Model Selection for Spatially Correlated Data

Andrés Riquelme*

May 30, 2017

Abstract

The aim of this paper is to compare several model selection techniques in detrending the error process under spatially dependent data with correlated regressors. Among the available model selection methods I compare the Adaptive Lasso (ALASSO) estimator (Zou, 2006), the smoothly clipped absolute deviation (SCAD) estimator (Fan and Li, 2001), the least absolute shrinkage and selection (LASSO) operator (Tibshirani, 1996) and elastic net estimator (Zou and Hastie, 2005). Among the various error processes I focus in fitting the Exponential covariance functions for the variogram estimation.

I found that the adaptive lasso is the best among the analyzed techniques.

This work is supported by the Chilean Science and Technology Fund (FONDECYT) under grant Number 11160948.

Keywords and phrases: Shrinkage, Spatially dependent data, lasso variogram.

JEL classification: C52, C53, C63.

*Universidad de Talca. Email: juriquelme@ncsu.edu.

1 Introduction

A challenging task for the applied researcher is to fit a variogram model that generates an accurate representation of the true spatial data generation process. Several authors have proposed and applied different methods such as Ordinary Least Squares (OLS), Weighted OLS (Cressie, 1985), Nonlinear Least Squares (NLLS), Weighted NLLS (Anderson et al., 2005), Generalized NLLS (Genton, 1998), Maximum Likelihood (Marchant and Lark, 2007) and Method of Moments estimators (Lark, 2000) among others. All of this techniques have several advantages (such as the inclusion of information of the covariance structure) and disadvantages (such as the high computational requirements involved.)

The richness of the new geo–referenced data sets pose a new challenge: to answer the question of which variables include in the model apart from the traditional latitude and longitude. For example, the Chilean National System of Air Quality Information provides information for several air pollutants in a per-minute frequency. If the research objective is interpolation or prediction of the air quality, the choice of an appropriate model selection technique becomes a nontrivial task. Suppose for the sake of this argument that we want to use 30 covariates in the estimation (including pollutants and environmental dynamics variables such as wind, atmospheric pressure, temperature etc.) and fit a second order polynomial with them. Then, we will end up having 497 variables in the model (including geo-localization variables and a constant term). If we want to check which candidate model gives the better prediction we will have to choose among $2^{497} - 1$ different specifications if we use a brute-force approach. Note that in this case the traditional significance test can be applied for model selection, but the method do not guarantee the accuracy of the prediction.

In this paper I address the model selection problem by using several pena-

lized regression techniques: the RIDGE regression (Hoerl and Kennard, 1970), the least absolute shrinkage and selection operator (LASSO) estimator (Tibshirani, 1996), the Adaptive Lasso (ALASSO) estimator (Zou, 2006), the smoothly clipped absolute deviation (SCAD) estimator (Fan and Li, 2001) and the elastic net (ENET) estimator (Zou and Hastie, 2005) to fit empirical variograms from simulated and actual data with a large number of covariates and different spatial correlation structures.

The results help applied researchers to know the relative performance of this available tools to tackle the problem of high dimensional modeling in several relevant spatial scenarios.

2 Literature Review

The main goal of this paper is to identify which model selection techniques perform better in selecting the correct model to estimate the error correlation process before fitting a variogram. This process is known as "trend removal". In this paper am fitting the Exponential variogram, but the methodology can be extended to Spherical, Gaussian and the Matérn covariance functions for the variogram estimation. The main contribution of paper is to find the best tool that makes (1) model selection and (2) variogram fitting in a single step.

To address the model selection problem several authors have used various approaches based on bayesian methods, like the Stochastic Search Variable Selection (George and McCulloch, 1993) or on likelihood-based information criteria (AIC, BIC, DIC, BF). For instance, Hoeting et al. (2006) derive an AIC information criteria from spatially dependent data and Huang and Chen (2007) propose a model selection criterion for geostatistical data, but with an emphasis in the kriging rather than in model selection problem itself.

