

Research Statement

Wesley Brooks

Introduction

Varying coefficient regression (VCR) is a data analysis method that relates the response to its covariates through a linear equation of coefficients $\boldsymbol{\beta}$. The response $Y(s)$, covariates $\mathbf{X}(s)$, and coefficients $\boldsymbol{\beta}(s)$ in a VCR model are indexed by the location parameter s . Let d be the dimension of the location parameter (e.g., $d = 1$ for a coefficients that vary with time). Assume n observations of the response and the covariates at locations s_1, \dots, s_n and let $Y_i = Y(s_i)$, $\mathbf{X}_i = \mathbf{X}(s_i)$, and $\boldsymbol{\beta}_i = \boldsymbol{\beta}(s_i)$. The generalized linear model (GLM) with varying coefficients is written

$$\begin{aligned} E[Y(s)|\mathbf{X}(s)] &= \mu(s) \\ \eta(s) = g(\mu(s)) &= \mathbf{X}'(s)\boldsymbol{\beta}(s) \\ \text{var}[Y(s)|\mathbf{X}(s)] &= \phi V(\mu(s)) \end{aligned}$$

where $g(\cdot)$, $V(\cdot)$, and ϕ are, respectively, the link function, variance function, and dispersion parameter of the response family. For simplicity of notation, assume the canonical link function.

Local adaptive grouped regularization (LAGR) is a method for local variable selection and estimation in varying coefficient regression (VCR) models (Brooks et al., 2014). The method of LAGR is an \mathcal{L}_1 regularization technique applied to local polynomial regression in which the coefficient functions are estimated pointwise by fitting a local model at each point of interest s in the domain \mathcal{S} (Fan and Gijbels, 1996). Estimating a model by LAGR entails selection of a global bandwidth parameter that applies to all of the local models and local smoothing parameters, one for each local model. Because the \mathcal{L}_1 regularization produces nonlinear estimates of the coefficients, the confidence intervals for local coefficient estimates in LAGR are complicated. One goal of inference in a VCR model estimated by LAGR is to calculate the confidence intervals.

The local variable selection action of LAGR estimates which covariates are relevant to the local fit at each location s . Another goal of inference is to state the confidence of local variable selection.

It is common to estimate confidence intervals and variable selection confidence in \mathcal{L}_1 regularization methods via the bootstrap (?). Due to the varying coefficients, observations in a VCR model are not exchangeable.

1 Methods

1.1 Information theory

Data arises from truth f , an unknowable function of infinite parameters. Statistical estimation and inference are directed toward identifying a simple and parsimonious approximation to f , called a model. Typically, the family of candidate models G is picked *a priori*. We assume an *a priori* choice to approximate f with a VCR model estimated via LAGR. Given a metric $I(\cdot, \cdot)$ that computes a distance between two function-like objects, there is some element g_0 of the family G that minimizes the distance $I(g, f)$. The minimizer g_0 may not be unique.

The Kullback-Leibler (KL) divergence is a common choice for the distance to minimize in estimation (?). In the context of a parametric model, minimizing the KL divergence is the same as maximizing the likelihood, so estimation by minimizing the KL divergence is a generalization of maximum likelihood estimation (?).

Kullback-Leibler divergence is calculated by the equation

$$KL = \int f(x) \log \frac{f(x)}{g(x)} dx$$

which depends on truth f , and is therefore unknowable. The Akaike Information Criterion (AIC) is an estimate of the KL divergence that can be calculated from the observed data. The AIC of model g is

$$AIC(g) = -2 \log \mathcal{L}(g) + 2df(g)$$

where $\log \mathcal{L}(g)$ is the log likelihood of the observed data under model g , and $df(g)$ is the degrees of freedom used by the model g .

1.2 Degrees of freedom of a local model

The degrees of freedom for local model \hat{g} are given

1.3 AIC-averaging of local model

The method of LAGR uses ℓ_1 regularization of an adaptive group lasso type for variable selection. The challenge of variable selection is that there are 2^p possible combinations of p candidate variables in a regression model. The number of combinations grows so quickly that it is impractical to test every possible combination except in the simplest cases. A key property of the adaptive lasso is that it sorts the covariates in a principled way. The sorting property is described by the following propositions:

$\lambda(s)$ is the tuning parameter that sets the amount of regularization in the local model. For an

