TECHNOMETRICS
2020, VOL. 62, NO. 4, 447–450
https://doi.org/10.1080/00401706.2020.1805021

Taylor & Francis
Taylor & Francis Group

Comment: From Ridge Regression to Methods of Regularization

Ming Yuan

Department of Statistics, Columbia University, New York, NY

1. Methods of Regularization

Ridge regression and the idea of regularization that it comes to symbolize are ubiquitous in modern data analysis since its formal introduction to statistics about half a century ago by Hoerl and Kennard (1970a, 1970b). In this short discussion, we shall focus in particular on their influence and connections with some of the popular techniques in nonparametric statistics and machine learning, and how the general perspective of methods of regularization allows us to view a plethora of seemingly different methods as variants of ridge regression.

When considering a multiple linear regression model:

$$y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1},$$

the ridge estimator of β is given by

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}}\{\|y - X\beta\|^2 + \lambda\|\beta\|^2\}.$$

The first term of the objective function on the right-hand side measures for a given $\beta \in \mathbb{R}^p$, how well it fits the data whereas the second term explicitly encourages for an estimate close to the origin. The tuning parameter $\lambda > 0$ balances the tradeoff between the fidelity to the data and preference toward simpler estimates.

Ridge regression embodies the idea of regularization which is ubiquitous in modern data analysis. More broadly, suppose that we are interested in estimating a parameter θ, be it the conditional mean, a quantile or the conditional density among others, from data; then following the spirit of ridge regression, a method of regularization proceeds to do so via

$$\hat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmin}}\left\{L_n(\theta; \text{data}) + \lambda\|\theta\|^2\right\}. \tag{1}$$

There are two main ingredients to this general recipe:

- *Parameter space*—$(\Theta, \|\cdot\|)$ is a normed space;
- *Loss function*—$L_n(\cdot; \text{data})$ is a loss function that measures the goodness of fit.

In particular, taking

$$\Theta^{\text{lin}} = \{\theta(x) = x^\top\beta : \beta \in \mathbb{R}^p\}$$

the class of linear functions endowed with norm $\|\theta\| = \|\beta\|_{\ell_2}$ and L_n to be the least squares loss yields the usual ridge regression. Despite of its simplicity, the flexibility in choosing the parameter space and loss function in (1) allows us to extend the idea beyond linear models or regression. In the rest of this discussion, we shall illustrate how a couple of simple guiding principles in doing so bring together many different models and estimates, and how each of them can be traced back to the idea of ridge regression.

2. Parameter Space—RKHS

The parameter space Θ^{lin} defined above is an example of the so-called reproducing kernel Hilbert space (RKHS)—a Hilbert space with continuous evaluation functionals (see, e.g., Aronszajn 1950). Another notable example of RKHS is the periodic Sobolev spaces, say, the collection of periodic twice differentiable functions defined over the unit interval $[0, 1]$:

$$\Theta^{\text{Sob}} := \Big\{f(x) = \sqrt{2}\sum_{k\geq 0}(a_k\sin(2\pi kx) + b_k\cos(2\pi kx)) : a_k, b_k \in \mathbb{R}, \int (f'')^2 < \infty\Big\}.$$

This particular choice of parameter space can be identified with the so-called (periodic) smoothing splines (see, e.g., Wahba 1990). More generally, RKHS proves to be an extremely useful and versatile option for the parameter space.

2.1. RKHS and Similarity Measures

An RKHS $\mathcal{H}$ of functions defined over a domain $\mathcal{X}$ is endowed with an inner product $\langle\cdot,\cdot\rangle$, and complete under the inner product such that for any $x \in \mathcal{X}$ the evaluation functional $L_x : \mathcal{H} \to \mathbb{R}, f \mapsto f(x)$ is continuous. By Reisz theorem, the continuity of L_x entails that there exists a $K_x \in \mathcal{H}$ such that

$$L_x f = \langle f, K_x\rangle, \quad \text{for all } f \in \mathcal{H}.$$

Now consider a bivariate function $K : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ such that $K(x, y) = \langle K_x, K_y\rangle$ for all $x, y \in \mathcal{X}$. It is clear that K is symmetric and positive definite in that for any $a_i \in \mathbb{R}$ and $x_i \in \mathcal{X}$,

$$\sum a_i a_j K(x_i, x_j) = \|\sum a_i K_{x_i}\|^2 \geq 0.$$

We shall call a symmetric and positive definite function as such a *kernel*, and in particular, K defined above the *reproducing kernel*

