Research statement

Wesley Brooks

Introduction

My general research interests are in the field of spatial statistics with natural science applications. Specifically, my thesis research introduces local variable selection in spatially varying coefficient regression models, and I am also working on the problem of spatial confounding. In the sections below, I expand on these research problems, my work on them, and my plans going forward.

Local variable selection

Whereas the coefficients in a typical spatial regression model are constant over the model's spatial domain, the coefficients in a varying coefficient regression (VCR) model are functions - here we'll assume smooth functions - of the location s [@Hastie-Tibshirani-1993]. The coefficient functions in a VCR model may be estimated by local polynomial regression, a variant of kernel smoothing that uses Taylor's expansion to represents the coefficient functions $\boldsymbol{\beta}(s)$ as polynomials in a neighborhood of s_0. For instance, a local polynomial model of order one estimates the value $\boldsymbol{\beta}(s_0)$, and the slope $\boldsymbol{\beta}'(s_0)$ of the coefficient functions at s_0 [@Fan-Gijbels-1996; @Fan-Zhang-1999].

Prior research for VCR models has focused on global variable selection for VCR models [@Wang-Li-Huang-2008; @Wang-Xia-2009; @Wei-Huang-Li-2011]. Global variable selection identifies the covariates with nonzero coefficient functions and uses the same set of covariates on the entire spatial domain. In contrast, my work shows how to estimate which variables in a VCR model have nonzero coefficients at any location in the model's domain. The method I developed is an $\mathcal{L}_1$ regularization procedure akin to the adaptive group lasso (Wang and Leng, 2008), so I call it local adaptive grouped regularization (LAGR). Its "oracle" properties are the subject of a manuscript that has been submitted to the Journal of the Royal Statistical Society (Series

B). In the manuscript and the accompanying software package, the response of the regression model can follow any exponential family distribution.

The estimation properties of the LAGR method are appealing, but due to the $\mathcal{L}_1$ regularization, the resulting estimator is a nonlinear function of of the data and thus has complicated confidence intervals [@Knight-Fu-2000]. A second manuscript, intended for submission to the Journal of Agricultural, Biological, and Environmental Statistics, demonstrates inference in the context of a VCR model estimated by LAGR. Topics covered there are the degrees of freedom used in estimation, estimating the AIC-optimal bandwidth and tuning parameters, model averaging with the AIC, and using a parametric bootstrap procedure to summarize quantities like confidence intervals for local coefficients and the confidence that a local coefficient is nonzero. Another manuscript, describing the R package lagr, is in preparation and intended for submission to the Journal of Statistical Software.

There are several outstanding research problems in this area, such as developing local variable selection in partially linear regression, where some coefficients are constant and others vary. Another is estimation and inference when the coefficient functions have different degrees of smoothness, and when the coefficients are smoother in one part of the domain than another.

Spatial confounding

A basic principle of spatial statistics is that nearby observations are more more alike than distant ones. In regression models, this is true of both the covariates and the response. In this setting, it can be unclear whether the observed regression relationship is due to a genuine relationship between the covariate and the response, or because they both have spatial structure on the same observational units [@Hodges-Reich-2010; @Paciorek-2011]. This phenomenon is called spatial confounding.

@Paciorek-2011 focuses on the roles of left-out confounding variables and the spatial scale of variation in geostatistical models, concluding that estimators can be biased by confounders that vary on a smaller spatial scale than the observed covariates. On the other hand, @Hodges-Reich-2010 argue that confounding can result from the arrangement of observational units in an intrinsic conditionally autoregressive (ICAR) model [Besag-York-Mollie-1991]. They show that effects reported as significant in the scientific literature may disappear completely when the response is projected orthogonal to the ICAR adjacency matrix.

Disagreement in the literature and the relevance to fields employing spatial statistical methods suggest that

there is productive research to be done here. For instance, I am currently studying methods to decompose variation in the response of an ICAR model into unconfounded, residual, and potentially confounded components. Another track I am exploring is how to discern confounding by aggregating or subdividing the data to vary the spatial scale of variation.

References