==

1. Introduction

Whereas the coefficients in traditional linear regression are scalar constants, the coefficients in a varying coefficients regression (VCR) model are functions - often *smooth* functions - of some effect-modifying variable (Hastie and Tibshirani, 1993; Cleveland and Grosse, 1991). Current practice for VCR models relies on global model selection to decide which variables should be included in the model, meaning that predictors are identified as relevant or irrelevant over the entire domain. Antoniadas et al. (2012) describe a method for globally selecting the relevant predictors in a VCR model where the coefficient functions are estimated with P-splines. Wang et al. (2008) show a method for doing global variable selection in a VCR model where the coefficient functions are estimated by basis expansion. Local adaptive grouped regularization (LAGR) is developed here as a method to select only the locally relevant predictors at any specific location $\boldsymbol{s}$ in the domain $\mathcal{D}$ of a VCR model. The method of LAGR applies to VCR models where the coefficients are estimated using locally linear kernel smoothing. Using kernel smoothing for nonparametric regression is described in detail in Fan and Gijbels (1996). The extension to estimating VCR models is made by Fan and Zhang (1999) for a VCR a univariate effect-modifying variable, and by Sun et al. (2014) for two-dimensional effect-modifying variable and autocorrelation among the obverved response. These methods minimize the boundary effect (Hastie and Loader, 1993) by estimating the coefficients as local polynomials of odd degree (usually locally linear). For linear regression models, the adaptive lasso (AL) produces consistent estimates of the coefficients and has been shown to have appealing properties for variable selection which, under suitable conditions, include the "oracle" property of asymptotically including exactly according to a density $f(\boldsymbol{s})$ with $\boldsymbol{s} \in \mathcal{D}$. For $i = 1, \ldots, n$, let $y(\boldsymbol{s}_i)$ and $\boldsymbol{x}(\boldsymbol{s}_i)$ denote, respectively, the univariate response and the $(p+1)$-variate vector of covariates measured at location $\boldsymbol{s}_i$. At each location $\boldsymbol{s}_i$, assume that the outcome is related to the covariates by a linear regression where the coefficients $\boldsymbol{\beta}(\boldsymbol{s}_i)$ may be spatially-varying and $\varepsilon(\boldsymbol{s}_i)$ is random error at location $\boldsymbol{s}_i$. That is,

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

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

$$\boldsymbol{\varepsilon} \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right). \tag{2}$$

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, to prepare for the estimation of locally linear coefficients, we augment the local design matrix with covariate-by-location interactions in two dimensions (Wang et al., 2008). The augmented local design matrix at location $\boldsymbol{s}_i$ is

$$\boldsymbol{Z}(\boldsymbol{s}_i) = (\boldsymbol{X} \;\; L_i\boldsymbol{X} \;\; M_i\boldsymbol{X}) \tag{3}$$

where $\boldsymbol{X}$ is the unaugmented matrix of covariates, $\boldsymbol{L}_i = \text{diag}\{s_{i',1} - s_{i,1}\}$ and $\boldsymbol{M}_i = \text{diag}\{s_{i',2} - s_{i,2}\}$ for $i' = 1, \ldots, n$. Now we have that $Y(\boldsymbol{s}_i) = \{\boldsymbol{Z}(\boldsymbol{s}_i)\}_i^T \boldsymbol{\zeta}(\boldsymbol{s}_i) + \varepsilon(\boldsymbol{s}_i)$, where $\{\boldsymbol{Z}(\boldsymbol{s}_i)\}_i^T$ is the ith row of the matrix $\boldsymbol{Z}(\boldsymbol{s}_i)$ as a row vector, and $\boldsymbol{\zeta}(\boldsymbol{s}_i)$ is the vector of local coefficients at location $\boldsymbol{s}_i$, augmented with the local gradients of the coefficient surfaces in the two spatial dimensions, indicated by ∇_u and ∇_v:

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

2.2. Local Likelihood and Coefficient Estimation

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

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

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 when the coefficient functions are smooth, the coefficients at location $\boldsymbol{s}$ can approximate the coefficients within some neighborhood of $\boldsymbol{s}$, with the quality of the approximation declining as the distance from $\boldsymbol{s}$ increases. This intuition is formalized by the local (log-)likelihood, which is maximized at location $\boldsymbol{s}$ to estimate the local coefficients $\boldsymbol{\zeta}(\boldsymbol{s})$:

$$\ell\left(\boldsymbol{\zeta}(\boldsymbol{s})\right) = -(1/2)\sum_{i=1}^{n} K_h(\|\boldsymbol{s}-\boldsymbol{s}_i\|)\left[\log\sigma^2 + \sigma^{-2}\left\{y(\boldsymbol{s}_i) - \boldsymbol{z}'(\boldsymbol{s}_i)\boldsymbol{\zeta}(\boldsymbol{s})\right\}^2\right] \tag{5}$$

where h is a bandwidth parameter and the $K_h(\|\boldsymbol{s}-\boldsymbol{s}_i\|)$ for $i = 1, \ldots, n$ are local weights from a kernel function. For instance, the Epanechnikov kernel is defined as (Samiuddin and el Sayyad, 1990):

$$\begin{aligned} K_h(\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|) &= h^{-2}K\left(h^{-1}\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|\right) \\ K(x) &= \begin{cases} (3/4)(1-x^2) & \text{if } x < 1, \\ 0 & \text{if } x \geq 1. \end{cases} \end{aligned} \tag{6}$$

