inference

Wesley Brooks

Introduction

It has often been noted that the estimation error underestimates the prediction error over novel data generated from the same process. Stein's unbiased risk estimation (SURE) is a framework for estimating the prediction error of a model. The so-called covariance penalty estimate arises from the SURE framework as an estimate of the degees of freedom used in estimating the model.

Equivalent computations arise from the framework of Akaike's information criterion (AIC). The AIC arises from the framework of estimating the likelihood of a fitted model with respect to the unknowable truth that gave rise to the data.

Our focus is on local variable selection and estimating the local coefficients.

Methods

Definition of a LAGR model

Assume that we observe data $\boldsymbol{y} = (y_1, \ldots, y_n)$ and covariates $\boldsymbol{X} =$

such that $y_i = m(\mu_i)$ where $\mu_i = Ey_i$.

Degrees of freedom for a LAGR model

The degrees of freedom used in estimating the model are

$$\hat{df} = \sum_{i=1}^{n} cw_{ii} / \mathrm{tr}(\mathrm{W(s_i)})$$

where $v$is the number of covariates selected at site i for local LAGR tuning parameter λ, $\boldsymbol{W}(\boldsymbol{s}_i)$ is the matrix of local kernel weights at location $\boldsymbol{s}_i$, and $w_{ii} = \boldsymbol{W}(\boldsymbol{s}_i)_{i,i}$.

By the covariance penalty, the degrees of freedom used in estimating a model is given by

$$df = \sum_{i=1}^{n} cov(y_i, \mu_i) = \sum_{i=1}^{n} \frac{d}{d\mu_i} m(\mu_i).$$

But it has been noted that the degrees of freedom is also given by the number of covariates with nonzero coefficients.

The divergence formula of (Zou, Hastie, and Tibshirani 2007) states that

$$\left(\frac{\partial \hat{\boldsymbol{\mu}}}{\partial \boldsymbol{y}} \right)_{i,j} = \frac{\partial \hat{\mu}_i}{\partial y_j}, i, j = 1, 2, \ldots, n.$$

Thus, the divergence

$$\nabla \cdot \hat{\boldsymbol{\mu}} = \mathrm{tr} \left(\frac{\partial \hat{\boldsymbol{\mu}}}{\partial \mathrm{y}} \right) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{\partial \hat{\mu}_{\mathrm{i}}}{\partial \mathrm{y_i}}.$$

Now in the case of a model estimated by LAGR, there is an independent model for each observation, so by (4.21) of Zou, Hastie, and Tibshirani (2007), the degrees of freedom used in fitting the ith observation are

$$\partial\hat{\mu}_i/\partial y_i = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{s})\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{W}(\boldsymbol{s})s_{i,i}.$$

And since the total degrees of freedom for the ith model are $\sum_{j=1}^{n}\partial\hat{\mu}_j/\partial y_j = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{s})\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{W}(\boldsymbol{s})s_{i,i} = |\mathcal{B}_\lambda|$, the result follows.

Bagging estimate of the bandwidth parameter

Calculating the AIC of a LAGR model is done conditional on the selected local tuning parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$. Selecting the bandwidth parameter to minimize the AIC may imply greater precision than is actually warranted. Bootstrap aggregation (bagging) is a method of incorporating the effects of model selection into parameter estimates and model predictions.

In the case of a LAGR model, bagging

Zou, Hui, Trevor Hastie, and Robert Tibshirani. 2007. “On the “degrees of Freedom" of the Lasso.” *Annals of Statistics* 35: 2173–92.