A relatively new line of research has been provided by the shrinkage estimation methods via penalized estimation in the context of spatially dependent data. Wang and Zhu (2009) consider the hard thresholding (Antoniadis and Fan, 2001), the SCAD estimator (Fan and Li, 2001), the LASSO estimator (Tibshirani, 1996) and the ridge estimators (Hoerl and Kennard, 1970) for errors from a strong mixing without assuming a Gaussian process. They discuss the asymptotic properties of the selected methods and perform a simulation exercise over grids of size 6, 12 and 24 with 100 replications with covariates with fixed cross-correlation matrix, but the numbers are not large enough to support an asymptotic analysis and the increase in the sampling points is a mayor contribution of this paper. Huang et al. (2010) propose the spatial LASSO with applications to GIS model selection to get variable selection, spatial neighborhood selection, and parameter estimation in a single step. Chu et al. (2011) use penalized maximum likelihood estimation in spatial linear models with Gaussian process errors. Their work is in a different line of this one since I will focus only on regularized estimators, not in their likelihood counterpart.

There are some alternative penalized estimators that have not been analyzed, particularly the Adaptive Lasso and the Elastic Net. This techniques have the advantage of having the oracle property and that can handle correlated covariates better than other penalized estimators such as the RIDGE or the LASSO. This is an important feature because with spatially correlated processes the covariates are likely to be correlated too (both among them and spatially). Also, the LARS algorithm (Efron et al., 2004) can be used to obtain these estimators and reduce the computational burden from $2^k - 1$ to k when k covariates are used.

2.1 Spatial Regression and Penalized Regression

Consider the spatial process $\{Z(s)\colon s \in \mathbb{R}\}$ and the linear regression

$$Z(s) = X(s)'\boldsymbol{\beta} + \varepsilon(s)$$

were $X(s) = (x_1(s), \ldots, x_k(s))'$ is a $k \times 1$ vector of covariates at location s, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients and $\varepsilon(s)$ is assumed to be second order stationary random variable with mean zero. For a sample of n locations, let $\boldsymbol{s} = (s_1, \ldots, s_n)'$ and define $\boldsymbol{Z} = (Z(s_1), \ldots Z(s_n))'$ and $\boldsymbol{X} = (X(s_1), \ldots, X(s_n))'$. The OLS estimator can be obtained by minimizing the quadratic function

$$\phi(\boldsymbol{\beta}) = (\boldsymbol{Z} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Z} - \boldsymbol{X}\boldsymbol{\beta})$$

with respect to $\boldsymbol{\beta}$. Hastie et al. (2009) identify two reasons to use alternative methods to the full OLS solution: *prediction accuracy* and *interpretation*. The penalized estimators can improve on this two aspects. The general form of the penalized least squares objective function is:

$$\phi(\boldsymbol{\beta}) = (Z - X\boldsymbol{\beta})'(Z - X\boldsymbol{\beta}) + \lambda_1 p_1(\boldsymbol{\beta}) + \lambda_2 p_2(\boldsymbol{\beta}).$$

for any *penalty functions* $p_i(\cdot)$ and a *regularization parameters* λ_i, $i = \{1, 2\}$. A family of penalized least squares estimators can be indexed by the penalizing function:

$$p_q(\boldsymbol{\beta}) = \sum_{j=1}^{m} w_j \, |\beta_j|^q.$$

This penalizing function bounds the L_q-norm of the parameters models as $\sum_j |\beta_j|^q < t$ for some $t > 0$.

