Functional gradient descent boosting for additive non-linear spatial autoregressive model

Ghislain Geniaux

Functional gradient descent boosting for additive non-linear spatial autoregressive model

Ghislain Geniaux, INRAE Ecodéveloppement UR767

Main objectives:

An extension of spatial autoregressive models (SAR, SDM, SEM and SARAR) to non linear semi-parametric models (additive model with penalized smoothing spline) using boosting algorithms (Friedman, 2001; Bühlmann et al., 2007)

This extension to gradient boosting is mainly based on an estimation of the spatial parameter by QML following the example of Basile and Gress (2004) and Su and Jin (2010),

With two additional extensions:

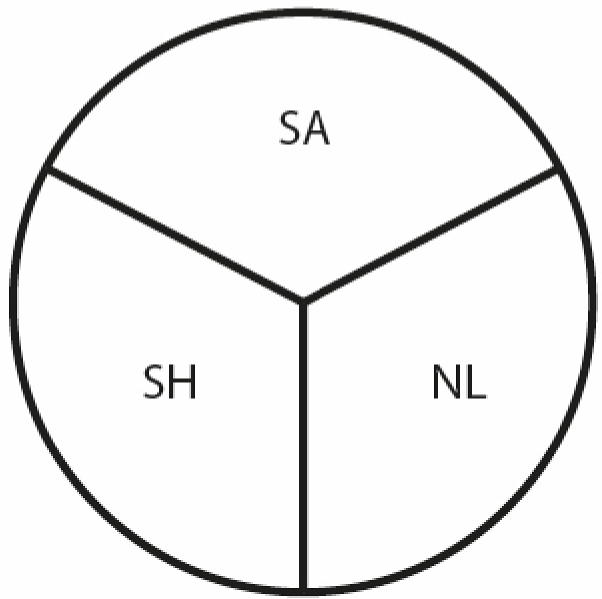
- a Closed Form Estimator (CFE, Smirnov 2020) for SAR and SEM models and,
- a Flexible Instrumental Variable Approach (FIVA, Marra and Radice 2010) for SAR models.

Outline:

1. Spatial dependence, spatial heterogeneity and non-linearity
2. Introducing gradient Boosting
3. Estimators
4. Monte Carlo Results
5. Empirical study on the sale prices of houses in France.

LINEAR SAR / SEM /SDM/SARAR

$Y = \lambda W_1 Y + \theta W_3 X + X\beta + (I - \rho W_2)^{-1}\varepsilon$



Local Linear Model (GWR) and spatial regime

$Y = X\beta(x,y) + \varepsilon$

NON LINEAR NP/SP

$Y = m(X) + \varepsilon$

$Y = \sum m_j(X_j) + \varepsilon$ **(ADDITIVE)**

Diagnostic
Anselin (1988)
Anselin & Griffith (1988)
Anselin & Bera (1998)

Estimation Method
LOCAL GWR-SAR (MV), GWR-SAC (MV)
MGWR-SAR (2SLS), SALE (MV)

Brundston et al. (1996)
Paez et al. (2002)
LeSage & Pace (2004)
Geniaux & Martinetti (2017)+NL

$Y = X\beta + \varepsilon(lon,lat)$
$Y = \lambda WY + Z\theta + X\beta(lon,lat) + \varepsilon$
$Y = \lambda(lon,lat)WY + X\beta(lon,lat)$

Diagnostic
Paelinck & Klaassen (1979)
Baltagi & Li (2001)
Graaff et al. (2001)
McMillen (2003)
Su & Qu (2016)
Lee et al. (2021)

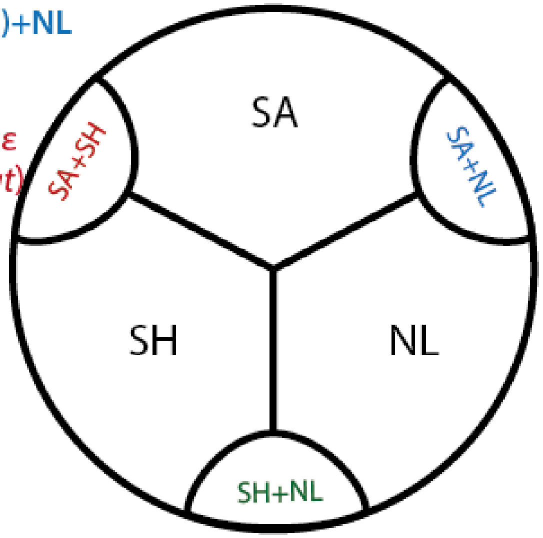
Estimation Method
Gres 2004 (?)
Basile & Gres 2005 (SP-SAR +QML)
Basile 2009 (SPA-SAR+QML)
Basile et al. 2014 (SPA-SAR+IV)+HS

LINEAR SAR / SEM /SDM/SARAR

$Y = \lambda W_1 Y + \theta W_3 X + X\beta + (I - \rho W_2)^{-1}\varepsilon$

Partially Linear AR
Su et Jin (2010) (QML)
Su (2012) (QML+GMM)
Zang (2013) (PDE)
Sun (2016) (+NL WY) ,
Sun et Malikov (2017) +HS

$Y = \lambda W_1 Y + m(X) + (I - \rho W_2)^{-1}\varepsilon$
$Y = \lambda W_1 Y + \Sigma m_j(X_j) + (I - \rho W_2)^{-1}\varepsilon$



