

# Oracle properties of local adaptive grouped regularization

Wesley Brooks

---

---

## 1. Spatially varying coefficients regression

### 1.1. Model

Consider  $n$  data points, observed at sampling locations  $\mathbf{s}_i = (s_{i,1} \ s_{i,2})^T$  for  $i = 1, \dots, n$ , which are distributed in a spatial domain  $D \subset \mathbb{R}^2$  according to a density  $f(\mathbf{s})$ . For  $i = 1, \dots, n$ , let  $y(\mathbf{s}_i)$  and  $\mathbf{x}(\mathbf{s}_i)$  denote, respectively, the univariate response and the  $(p + 1)$ -variate vector of covariates measured at location  $\mathbf{s}_i$ . At each location  $\mathbf{s}_i$ , assume that the outcome is related to the covariates by a linear model where the coefficients  $\boldsymbol{\beta}(\mathbf{s}_i)$  may be spatially-varying and  $\varepsilon(\mathbf{s}_i)$  is random error at location  $\mathbf{s}_i$ . That is,

$$y(\mathbf{s}_i) = \mathbf{x}(\mathbf{s}_i)' \boldsymbol{\beta}(\mathbf{s}_i) + \varepsilon(\mathbf{s}_i). \quad (1)$$

Further assume that the error term  $\varepsilon(\mathbf{s}_i)$  is normally distributed with zero mean and variance  $\sigma^2$ , and that  $\varepsilon(\mathbf{s}_i)$ ,  $i = 1, \dots, n$  are independent. That is,

$$\boldsymbol{\varepsilon} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2). \quad (2)$$

In the context of nonparametric regression, the boundary-effect bias can be reduced by local polynomial modeling, usually in the form of a locally linear model (Fan and Gijbels, 1996). Here, locally linear coefficients are estimated by augmenting the local design matrix with covariate-by-location interactions in two dimensions as proposed by Wang et al. (2008). The augmented local design

matrix at location  $\mathbf{s}_i$  is

$$\mathbf{Z}(\mathbf{s}_i) = (\mathbf{X} \ L_i \mathbf{X} \ M_i \mathbf{X}) \quad (3)$$

where  $\mathbf{X}$  is the unaugmented matrix of covariates,  $L_i = \text{diag}\{s_{i'_1} - s_{i_1}\}$  and  $M_i = \text{diag}\{s_{i'_2} - s_{i_2}\}$  for  $i' = 1, \dots, n$ .

Now we have that  $Y(\mathbf{s}_i) = \{\mathbf{Z}(\mathbf{s}_i)\}_i^T \boldsymbol{\zeta}(\mathbf{s}_i) + \varepsilon(\mathbf{s}_i)$ , where  $\{\mathbf{Z}(\mathbf{s}_i)\}_i^T$  is the  $i$ th row of the matrix  $\mathbf{Z}(\mathbf{s}_i)$  as a row vector, and  $\boldsymbol{\zeta}(\mathbf{s}_i)$  is the vector of local coefficients at location  $\mathbf{s}_i$ , augmented with the local gradients of the coefficient surfaces in the two spatial dimensions, indicated by  $\nabla_u$  and  $\nabla_v$ :

$$\boldsymbol{\zeta}(\mathbf{s}_i) = (\boldsymbol{\beta}(\mathbf{s}_i)^T \ \nabla_u \boldsymbol{\beta}(\mathbf{s}_i)^T \ \nabla_v \boldsymbol{\beta}(\mathbf{s}_i)^T)^T$$

## 1.2. Estimation

The total log-likelihood of the observed data is the sum of the log-likelihood of each individual observation:

$$\ell \{\boldsymbol{\zeta}\} = -(1/2) \sum_{i=1}^n \left[ \log \sigma^2 + \sigma^{-2} \{y(\mathbf{s}_i) - \mathbf{z}'(\mathbf{s}_i) \boldsymbol{\zeta}(\mathbf{s}_i)\}^2 \right]. \quad (4)$$

Since there are a total of  $n \times 3(p+1) + 1$  parameters for  $n$  observations, the model is not identifiable and it is not possible to directly maximize the total likelihood. But since the coefficient functions are smooth, the coefficients at location  $\mathbf{s}$  can approximate the coefficients within some neighborhood of  $\mathbf{s}$ , with the quality of the approximation declining as the distance from  $\mathbf{s}$  increases.