CONTACT Ming Yuan ming.yuan@columbia.edu Department of Statistics, Columbia University, New York, NY 10027.

of $(\mathcal{H}, \langle\cdot,\cdot\rangle)$. In the case of Θ^{lin} endowed with Euclidean norm on the coefficient vectors, K is simply the Euclidean inner product, that is, $K(x_i, x_j) = x_i^\top x_j$. The Moore–Aronszjan theorem indicates that there is a one-to-one correspondence between an RKHS and a kernel. To signify this correspondence, we often write $\mathcal{H}(K)$ as the RKHS $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ with reproducing kernel K.

Although the use of RKHS in statistics can be traced back at least to the early 1960s, it gained tremendous popularity in the late 1990s and early 2000s with the rise of kernel methods in machine learning (see, e.g., Scholkopf and Smola 2001 and references therein). While the concept of RKHS may be abstract to many, kernels on the other hand are more intuitive—they are generalizations of the inner product in Euclidean spaces to a generic domain and can be viewed as a *similarity measure* between a pair of objects from the same domain. Fundamentally, any sensible prediction is based on the idea that *similar* inputs result in *similar* outcomes. While the similarity between outcomes is determined by the choice of a loss function, the similarity between inputs now rests on the choice of a kernel or similarity measure.

2.2. Statistical Modeling via Choices of Kernel

Traditionally in statistics, the choice of the parameter space Θ is the focus of modeling. The connection with kernels means we can instead focus on defining the appropriate similarity measures. These two angles, in the case of RKHS, turn out to be equivalent. We now briefly describe several examples that demonstrate the intimate yet subtle connections between a number of different models.

The first example is the additive models (see, e.g., Hastie and Tibshirani 1990). Under the additive model, a p-variate function f can be represented by

$$f(\mathbf{x}) = f_1(x_1) + \cdots + f_p(x_p),$$

where x_j is the jth coordinate of $\mathbf{x}$. Note that such an additive representation may not be unique and usually side conditions such as $f_j(0) = 0$ are imposed to ensure identifiability. Assuming that f_j comes from an RKHS $\mathcal{H}(K_j)$ identified with a kernel K_j, then f resides in an RKHS with kernel $K_1 + \cdots + K_p$. This observation immediately relates additive models to the so-called multiple kernel learning (see, e.g., Gönen and Alpaydin 2011 and references therein).

Another example where the connection is a little less apparent is the so-called varying-coefficient models (see, e.g., Hastie and Tibshirani 1993). The varying-coefficient model posits the following structure of the conditional mean of a response given covariates $t \in \mathbb{R}$ and $\mathbf{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p$:

$$f(\mathbf{x}) = f_1(t)x_1 + f_2(t)x_2 + \cdots + f_p(t)x_p.$$

Similar to additive models, side conditions can be imposed to ensure the identifiability of the representation on the right-hand side. Assuming that f_j's come from a common RKHS characterized by a kernel K_0, then f resides in an RKHS identified with the kernel:

$$K((t,\mathbf{x}),(s,\mathbf{y})) = \left(\sum_{k=1}^{p} x_k y_k\right) K_0(t,s),$$

which can be viewed as the tensor product of the linear kernel in $\mathbb{R}^p$ and K_0. This draws a connection between varying-coefficient models and another popular yet seemingly different class of statistical models based upon the so-called tensor product RKHSs (see, e.g., Wahba 1990).

Our last example is the so-called functional linear regression where the covariate X is a square-integrable stochastic process defined over a certain domain $\mathcal{T}$:

$$Y = \int_{\mathcal{T}} X(t)\beta(t)dt + \varepsilon.$$

Assuming that β is a member of the RKHS with kernel K_0, under suitable technical conditions, the regression function can be identified with an RKHS with kernel

$$K(x_1, x_2) = \int_{\mathcal{T}\times\mathcal{T}} x_1(t)x_2(s)K_0(s,t)dsdt.$$

As we can see from these examples, RKHS provides a unified framework for many seemingly different statistical models, all of which when put in the context of methods of regularization share the same spirit as ridge regression. The prowess of such a framework, however, goes beyond such a conceptual synthesis. It also allows for a unified treatment from an operational level thanks to the so-called representer lemma, or kernel trick.