Proposition 1.1 *The sequence of adaptive lasso tuning parameters $\infty, \lambda_m, \dots, \lambda_1, 0$ corresponds to a sequence of models $g_\infty, g_m, \dots, g_1, g_0$ where g_∞ is the intercept-only model and g_0 is the full model. As $n \rightarrow \infty$, the sequence matches the correct ordering of the models, with a $p \in 1, \dots, m$ such that all of the variables in g_m, \dots, g_{p+1} are relevant, no relevant variables are added in g_{p-1}, \dots, g_0 , and of all the candidate models g_p is closest to f in a KL sense.*

Proposition 1.2 *The AIC-weighted model average \hat{g} converges to g_0 as $n \rightarrow \infty$. In other words, given enough data, we can correctly estimate which model is closest to truth.*

The difficulty is to approximate f with observed data, in which case we cannot know whether the AIC-best estimate \hat{g} is the same as g_0 . Therefore the AIC is used to make a weighted average of the model sequence $g_\infty, g_m, \dots, g_1, g_0$.

Proposition 1.3 *The AIC-optimal model average \bar{g} converges to the KL-optimal model g_0 .*

1.4 Degrees of freedom of the whole model

In the context of Steins Unbiased Risk Estimation (SURE), the degrees of freedom are given by

$$df = \sum_{i=1}^n \text{cov}(y_i, \mu_i)$$

For a VCR model estimated by LAGR, then, the degrees of freedom are

$$df = \sum_{i=1}^n df_i$$
$$df_i =$$

1.5 Profile AIC of bandwidth parameter

For any value of the bandwidth parameter h , the AIC is given by

$$AIC = -2 \log \mathcal{L}(h) + 2df(h).$$

The profile AIC is the AIC as a function of h , $AIC(h)$.

Proposition 1.4 *The bandwidth \hat{h} that minimizes the profile AIC is a good estimate of the KL-optimal bandwidth.*

1.6 Bandwidth model averaging

Even if the AIC minimizer \hat{h} is a good estimate of the KL-optimal bandwidth, basing inference on a single estimated value of the bandwidth parameter risks model-selection bias. Model averaging over the bandwidth is intended to eliminate model-selection bias. The model weighting is by the AIC, based on the idea that the AIC is an unbiased estimate of the log-likelihood of the data under the true data-generating mechanism

Proposition 1.5 *The AIC-weighted model average converges to the KL-optimal g_0 , and the resulting parameter estimates are less biased than under the point-modeled bandwidth.*

1.7 Parametric bootstrap draws

The methods described to now relate to averaging over uncertainty in model selection. However, even if the model were *a priori* known, there would be uncertainty in the model estimation. The adaptive group lasso is an L1 regularization method, which results in nonlinear estimates of the model parameters. Thus, the distribution of the estimates is complicated. The parametric bootstrap is used here in preference to attempting an analytical expression of the estimation distribution.

Our parametric bootstrap procedure works by drawing $\beta^*(s)$ from the joint distribution of the VCR coefficients, as estimated by local polynomial regression (no regularization). These draws are used to generate a resampled response Y^* , from which we estimate $\beta^{**}(s)$ by LAGR. The $\beta^{**}(s)$ form our parametric bootstrap estimates of $\beta(s)$.

1.8 Smoothing residuals

There is bias in the local coefficient estimates, due to curvature in the coefficient functions. This results in clustering of the residuals. In order to justify the parametric bootstrap we want the resampled data to be from the same distribution as the original data.

In order to correct for the bias in estimating the coefficient surfaces, the clustered residuals are modeled with an spline, hopefully leaving only uncorrelated residuals.

1.9 Inference from parametric bootstrap draws

The draws from the parametric bootstrap are used for inference. The nonparametric delta method of Efron (2014) is used to calculate the distribution of those quantities that we wish to measure.

Proposition 1.6 *The confidence intervals calculated in this way achieve the nominal coverage.*

Proposition 1.7 *A chi-square test based on the nonparametric delta method achieves the nominal level for testing the null hypothesis that a linear combination of local coefficients is zero.*

References

- Wesley Brooks, Jun Zhu, and Zudi Lu. Local adaptive grouped regularization and its oracle properties for varying coefficient regression. In prep, 2014.
- Bradley Efron. Estimation and accuracy after model selection. *Journal of the American Statistical Association*, 109(507):991–1007, 2014.
- Jianqing Fan and Irene Gijbels. *Local Polynomial Modeling and its Applications*. Chapman and Hall, London, 1996.