This intuition is formalized by the local likelihood, which is maximized at location  $\mathbf{s}$  to estimate

the local coefficients  $\boldsymbol{\zeta}(\mathbf{s})$ :

$$\mathcal{L}\{\boldsymbol{\zeta}(\mathbf{s})\} = \prod_{i=1}^n \left\{ (2\pi\sigma^2)^{-1/2} \exp \left[ -(1/2)\sigma^{-2} \{y(\mathbf{s}_i) - \mathbf{z}'(\mathbf{s}_i)\boldsymbol{\zeta}(\mathbf{s})\}^2 \right] \right\}^{K_h(\|\mathbf{s} - \mathbf{s}_i\|)}, \quad (5)$$

The weights are computed from a kernel function  $K_h(\cdot)$  such as the Epanechnikov kernel:

$$K_h(\|\mathbf{s}_i - \mathbf{s}_{i'}\|) = h^{-2} K(h^{-1}\|\mathbf{s}_i - \mathbf{s}_{i'}\|)$$

$$K(x) = \begin{cases} (3/4)(1 - x^2) & \text{if } x < 1, \\ 0 & \text{if } x \geq 1. \end{cases} \quad (6)$$

Thus, the local log-likelihood function is, up to an additive constant:

$$\ell\{\boldsymbol{\zeta}(\mathbf{s})\} = -(1/2) \sum_{i=1}^n K_h(\|\mathbf{s} - \mathbf{s}_i\|) \left[ \log \sigma^2 + \sigma^{-2} \{y(\mathbf{s}_i) - \mathbf{z}'(\mathbf{s}_i)\boldsymbol{\zeta}(\mathbf{s})\}^2 \right]. \quad (7)$$

Letting  $\mathbf{W}(\mathbf{s})$  be a diagonal weight matrix where  $W_{ii}(\mathbf{s}) = K_h(\|\mathbf{s} - \mathbf{s}_i\|)$ , the local likelihood is maximized by weighted least squares:

$$\begin{aligned} \mathcal{S}\{\boldsymbol{\zeta}(\mathbf{s})\} &= (1/2) \{\mathbf{Y} - \mathbf{Z}(\mathbf{s})\boldsymbol{\zeta}(\mathbf{s})\}^T \mathbf{W}(\mathbf{s}) \{\mathbf{Y} - \mathbf{Z}(\mathbf{s})\boldsymbol{\zeta}(\mathbf{s})\}^T \\ \therefore \tilde{\boldsymbol{\zeta}}(\mathbf{s}) &= \{\mathbf{Z}^T(\mathbf{s})\mathbf{W}(\mathbf{s})\mathbf{Z}(\mathbf{s})\}^{-1} \mathbf{Z}^T(\mathbf{s})\mathbf{W}(\mathbf{s})\mathbf{Y} \end{aligned} \quad (8)$$

## 2. Local variable selection and parameter estimation

### 2.1. Local variable selection

Local adaptive grouped regularization (LAGR) is explored as a method of local variable selection and coefficient estimation in SVCR models. The proposed LAGR penalty is an adaptive  $\ell_1$  penalty akin to the adaptive group lasso (Wang and Leng, 2008; Zou, 2006).

Grouped variables are selected together for inclusion in the model. Each group in a LAGR model

consists of one covariate and its gradients on the two dimensions of spatial location. That is,

$$\boldsymbol{\zeta}_j(\mathbf{s}) = (\beta_j(\mathbf{s}) \quad \nabla_u \beta_j(\mathbf{s}) \quad \nabla_v \beta_j(\mathbf{s}))^T \text{ for } j = 1, \dots, p.$$

The objective function for the LAGR at location  $\mathbf{s}$  is the penalized local sum of squares:

$$\begin{aligned} Q\{\boldsymbol{\zeta}(\mathbf{s})\} &= \mathcal{S}\{\boldsymbol{\zeta}(\mathbf{s})\} + \mathcal{J}\{\boldsymbol{\zeta}(\mathbf{s})\} \\ &= (1/2) \{\mathbf{Y} - \mathbf{Z}(\mathbf{s})\boldsymbol{\zeta}(\mathbf{s})\}^T \mathbf{W}(\mathbf{s}) \{\mathbf{Y} - \mathbf{Z}(\mathbf{s})\boldsymbol{\zeta}(\mathbf{s})\} + \sum_{j=1}^p \phi_j(\mathbf{s}) \|\boldsymbol{\zeta}_j(\mathbf{s})\| \end{aligned} \quad (9)$$

which is the sum of the weighted sum of squares  $\mathcal{S}\{\boldsymbol{\zeta}(\mathbf{s})\}$  and the LAGR penalty  $\mathcal{J}\{\boldsymbol{\zeta}(\mathbf{s})\}$ .