2.3. Representer Lemma and Kernel Trick

In the case of ridge regression, Θ^{lin} is finite-dimensional and the ridge estimate can be computed explicitly. This is not always possible for general and possibly infinite-dimensional RKHSs. Nonetheless, one of the key practical advantages of using RKHS in the framework of the method of regularization (1) is that it allows for efficient computation even if Θ is infinite-dimensional thanks to the so-called representer lemma (Kimeldorf and Wahba 1971). The renowned lemma states that if the loss function $L_n(\theta; \text{data})$ depends on θ only through its evaluations $\theta(x_1), \ldots, \theta(x_n)$, then the solution $\hat\theta$ to (1) can be expressed as

$$\hat\theta(\cdot) = \sum_{i=1}^{n} \alpha_i K(x_i, \cdot)$$

for some $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, where K is the reproducing kernel of the RKHS $(\Theta, \|\cdot\|)$. We can then plug it back to (1) to get

$$(\alpha_1, \ldots, \alpha_n) = \operatorname*{argmin}_{\eta_1,\ldots,\eta_n\in\mathbb{R}} \left\{ L_n\left(\sum_{i=1}^{n} \eta_i K(x_i,\cdot); \text{data}\right) + \lambda\left(\sum_{i,j=1}^{n} \eta_i\eta_j K(x_i,x_j)\right)\right\}. \quad (2)$$

A couple of important observations can be made here. First of all, the representer lemma shows that there is no loss of generality to estimate a function $\theta \in \Theta$ by a linear combination of up to n *basis* functions. As such, it allows us to estimate a possibly infinite-dimensional parameter θ by solving a finite-dimensional optimization problem (2). Moreover, as we assumed, L_n depends on θ only through $\theta(x_1), \ldots, \theta(x_n)$. As such, the objective function on the right-hand side will only

involve K via $G = [K(x_i, x_j)]_{1\le i,j\le n}$, the Gram matrix. This has an important practical appeal—if we can define a similarity measure on the observed inputs, then we can go ahead and estimate the function θ.

In the case of ridge regression, L_n is the least squares loss, and we can further derive

$$(\alpha_1, \ldots, \alpha_n) = (G^\top G + \lambda G)^{-1} G y.$$

This explicit derivation is often referred to as the so-called kernel trick, and the above formula applies beyond the usual ridge regression where, as mentioned before, $K(x, y) = x^\top y$, to the case where $(\Theta, \|\cdot\|)$ is an arbitrary RKHS.

3. Loss Function—Fisher Consistency

The other ingredient in the method of regularization (1) is the loss function. The choice of the loss function oftentimes is clear for a given statistical problem. For example, from the consideration of statistical efficiency alone, we almost always take L_n to be the negative log-likelihood whenever possible. The difficulty with this choice, however, is that the corresponding optimization problem in either (1) or (2) may become computationally intractable. The most notable example is perhaps classification where the computational challenge in direct minimization of the misclassification loss is widely recognized and has stimulated the proposal of numerous *surrogate* loss functions. In general, a reasonable loss function needs to be *Fisher consistent* in that the "true" parameter θ_* can be identified with

$$\theta_* = \operatorname*{argmin}_\theta \mathbb{E} L_n(\theta; \text{data}).$$

We now give a few examples illustrating this guiding principle in choosing appropriate loss functions.

3.1. Regression

Consider, for example, a regression problem

$$Y = \theta_*(X) + \varepsilon, \tag{3}$$

where ε is the idiosyncratic noise independent of X. In the usual mean regression setting, ε is assumed to be centered with mean zero and finite variance, and our goal is to estimate the conditional expectation: $\theta_*(x) = \mathbb{E}(Y|X = x)$. A natural choice of the loss function is the least squares:

$$L_n(\theta; \text{data}) = \frac{1}{n}\sum_{i=1}^n \left[y_i - \theta(x_i)\right]^2.$$

It is not hard to see that in this case L_n is Fisher consistent.