Letting $\boldsymbol{W}(\boldsymbol{s}) = diag\left\{K_h(\|\boldsymbol{s}-\boldsymbol{s}_i\|)\right\}$ be a diagonal matrix of kernel weights, the local likelihood is maximized by weighted least squares:

$$\mathcal{S}\left\{\boldsymbol{\zeta}(\boldsymbol{s})\right\} = (1/2)\left\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\}^T \boldsymbol{W}(\boldsymbol{s})\left\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\}^T$$

Thus, we have

$$\tilde{\boldsymbol{\zeta}}(\boldsymbol{s}) = \left\{\boldsymbol{Z}^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}(\boldsymbol{s})\right\}^{-1}\boldsymbol{Z}^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Y}$$

Now Theorem 3 of Sun et al. (2014) says that, for any given $\boldsymbol{s}$

$$\sqrt{nh^2 f(\boldsymbol{s})}\left[\hat{\boldsymbol{\beta}}(\boldsymbol{s}) - \boldsymbol{\beta}(\boldsymbol{s}) - (1/2)\kappa_0^{-1}\kappa_2 h^2\left\{\boldsymbol{\beta}_{uu}(\boldsymbol{s}) + \boldsymbol{\beta}_{vv}(\boldsymbol{s})\right\}\right] \xrightarrow{D} N\left(\boldsymbol{0}, \kappa_0^{-2}\nu_0\sigma^2\Psi^{-1}\right)$$

3. Local Variable Selection with LAGR

3.1. The LAGR-Penalized Local Likelihood

Estimating the local coefficients by (**??**) relies on *a priori* variable selection. A new method of penalized regression to simultaneously select the locally relevant predictors and estimate the local coefficients. For this purpose, each raw covariate is grouped with its covariate-by-location interactions. That is, $\boldsymbol{\zeta}_j(\boldsymbol{s}) = \left(\beta_j(\boldsymbol{s}) \quad \nabla_u\beta_j(\boldsymbol{s}) \quad \nabla_v\beta_j(\boldsymbol{s})\right)^T$ for $j = 1, \ldots, p$. By the mechanism of the group lasso, variables within the same group are included in or dropped from the model together. The intercept group is left unpenalized. The proposed LAGR penalty is an adaptive ℓ_1 penalty akin to the adaptive group lasso (Wang and Leng, 2008; Zou, 2006). More specifically, we consider the penalized local sum of squares at location $\boldsymbol{s}$:

$$\mathcal{J}\left(\boldsymbol{\zeta}(\boldsymbol{s})\right) = \mathcal{S}\left(\boldsymbol{\zeta}(\boldsymbol{s})\right) + \mathcal{P}\left(\boldsymbol{\zeta}(\boldsymbol{s})\right)$$

where $\mathcal{S}\left(\boldsymbol{\zeta}(\boldsymbol{s})\right) = (1/2)\left\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\}^T \boldsymbol{W}(\boldsymbol{s})\left\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\}^T$ is the locally weighted sum of squares, $\mathcal{P}\left(\boldsymbol{\zeta}(\boldsymbol{s})\right) = \sum_{j=1}^{p} \phi_j(\boldsymbol{s})\|\boldsymbol{\zeta}_j(\boldsymbol{s})\|$ is a local adaptive grouped regularization (LAGR) penalty, and $\|\cdot\|$ is the L_2-norm. The LAGR penalty for the jth group of coefficients $\boldsymbol{\zeta}_j(\boldsymbol{s})$ at location $\boldsymbol{s}$ is $\phi_j(\boldsymbol{s}) = \lambda_n(\boldsymbol{s})\|\tilde{\boldsymbol{\zeta}}_j(\boldsymbol{s})\|^{-\gamma}$, where $\lambda_n(\boldsymbol{s}) > 0$ is a local tuning parameter applied to all coefficients at location $\boldsymbol{s}$ and $\tilde{\boldsymbol{\zeta}}_j(\boldsymbol{s})$ is the vector of unpenalized local coefficients from (**??**). *3.2. Oracle Properties*

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

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

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

Remarks. Together, Theorem 1 and Theorem 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 establish 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, $h^{-1}n^{-1/2}\lambda_n(\boldsymbol{s}) \to 0$ for $j \leq p_0$ and $hn^{-1/2}\lambda_n(\boldsymbol{s})\|\boldsymbol{\zeta}_j(\boldsymbol{s})\|^{-\gamma} \to \infty$ for $j > p_0$. We require that $\lambda_n(\boldsymbol{s})$ can satisfy both assumptions. Suppose $\lambda_n(\boldsymbol{s}) = n^{\alpha}$, and recall that $h = O(n^{-1/6})$ and $\|\tilde{\boldsymbol{\zeta}}_p(\boldsymbol{s})\| = O(h^{-1}n^{-1/2})$. Then $h^{-1}n^{-1/2}\lambda_n(\boldsymbol{s}) = O(n^{-1/3+\alpha})$ and $hn^{-1/2}\lambda_n(\boldsymbol{s})\|\tilde{\boldsymbol{\zeta}}_p(\boldsymbol{s}\|^{-\gamma} = O(n^{-2/3+\alpha+\gamma/3})$. Thus, $(2-\gamma)/3 < \alpha < 1/3$, which can only be satisfied for $\gamma > 1$.

