inference

Wesley Brooks

Introduction

What are the goals of inference?

- Select the model
- Estimate the model parameters
- Estimate confidence intervals for the parameters

What challenges are unique to inference in LAGR models?

- Selecting the bandwidth
- LAGR is not a linear smoother, so estimating degrees of freedom is difficult
- LAGR uses lasso, so estimating distributions of coefficient estimates is difficult

Why use AIC? - Likelihood is the basis for most statistical inference - The likelihood can be used for model selection and estimation - Using the same data for selection and estimation produces a downward-biased estimate of the likelihood - The AIC is bias-corrected likelihood - Computing AIC requires an expression for the degrees of freedom used in estimating a model

Why use the bootstrap?

Local adaptive grouped regularization (LAGR) is a method for local variable selection and local coefficient estimation in a varying coefficient regression (VCR) model (Brooks, Zhu, and Lu 2014). Estimating a model via LAGR is straightforward. This paper addresses inference for a VCR model estimated by LAGR, focusing on the estimation of confidence intervals for the local coefficient estimates and estimation of which local coefficients should be shrunk to exactly zero.

The method of LAGR possesses the oracle property of asymptotically selecting exactly the correct variables and estimating them as accurately as if their identities were known in advance. For local selection and estimation, LAGR relies on a version of the adaptive group Lasso (Yuan and Lin 2006, Wang and Leng (2008)). Thus, local coefficients estimated by LAGR asymptotically acheive the distributions given in Sun et al. (2014) and Cai, Fan, and Li (2000).

However, the asymptotic case is never realized in actual data analysis. Given a finite quantity of data, the set of covariates selected by LAGR is subject to uncertainty. Since coefficient estimates in a model estimated by LAGR are conditional on the selected covariates, the asymptotic expression for the distribution of the local coefficients is not useful for inference.

Further, while the local coefficient estimation is conditional on the local covariate selection, the local covariate selection is itself conditional on the bandwidth parameter h. In order to acheive the oracle properties, the optimal bandwidth for a LAGR model was shown to be $h_n = O(n^{-1/6})$ (Brooks, Zhu, and Lu 2014). However, the optimal rate is not enough information to determine the bandwidth, and is anyhow irrelevant when n is fixed, as is the case in most practical data analysis.

Statistical properties of estimators are typically established assuming *a priori* model selection. The properties of estimates that are computed from the same data as was used for model selection are

The properties of estimators after model selection is an area of active research. A typical approach in applications is to select a model that minimizes a selection criterion such as the AIC or BIC, and then proceed

with estimation as though the model had been selected in advance. However, this leads to unreliable inference from the selected model (Leeb and Pötscher 2006). It also ignores the discontinuous nature of estimation after model selection, which is well illustrated by Figures 8 and 9 of Efron (2014).

Model averaging is a technique that attempts to estimate the model parameters without conditioning on the selected model. When prior distributions can be established for the candidate models and their parameters, Bayesian model averaging (BMA) is a principled approach to multimodel inference (Hoeting et al. 1999). However, the establishing the prior distributions is not a trivial step, especially in a setting with many parameters, as VCR has. There is an analagous procedure called frequentist model averaging (FMA) that avoids the need to specify prior distributions. The asyptotic distribution of estimators derived from FMA has been worked out in the framework of "dwindling confidence", where coefficients decrease at a $\sqrt{n}$ rate as the sample size increases (Hjort and Claeskens 2003). The framework of dwindling confidence is similar to the framework of "moving parameter" asymptotics, under which the adaptive Lasso estimator fails to be consistent (Potscher and Schneider 2009).

It is impossible to use maximum likelihood to estimate the bandwidth parameter, as reveled by a simple example: given a data set $(\boldsymbol{X}, Y, \boldsymbol{S})$, let $h = 0$. Then $\hat{Y} = Y$ trivially, which results in the maximum possible likelihood. This is clearly an example of overfitting, because the model can tell us nothing about any future observations.

Methods

AIC step

We work within an information-theoretic framework where the optimal model is the one closest to truth f in the sense of Kullback-Leibler (KL) distance (Kullback and Leibler 1951). Since the truth f is unknown, we are left to estimate the KL distance, which we do by means of the Akaike Information Criterion (AIC) (Burnham and Anderson 2002, Akaike (1973)). A model's AIC is an estimate of its expected KL distance from the truth. In fact, the AIC is an estimate of the log likelihood of an independent realization of the response, conditional on the observed covariates. The key to AIC is that a model's log likelihood is penalized by a factor equal to the degrees of freedom used in estimating the model.

Which model minimizes the AIC is not the only consideration. For model selection via LAGR, the AIC is seems to be quite discontiuous. What's more, small differences in the AIC are indicative of ambiguity in model selection. In this work we consider model averaging with model weights based on their AIC values. The smoothed AIC estimate is (Burnham and Anderson 2002)

$$\breve{\boldsymbol{\beta}}(\boldsymbol{s}) = \sum_{j=1}^{M} w_j \hat{\boldsymbol{\beta}}_j(\boldsymbol{s}) / \sum_{k=1}^{M} w_k \tag{1}$$

$$w_j = \exp(-\Delta_j/2) \tag{2}$$

$$\Delta_j = \mathrm{AIC}_j - \min_k \mathrm{AIC}_k \tag{3}$$

The smoothed AIC can be applied to any function of the estimated coefficients. In particular, the smoothed AIC can be used for prediction and to estimate whether a given coefficient is exactly zero.

