# Research Statement

Wesley Brooks

# Introduction

# 1 Methods

## 1.1 Seeking KL-optimal model

Truth is an unknowable, infinite-parameter function $f$. We approximate $f$ with a model $\hat{g}$, from the family $G$. Our goal in estimation is to find $g_0 \in G$ that is the closest element of $G$ to the truth $f$. Closest is a statement of relative distance in terms of the Kullbach-Leibler (KL) divergence. The goal of inference is to describe properties of the range of elements in $G$ that lie closest to $f$ with confidence $1 - \alpha$.

Kullbach-Leibler divergence is calculated by the equation

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

which depends on the truth $f$, and is therefore unknowable. The Akaike Information Criterion (AIC) is an estimate of the KL divergence, calculated by the equation

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

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

## 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 $\ell_1$ 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: the sequence of adaptive lasso tuning parameters $\infty, \lambda_m, \ldots, \lambda_1, 0$ corresponds to a sequence of models $g_\infty, g_m, \ldots, g_1, g_0$ where $g_\infty$ is the intercept-only model and $g_0$ is the full model. As $n \to \infty$, the sequence matches the correct ordering of the models, with a $p \in 1, \ldots, m$ such that all of the variables in $g_m, \ldots, g_{p+1}$ are relevant, no relevant variables are added in $g_{p-1}, \ldots, g_0$, and of all the candidate models $g_p$ is closest to $f$ in a KL sense.

Proposition 2: the AIC-optimal model $\hat{g}$ converges to $g_p$ as $n \to \infty$. 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 $g_p$. Therefore the AIC is used to make a weighted average of the model sequence $g_\infty, g_m, \ldots, g_1, g_0$.

Proposition 3: the AIC-optimal model average $\bar{g}$ converges to the KL-optimal model $\hat{g}$.

## 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

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

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: 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: the AIC-weighted model average converges to the KL-optimal $\hat{g}$, 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 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 $\boldsymbol{\beta}^*(s)$ 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 $\boldsymbol{\beta}^{**}(s)$ by LAGR. The $\boldsymbol{\beta}^{**}(s)$ form our parametric bootstrap estimates of $\boldsymbol{\beta}(s)$.

## 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 **?** is used to calculate the distribution of those quantities that we wish to measure.

Proposition: the confidence intervals calculated in this way achieve the nominal coverage.

Proposition: 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.