3.3. Selecting the tuning parameter $\lambda_n(\boldsymbol{s})$

In practical application, it is necessary to select the LAGR tuning parameter $\lambda_n(\boldsymbol{s})$ for each local model. A popular approach in other lasso-type problems is to select the tuning parameter that maximizes a criterion that approximates the expected log-likelihood of a new, independent data set drawn from the same distribution. This is the framework of Mallows' Cp (Mallows, 1973), Stein's unbiased risk estimate (SURE) (Stein, 1981) and Akaike's information criterion (AIC) (Akaike, 1973).

These criteria use a so-called covariance penalty to estimate the bias due to using the same data set to select a model and to estimate its parameters (Efron, 2004). We adopt the approximate degrees of freedom for the adaptive group lasso from Yuan and Lin (2006) and minimize the AICc to select the tuning parameter $\lambda_n(\boldsymbol{s})$ Hurvich et al. (1998):

$$
\begin{aligned}
\hat{df}(\lambda;\boldsymbol{s}) =& \sum_{j=1}^{p} I\left(\|\hat{\boldsymbol{\zeta}}(\lambda;\boldsymbol{s})\| > 0\right) + \sum_{j=1}^{p} \frac{\|\hat{\boldsymbol{\zeta}}(\lambda;\boldsymbol{s})\|}{\|\tilde{\boldsymbol{\zeta}}(\boldsymbol{s})\|}(p_j - 1) \\
\text{AIC}_c(\lambda;\boldsymbol{s}) =& \sum_{i=1}^{n} K_h(\|\boldsymbol{s} - \boldsymbol{s}_i\|)\sigma^{-2}\left\{y(\boldsymbol{s}_i) - z'(\boldsymbol{s}_i)\hat{\boldsymbol{\zeta}}(\lambda;\boldsymbol{s})\right\}^2 \\
&+ 2\hat{df}(\lambda;\boldsymbol{s}) + \frac{2\hat{df}(\lambda;\boldsymbol{s})\left\{\hat{df}(\lambda;\boldsymbol{s}) + 1\right\}}{\sum_{i=1}^{n} K_h(\|\boldsymbol{s} - \boldsymbol{s}_i\|) - \hat{df}(\lambda;\boldsymbol{s}) - 1}
\end{aligned}
$$

where the local coefficient estimate is written $\hat{\boldsymbol{\zeta}}(\lambda;\boldsymbol{s})$ to emphasize that it depends on the tuning parameter.

4. Simulation Study

4.1. Simulation setup

A simulation study was conducted to assess the performance of the method described in Sections 2–3. Data were simulated on the domain $[0,1]^2$, which was divided into a 30×30 grid. Each of $p = 5$ covariates $X_1, \ldots, X_5$ was simulated by a Gaussian random field with mean zero and exponential covariance function $\text{Cov}\left(X_{ji}, X_{ji'}\right) = \sigma_x^2 \exp\left(-\tau_x^{-1}\delta_{ii'}\right)$ where $\sigma_x^2 = 1$ is the variance, $\tau_x = 0.1$ is the range parameter, and $\delta_{ii'} = \|\boldsymbol{s} - \boldsymbol{s}_i\|$ is the Euclidean distance between $\boldsymbol{s}$ and $\boldsymbol{s}_i$.

Correlation was induced between the covariates by multiplying the design matrix $\boldsymbol{X}$ by $\boldsymbol{R}$, where $\boldsymbol{R}$ is the Cholesky decomposition of the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{R}'\boldsymbol{R}$. The covariance matrix $\boldsymbol{\Sigma}$ is a 5×5 matrix that has ones on the diagonal and ρ for all off-diagonal entries, where ρ is the between-covariate correlation.

The simulated response was $y_i = \boldsymbol{x}_i' \boldsymbol{\beta}_i + \varepsilon_i$ for $i = 1, \ldots, n$ where $n = 900$ and the ε_i's were iid Gaussian with mean zero and variance σ_ε^2. The simulated data included the response y and five covariates $x_1, \ldots, x_5$. The true data-generating model uses only x_1. The variables $x_2, \ldots, x_5$ are included to assess performance in model selection.

Three different functions were used for the coefficient surface $\beta_1(\boldsymbol{s})$:

$$\beta_{step}(\boldsymbol{s}) = \begin{cases} 1 & if \ s_x > 0.6 \\ 5s_x - 2 & if \ 0.4 < s_x \leq 0.6 \\ 0 & o.w. \end{cases},$$

$\beta_{gradient}(\boldsymbol{s}) = s_x$, and $\beta_{parabola}(\boldsymbol{s}) = 1 - 2\left\{(s_x - 0.5)^2 + (s_y - 0.5)^2\right\}$ (Figure 1). The first is a step function, which is equal to zero in 40% of the spatial domain, equal to one in a different 40% of the spatial domain, and increases linearly in the middle 20% of the domain. The second is a gradient function, which increases linearly from zero at one end of the domain to one at the other. The final coefficient function is a parabola taking its maximum value of 1 at the center of the domain and falling to zero at each corner of the domain.