The LAGR penalty for the  $j$ th group of coefficients  $\boldsymbol{\zeta}_j(\mathbf{s})$  at location  $\mathbf{s}$  is  $\phi_j(\mathbf{s}) = \lambda_n(\mathbf{s}) \|\tilde{\boldsymbol{\zeta}}_j(\mathbf{s})\|^{-\gamma}$ , where  $\lambda_n(\mathbf{s}) > 0$  is a the local tuning parameter applied to all coefficients at location  $\mathbf{s}$  and  $\tilde{\boldsymbol{\zeta}}_j(\mathbf{s})$  is the vector of unpenalized local coefficients from (8).

## 2.2. Computation

### 2.2.1. Tuning parameter selection

Implementing LAGR requires the selection of local tuning parameters. The criteria commonly used for selecting tuning parameters in lasso-type models are appropriate here, including GCV (Wahba, 1990), Cp (Mallows, 1973), AIC (Akaike, 1973), and BIC (Schwarz, 1978). All of the examples and simulations presented herein used the corrected AIC (AICc) (Hurvich et al., 1998) for tuning parameter selection:

$$\text{AIC}_c(\mathbf{s}) = \hat{\sigma}^{-2}(\mathbf{s}) \|\mathbf{Y} - \mathbf{Z}(\mathbf{s})\hat{\boldsymbol{\zeta}}(\mathbf{s})\|^2 + 2df + 2 \frac{df(df+1)}{\sum_{i=1}^n K_h(\|\mathbf{s} - \mathbf{s}_i\|) - df - 1} \quad (10)$$

where  $df$  is the degrees of freedom as defined in Yuan and Lin (2006):

$$df = \sum_{j=1}^p I \left\{ \|\hat{\zeta}_j(\mathbf{s})\| \right\} + 2 \sum_{j=1}^p \frac{\|\hat{\zeta}_j(\mathbf{s})\|}{\|\tilde{\zeta}_j(\mathbf{s})\|} \quad (11)$$

### 3. Asymptotic properties

#### 3.1. Notation and assumptions

Consider the local model at location  $\mathbf{s}$  where there are  $p_0 < p$  covariates  $\mathbf{X}_{(a)}(\mathbf{s})$  with nonzero local regression coefficients, indicated by  $\beta_{(a)}(\mathbf{s})$ . The remaining covariates  $\mathbf{X}_{(b)}(\mathbf{s})$  have true coefficients equal to zero, indicated by  $\beta_{(b)}(\mathbf{s})$ . Without loss of generality, assume these are covariates  $1, \dots, p_0$ .

Let  $\Psi = E \{ \mathbf{X}(\mathbf{s}_1) \mathbf{X}^T(\mathbf{s}_1) \}$ .

Let  $h = O(n^{-1/6})$ .

Let  $a_n = \max\{\phi_j(\mathbf{s}), j \leq p_0\}$  be the largest penalty applied to a covariate group whose true coefficient norm is nonzero, and  $b_n = \min\{\phi_j(\mathbf{s}), j > p_0\}$  be the smallest penalty applied to a covariate group whose true coefficient norm is zero.

Let  $\mathbf{Z}_k(\mathbf{s})$  be the design matrix for covariate group  $k$ , and  $\mathbf{Z}_{-k}(\mathbf{s})$  be the design matrix for all the data except covariate group  $k$ , respectively. Similarly, let  $\zeta_k(\mathbf{s})$  be the coefficients for covariate group  $k$  and  $\zeta_{-k}(\mathbf{s})$  be the coefficients for all covariate groups except  $k$ .