Local Linear Model (GWR) and spatial regime
$Y = X\beta(x,y) + \varepsilon$

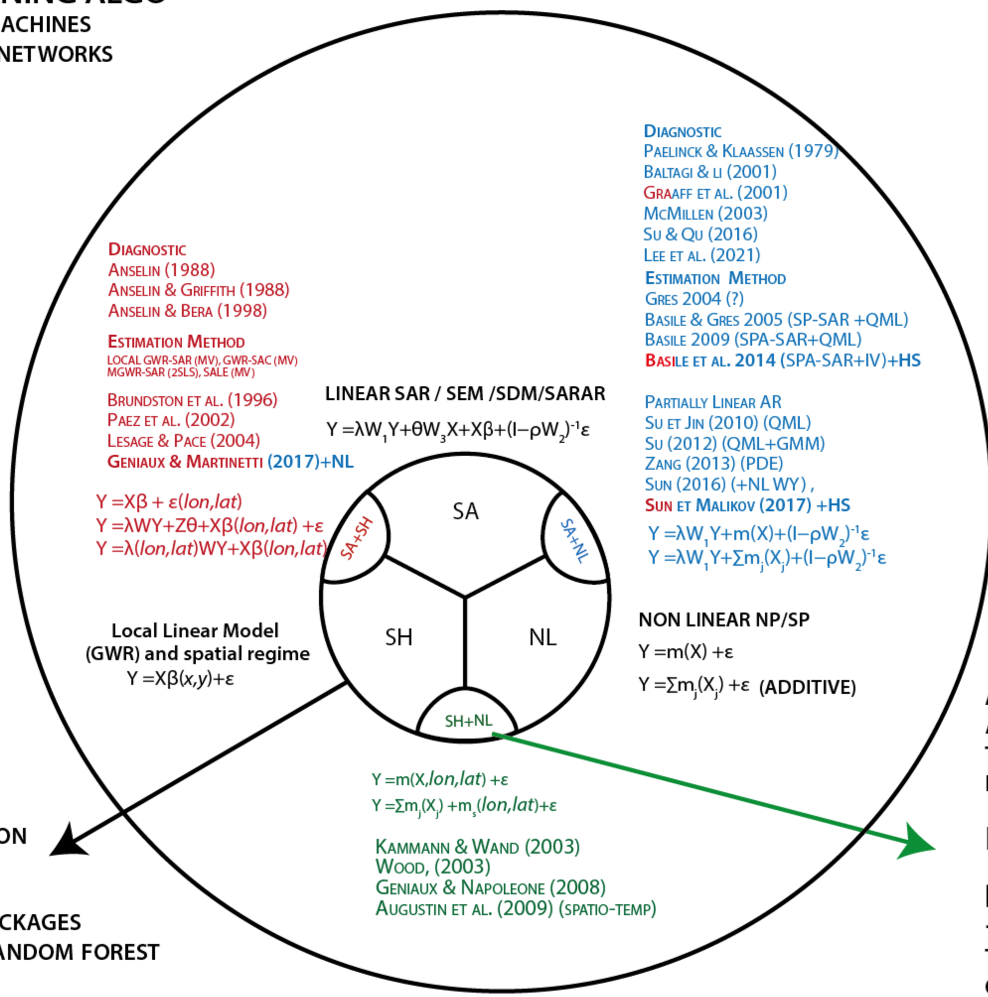
NON LINEAR NP/SP
$Y = m(X) + \varepsilon$
$Y = \Sigma m_j(X_j) + \varepsilon$ **(ADDITIVE)**

$Y = m(X, lon, lat) + \varepsilon$
$Y = \Sigma m_j(X_j) + m_s(lon, lat) + \varepsilon$

Kammann & Wand (2003)
Wood, (2003)
Geniaux & Napoleone (2008)
Augustin et al. (2009) (spatio-temp)

MACHINE LEARNING ALGO

Support-vector machines
Artificial neural networks
Decision tree
Knn
Boosting



Locally weighted version of ML algo :

Example: spatialML R packages
=> locally weighted random forest

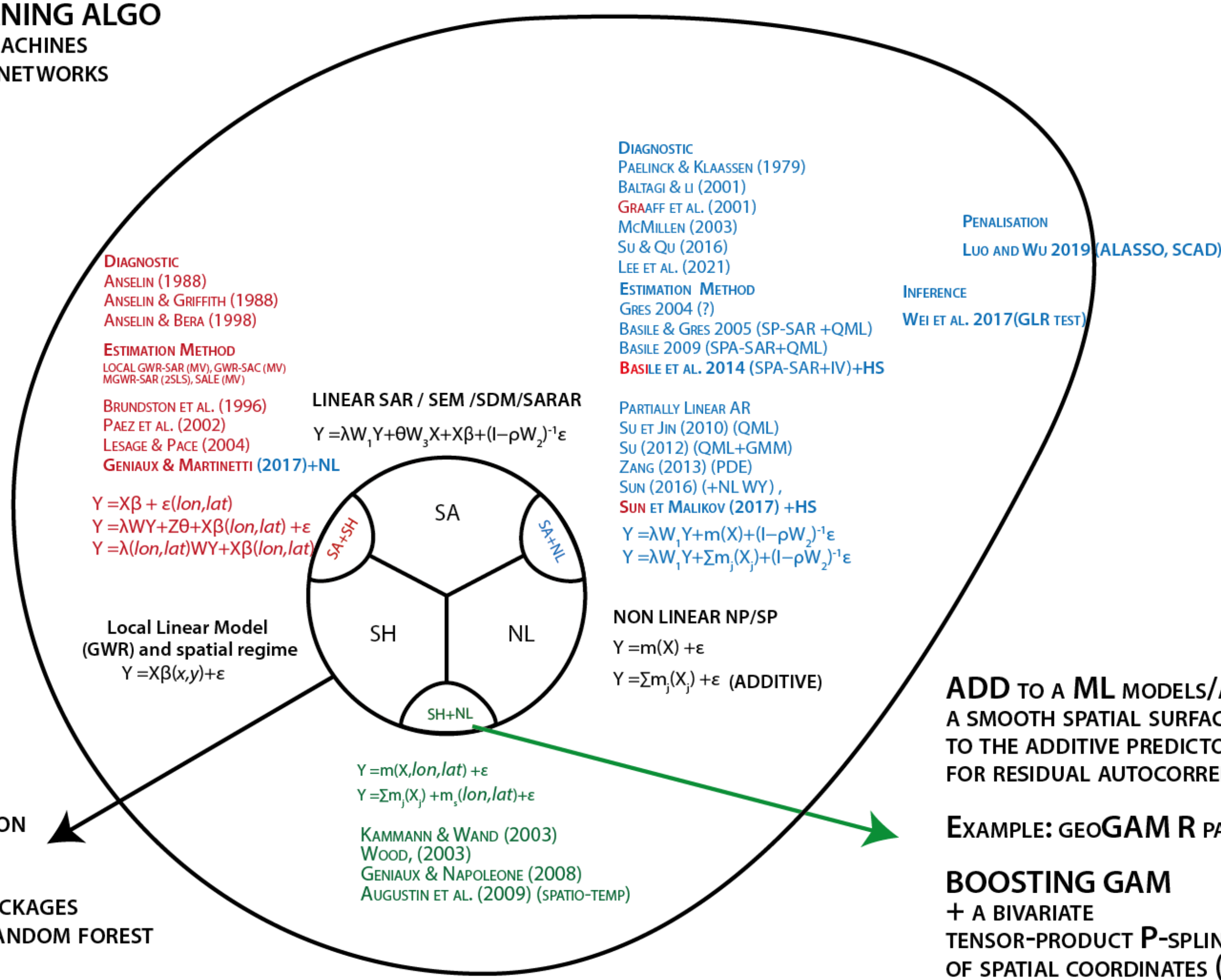
ADD to a ML models/algo
a smooth spatial surface
to the additive predictor to account
for residual autocorrelation