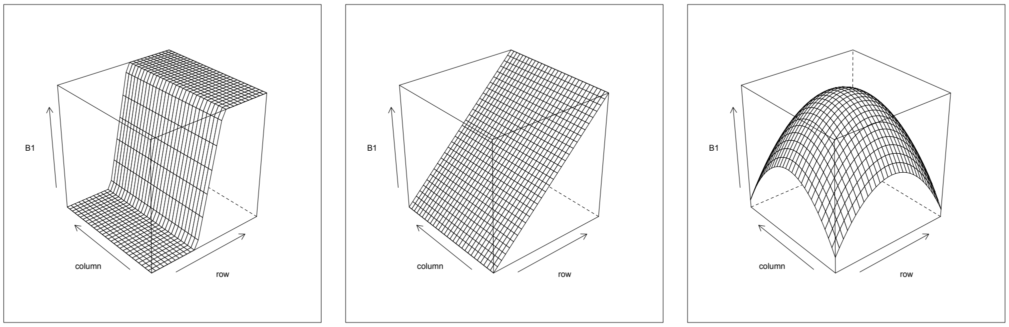

Figure 1: These are, respectively, the step, gradient, and parabola functions that were used for the coefficient function $\beta_1(\boldsymbol{s})$ in the VCR model $y(\boldsymbol{s}_i) = x_1(\boldsymbol{s}_i)\beta_1(\boldsymbol{s}_i) + \varepsilon(\boldsymbol{s}_i)$ when generating the data for the simulation study.

$$\beta_{gradient}(\boldsymbol{s}) = \ s_x \tag{7}$$

In total, three parameters were varied to produce 18 settings, each of which was simulated 100 times. There were three functional forms for the coefficient surface $\beta_1(\boldsymbol{s})$; data was simulated both with low ($\rho = 0$), medium ($\rho = 0.5$), and high ($\rho = 0.9$) correlation between the covariates; and simulations were made with low ($\sigma^2_\varepsilon = 0.25$) and high ($\sigma^2_\varepsilon = 1$) variance for the random error term. The simulation settings are enumerated in Table 1.

The results are presented in terms of the mean integrated squared error

Setting	$\beta_1(\boldsymbol{s})$	ρ	σ^2_ε
1	step	0	0.25
2			1
3		0.5	0.25
4			1
5		0.9	0.25
6			1
7	gradient	0	0.25
8			1
9		0.5	0.25
10			1
11		0.9	0.25
12			1
13	parabola	0	0.25
14			1
15		0.5	0.25
16			1
17		0.9	0.25
18			1

Table 1: Listing of the simulation settings used to assess the performance of LAGR models versus oracle selection and no selection.

(MISE) of the coefficient surface estimates $\hat{\beta}_1(\boldsymbol{s}), \ldots, \hat{\beta}_5(\boldsymbol{s})$, the MISE of the fitted response $\hat{y}(\boldsymbol{s})$, and the frequency with which the coefficient surface estimates $\hat{\beta}_1(\boldsymbol{s}), \ldots, \hat{\beta}_5(\boldsymbol{s})$ in the LAGR model were zero. The performance of LAGR was compared to that of a VCR model without variable selection, and to a VCR model with oracular selection. Oracular selection means that exactly the correct set of covariates was used to fit each local model.

4.2. Simulation Results

The MISE of the estimates of $\beta_1(\boldsymbol{s})$ are in Table 2. Recall that $\beta_2(\boldsymbol{s}), \ldots, \beta_5(\boldsymbol{s})$ are exactly zero across the entire domain. Oracle selection will estimate these coefficients perfectly, so we focus on the comparison between estimation by LAGR and by the VCR model with no selection. These results show that for every simulation setting, LAGR estimation is more accurate than the standard VCR model (Table 3).

From Table 4 we see that LAGR has good ability to identify zero-coefficient covariates. The frequency with which $\beta_2(\boldsymbol{s}), \ldots, \beta_5(\boldsymbol{s})$ were dropped from the LAGR models ranged from 0.78 to 0.97. The MISE of the fitted $\hat{y}(\boldsymbol{s})$ is listed in Table 5, where the highlighting is based on which methods estimate an error variance that is closest to the known truth for the simulation. The results are all very similar to each other, indicating that no method was consistently better than the others in this simulation at fitting the model output

	LAGR	VCR	oracle
1	*0.02*	0.02	**0.01**
2	*0.03*	0.03	**0.02**
3	*0.02*	0.02	**0.01**
4	*0.03*	0.05	**0.02**
5	*0.03*	0.05	**0.01**
6	*0.12*	0.17	**0.02**
7	0.01	*0.01*	**0.00**
8	0.03	*0.02*	**0.01**
9	0.01	*0.01*	**0.00**
10	0.04	*0.03*	**0.01**
11	*0.03*	0.04	**0.00**
12	*0.14*	0.14	**0.01**
13	0.01	*0.01*	**0.01**
14	0.03	*0.02*	**0.02**
15	0.01	*0.01*	**0.01**
16	0.03	*0.03*	**0.02**
17	*0.02*	0.04	**0.01**
18	0.17	*0.14*	**0.02**

Table 2: The MISE for the estimates of $\beta_1(\boldsymbol{s})$ in each simulation setting, under variable selection via LAGR, no variable selection, and oracular variable selection. Highlighting indicates the **lowest** and *next-lowest* MISE.