Finally, let  $\kappa_0 = \int_{R^2} K(\|\mathbf{s}\|) ds$ ,  $\kappa_2 = \int_{R^2} [(1, 0)\mathbf{s}]^2 K(\|\mathbf{s}\|) ds = \int_{R^2} [(0, 1)\mathbf{s}]^2 K(\|\mathbf{s}\|) ds$ , and  $\nu_0 = \int_{R^2} K^2(\|\mathbf{s}\|) ds$ .

#### 3.2. Results

*Asymptotic normality.*

**Theorem 3.1.** *If  $h^{-1}n^{-1/2}a_n \xrightarrow{p} 0$  and  $hn^{-1/2}b_n \xrightarrow{p} \infty$  then*

$$h\sqrt{n} \left[ \hat{\beta}_{(a)}(\mathbf{s}) - \beta_{(a)}(\mathbf{s}) - \frac{\kappa_2 h^2}{2\kappa_0} \{ \nabla_{uu}^2 \beta_{(a)}(\mathbf{s}) + \nabla_{vv}^2 \beta_{(a)}(\mathbf{s}) \} \right] \xrightarrow{d} N(0, f(\mathbf{s})^{-1} \kappa_0^{-2} \nu_0 \sigma^2 \Psi^{-1})$$

*Selection.*

**Theorem 3.2.** *If  $h^{-1}n^{-1/2}a_n \xrightarrow{p} \infty$  and  $hn^{-1/2}b_n \xrightarrow{p} \infty$  then  $P \left\{ \|\hat{\zeta}_j(\mathbf{s})\| = 0 \right\} \rightarrow 0$  if  $j \leq p_0$  and  $P \left\{ \|\hat{\zeta}_j(\mathbf{s})\| = 0 \right\} \rightarrow 1$  if  $j > p_0$ .*

*Remarks.* Together, Theorem 3.1 and Theorem 3.2 indicate that the LAGR estimates have the same asymptotic distribution as a local regression model where the nonzero coefficients are known in advance (Sun et al., 2014), and that the LAGR estimates of true zero coefficients go to zero with probability one. Thus, selection and estimation by LAGR has the oracle property.

*A note on rates.* To prove the oracle properties of LAGR, we assumed that  $h^{-1}n^{-1/2}a_n \xrightarrow{p} 0$  and  $hn^{-1/2}b_n \xrightarrow{p} \infty$ . Therefore,  $h^{-1}n^{-1/2}\lambda_n(\mathbf{s}) \rightarrow 0$  for  $j \leq p_0$  and  $hn^{-1/2}\lambda_n(\mathbf{s})\|\zeta_j(\mathbf{s})\|^{-\gamma} \rightarrow \infty$  for  $j > p_0$ .

We require that  $\lambda_n(\mathbf{s})$  can satisfy both assumptions. Suppose  $\lambda_n(\mathbf{s}) = n^\alpha$ , and recall that  $h = O(n^{-1/6})$  and  $\|\tilde{\zeta}_p(\mathbf{s})\| = O(h^{-1}n^{-1/2})$ . Then  $h^{-1}n^{-1/2}\lambda_n(\mathbf{s}) = O(n^{-1/3+\alpha})$  and  $hn^{-1/2}\lambda_n(\mathbf{s})\|\tilde{\zeta}_p(\mathbf{s})\|^{-\gamma} = O(n^{-2/3+\alpha+\gamma/3})$ .

So  $(2 - \gamma)/3 < \alpha < 1/3$ , which can only be satisfied for  $\gamma > 1$ .