If $\lambda_1 = \lambda_2 = 0$ we have a regular OLS estimation. If $\lambda_1 = 0$, $\lambda_2 > 0$ and

$w_j = 1$ we get the ridge estimator. Note that neither OLS or ridge shrink any parameter to zero because they generate a concave penalty. If $\lambda_1 > 0$, $\lambda_2 = 0$ and $w_j = 1$ we have the LASSO estimator and if $\lambda_1 > 0$, $\lambda_2 = 0$ and $w_j = 1/|\hat{\beta}_j|^\gamma$, with $\hat{\beta}_j$ being a preliminary estimator of β_j we get the ALASSO estimator. If $\lambda_1, \lambda_2 > 0$ and $w_j = 1$ we get the ENET estimator. The SCAD estimator has a different functional form for the penalty function: $\lambda_2 = 0$, $w_j = 1$ and λ_1 takes different values depending on the value of β:

$$p_{\lambda_1}(\beta) = \begin{cases} \lambda |\beta|, & \text{if } |\beta| \leq \lambda \\ \lambda^2 + (\alpha - 1)^{-1} \left(\alpha\lambda |\beta| - \frac{\beta^2}{2} - \alpha\lambda^2 + \frac{\lambda^2}{2} \right), & \text{if } \lambda < |\beta| \leq \alpha\lambda \\ \frac{(\alpha+1)\lambda^2}{2}, & \text{if } |\beta| > \alpha\lambda \end{cases}$$

for some $\alpha > 0$ (Fan and Li, 2001).

Note that RIDGE, LASSO, ALASSO and ENET are biased estimators and the ALASSO has the oracle property Zou (2006). The SCAD have the oracle property and it is unbiased. Among this methods, the LASSO, ALASSO and ENET are preferred by the researchers due to the computational advantage of its estimation using the LARS algorithm, but their performance under spatially correlated data has not been tested.

2.2 Simulation Exercise

The simulated spatial domain correspond to a square grid on $\mathbb{R}^2$ of side l equal to 10, 25 and 50, which generate sample sizes n equal to 100, 625 and 2500. For the trend I consider a third order polynomial and a set of $\sqrt{n}$ additional regressors. These regressors are correlated, with a non-fixed correlation matrix $\Omega = \rho^{|i-j|}$ for the i-th row and j-th column respectively with $\rho = 0.8$. The error term follows a multivariate normal distribution with covariance matrix drawn

from an exponential variogram of the form $\gamma(h) = (\theta_1 - \theta_0)\exp(1 - h/\theta_2)$, where h is the distance between a pair of points. The range is $\theta_2 = l \times 3/5$ and the sill is $\theta_1 = \theta_2/2$. The nugget effect θ_0 is chosen to be 0 and $\theta_1/2$. The $\boldsymbol{\beta}$ coefficients are constructed in the following way: only the second order polynomial terms and the 20% most and less correlated covariates are non zero. For example, with 100 additional regressors the second order polynomial and the first 20 and the last 20 of the additional covariates will have have nonzero coefficients. All the specifications include a constant equal to -1. The number of covariates for each grid size are 20, 35 and 70 respectively and each scenario is simulated 1000 times. An example of the simulated scenarios is shown in figure 1

To select the tuning parameter for the LASSO, ALASSO, SCAD and ENET methods I use Mallow's Cp criteria as recommended by Efron et al. (2004). This criteria requires to choose only among k estimates when using the LARS algorithm. In the ALASSO estimation $\gamma = 1$ and $\hat{\beta}_j$ are OLS estimators for the penalty weights. $\alpha = 1$ in the SCAD penalty.

The empirical variograms are estimated using a maximum distance of $3/5 \times l$ and the exponential parameters are estimated using non linear least squares.