	LAGR	VCR
1	**0.000**	0.005
2	**0.001**	0.019
3	**0.000**	0.008
4	**0.002**	0.030
5	**0.003**	0.041
6	**0.017**	0.150
7	**0.000**	0.005
8	**0.001**	0.018
9	**0.000**	0.008
10	**0.002**	0.032
11	**0.004**	0.037
12	**0.018**	0.147
13	**0.000**	0.005
14	**0.001**	0.018
15	**0.000**	0.008
16	**0.002**	0.031
17	**0.004**	0.038
18	**0.027**	0.146

Table 3: The MISE for the estimates of $\beta_2(\boldsymbol{s}), \ldots, \beta_5(\boldsymbol{s})$ in each simulation setting, under variable selection via LAGR and no variable selection. Highlighting indicates the **lowest** MISE.

	Frequency of exact zero
1	0.97
2	0.96
3	0.96
4	0.92
5	0.86
6	0.85
7	0.96
8	0.95
9	0.94
10	0.92
11	0.80
12	0.85
13	0.97
14	0.94
15	0.95
16	0.88
17	0.79
18	0.78

Table 4: Proportion of local models under each setting in which the coefficients $\beta_2(\boldsymbol{s}), \ldots, \beta_5(\boldsymbol{s})$ are estimated as exactly zero.

	LAGR	VCR	oracle
1	*0.25*	0.26	**0.25**
2	*1.00*	**1.00**	0.99
3	*0.26*	0.26	**0.25**
4	*0.99*	**1.00**	0.98
5	*0.27*	0.30	**0.25**
6	*1.08*	1.14	**0.98**
7	*0.25*	**0.25**	0.25
8	**0.99**	*0.99*	0.97
9	*0.25*	**0.25**	0.24
10	*1.00*	**1.00**	0.97
11	*0.27*	0.28	**0.24**
12	*1.09*	1.12	**0.97**
13	*0.25*	**0.25**	0.25
14	**1.00**	*1.00*	0.98
15	**0.25**	0.25	*0.25*
16	**1.00**	*1.00*	0.97
17	*0.26*	0.28	**0.24**
18	1.13	*1.12*	**0.98**

Table 5: The MISE for the fitted output in each simulation setting, under variable selection via LAGR, no variable selection, and oracular variable selection. Highlighting indicates the **closest** and *next-closest* to the actual error variance σ_ε^2 for that setting.

The proposed LAGR method was accurate in selection and estimation, with estimation accuracy for $\beta_1(\boldsymbol{s})$ about equal to that of the VCR model with no selection, and with consistently better accuracy for estimating $\beta_2(\boldsymbol{s}), \ldots, \beta_5(\boldsymbol{s})$.

There was minimal difference in the performance of the proposed LAGR method between low ($\sigma_\varepsilon = 0.5$) and high ($\sigma_\varepsilon = 1$) error variance, and between no ($\rho = 0$) and moderate ($\rho = 0.5$) correlation among the predictor variables. But the selection and estimation accuracy did decline when there was high ($\rho = 0.9$) correlation among the predictor variables.

5. Data Example

The proposed LAGR estimation method was used to estimate the coefficients in a VCR model of the effect of some covariates on the price of homes in Boston. The data source is the Boston house price data set, which is based on the 1970 U.S. census Harrison and Rubinfeld (1978); Gilley and Pace (1996); Pace and Gilley (1997). In the data, we have the median price of homes sold in 506 census tracts (MEDV), along with the following potential covariates: CRIM (the per-capita crime rate in the tract), RM (the mean number of rooms for houses sold in the tract), RAD (an index of how accessible the tract is from Boston's radial roads), TAX (the property tax per $10,000 of property value), and LSTAT (the percentage of the tract's residents who are considered "lower status"). The bandwidth parameter was set to 0.2 for a nearest neighbors-type bandwidth, meaning that the sum of kernel weights for each local model was 20% of the total number of observations. The kernel used was the Epanechnikov kernel.

5.1. Results

Estimates of the regression coefficients are plotted in Figure 2. One interesting result is that LAGR indicates that the TAX variable was nowhere an important predictor of the median house price. Another is that the coefficients of CRIM and LSTAT are everywhere negative or zero (meaning that the increasing the crime rate or proportion of lower-status individuals reduces the median house price where the effect is discernable) and that of RM is positive (meaning that when the average house in a tract has more rooms, the median house will be more expensive). The coefficient of RAD is positive in some areas and negative in others. This indicates that there are parts of Boston where access to radial roads is positively associated with an increase in the median house price and parts where the association is negative.

There is not an obvious spatial pattern to the local coefficients for RAD - there are more tracts with negative coefficients than positive, and the positive coefficients do appear to be clustered, but the tracts with positive coefficients are also adjacent to tracts with negative coefficients.