## A. Proofs of theorems

*Proof of theorem 3.1.* Define  $V_4^{(n)}(\mathbf{u})$  to be the

$$\begin{aligned}
V_4^{(n)}(\mathbf{u}) &= Q \left\{ \zeta(\mathbf{s}) + h^{-1}n^{-1/2}\mathbf{u} \right\} - Q \left\{ \zeta(\mathbf{s}) \right\} \\
&= (1/2) \left[ \mathbf{Y} - \mathbf{Z}(\mathbf{s}) \left\{ \zeta(\mathbf{s}) + h^{-1}n^{-1/2}\mathbf{u} \right\} \right]^T \mathbf{W}(\mathbf{s}) \left[ \mathbf{Y} - \mathbf{Z}(\mathbf{s}) \left\{ \zeta(\mathbf{s}) + h^{-1}n^{-1/2}\mathbf{u} \right\} \right] \\
&\quad + \sum_{j=1}^p \phi_j(\mathbf{s}) \|\zeta_j(\mathbf{s}) + h^{-1}n^{-1/2}\mathbf{u}_j\| \\
&\quad - (1/2) \left\{ \mathbf{Y} - \mathbf{Z}(\mathbf{s})\zeta(\mathbf{s}) \right\}^T \mathbf{W}(\mathbf{s}) \left\{ \mathbf{Y} - \mathbf{Z}(\mathbf{s})\zeta(\mathbf{s}) \right\} - \sum_{j=1}^p \phi_j(\mathbf{s}) \|\zeta_j(\mathbf{s})\| \\
&= (1/2) \mathbf{u}^T \left\{ h^{-2}n^{-1} \mathbf{Z}^T(\mathbf{s}) \mathbf{W}(\mathbf{s}) \mathbf{Z}(\mathbf{s}) \right\} \mathbf{u} - \mathbf{u}^T \left[ h^{-1}n^{-1/2} \mathbf{Z}^T(\mathbf{s}) \mathbf{W}(\mathbf{s}) \left\{ \mathbf{Y} - \mathbf{Z}(\mathbf{s})\zeta(\mathbf{s}) \right\} \right] \\
&\quad + \sum_{j=1}^p n^{-1/2} \phi_j(\mathbf{s}) n^{1/2} \left\{ \|\zeta_j(\mathbf{s}) + h^{-1}n^{-1/2}\mathbf{u}_j\| - \|\zeta_j(\mathbf{s})\| \right\} \tag{12}
\end{aligned}$$

Note the different limiting behavior of the third term between the cases  $j \leq p_0$  and  $j > p_0$ :

*Case  $j \leq p_0$ .* If  $j \leq p_0$  then  $n^{-1/2}\phi_j(\mathbf{s}) \rightarrow n^{-1/2}\lambda_n(\mathbf{s})\|\zeta_j(\mathbf{s})\|^{-\gamma}$  and  $|\sqrt{n} \{ \|\zeta_j(\mathbf{s}) + h^{-1}n^{-1/2}\mathbf{u}_j\| - \|\zeta_j(\mathbf{s})\| \}| \leq h^{-1}\|\mathbf{u}_j\|$  so

$$\lim_{n \rightarrow \infty} \phi_j(\mathbf{s}) \left( \|\zeta_j(\mathbf{s}) + h^{-1}n^{-1/2}\mathbf{u}_j\| - \|\zeta_j(\mathbf{s})\| \right) \leq h^{-1}n^{-1/2}\phi_j(\mathbf{s})\|\mathbf{u}_j\| \leq h^{-1}n^{-1/2}a_n\|\mathbf{u}_j\| \rightarrow 0$$

*Case  $j > p_0$ .* If  $j > p_0$  then  $\phi_j(\mathbf{s}) (\|\zeta_j(\mathbf{s}) + h^{-1}n^{-1/2}\mathbf{u}_j\| - \|\zeta_j(\mathbf{s})\|) = \phi_j(\mathbf{s})h^{-1}n^{-1/2}\|\mathbf{u}_j\|$ .

And note that  $h = O(n^{-1/6})$  so that if  $hn^{-1/2}b_n \xrightarrow{p} \infty$  then  $h^{-1}n^{-1/2}b_n \xrightarrow{p} \infty$ .

Now, if  $\|\mathbf{u}_j\| \neq 0$  then

$$h^{-1}n^{-1/2}\phi_j(\mathbf{s})\|\mathbf{u}_j\| \geq h^{-1}n^{-1/2}b_n\|\mathbf{u}_j\| \rightarrow \infty$$

. On the other hand, if  $\|\mathbf{u}_j\| = 0$  then  $h^{-1}n^{-1/2}\phi_j(\mathbf{s})\|\mathbf{u}_j\| = 0$ .