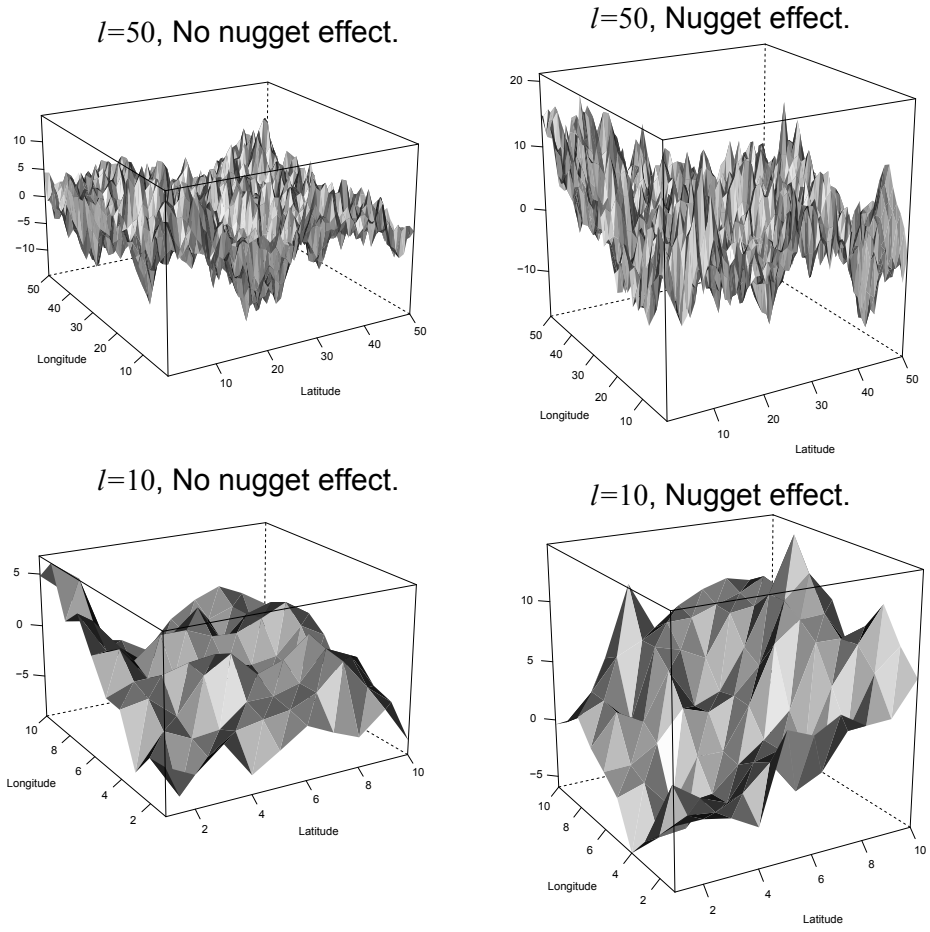

Figure 1: Example of the simulated samples for $l = \{10, 50\}$ with and without nugget effect.

3 Results

To asses the performance of the estimated methods in selecting the right covariates I compute the proportion of two relevant cases: *Correct Only* corresponds to the case were none of the nonzero coefficients has been shrunk to zero. In this case the shrinkage bias can be eliminated by using the two-step estimation proposed by Belloni et al. (2012): use a penalized regression to select the valid covariates and then use them as regressors in a simple OLS regression. The second case is *Any Incorrect* and corresponds to the alternative case in which at least one relevant covariate was not selected. The performance of the penalized estimations are presented in the tables 1 and 2 for the cases with and without nugget effect respectively, each of them for grids of side 10, 25 and 50. The average number selected of regressors is presented for reference.

From the results we can see that the ALASSO technique consistently has the best rate of selecting the valid covariates. We find results in line with Wang and Zhu (2009) in the sense that SCAD outperform LASSO, except in the smaller grid. Without the nugget effect all the methods perform well as the sample increases, except for the ENET that selects all the valid covariates with probability 0.8 with $l = 50$. The best model selection is the ALASSO. The results are similar when a nugget effect is added except by the LASSO, which performs poorly even with large sample sizes it has a probability of selecting the valid covariates of 0.4 when $l = 50$. This is not surprising given the fact that I introduced correlation between the predictors and the LASSO will tend to select only one of the highly correlated predictors. ALASSO can solve this problem (Zou, 2006).

To compare the model selection techniques in estimating the variogram parameters the mean estimates of $\theta_0 + \theta_1$ and θ_2 and their 95% confidence intervals are presented in the tables 3 and 4. These parameters were obtained using non

linear least squares from the residuals after removing the trend using the corresponding penalized estimation. The estimates from OLS and the mean squared error are presented for reference purposes.