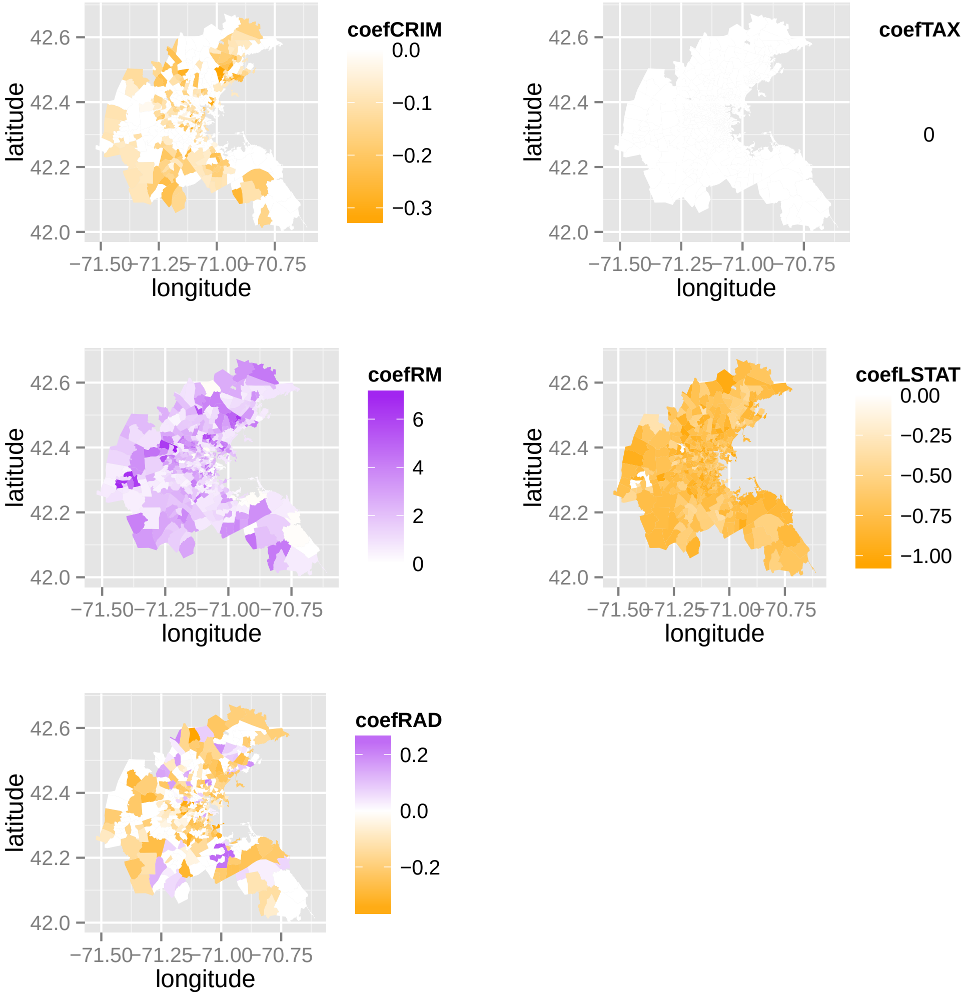

Figure 2: The LAGR estimates of coefficients for the Boston house price data.

	Mean	SD	Prop. zero
CRIM	-0.07	0.08	0.49
RM	1.92	1.43	0.02
RAD	-0.08	0.13	0.37
TAX	0.00	0.00	1.00
LSTAT	-0.72	0.16	0.01

Table 6: The mean, standard deviation, and proportion of zeros among the local coefficients in a model for the median house price in census tracts in Boston, with coefficients selected and fitted by LAGR.

Indeed, there is not an obvious spatial pattern to any of the coefficient surfaces except for TAX, which is zero everywhere.

A summary of the local coefficients is in Table 6. It indicates that RM is the only predictor variable with a positive mean of the local coefficients, but also that the mean of the local coefficients of RM is the largest coefficient - at 1.92, it is more than twice as large in magnitude as the mean local coefficient of LSTAT (-0.72), which is second-largest.

The coefficient of the CRIM variable was estimated to be exactly zero at 49% of the locations. The percentage for the RAD variable was 37%.

In their example using the same data, Sun et al. (2014) estimated that the coefficients of RAD annd LSTAT should be constant, at 0.36 and -0.45, respectively. That conclusion differs from our result, which says that the mean local coefficient of RAD is actually negative (-0.08), while our mean fitted local coefficient for LSTAT was more negative than the estimate of Sun et al. (2014).

AppendixA. Proofs of Theorems

Proof of Theorem 1. □

Let $V_4^{(n)}(\boldsymbol{u}) = \mathcal{J}\left(\boldsymbol{\zeta}(\boldsymbol{s}) + h^{-1}n^{-1/2}\boldsymbol{u}\right) - \mathcal{J}\left(\boldsymbol{\zeta}(\boldsymbol{s})\right)$. Then we have that

$$
\begin{aligned}
V_4^{(n)}(\boldsymbol{u}) =&(1/2)\left[\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\left\{\boldsymbol{\zeta}(\boldsymbol{s}) + h^{-1}n^{-1/2}\boldsymbol{u}\right\}\right]^T \boldsymbol{W}(\boldsymbol{s})\left[\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\left\{\boldsymbol{\zeta}(\boldsymbol{s}) + h^{-1}n^{-1/2}\boldsymbol{u}\right\}\right] \\
&+ \sum_{j=1}^{p} \phi_j(\boldsymbol{s})\|\boldsymbol{\zeta}_j(\boldsymbol{s}) + h^{-1}n^{-1/2}\boldsymbol{u}_j\| \\
&- (1/2)\left\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\}^T \boldsymbol{W}(\boldsymbol{s})\left\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\} - \sum_{j=1}^{p}\phi_j(\boldsymbol{s})\|\boldsymbol{\zeta}_j(\boldsymbol{s})\| \\
=& (1/2)\boldsymbol{u}^T\left\{h^{-2}n^{-1}\boldsymbol{Z}^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}(\boldsymbol{s})\right\}\boldsymbol{u} \\
&- \boldsymbol{u}^T\left[h^{-1}n^{-1/2}\boldsymbol{Z}^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\left\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\}\right] \qquad \text{(A.1)} \\
&+ \sum_{j=1}^{p} n^{-1/2}\phi_j(\boldsymbol{s})n^{1/2}\left\{\|\boldsymbol{\zeta}_j(\boldsymbol{s}) + h^{-1}n^{-1/2}\boldsymbol{u}_j\| - \|\boldsymbol{\zeta}_j(\boldsymbol{s})\|\right\} \qquad \text{(A.2)}
\end{aligned}
$$

The limiting behavior of the third term differs between the cases $j \leq p_0$ and $j > p_0$.