Thus, the limit of  $V_4^{*(n)}(\mathbf{u})$  is the same as the limit of  $V_4^{*(n)}(\mathbf{u})$  where

$$V_4^{*(n)}(\mathbf{u}) = \begin{cases} (1/2)\mathbf{u}^T \{h^{-2}n^{-1}\mathbf{Z}^T(\mathbf{s})\mathbf{W}(\mathbf{s})\mathbf{Z}(\mathbf{s})\} \mathbf{u} - \mathbf{u}^T [h^{-1}n^{-1/2}\mathbf{Z}^T(\mathbf{s})\mathbf{W}(\mathbf{s})\{\mathbf{Y} - \mathbf{Z}(\mathbf{s})\boldsymbol{\zeta}(\mathbf{s})\}] & \text{if } \|\mathbf{u}_j\| = 0 \ \forall j > p_0 \\ \infty & \text{otherwise} \end{cases}.$$

From which it is clear that  $V_4^{*(n)}(\mathbf{u})$  is convex and its unique minimizer is  $\hat{\mathbf{u}}^{(n)}$ :

$$\begin{aligned} 0 &= \{h^{-2}n^{-1}\mathbf{Z}^T(\mathbf{s})\mathbf{W}(\mathbf{s})\mathbf{Z}(\mathbf{s})\} \hat{\mathbf{u}}^{(n)} - [h^{-1}n^{-1/2}\mathbf{Z}^T(\mathbf{s})\mathbf{W}(\mathbf{s})\{\mathbf{Y} - \mathbf{Z}(\mathbf{s})\boldsymbol{\zeta}(\mathbf{s})\}] \\ \therefore \hat{\mathbf{u}}^{(n)} &= \{n^{-1}\mathbf{Z}^T(\mathbf{s})\mathbf{W}(\mathbf{s})\mathbf{Z}(\mathbf{s})\}^{-1} [hn^{-1/2}\mathbf{Z}^T(\mathbf{s})\mathbf{W}(\mathbf{s})\{\mathbf{Y} - \mathbf{Z}(\mathbf{s})\boldsymbol{\zeta}(\mathbf{s})\}] \end{aligned} \quad (13)$$

By the epiconvergence results of Geyer (1994) and Knight and Fu (2000), the minimizer of the limiting function is the limit of the minimizers  $\hat{\mathbf{u}}^{(n)}$ . And since, by Lemma 2 of Sun et al. (2014),

$$\hat{\mathbf{u}}^{(n)} \xrightarrow{d} N \left( \frac{\kappa_2 h^2}{2\kappa_0} \{\nabla_{uu}^2 \boldsymbol{\zeta}_j(\mathbf{s}) + \nabla_{vv}^2 \boldsymbol{\zeta}_j(\mathbf{s})\}, f(\mathbf{s})\kappa_0^{-2}\nu_0\sigma^2\Psi^{-1} \right) \quad (14)$$

the result is proven.  $\square$

*Proof of theorem 3.2.* We showed in Theorem 3.1 that  $\hat{\boldsymbol{\zeta}}_j(\mathbf{s}) \xrightarrow{p} \boldsymbol{\zeta}_j(\mathbf{s}) + \frac{\kappa_2 h^2}{2\kappa_0} \{\nabla_{uu}^2 \boldsymbol{\zeta}_j(\mathbf{s}) + \nabla_{vv}^2 \boldsymbol{\zeta}_j(\mathbf{s})\}$ , so to complete the proof of selection consistency, it only remains to show that  $P \left\{ \hat{\boldsymbol{\zeta}}_j(\mathbf{s}) = 0 \right\} \rightarrow 1$



if  $j > p_0$ .

The proof is by contradiction. Without loss of generality we consider only the case  $j = p$ .

Assume  $\|\hat{\zeta}_p(\mathbf{s})\| \neq 0$ . Then  $Q\{\zeta(\mathbf{s})\}$  is differentiable w.r.t.  $\zeta_p(\mathbf{s})$  and is minimized where

