbb{E}\big(\hat{f}_n(x_1) \,|\, X\big) - f(x_1)\Big)^2}_{B} + \underbrace{\mathbb{E}_X \text{Var}\big(\hat{f}_n(x_1) \,|\, X\big)}_{V}.$$

The last two terms on the right-hand side above are integrated bias and variance terms associated with $\hat{f}_n$, which we denote by B and V, respectively. Importantly, whenever the Fixed-X variance of the estimator $\hat{f}_n$ in question is unaffected by the form of $f(x) = \mathbb{E}(y|x)$ (e.g., as is the case in least squares regression), then so is the integrated variance V.

For Random-X, we can condition on $x_1, \ldots, x_n$ and x_0, and then use similar arguments to yield the decomposition

$$\text{ErrR} = \sigma^2 + \mathbb{E}_{X,x_0}\Big(\mathbb{E}\big(\hat{f}_n(x_0) \,|\, X, x_0\big) - f(x_0)\Big)^2 + \mathbb{E}_{X,x_0}\text{Var}\big(\hat{f}_n(x_0) \,|\, X, x_0\big).$$

For reasons that will become clear in what follows, it suits our purpose to rearrange this as

$$\begin{aligned}
\text{ErrR} &= \sigma^2 + B + V && (3)\\
&\quad + \underbrace{\mathbb{E}_{X,x_0}\Big(\mathbb{E}\big(\hat{f}_n(x_0) \,|\, X, x_0\big) - f(x_0)\Big)^2 - \mathbb{E}_X\Big(\mathbb{E}\big(\hat{f}_n(x_1) \,|\, X\big) - f(x_1)\Big)^2}_{B^+} && (4)\\
&\quad + \underbrace{\mathbb{E}_{X,x_0}\text{Var}\big(\hat{f}_n(x_0) \,|\, X, x_0\big) - \mathbb{E}_X\text{Var}\big(\hat{f}_n(x_1) \,|\, X\big)}_{V^+}. && (5)
\end{aligned}$$

We call the quantities in (4), (5) the *excess bias* and *excess variance* of $\hat{f}_n$ ("excess" here referring to the extra amount of bias and variance that can be attributed to the randomness of x_0), denoted