Example: geoGAM R package

BOOSTING GAM
+ a bivariate
tensor-product P-spline
of spatial coordinates (Wood 2006)

MACHINE LEARNING ALGO

Support-vector machines
Artificial neural networks
Decision tree
Knn
Boosting



Locally weighted version
of ML algo :

EXAMPLE: SPATIALML R PACKAGES
=> LOCALLY WEIGHTED RANDOM FOREST

ADD to a ML models/algo
A SMOOTH SPATIAL SURFACE
TO THE ADDITIVE PREDICTOR TO ACCOUNT
FOR RESIDUAL AUTOCORRELATION

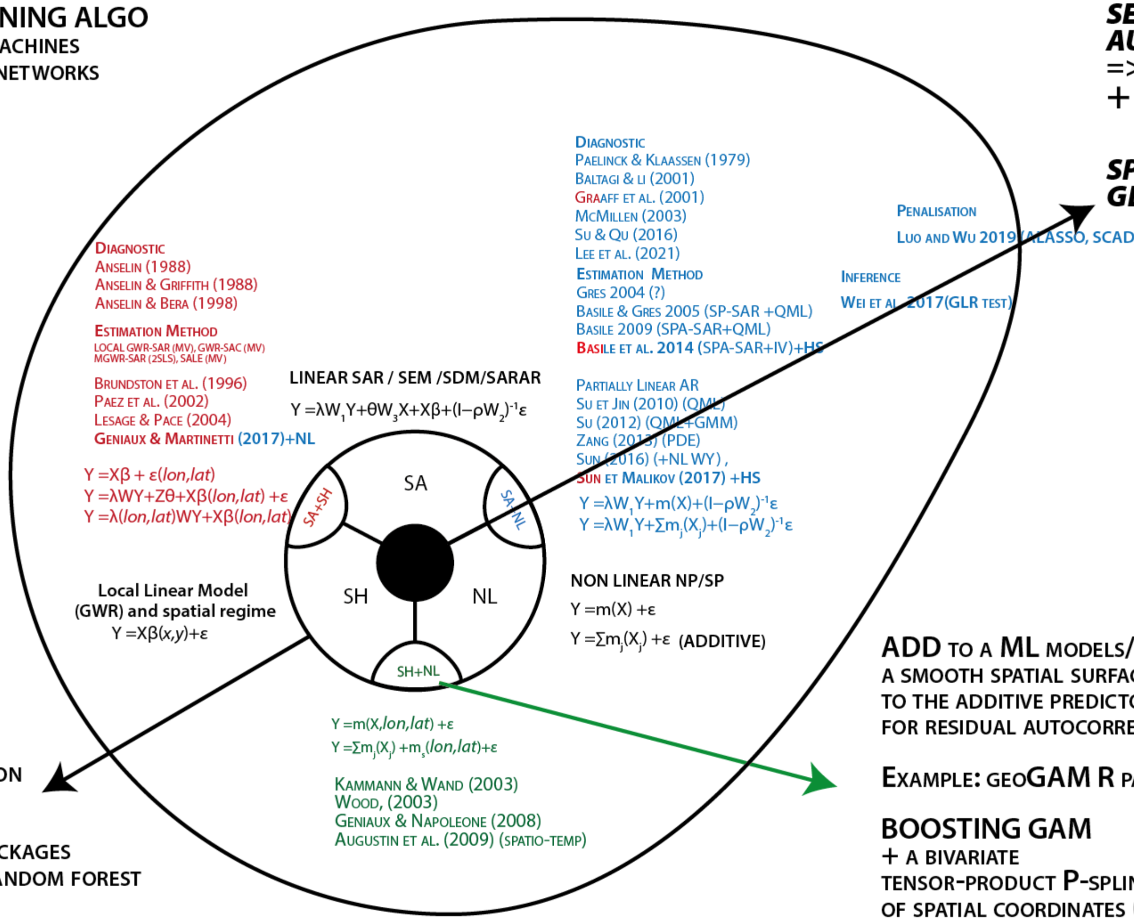
Example: geoGAM R package

BOOSTING GAM
+ A BIVARIATE
TENSOR-PRODUCT P-SPLINE
OF SPATIAL COORDINATES (Wood 2006)

MACHINE LEARNING ALGO
SUPPORT-VECTOR MACHINES
ARTIFICIAL NEURAL NETWORKS
DECISION TREE
KNN
BOOSTING

SEMIPARAMETRIC SPATIAL AUTOREGRESSIVE MODELS
=> BOOSTING GAM
+ SPATIAL DEPENDENCE

SPBOOST R PACKAGE
GENIAUX 2021



LOCALLY WEIGHTED VERSION OF ML ALGO :

EXAMPLE: SPATIALML R PACKAGES
=> LOCALLY WEIGHTED RANDOM FOREST

ADD TO A ML MODELS/ALGO
A SMOOTH SPATIAL SURFACE
TO THE ADDITIVE PREDICTOR TO ACCOUNT
FOR RESIDUAL AUTOCORRELATION

EXAMPLE: GEOGAM R PACKAGE

BOOSTING GAM
+ A BIVARIATE
TENSOR-PRODUCT P-SPLINE
OF SPATIAL COORDINATES (WOOD 2006)

A simple definition of boosting

- Fit an additive model Y=$\sum_t h_t(x_{j(t)}) + \epsilon_t$ in a forward stage-wise manner,
- In each stage t, introduce a weak learner $h_t(x_j(t))$ to compensate the shortcomings of existing weak learners: $\hat{h}_{t+1}(.) = \hat{h}_t(.) + \alpha \hat{h}_t(x_j(t))$
- α is a learning rate that corresponds to the step size used in gradient descent: $0 < \alpha < 1$. In pratice, $\alpha = 0.1$
- In Gradient Boosting, “shortcomings” (choice of $j(t)$) are identified by gradients.
- In Adaboost,“shortcomings” are identified by high-weight data points.
- Both high-weight data points and gradients tell us how to improve our model.