Case $j \leq p_0$:. If $j \leq p_0$, then $n^{-1/2}\phi_j(\boldsymbol{s}) \to n^{-1/2}\lambda_n(\boldsymbol{s})\|\boldsymbol{\zeta}_j(\boldsymbol{s})\|^{-\gamma}$ and $|\sqrt{n}\left\{\|\boldsymbol{\zeta}_j(\boldsymbol{s}) + h^{-1}n^{-1/2}\boldsymbol{u}_j\| - \|\boldsymbol{\zeta}_j(\boldsymbol{s})\|\right\}| \leq h^{-1}\|\boldsymbol{u}_j\|$. Thus,

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

Case $j > p_0$:. If $j > p_0$, then $\phi_j(\boldsymbol{s})\left\{\|\boldsymbol{\zeta}_j(\boldsymbol{s}) + h^{-1}n^{-1/2}\boldsymbol{u}_j\| - \|\boldsymbol{\zeta}_j(\boldsymbol{s})\|\right\} = \phi_j(\boldsymbol{s})h^{-1}n^{-1/2}\|\boldsymbol{u}_j\|$.

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

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

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

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

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

$$V_4^{*(n)}(\boldsymbol{u}) = (1/2)\boldsymbol{u}^T\left\{h^{-2}n^{-1}\boldsymbol{Z}^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}(\boldsymbol{s})\right\}\boldsymbol{u} - \boldsymbol{u}^T\left[h^{-1}n^{-1/2}\boldsymbol{Z}^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\left\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\}\right].$$

if $\|\boldsymbol{u}_j\| = 0\ \forall j > p_0$ and $V_4^{*(n)}(\boldsymbol{u}) = \infty$ otherwise. It follows that $V_4^{*(n)}(\boldsymbol{u})$ is convex and its unique minimizer $\hat{\boldsymbol{u}}^{(n)}$ is found by solving the equation:

$$\boldsymbol{0}_{3p} = \left\{h^{-2}n^{-1}\boldsymbol{Z}(\boldsymbol{s})^T\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}(\boldsymbol{s})\right\}\hat{\boldsymbol{u}}^{(n)} - \left[h^{-1}n^{-1/2}\boldsymbol{Z}(\boldsymbol{s})^T\boldsymbol{W}(\boldsymbol{s})\left\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\}\right]. \tag{A.3}$$

That is,

$$\hat{\boldsymbol{u}}^{(n)} = \left\{n^{-1}\boldsymbol{Z}(\boldsymbol{s})^T\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}(\boldsymbol{s})\right\}^{-1}\left[hn^{-1/2}\boldsymbol{Z}(\boldsymbol{s})^T\boldsymbol{W}(\boldsymbol{s})\left\{\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s})\right\}\right].$$

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

$$\hat{\boldsymbol{u}}^{(n)} \xrightarrow{d} N\left(\frac{\kappa_2 h^2}{2\kappa_0}\{\nabla^2_{uu}\boldsymbol{\zeta}_j(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\zeta}_j(\boldsymbol{s})\}, f(\boldsymbol{s})\kappa_0^{-2}\nu_0\sigma^2\Psi^{-1}\right) \qquad \text{(A.4)}$$

the result of Theorem **??** follows.

Proof of Theorem 2. We showed in Theorem 1 that $\hat{\boldsymbol{\zeta}}_j(\boldsymbol{s}) \xrightarrow{p} \boldsymbol{\zeta}_j(\boldsymbol{s}) + \frac{\kappa_2 h^2}{2\kappa_0}\{\nabla^2_{uu}\boldsymbol{\zeta}_j(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\zeta}_j(\boldsymbol{s})\}$, so to complete the proof of selection consistency, it only remains to show that $P\left\{\hat{\boldsymbol{\zeta}}_j(\boldsymbol{s}) = 0\right\} \to 1$ if $j > p_0$. □

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

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

$$
\begin{aligned}
0 &= \boldsymbol{Z}_p^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\Big\{\boldsymbol{Y} - \boldsymbol{Z}_{-p}(\boldsymbol{s})\hat{\boldsymbol{\zeta}}_{-p}(\boldsymbol{s}) - \boldsymbol{Z}_p(\boldsymbol{s})\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})\Big\} - \phi_p(\boldsymbol{s})\frac{\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})}{\|\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})\|} \\
&= \boldsymbol{Z}_p^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\left[\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s}) - \frac{h^2\kappa_2}{2\kappa_0}\left\{\nabla^2_{uu}\boldsymbol{\zeta}(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\zeta}(\boldsymbol{s})\right\}\right] \\
&\qquad + \boldsymbol{Z}_p^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}_{-p}(\boldsymbol{s})\left[\boldsymbol{\zeta}_{-p}(\boldsymbol{s}) + \frac{h^2\kappa_2}{2\kappa_0}\left\{\nabla^2_{uu}\boldsymbol{\zeta}_{-p}(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\zeta}_{-p}(\boldsymbol{s})\right\} - \hat{\boldsymbol{\zeta}}_{-p}(\boldsymbol{s})\right] \\
&\qquad + \boldsymbol{Z}_p^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}_p(\boldsymbol{s})\left[\boldsymbol{\zeta}_p(\boldsymbol{s}) + \frac{h^2\kappa_2}{2\kappa_0}\left\{\nabla^2_{uu}\boldsymbol{\zeta}_p(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\zeta}_p(\boldsymbol{s})\right\} - \hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})\right] \\
&\qquad - \phi_p(\boldsymbol{s})\frac{\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})}{\|\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})\|}
\end{aligned}
\tag{A.5}
$$