$$\begin{aligned}
0 &= \mathbf{Z}_p^T(\mathbf{s})\mathbf{W}(\mathbf{s}) \left\{ \mathbf{Y} - \mathbf{Z}_{-p}(\mathbf{s})\hat{\zeta}_{-p}(\mathbf{s}) - \mathbf{Z}_p(\mathbf{s})\hat{\zeta}_p(\mathbf{s}) \right\} - \phi_p(\mathbf{s}) \frac{\hat{\zeta}_p(\mathbf{s})}{\|\hat{\zeta}_p(\mathbf{s})\|} \\
&= \mathbf{Z}_p^T(\mathbf{s})\mathbf{W}(\mathbf{s}) \left[ \mathbf{Y} - \mathbf{Z}(\mathbf{s})\zeta(\mathbf{s}) - \frac{h^2\kappa_2}{2\kappa_0} \{ \nabla_{uu}^2 \zeta(\mathbf{s}) + \nabla_{vv}^2 \zeta(\mathbf{s}) \} \right] \\
&\quad + \mathbf{Z}_p^T(\mathbf{s})\mathbf{W}(\mathbf{s})\mathbf{Z}_{-p}(\mathbf{s}) \left[ \zeta_{-p}(\mathbf{s}) + \frac{h^2\kappa_2}{2\kappa_0} \{ \nabla_{uu}^2 \zeta_{-p}(\mathbf{s}) + \nabla_{vv}^2 \zeta_{-p}(\mathbf{s}) \} - \hat{\zeta}_{-p}(\mathbf{s}) \right] \\
&\quad + \mathbf{Z}_p^T(\mathbf{s})\mathbf{W}(\mathbf{s})\mathbf{Z}_p(\mathbf{s}) \left[ \zeta_p(\mathbf{s}) + \frac{h^2\kappa_2}{2\kappa_0} \{ \nabla_{uu}^2 \zeta_p(\mathbf{s}) + \nabla_{vv}^2 \zeta_p(\mathbf{s}) \} - \hat{\zeta}_p(\mathbf{s}) \right] \\
&\quad - \phi_p(\mathbf{s}) \frac{\hat{\zeta}_p(\mathbf{s})}{\|\hat{\zeta}_p(\mathbf{s})\|}
\end{aligned} \tag{15}$$

So

$$\begin{aligned}
\frac{h}{\sqrt{n}}\phi_p(\mathbf{s}) \frac{\hat{\zeta}_p(\mathbf{s})}{\|\hat{\zeta}_p(\mathbf{s})\|} &= \mathbf{Z}_p^T(\mathbf{s})\mathbf{W}(\mathbf{s}) \frac{h}{\sqrt{n}} \left[ \mathbf{Y} - \mathbf{Z}(\mathbf{s})\zeta(\mathbf{s}) - \frac{h^2\kappa_2}{2\kappa_0} \{ \nabla_{uu}^2 \zeta(\mathbf{s}) + \nabla_{vv}^2 \zeta(\mathbf{s}) \} \right] \\
&\quad + \{ n^{-1} \mathbf{Z}_p^T(\mathbf{s})\mathbf{W}(\mathbf{s})\mathbf{Z}_{-p}(\mathbf{s}) \} h\sqrt{n} \left[ \zeta_{-p}(\mathbf{s}) + \frac{h^2\kappa_2}{2\kappa_0} \{ \nabla_{uu}^2 \zeta_{-p}(\mathbf{s}) + \nabla_{vv}^2 \zeta_{-p}(\mathbf{s}) \} - \hat{\zeta}_{-p}(\mathbf{s}) \right] \\
&\quad + \{ n^{-1} \mathbf{Z}_p^T(\mathbf{s})\mathbf{W}(\mathbf{s})\mathbf{Z}_p(\mathbf{s}) \} h\sqrt{n} \left[ \zeta_p(\mathbf{s}) + \frac{h^2\kappa_2}{2\kappa_0} \{ \nabla_{uu}^2 \zeta_p(\mathbf{s}) + \nabla_{vv}^2 \zeta_p(\mathbf{s}) \} - \hat{\zeta}_p(\mathbf{s}) \right]
\end{aligned} \tag{16}$$

From Lemma 2 of Sun et al. (2014),  $\{n^{-1} \mathbf{Z}_p^T(\mathbf{s})\mathbf{W}(\mathbf{s})\mathbf{Z}_{-p}(\mathbf{s})\} = O_p(1)$  and  $\{n^{-1} \mathbf{Z}_p^T(\mathbf{s})\mathbf{W}(\mathbf{s})\mathbf{Z}_p(\mathbf{s})\} = O_p(1)$ .