Additive function

It is easy to see that the structure of this function is equivalent to the structure of the additive predictor of a GAM (see Hastie and Tibshirani 1990).

$$\hat{h} = \hat{h_1} + \hat{h_2} + ... + \hat{h_P}$$

where $\hat{h_1}, \hat{h_2}, ..., \hat{h_P}$ correspond to the functions specified by the base-learners. Consequently, $\hat{h_1}, \hat{h_2}, ..., \hat{h_P}$ depend on the predictor variables that were used as input variables of the respective base-learners.

Note that a base-learner can be selected multiple times in the course of the boosting algorithm.

Spatial autogressive non-linear model: a semi-parametri additive estimator

We will consider the two following spatial autoregressives models:

$$Y = \rho WY + \sum_{j=1}^{p} h_j(X_j) + \epsilon \ \ (SAR)$$

$$Y = \sum_{j=1}^{p} h_j(X_j) + (I - \lambda M)^{-1}\epsilon \ \ (SEM)$$

Spatial autogressive non-linear model: a semi-parametri additive estimator

We have the following concentrated loglikelihood:

$$lnL(\rho) = C + |I - \rho W| \tag{1}$$

$$- \frac{n}{2} ln\Big(\frac{(Y - \rho WY - \sum h_j(X_j))'(Y - \rho WY - \sum h_j(X_j))}{n}\Big)$$

Where $C = -(n/2)ln(2\pi) - (n/2)$.

Spatial autogressive non-linear model: a semi-parametri additive estimator

ρ can be estimated by minimizing numerically the concentrated loglikelihood, using the filtered model with $\tilde{Y} = (I - \rho WY)$

While $\sum h_j(X_j)$ can be estimated for a given ρ using semi-parametric additive models with penalized Spline or using boosting with splines as base-learner.

Spatial autogressive non-linear model: a semi-parametri additive estimator

BUT gradient boosting allows a more flexible and general approach by customizing the gradients and the loss function.

Table 1
Gradients, Loss functions and concentrated likelihood

	Aspatial	SAR	SEM	SARAR
Gradient	$(y-f)$	$(y-\rho Wy-f)$	$(I-\lambda W)(y-f)$	$(I-\lambda W)(y-\rho Wy-f)$
Loss Function	$\sum(y-f)^2$	$\sum(y-\rho Wy-f)^2$	$\sum(((I-\lambda W)(y-f))^2)$	$\sum(((I-\lambda W)(y-\rho Wy-f))^2)$

Algorithm 1. BSPA_SAR_ML and BSPA_SEM_ML algorithm

initialize learning rate α (and eventually m_stop).

initialize init $i = 0$, $\eta = 10^{-8}$ and ρ^i with $-1 < \rho^i < 1$

while $\Delta lnL(\rho) < \eta$ **do**

    if *m_stop is NULL* **then**

        Minimize $SSE(\rho^i; \alpha, m\_stop)$ with respect to m_stop using k folds cross validation.

    end

    Estimate $S\hat{S}E(\rho^i; \alpha, m\_stop)$ using functional gradient descent boosting with customized (SAR or SEM) gradient and loss function (function gamboost of R package *mboost*),

    compute $l\hat{n}L(\rho^i; \alpha, m\_stop)$

    compute $\hat{\Delta}^i lnL(\rho^i; \alpha, m\_stop) = l\hat{n}L(\rho^i; \alpha, m\_stop) - l\hat{n}L(\rho^{i-1}; \alpha, m\_stop)$

    chose ρ^{i+1} using golden search rule

end

if *statistical significance is required* **then**

    Rewrite the model as a additive model with convenient base-learner for each selected covariate

    Estimate the model using IPRLS with GCV (function gam of R package *mgcv*) to obtain p-ratio test for each covariate

    Compute a Likelihood-ratio test as $LR = -2 * (lnL(0; \alpha, m\_stop) - lnL(\hat{\rho}^i; \alpha, m\_stop))$

end

A Closed Form Estimator (CFE) approach (Smirnov 2020)

With SAR model, the calculation of the parameter ρ using CFE approach is given by:

$$\rho = \frac{b - \sqrt{D}}{2a}$$

where $D = b^2 - 4ac, a = e_1'e_2, b = e_0'e_2 + e_1'e1, c = e_0'e_1.$

With non linear terms, we have:

$$e_0 = Y - \sum \hat{h}_0(x)$$

$$e_1 = WY - \sum \hat{h}_1(x)$$

$$e_2 = W^2Y - \sum \hat{h}_2(x)$$

A FIVA/control function approach

Instruments:

$$Z = [X, WX, WWX, WWWX]$$

First Stage:

$$Wy = \sum_j' g_j(Z_j') + u$$

Second Stage:

$$y = \sum_j h_j(X_j) + \hat{W}y + \hat{u} + \epsilon$$

Monte Carlo Design

Table 2
Monte Carlo Design

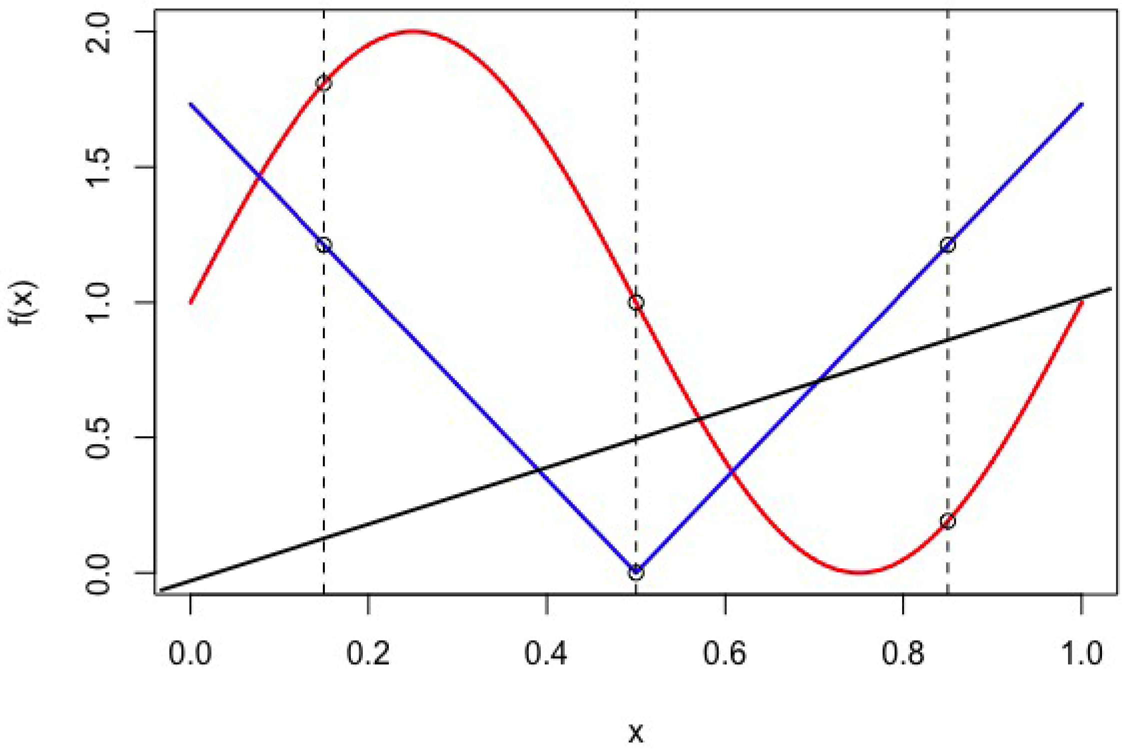
	E1	E2	E3
DGP	SAR, SEM		
ρ	(0,0.2,0.6,0.9)		
Number of repetitions nr	1000		
Locations (u_i, v_i)	Unit square $[0,1]^2$		
W	equal weight, 4 first neighbors		
n	n=(1000,2000,5000)	n=2000	
x_j	$x_j \sim U(0,1)$ $j = 1,2,3$		
SNR	SNR=(0.7,0.9)	0.7	
Non-Linear	$f_1(x_1) = x_1$; $f_2(x_2) = \sqrt{(3(2x_2-1)^2)}$; $f_3(x_3) = sin(2\pi x_3) + 1$		
Spatial Heterogeneity		$\beta_0 = \beta_0 + sqrt(abs(u_i - 0.35) + abs(v_i - 0.65))$ AND/OR $\tilde{x}_2 = (I - 0.2\tilde{W})^{-1}x_2$ AND/OR $\tilde{f}_2(x_2) = \sqrt{(3(2x_2-1)^2)} + (1 + u_i - v_i)^2$	
Correlation between true covariates x_{ij} and false covariates $(x_{ij'}, u_i, v_i)$			$x_{j'} \sim N(0,\Sigma)$, $j' = (4,5,6)$ $cor(x_4, x_1) = (0, 0.1)$ $cor(x_5, x_1) = (0, 0.2)$ $cor(x_6, x_1) = (0, 0)$
Number of different cases	48	18	6

Monte Carlo Design

$f_1(x_1) = x_1$

$f_2(x_2) = \sqrt{(3(2x_2 - 1)^2)}$

$f_3(x_3) = sin(2\pi x_3) + 1$



Monte Carlo Design

List of estimators used in Monte Carlo experiments

Estimators / DGP	SAR NON LINEAR	SEM NON LINEAR	R Packages
BSPA_SAR_ML	x		spboost
BSPA_SAR_CFE	x		spboost
GAM_SAR_ML	x		spboost
GAM_SAR_CFE	x		spboost
GAM_SAR_FIVA	x		spboost
GAM_SAR_2SLSA	x		spboost
BSPA_SEM_ML		x	spboost
BSPA_SEM_CFE		x	spboost
gam	x		mgcv
gamboost	x		mboost
lagsarlm	x		spatialreg
errorsarlm		x	spatialreg