Thus,

$$
\frac{h}{\sqrt{n}}\phi_p(\boldsymbol{s})\frac{\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})}{\|\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})\|} = \tag{A.6}
$$

$$
\boldsymbol{Z}_p^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\frac{h}{\sqrt{n}}\left[\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s}) - \frac{h^2\kappa_2}{2\kappa_0}\left\{\nabla^2_{uu}\boldsymbol{\zeta}(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\zeta}(\boldsymbol{s})\right\}\right] \tag{A.7}
$$

$$
+ \left\{n^{-1}\boldsymbol{Z}_p^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}_{-p}(\boldsymbol{s})\right\} h\sqrt{n}\left[\boldsymbol{\zeta}_{-p}(\boldsymbol{s}) + \frac{h^2\kappa_2}{2\kappa_0}\left\{\nabla^2_{uu}\boldsymbol{\zeta}_{-p}(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\zeta}_{-p}(\boldsymbol{s})\right\} - \hat{\boldsymbol{\zeta}} \tag{A.8}
$$

$$
+ \left\{n^{-1}\boldsymbol{Z}_p^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}_p(\boldsymbol{s})\right\} h\sqrt{n}\left[\boldsymbol{\zeta}_p(\boldsymbol{s}) + \frac{h^2\kappa_2}{2\kappa_0}\left\{\nabla^2_{uu}\boldsymbol{\zeta}_p(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\zeta}_p(\boldsymbol{s})\right\} - \hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})\right] \tag{A.9}
$$

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

and $n^{-1}\boldsymbol{Z}_p^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\boldsymbol{Z}_p(\boldsymbol{s}) = O_p(1)$.

From Theorem 3 of Sun et al. (2014), we have that $h\sqrt{n}\left[\hat{\boldsymbol{\zeta}}_{-p}(\boldsymbol{s}) - \boldsymbol{\zeta}_{-p}(\boldsymbol{s}) - \frac{h^2\kappa_2}{2\kappa_0}\left\{\nabla^2_{uu}\zeta_{-p}(\boldsymbol{s}) + \nabla^2_{vv}\zeta_{-p}(\boldsymbol{s})\right\}\right] = O_p(1)$ and $h\sqrt{n}\left[\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s}) - \boldsymbol{\zeta}_p(\boldsymbol{s}) - \frac{h^2\kappa_2}{2\kappa_0}\left\{\nabla^2_{uu}\zeta_p(\boldsymbol{s}) + \nabla^2_{vv}\zeta_p(\boldsymbol{s})\right\}\right] = O_p(1)$.

Thus, the second and third terms of the sum in (A.6) are $O_p(1)$, and we showed in the proof of Theorem 1 that

$$h\sqrt{n}\boldsymbol{Z}_p^T(\boldsymbol{s})\boldsymbol{W}(\boldsymbol{s})\left[\boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{s})\boldsymbol{\zeta}(\boldsymbol{s}) - \frac{h^2\kappa_2}{2\kappa_0}\left\{\nabla^2_{uu}\boldsymbol{\zeta}(\boldsymbol{s}) + \nabla^2_{vv}\boldsymbol{\zeta}(\boldsymbol{s})\right\}\right] = O_p(1).$$

The three terms of the sum to the right of the equals sign in (A.6) are $O_p(1)$, so for $\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})$ to be a solution, we must have that $hn^{-1/2}\phi_p(\boldsymbol{s})\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})/\|\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})\| = O_p(1)$.

But since by assumption $\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s}) \neq 0$, there must be some $k \in \{1, 2, 3\}$ such that $|\hat{\zeta}_{p_k}(\boldsymbol{s})| = \max\{|\hat{\zeta}_{p_{k'}}(\boldsymbol{s})| : 1 \leq k' \leq 3\}$. And for this k, we have that $|\hat{\zeta}_{p_k}(\boldsymbol{s})|/\|\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})\| \geq 1/\sqrt{3} > 0$.

Now since $hn^{-1/2}b_n \to \infty$, we have that $hn^{-1/2}\phi_p(\boldsymbol{s})\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})/\|\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s})\| \geq hb_n/\sqrt{3n} \to \infty$ and therefore the term to the left of the equals sign dominates the sum to the right of the equals sign in (A.6). Thus, for large enough n, $\hat{\boldsymbol{\zeta}}_p(\boldsymbol{s}) \neq 0$ cannot maximize $\mathcal{J}$.

Thus, $P\left\{\hat{\boldsymbol{\zeta}}_{(b)}(\boldsymbol{s}) = 0\right\} \to 1$.