More generally, one may consider generalized regression where the conditional distribution of $Y|X$ belongs to an exponential family:

$$Y|X \sim h(Y)\exp[Y\theta_*(X) - b(\theta_*(X))],$$

for some strictly convex function $b : \mathbb{R} \mapsto \mathbb{R}$ (see, e.g., McCullagh and Nelder 1989). Our goal is to estimate the canonical parameter θ_* that can be identified by

$$b'(\theta_*(x)) = \mathbb{E}(Y|X = x).$$

In this case, we can take L_n to be the negative (conditional) log-likelihood:

$$L_n(\theta; \text{data}) = \frac{1}{n}\sum_{i=1}^n \left[b(\theta(x_i)) - y_i\theta(x_i)\right].$$

The convexity of $b(\cdot)$ implies again that L_n is Fisher consistent.

A third example is quantile regression, where the τ quantile of ε in (3) is assumed to be zero, that is,

$$\theta_*(x) = \inf\left\{y : \mathbb{P}(Y \ge y|X = x) \ge \tau\right\}.$$

See, for example, Koenker and Hallock (2001). Write

$$\rho_\tau(u) = \tau(u)_- + (1-\tau)(u)_+,$$

where $(\cdot)_-$ and $(\cdot)_+$ represent the negative and positive parts, respectively. Then the following loss function is commonly adopted:

$$L_n(\theta) = \frac{1}{n}\sum_{i=1}^n \rho_\tau(y_i - \theta(x_i)).$$

This is another example of Fisher consistent loss function as it can be shown that θ_* minimizes $\mathbb{E}L_n(\theta; \text{data})$.

3.2. Density Estimation

The goal of density estimation is to estimate a density function $p(\cdot)$ after observing

$$x_1, x_2, \ldots, x_n \sim_{\text{iid}} p(\cdot).$$

To incorporate natural constraints for density, that it is nonnegative and integrates to one, it is often customary to reparameterize a density function in terms of its logarithm:

$$p(\cdot) = \exp(\theta(\cdot))/\int \exp(\theta(x))dx.$$

An obvious choice of loss function for estimating θ is the negative log-likelihood:

$$L_n(\theta; \text{data}) = -\frac{1}{n}\sum_{i=1}^n \theta(x_i) + \log\left(\int \exp(\theta(x))dx\right).$$

It is fairly straightforward to show that L_n is indeed Fisher consistent. Using this loss function in (1), however, incurs a couple of computational challenges. First, θ and $\theta + c$ for any constant $c \in \mathbb{R}$ correspond to the same density function. To avoid such ambiguity, side conditions on θ, for example, $\int \theta = 0$, is often imposed. Another challenge when using L_n is that it is not *convex* in θ, which makes the optimization problem involved hard to solve.

An ingenious solution is provided by Silverman (1982) who suggested an alternative loss function:

$$L_n(\theta; \text{data}) = -\frac{1}{n}\sum_{i=1}^n \theta(x_i) + \left(\int \exp(\theta(x))dx\right).$$

This loss function is strictly convex. Furthermore, it can be shown that it is also Fisher consistent in that

$$\log p = \operatorname*{argmin}_{\theta} \mathbb{E} L_n(\theta).$$

3.3. Classification

Now consider the problem of classification. To fix ideas, we shall focus on binary classification where (X, Y) is a random couple where $X \in \mathcal{X} = \mathbb{R}^p$, $Y \in \mathcal{Y} = \{\pm 1\}$. The goal is to construct a good classifier that maps from $\mathcal{X}$ to $\mathcal{Y}$, so we can predict Y after observing X in the future. A natural choice of the loss function in this context is the misclassification loss:

$$L_n(\theta; \text{data}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(y_i \neq \theta(x_i)),$$

and the so-called Bayes rule provides the *optimal* classifier with the smallest misclassification error:

$$\operatorname*{argmin}_{\theta} \mathbb{E} L_n(\theta; \text{data}) = \theta_*(\cdot)$$
$$:= \begin{cases} +1 & \text{if } \mathbb{P}(Y = +1 | X = \cdot) > 1/2, \\ -1 & \mathbb{P}(Y = -1 | X = \cdot) > 1/2. \end{cases}$$

The challenge, however, is again computation. Optimizing $L_n(\cdot)$ is practically infeasible even for moderate size problems.