Bootstrap step

Even if the model were selected in advance, we lack an expression for the finite-sample distribution of the coefficients of a VCR model estimated by LAGR. We therefore turn to the bootstrap to estimate the finite-sample distribution of the coefficients. As it happens, we can capitalize on the bootstrap draws to smooth the discontinutiy in model selection as well.

We use the parametric bootstrap to draw from the full model at each location, then apply the method of LAGR to the bootstrap draw to get a new estimate $\breve{\boldsymbol{\beta}}^*(\boldsymbol{s})$ of the coefficients.

Sorting property of adaptive Lasso

Given the form of a model and the p covariates for consideration, there are 2^p possible regression models to consider. It is computationally impractical to test all of these models. The adaptive lasso provides a principled way of reducing the number of models to consider from $O(2^p)$ to $O(p)$ (proof?). Despite the sorting being "principled" it remains to show that the models are good, and that they are wisely sorted.

It has been shown that LAGR can asymptotically select exactly the "correct" covariates. In order to prove that property, we assumed that the tuning parameter $\lambda_n = n^\alpha$ for $\alpha \in ((2-\gamma)/3, 1/3)$. This range of rates, though, isn't enough to decide exactly how to set the tuning parameter. In our case, we've used the AIC to tune the local variable selection.

At location $\boldsymbol{s}$, the sequence $\mathcal{M}_n$ is the sequence of models as λ_n sweeps from $\lambda_{n,0}$ to 0. The first element of $\mathcal{M}_n$ is always $m_0 = \mathbf{0}$, the null (or intercept-only) model. The last element of $\mathcal{M}_n$ is always the full model. Proposition 0.1 suggests that with probability going to one, the KL-best model appears in the sequence.

Proposition 0.1 (Sorting Property) *As $n \to \infty$, the models $\mathcal{M}_n(\boldsymbol{s})$ on the LAGR solution path at $\boldsymbol{s}$ approach a well-sorted state, meaning that the models are are nested such that the $m_0, m_1, \ldots, m_{p_0}$ add one variable of the KL-best model at each step, and models $m_{p_0+1}, \ldots, m_p$ each add one variable that is not in the KL-best model.*

We give weights to the models in the sequence based on their AIC values. Proposition 0.2 indicates that the AIC weight vector converges to the weight vector that minimizes the KL distance between truth and the weighted model average.

Proposition 0.2 (Consistency) *Let $w(\boldsymbol{s})$ be a vector of model weights at $\boldsymbol{s}$ and $\hat{w}(\boldsymbol{s})$ be the vector of weights estimated by LAGR and the AIC. Then*

$$\frac{L(\hat{w}(\boldsymbol{s}))}{\inf_{w(\boldsymbol{s})\in\mathcal{H}} L(w(\boldsymbol{s}))} \xrightarrow{P} 1 \tag{4}$$

where Lw is the K-L distance between truth and the weighted model average indicated by w.

References

Akaike, Hirogotu. 1973. "Information Theory and an Extension of the Maximum Likelihood Principle." In *2nd International Symposium on Information Theory*, edited by B.N. Petrov and F. Csaki, 267–81.

Brooks, Wesley, Jun Zhu, and Zudi Lu. 2014. "Local Adaptive Grouped Regularization and Its Oracle Properties for Varying Coefficient Regression."

Burnham, Kenneth P., and David R. Anderson. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer New York.

Cai, Zongwu, Jianqing Fan, and Runze Li. 2000. "Efficient Estimation and Inferences for Varying-Coefficient Models." *Journal of the American Statistical Association* 95: 888–902.

Efron, Bradley. 2014. "Estimation and Accuracy After Model Selection." *Journal of the American Statistical Association* 109 (507): 991–1007.

Hjort, Nils Lid, and Gerda Claeskens. 2003. "Ferquentist Model Averaging Estimators." *Journal of the American Statistical Association* 98 (464): 879–99.

Hoeting, Jennifer A., David Madigan, Adrian E. Raftery, and Chris T. Volinsky. 1999. "Bayesian Model Averaging: A Tutorial." *Statistical Science* 14 (4): 382–417.

Kullback, Solomon, and Richard Leibler. 1951. "On Information and Sufficiency." *Annals of Mathematical Statistics* 22: 79–86.

Leeb, Hannes, and Benedikt M. Pötscher. 2006. "Can One Estimate the Conditional Distribution of Post-Model-Selection Estimators?" *Annals of Statistics* 34 (5): 2554–91.

Potscher, Benedikt M., and Ulrike Schneider. 2009. "On the Distribution of the Adaptive LASSO Estimator." *Journal of Statistical Planning and Inference* 139: 2775–90.

Sun, Yan, Hongjia Yan, Wenyang Zhang, and Zudi Lu. 2014. "A Semiparametric Spatial Dynamic Model." *Annals of Statistics* 42: 700–727.

Wang, Hansheng, and Chenlei Leng. 2008. "A Note on Adaptive Group Lasso." *Computational Statistics and Data Analysis* 52: 5277–86.

Yuan, Ming, and Yi Lin. 2006. "Model Selection and Estimation in Regression with Grouped Variables." *Journal of the Royal Statistical Society Series B* 68: 49–67.