The lagr Package: local adaptive grouped regularization in R

by Wesley Brooks

Abstract An abstract of less than 150 words.

A basic fact of varying coefficient regression (VCR) models is that the coefficients vary over the model's domain. It is natural, then, to allow that a coefficient function may be nonzero in part of the domain and exactly zero elsewhere. To date, methods of estimating VCR models could allow only global variable selection, where a variable is either in or out of the model over the entire domain. The method of local adaptive grouped regularization (LAGR) is an estimation method for VCR models that allows for local variable selection (**?**).

Some R packages that estimate VCR models are **spgwr**, **mgcv**, **np**, **dlm**.

Method

Local polynomial regression

Models estimated by LAGR are of the local polynomial regression type, which is a form of kernel smoothing.

Penalized local log likelihood

Each local model is fit by maximum penalized local likelihood. The penalized local log likelihood function is

$$\ell_i = \sum_{j=1}^{n} w_{ij} \log \mathcal{L}(\boldsymbol{\beta}_i, Y_j) + \lambda_i \sum_{k=1}^{p} \phi_k \|\boldsymbol{\beta}_k\| \tag{1}$$

where ...

Local degrees of freedom

In the context of Stein's unbiased risk estimation, the degrees of freedom used in fitting a model are defined as

$$df = \sum_{i=1}^{n} \frac{\text{cov}(y_i, \hat{y}_i)}{\sigma^2} \tag{2}$$

where y_i is the observed value of the ith response, $\hat{y}_i$ is the corresponding fitted value, and σ^2 is the exponential family dispersion parameter (**?**). The degrees of freedom for an adaptive group lasso estimate are ... (**?**).

Model averaging

A weighted average of the candidate models is used to acknowledge uncertainty in the variable selection.

$$\hat{g} = \text{argmin}\, w_j g_j \tag{3}$$

$$w_j \geq 0 \forall j \in 1, \ldots, m \tag{4}$$

Bandwidth parameter

In order to estimate a VCR model by a local polynomial method like LAGR, we need to set the bandwidth parameter. In the `lagr` function, the bandwidth can be specified in terms of distance or k-nearest neighbors. The k-nearest neighbors method is a type of adaptive bandwidth that specifies a value for $\sum_{j=1}^{n} w_{ij}/n$, while the distance method specifies an identical h .

To estimate the bandwidth parameter, we profile it with our favorite model selection criterion. The optimal value of the bandwidth parameter is the one that minimizes the selection criterion. However, selecting this bandwidth and treating it as known truth would introduce model-selection bias (**?**). We

average over the implicit distribution of the bandwidth parameter based on the profile AIC that was calculated in

Total degrees of freedom

In the typical local polynomial regression model, the degrees of freedom are calculated as the trace of the projection matrix. Because LAGR is an $\mathcal{L}_1$ regularization method, though, it is nonlinear and thus generates no projection matrix. Recall the definition of degrees of freedom (**??**).

For estimating the bandwidth, each observation is estimated with a local model. Only observations colocated with the local model are affected by the local fit, so the total degrees of freedom are the sum of the colocated degrees of freedom from each local model. An unbiased(?) approximation of the colocated degrees of freedom is $df_i / sum_{j=1}^{n} w_{ij}$.

Code, explained

Package

The R package **lagr** (https://github.com/wrbrooks/lagr) provides functions for estimation, inference, and plotting in a model estimated by LAGR. Its primary functions are `lagr` and `lagr.tune`. The `lagr` function estimates a model by LAGR, while the `lagr.tune` function estimates profiles the bandwidth parameter with respect to a model selection criterion (AIC, BIC, or GCV).

Estimation of a model

Estimation is carried out by blockwise coordinate descent. This is an iterative process and carrying out the coordinate descent algorithm in a compiled C++ library is considerably faster than doing so in R. The **Rcpp11** is used to integrate C++ code into the **lagr** package.

The model weighting is a constrained quadratic programming problem. The solution is found via the **quadprog** package.

Response family

R provides several `family` objects representing exponential family distributions (e.g., `gaussian()`, `binomial()`, `poisson()`). In the **lagr** package, these objects supply the link and variance functions for fitting the local GLM models. Because the **Rcpp11** package provides "seamless R and C++ integration", we can use objects of type `function` within the C++ code that can represent either R or C++ functions. This capability allows us to call the `link()` and `varfun()` functions of any `family` object. As a result, the user can write their own `family` object in either R or C/C++ and use it as the response distribution of a VCR model estimated by LAGR.

Bandwidth estimation

Bandwidth profiling in the `lagr.tune` function is carried out using the `optim` function.

Plotting

The function `lagr.plot.shapefile` plots the inputs or outputs of a `lagr.model` object, provided that the data was specified via an `sp` object (see package **sp**).

Examples

Wisconsin poverty data?

Wesley Brooks
Department of Statistics, University of Wisconsin-Madison
1300 University Ave. Madison, WI 53706
USA wrbrooks@uwalumni.com