To overcome such a difficulty, one often resorts to convex relaxations. Instead of seeking directly a classifier that maps from $\mathcal{X}$ to $\mathcal{Y}$, we consider recovering a *discriminant function*. A discriminant function g maps from $\mathcal{X}$ to $\mathbb{R}$, and it can be translated to a classifier $\theta = \text{sign}(g)$. Loss functions of the following type are often entertained:

$$L_{n,\phi}(g; \text{data}) = \frac{1}{n} \sum_{i=1}^{n} \phi(y_i g(x_i)),$$

where $\phi : \mathbb{R} \mapsto \mathbb{R}$ is a convex function. Notable examples of ϕ, include the exponential loss ($\phi(u) = \exp(-u)$) related to boosting, and hinge loss ($\phi(u) = (1-u)_+$) associated with support vector machines, among numerous others.

For such a convex relaxation approach to work, we need to first make sure that the loss function $L_{n,\phi}$ is Fisher consistent in that $\theta_* = \text{sign}(g_*)$ where

$$g_* = \operatorname{argmin} \mathbb{E} L_{n,\phi}(g; \text{data}).$$

It turns out that such a consistency property holds under fairly general conditions: if ϕ is convex, then $L_{n,\phi}$ is Fisher consistent *if and only if* ϕ is differentiable at 0 and $\phi'(0) < 0$ (see, e.g., Lin 2002; Zhang 2004).

4. Summary

Since its formal introduction by Hoerl and Kennard (1970a, 1970b), ridge regression and the method of regularization embodied by ridge regression can be found in different corners of statistics, oftentimes under different disguises. Looking at them through the lens of the method of regularization reveals their conceptual connection with ridge regression, allows us to offer a unified treatment, and provides further insights into their operating characteristics. The legacy of ridge regression, on the other hand, lives through and beyond these essential tools in modern data analysis.

Funding

This research was supported by NSF grants DMS-1803450 and DMS-2015285. Part of the work was done while the author was visiting the Institute for Theoretical Studies at ETH Zürich, Switzerland, and he wish to thank the institute for their hospitality.

References

Aronszajn, N. (1950), "Theory of Reproducing Kernels," *Transactions of the American Mathematical Society*, 68, 337–404. [447]

Gönen, M., and Alpaydin, E. (2011), "Multiple Kernel Learning Algorithms," *Journal of Machine Learning Research*, 12, 2211–2268. [448]

Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized Additive Models* (Vol. 43), Boca Raton, FL: CRC Press. [448]

——— (1993), "Varying-Coefficient Models," *Journal of the Royal Statistical Society*, Series B, 55, 757–779. [448]

Hoerl, A. E., and Kennard, R. W. (1970a), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55–67. [447,450]

——— (1970b), "Ridge Regression: Applications to Nonorthogonal Problems," *Technometrics*, 12, 69–82. [447,450]

Kimeldorf, G., and Wahba, G. (1971), "Some Results on Tchebycheffian Spline Functions," *Journal of Mathematical Analysis and Applications*, 33, 82–95. [448]

Koenker, R. and Hallock, K. F. (2001), "Quantile Regression," *Journal of Economic Perspectives*, 15, 143–156. [449]

Lin, Y. (2002), "Support Vector Machines and the Bayes Rule in Classification," *Data Mining and Knowledge Discovery*, 6, 259–275. [450]

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman & Hall. [449]

Scholkopf, B., and Smola, A. J. (2001), *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, Cambridge, MA: MIT Press. [448]

Silverman, B. W. (1982), "On the Estimation of a Probability Density Function by the Maximum Penalized Likelihood Method," *The Annals of Statistics*, 10, 795–810. [449]

Wahba, G. (1990), *Spline Models for Observational Data* (Vol. 59), Philadelphia, PA: SIAM. [447,448]

Zhang, T. (2004), "Statistical Behavior and Consistency of Classification Methods Based on Convex Risk Minimization," *The Annals of Statistics*, 32, 56–85. [450]