From Theorem 3 of Sun et al. (2014), we have that  $h\sqrt{n} \left[ \hat{\zeta}_{-p}(\mathbf{s}) - \zeta_{-p}(\mathbf{s}) - \frac{h^2\kappa_2}{2\kappa_0} \{ \nabla_{uu}^2 \zeta_{-p}(\mathbf{s}) + \nabla_{vv}^2 \zeta_{-p}(\mathbf{s}) \} \right] =$

$$O_p(1) \text{ and } h\sqrt{n} \left[ \hat{\zeta}_p(\mathbf{s}) - \zeta_p(\mathbf{s}) - \frac{h^2 \kappa_2}{2\kappa_0} \{ \nabla_{uu}^2 \zeta_p(\mathbf{s}) + \nabla_{vv}^2 \zeta_p(\mathbf{s}) \} \right] = O_p(1).$$

So the second and third terms of the sum in (16) are  $O_p(1)$ .

We showed in the proof of 3.1 that  $h\sqrt{n} \mathbf{Z}_p^T(\mathbf{s}) \mathbf{W}(\mathbf{s}) \left[ \mathbf{Y} - \mathbf{Z}(\mathbf{s}) \zeta(\mathbf{s}) - \frac{h^2 \kappa_2}{2\kappa_0} \{ \nabla_{uu}^2 \zeta(\mathbf{s}) + \nabla_{vv}^2 \zeta(\mathbf{s}) \} \right] = O_p(1)$ .

The three terms of the sum to the right of the equals sign in (16) are  $O_p(1)$ , so for  $\hat{\zeta}_p(\mathbf{s})$  to be a solution, we must have that  $hn^{-1/2} \phi_p(\mathbf{s}) \hat{\zeta}_p(\mathbf{s}) / \|\hat{\zeta}_p(\mathbf{s})\| = O_p(1)$ .

But since by assumption  $\hat{\zeta}_p(\mathbf{s}) \neq 0$ , there must be some  $k \in \{1, \dots, 3\}$  such that  $|\hat{\zeta}_{p_k}(\mathbf{s})| = \max\{|\hat{\zeta}_{p_{k'}}(\mathbf{s})| : 1 \leq k' \leq 3\}$ . And for this  $k$ , we have that  $|\hat{\zeta}_{p_k}(\mathbf{s})| / \|\hat{\zeta}_p(\mathbf{s})\| \geq 1/\sqrt{3} > 0$ .

Now since  $hn^{-1/2} b_n \rightarrow \infty$ , we have that  $hn^{-1/2} \phi_p(\mathbf{s}) \hat{\zeta}_p(\mathbf{s}) / \|\hat{\zeta}_p(\mathbf{s})\| \geq hb_n / \sqrt{3n} \rightarrow \infty$  and therefore the term to the left of the equals sign dominates the sum to the right of the equals sign in (16). So for large enough  $n$ ,  $\hat{\zeta}_p(\mathbf{s}) \neq 0$  cannot maximize  $Q$ .

So  $P \left\{ \hat{\zeta}_{(b)}(\mathbf{s}) = 0 \right\} \rightarrow 1$ . □

## B. References

### References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle.  
 In B. Petrov and F. Csaki (Eds.), *2nd International Symposium on Information Theory*, pp. 267–281.
- Fan, J. and I. Gijbels (1996). *Local Polynomial Modeling and its Applications*. Chapman and Hall, London.

- Geyer, C. J. (1994). On the asymptotics of constrained m-estimation. *Annals of Statistics* 22, 1993–2010.
- Hurvich, C. M., J. S. Simonoff, and C.-L. Tsai (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society Series B* 60, 271–293.
- Knight, K. and W. Fu (2000). Asymptotics for lasso-type estimators. *Annals of Statistics* 28, 1356–1378.
- Mallows, C. (1973). Some comments on cp. *Technometrics*, 661–675.
- Schwarz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics* 6, 461–464.
- Sun, Y., H. Yan, W. Zhang, and Z. Lu (2014). A semiparametric spatial dynamic model. *Annals of Statistics*.
- Wahba, G. (1990). *Spline models for observational data*. Institute of Mathematical Sciences.
- Wang, H. and C. Leng (2008). A note on adaptive group lasso. *Computational Statistics and Data Analysis* 52, 5277–5286.
- Wang, N., C.-L. Mei, and X.-D. Yan (2008). Local linear estimation of spatially varying coefficient models: an improvement on the geographically weighted regression technique. *Environment and Planning A* 40, 986–1005.
- Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B* 68, 49–67.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101, 1418–1429.