Monte Carlo Results: E1 experiments

Table 5
Result for experiments E1, nonlinear SAR, SNR=0.9 n=2000

ρ	terms	gamboost Bias	gamboost RMSE	gam Bias	gam RMSE	lagsarlm Bias	lagsarlm RMSE	BSPA_SAR_ML Bias	BSPA_SAR_ML RMSE	BSPA_SAR_CFE Bias	BSPA_SAR_CFE RMSE	GAM_SAR_ML Bias	GAM_SAR_ML RMSE	GAM_SAR_CFE Bias	GAM_SAR_CFE RMSE	GAM_SAR_FIVA Bias	GAM_SAR_FIVA RMSE	GAM_SAR_2SLSA Bias	GAM_SAR_2SLSA RMSE
	$\hat{\rho}$					-0.0020	*0.0306*	-0.0021	*0.0276*	-0.0021	*0.0276*	-0.0022	*0.0276*	-0.0021	*0.0276*	0.0401	*0.0749*	0.0157	*0.0531*
0.0	$\hat{f}_1$		*0.0172*		*0.0142*		*0.5355*		*0.0172*		*0.0172*		*0.0142*		*0.0141*		*0.0140*		*0.0140*
	$\hat{f}_2$		*0.0459*		*0.0343*		*0.2885*		*0.0459*		*0.0459*		*0.0343*		*0.0343*		*0.0344*		*0.0344*
	$\hat{f}_3$		*0.0391*		*0.0202*		*0.5384*		*0.0391*		*0.0391*		*0.0202*		*0.0202*		*0.0203*		*0.0203*
	$\hat{\rho}$					-0.0022	*0.0275*	-0.0022	*0.0253*	-0.0020	*0.0255*	-0.0023	*0.0253*	-0.0020	*0.0255*	0.0511	*0.0787*	0.0215	*0.0533*
0.2	$\hat{f}_1$		*0.0164*		*0.0182*		*0.5376*		*0.0174*		*0.0174*		*0.0154*		*0.0154*		*0.0145*		*0.0145*
	$\hat{f}_2$		*0.0460*		*0.0371*		*0.2888*		*0.0466*		*0.0466*		*0.0366*		*0.0366*		*0.0366*		*0.0366*
	$\hat{f}_3$		*0.0379*		*0.0209*		*0.5385*		*0.0393*		*0.0393*		*0.0215*		*0.0216*		*0.0221*		*0.0221*
	$\hat{\rho}$					-0.0019	*0.0177*	-0.0019	*0.0169*	-0.0016	*0.0180*	-0.0020	*0.0169*	-0.0016	*0.0180*	0.0642	*0.0767*	0.0288	*0.0462*
0.6	$\hat{f}_1$		*0.0375*		*0.0605*		*0.5448*		*0.0187*		*0.0187*		*0.0193*		*0.0193*		*0.0155*		*0.0155*
	$\hat{f}_2$		*0.0533*		*0.0615*		*0.2896*		*0.0494*		*0.0495*		*0.0438*		*0.0438*		*0.0441*		*0.0441*
	$\hat{f}_3$		*0.0443*		*0.0444*		*0.5390*		*0.0399*		*0.0399*		*0.0258*		*0.0259*		*0.0298*		*0.0298*
	$\hat{\rho}$					-0.0009	*0.0060*	-0.0008	*0.0059*	-0.0007	*0.0076*	-0.0009	*0.0059*	-0.0007	*0.0076*	0.0332	*0.0365*	0.0161	*0.0213*
0.9	$\hat{f}_1$		*0.2243*		*0.2621*		*0.5549*		*0.0226*		*0.0226*		*0.0249*		*0.0250*		*0.0183*		*0.0183*
	$\hat{f}_2$		*0.2061*		*0.2188*		*0.2912*		*0.0542*		*0.0542*		*0.0532*		*0.0533*		*0.0540*		*0.0540*
	$\hat{f}_3$		*0.2319*		*0.2453*		*0.5396*		*0.0421*		*0.0421*		*0.0316*		*0.0317*		*0.0389*		*0.0389*

Monte Carlo Results: E1 experiments

Table 5
Result for experiments E1, nonlinear SAR, SNR=0.9 n=2000

		gamboost		gam		lagsarlm		BSPA_SAR_ML		BSPA_SAR_CFE		GAM_SAR_ML		GAM_SAR_CFE		GAM_SAR_FIVA		GAM_SAR_2SLSA	
ρ	terms	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
0.0	$\hat{\rho}$					-0.0020	*0.0306*	-0.0021	*0.0276*	-0.0021	*0.0276*	-0.0022	*0.0276*	-0.0021	*0.0276*	0.0401	*0.0749*	0.0157	*0.0531*
	$\hat{f}_1$		*0.0172*		*0.0142*		*0.5355*		*0.0172*		*0.0172*		*0.0142*		*0.0141*		*0.0140*		*0.0140*
	$\hat{f}_2$		*0.0459*		*0.0343*		*0.2885*		*0.0459*		*0.0459*		*0.0343*		*0.0343*		*0.0344*		*0.0344*
	$\hat{f}_3$		*0.0391*		*0.0202*		*0.5384*		*0.0391*		*0.0391*		*0.0202*		*0.0202*		*0.0203*		*0.0203*
0.2	$\hat{\rho}$					-0.0022	*0.0275*	-0.0022	*0.0253*	-0.0020	*0.0255*	-0.0023	*0.0253*	-0.0020	*0.0255*	0.0511	*0.0787*	0.0215	*0.0533*
	$\hat{f}_1$		*0.0164*		*0.0182*		*0.5376*		*0.0174*		*0.0174*		*0.0154*		*0.0154*		*0.0145*		*0.0145*
	$\hat{f}_2$		*0.0460*		*0.0371*		*0.2888*		*0.0466*		*0.0466*		*0.0366*		*0.0366*		*0.0366*		*0.0366*
	$\hat{f}_3$		*0.0379*		*0.0209*		*0.5385*		*0.0393*		*0.0393*		*0.0215*		*0.0216*		*0.0221*		*0.0221*
0.6	$\hat{\rho}$					-0.0019	*0.0177*	-0.0019	*0.0169*	-0.0016	*0.0180*	-0.0020	*0.0169*	-0.0016	*0.0180*	0.0642	*0.0767*	0.0288	*0.0462*
	$\hat{f}_1$		*0.0375*		*0.0605*		*0.5448*		*0.0187*		*0.0187*		*0.0193*		*0.0193*		*0.0155*		*0.0155*
	$\hat{f}_2$		*0.0533*		*0.0615*		*0.2896*		*0.0494*		*0.0495*		*0.0438*		*0.0438*		*0.0441*		*0.0441*
	$\hat{f}_3$		*0.0443*		*0.0444*		*0.5390*		*0.0399*		*0.0399*		*0.0258*		*0.0259*		*0.0298*		*0.0298*
0.9	$\hat{\rho}$					-0.0009	*0.0060*	-0.0008	*0.0059*	-0.0007	*0.0076*	-0.0009	*0.0059*	-0.0007	*0.0076*	0.0332	*0.0365*	0.0161	*0.0213*
	$\hat{f}_1$		*0.2243*		*0.2621*		*0.5549*		*0.0226*		*0.0226*		*0.0249*		*0.0250*		*0.0183*		*0.0183*
	$\hat{f}_2$		*0.2061*		*0.2188*		*0.2912*		*0.0542*		*0.0542*		*0.0532*		*0.0533*		*0.0540*		*0.0540*
	$\hat{f}_3$		*0.2319*		*0.2453*		*0.5396*		*0.0421*		*0.0421*		*0.0316*		*0.0317*		*0.0389*		*0.0389*

Monte Carlo Results: E1 experiments

Table 5

Result for experiments E1, nonlinear SAR, SNR=0.9 n=2000

		gamboost		gam		lagsarlm		BSPA_SAR_ML		BSPA_SAR_CFE		GAM_SAR_ML		GAM_SAR_CFE		GAM_SAR_FIVA		GAM_SAR_2SLSA	
ρ	terms	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
0.0	$\hat{\rho}$					-0.0020	*0.0306*	-0.0021	*0.0276*	-0.0021	*0.0276*	-0.0022	*0.0276*	-0.0021	*0.0276*	0.0401	*0.0749*	0.0157	*0.0531*
	$\hat{f}_1$		*0.0172*		*0.0142*		*0.5355*		*0.0172*		*0.0172*		*0.0142*		*0.0141*		*0.0140*		*0.0140*
	$\hat{f}_2$		*0.0459*		*0.0343*		*0.2885*		*0.0459*		*0.0459*		*0.0343*		*0.0343*		*0.0344*		*0.0344*
	$\hat{f}_3$		*0.0391*		*0.0202*		*0.5384*		*0.0391*		*0.0391*		*0.0202*		*0.0202*		*0.0203*		*0.0203*
0.2	$\hat{\rho}$					-0.0022	*0.0275*	-0.0022	*0.0253*	-0.0020	*0.0255*	-0.0023	*0.0253*	-0.0020	*0.0255*	0.0511	*0.0787*	0.0215	*0.0533*
	$\hat{f}_1$		*0.0164*		*0.0182*		*0.5376*		*0.0174*		*0.0174*		*0.0154*		*0.0154*		*0.0145*		*0.0145*
	$\hat{f}_2$		*0.0460*		*0.0371*		*0.2888*		*0.0466*		*0.0466*		*0.0366*		*0.0366*		*0.0366*		*0.0366*
	$\hat{f}_3$		*0.0379*		*0.0209*		*0.5385*		*0.0393*		*0.0393*		*0.0215*		*0.0216*		*0.0221*		*0.0221*
0.6	$\hat{\rho}$					-0.0019	*0.0177*	-0.0019	*0.0169*	-0.0016	*0.0180*	-0.0020	*0.0169*	-0.0016	*0.0180*	0.0642	*0.0767*	0.0288	*0.0462*
	$\hat{f}_1$		*0.0375*		*0.0605*		*0.5448*		*0.0187*		*0.0187*		*0.0193*		*0.0193*		*0.0155*		*0.0155*
	$\hat{f}_2$		*0.0533*		*0.0615*		*0.2896*		*0.0494*		*0.0495*		*0.0438*		*0.0438*		*0.0441*		*0.0441*
	$\hat{f}_3$		*0.0443*		*0.0444*		*0.5390*		*0.0399*		*0.0399*		*0.0258*		*0.0259*		*0.0298*		*0.0298*
0.9	$\hat{\rho}$					-0.0009	*0.0060*	-0.0008	*0.0059*	-0.0007	*0.0076*	-0.0009	*0.0059*	-0.0007	*0.0076*	0.0332	*0.0365*	0.0161	*0.0213*
	$\hat{f}_1$		*0.2243*		*0.2621*		*0.5549*		*0.0226*		*0.0226*		*0.0249*		*0.0250*		*0.0183*		*0.0183*
	$\hat{f}_2$		*0.2061*		*0.2188*		*0.2912*		*0.0542*		*0.0542*		*0.0532*		*0.0533*		*0.0540*		*0.0540*
	$\hat{f}_3$		*0.2319*		*0.2453*		*0.5396*		*0.0421*		*0.0421*		*0.0316*		*0.0317*		*0.0389*		*0.0389*

Monte Carlo Results

Table 5

Result for experiments E1, nonlinear SAR, SNR=0.9 n=2000

		gamboost		gam		lagsarlm		BSPA_SAR_ML		BSPA_SAR_CFE		GAM_SAR_ML		GAM_SAR_CFE		GAM_SAR_FIVA		GAM_SAR_2SLSA	
ρ	terms	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
0.0	$\hat{\rho}$					-0.0020	*0.0306*	-0.0021	*0.0276*	-0.0021	*0.0276*	-0.0022	*0.0276*	-0.0021	*0.0276*	0.0401	*0.0749*	0.0157	*0.0531*
	$\hat{f}_1$		*0.0172*		*0.0142*		*0.5355*		*0.0172*		*0.0172*		*0.0142*		*0.0141*		*0.0140*		*0.0140*
	$\hat{f}_2$		*0.0459*		*0.0343*		*0.2885*		*0.0459*		*0.0459*		*0.0343*		*0.0343*		*0.0344*		*0.0344*
	$\hat{f}_3$		*0.0391*		*0.0202*		*0.5384*		*0.0391*		*0.0391*		*0.0202*		*0.0202*		*0.0203*		*0.0203*
0.2	$\hat{\rho}$					-0.0022	*0.0275*	-0.0022	*0.0253*	-0.0020	*0.0255*	-0.0023	*0.0253*	-0.0020	*0.0255*	0.0511	*0.0787*	0.0215	*0.0533*
	$\hat{f}_1$		*0.0164*		*0.0182*		*0.5376*		*0.0174*		*0.0174*		*0.0154*		*0.0154*		*0.0145*		*0.0145*
	$\hat{f}_2$		*0.0460*		*0.0371*		*0.2888*		*0.0466*		*0.0466*		*0.0366*		*0.0366*		*0.0366*		*0.0366*
	$\hat{f}_3$		*0.0379*		*0.0209*		*0.5385*		*0.0393*		*0.0393*		*0.0215*		*0.0216*		*0.0221*		*0.0221*
0.6	$\hat{\rho}$					-0.0019	*0.0177*	-0.0019	*0.0169*	-0.0016	*0.0180*	-0.0020	*0.0169*	-0.0016	*0.0180*	0.0642	*0.0767*	0.0288	*0.0462*
	$\hat{f}_1$		*0.0375*		*0.0605*		*0.5448*		*0.0187*		*0.0187*		*0.0193*		*0.0193*		*0.0155*		*0.0155*
	$\hat{f}_2$		*0.0533*		*0.0615*		*0.2896*		*0.0494*		*0.0495*		*0.0438*		*0.0438*		*0.0441*		*0.0441*
	$\hat{f}_3$		*0.0443*		*0.0444*		*0.5390*		*0.0399*		*0.0399*		*0.0258*		*0.0259*		*0.0298*		*0.0298*
0.9	$\hat{\rho}$					-0.0009	*0.0060*	-0.0008	*0.0059*	-0.0007	*0.0076*	-0.0009	*0.0059*	-0.0007	*0.0076*	0.0332	*0.0365*	0.0161	*0.0213*
	$\hat{f}_1$		*0.2243*		*0.2621*		*0.5549*		*0.0226*		*0.0226*		*0.0249*		*0.0250*		*0.0183*		*0.0183*
	$\hat{f}_2$		*0.2061*		*0.2188*		*0.2912*		*0.0542*		*0.0542*		*0.0532*		*0.0533*		*0.0540*		*0.0540*
	$\hat{f}_3$		*0.2319*		*0.2453*		*0.5396*		*0.0421*		*0.0421*		*0.0316*		*0.0317*		*0.0389*		*0.0389*

Monte Carlo Results: E1 experiments

In short for the first set of experiments,

- All of our estimators behave as consistent estimators,
- when $\rho > 0$, $\hat{\rho}$ is unbiased as well as non linear terms,
- when $\rho > 0$, the RMSE of non linear terms are slightly higher than their reference value without spatial dependence,
- CFE versions provide results almost identical to ML, and are fatest,
- FIVA version are more biased and very slow (several non linear models with J*4 variables to compute).

Monte Carlo Results: E2 experiments

In short for the second set of experiments, which introduces spatially autocorrelated covariates and unobserved spatial heterogeneity in the DGP:

The spatial autocorrelation of X_2 leads to a slight increase of the RMSEs of non linear terms across all estimators, even when there is no spatial dependence

The other form of spatial heterogeneity (unobserved spatial heterogeneity, spatial trend in $f(X_2)$) introduced in DGP have no consequence.

Monte Carlo Results: E3 experiments

In short, for the third set of experiments, in which we add unnecessary, possibly correlated covariates in the set of candidate variables:

- no effect on the estimation of ρ.
- The RMSEs of the non-linear terms are not affected as long as the unnecessary covariates are not correlated with the true variables of the DGP.

Some erroneous variables may be selected very rarely by the boosting algorithm , but this has no impact on the performance of the estimator regarding the estimates for the true variables.

Computing time

Average Computation time without paralellization, in seconds (MacBook Pro, 2.9 GHz Intel Core i9, 32 Go RAM)

n	k	gamboost	gam	lagsarlm	BSPA_SAR_ML	GAM_SAR_ML	BSPA_SAR_CFE	GAM_SAR_CFE	GAM_SAR_FIVA	GAM_SAR_2SLSA
1000	3	0.826	0.056	7.895	10.324	0.593	3.547	0.136	1.581	0.052
5000	3	1.499	0.129	1.090	18.582	0.849	5.693	0.197	6.406	0.102
10000	3	2.784	0.101	1.977	30.288	1.415	8.909	0.264	9.161	0.170
50000	3	10.930	0.313	8.633	131.356	5.119	45.061	0.938	29.671	0.605
250000	3	48.946	1.601	52.581	642.062	34.991	201.379	4.180	143.569	3.193
1000	10	2.455	0.476	7.890	34.703	6.195	10.041	1.065	95.525	0.450
5000	10	4.173	0.674	1.856	65.755	8.533	17.289	1.432	95.992	0.691
10000	10	6.709	0.677	3.312	97.239	10.016	25.342	1.510	80.806	0.894
1000	20	4.848	5.585	6.835	66.976	86.177	18.505	12.519	958.461	4.739
5000	20	8.374	6.437	3.860	125.432	76.284	34.229	11.743	1066.092	4.636
10000	20	12.455	5.899	6.678	192.953	74.235	48.912	12.752	925.779	5.720

Empirical study with house sales data

n=39673, 40 candidate covariates selected by stepAIC (forward+backward) for estimators without boosting.

	RMSE		MAPE		R2	
MODELS	IN	OUT	IN	OUT	IN	OUT
OLS	74836.17	66784.04	39.93 %	39.3 %	0.36	0.5
LMBOOST	73310.36	66348.64	42.53 %	41.9 %	0.39	0.5
GWR	57248.57	61860.45	32.27 %	34.64 %	0.65	0.58
SAR	68220.37	67591.72	36.18 %	43.23 %	0.47	0.48
GAM	56576.87	55603.84	32.56 %	32.74 %	0.64	0.64
GAM geoadditive	55472.67	54998.71	30.87 %	31.46 %	0.66	0.66
MGWRSAR (0,kc,kv)	56401.88	59412.69	31.51%	32.82 %	0.66	0.61
XGBOOST	49128.19	54064.72	26.26 %	30.94 %	0.73	0.66
XGBOOST geoadditive	51231.36	54197.7	27.31 %	**30.8 %**	0.71	0.66
BSPA_SAR_ML geoadditive	53875.23	54323.52	29.57 %	**30.8 %**	0.68	0.66
GAM_SAR_ML geoadditive	54943.05	**54173.09**	30.59 %	30.96 %	0.67	**0.67**

Table 17
In-sample and out-sample accuracy for Vaucluse district

Conclusion

Our boosting estimator, allow to obtain estimates that are:

- better than the linear SAR and SEM estimators,
- better or as well as other nonlinear semi-parametric geoadditive estimators without spatial dependence based on boosting.
- and are fast enough to be used on large samples

Our next research efforts will focus:

- on the use of boosting to evaluate a spatial weight matrix with a spatially varying weighting scheme (see Kostov 2010, 2013))
- The development of a boosting SAR Probit version, building on Approximate likelihood estimation (Martinetti and Geniaux, 2017) (it's already functional and being tested at the moment).

A 2SLS augmented approach instead of FIVA

Instruments:

$$Z = [\tilde{X}, W\tilde{X}, WW\tilde{X}, WWW\tilde{X}]$$

where $\tilde{X} = [X, X^2, X^3]$

Linear First Stage:

$$Wy = \theta Z + u$$

Second Stage:

$$y = \sum_j h_j(X_j) + \hat{W}y + \hat{u} + \epsilon$$