In the setup without nugget effect the there is not a clear difference between which model selection technique performs the best. In the case of no nugget effect, when $l = 10$ the smallest bias for $\theta_0+\theta_1$ is obtained by the ENET estimator, but for θ_2 the smallest bias is attained by the OLS. When $l = 25$ the smaller bias is achieved by OLS for all the parameters and when $l = 50$ the best method is the ALASSO. With a nugget effect the techniques are not able to estimate $\theta_0 + \theta_1$ properly, although the θ_2 parameter is correctly estimated at the 95% confidence using all of the methods. Perhaps this is due to the fixed rule for selecting the maximum lag size in the empirical variogram, but further exploration of this aspect is required to have a conclusive answer.

4 Conclusion

I show that the adaptive lasso method for model selection outperforms all its studied alternatives in selecting the correct model when the data used is spatially correlated and fit using an exponenetial variogram. Further research should focus on extend the analysis from the exponential variogram to a family of errors with different correlation structure and to other model selection techniques available.

Table 1: Performance of the Penalized Regression Techniques With Spatially Correlated Data Without Nugget Effect

	Correct Only	Any Incorrect	Average Regressors
$l = 10$, $k = 20$			
LASSO	0.757	0.243	14.029
ALASSO	0.988	0.012	10.410
SCAD	0.522	0.478	16.118
ENET	0.804	0.196	13.747
$l = 25$, $k = 35$			
LASSO	0.540	0.460	28.614
ALASSO	0.997	0.003	17.921
SCAD	0.846	0.154	25.127
ENET	0.731	0.269	25.928
$l = 50$, $k = 70$			
LASSO	0.908	0.092	12.790
ALASSO	1.000	0.000	28.962
SCAD	0.930	0.070	41.404
ENET	0.790	0.210	43.370

Results obtained after 1000 replications. See page 5 for details. *Correct Only* corresponds to the case were none of the nonzero coefficients has been shrunk to zero. *Any Incorrect* and corresponds to the alternative case in which at least one relevant covariate was not selected

Table 2: Performance of the Penalized Regression Techniques With Spatially Correlated Data With Nugget Effect

	Correct Only	Any Incorrect	Average Regressors
$l = 10$, $k = 20$			
LASSO	0.681	0.319	14.496
ALASSO	0.980	0.020	10.898
SCAD	0.484	0.516	16.016
ENET	0.748	0.251	14.141
$l = 25$, $k = 35$			
LASSO	0.453	0.547	29.017
ALASSO	0.978	0.022	18.991
SCAD	0.719	0.281	26.448
ENET	0.592	0.408	26.349
$l = 50$, $k = 70$			
LASSO	0.407	0.593	50.944
ALASSO	0.993	0.007	32.571
SCAD	0.816	0.184	43.857
ENET	0.652	0.348	43.061

Results obtained after 1000 replications. See page 5 for details. *Correct Only* corresponds to the case were none of the nonzero coefficients has been shrunk to zero. *Any Incorrect* and corresponds to the alternative case in which at least one relevant covariate was not selected

Table 3: Variogram Parameter Estimation Using Non Linear Least Squares from the Residuals of each Penalized Regression Techniques Without Nugget Effect

	mse	$\theta_0 + \theta_1$	θ_2	θ_1: 95% C.I.	θ_2: 95% C.I.
$l = 10$, $k = 20$					
TRUE		3	6		
OLS	3.198	2.708	7.423	(1.62; 4.10)	(1.06; 13.89)
LASSO	3.185	2.903	5.604	(1.68; 4.59)	(1.07; 13.78)
ALASSO	3.065	2.905	5.887	(1.71; 4.41)	(1.08; 13.82)
SCAD	3.267	2.778	6.425	(1.66; 4.20)	(1.06; 13.87)
ENET	3.399	2.920	5.281	(1.72; 4.55)	(1.05; 13.79)
$l = 25$, $k = 35$					
TRUE		7.5	15		
OLS	7.733	7.472	15.537	(5.23; 11.49)	(10.67; 22.71)
LASSO	7.732	7.559	15.828	(5.24; 11.70)	(10.91; 23.22)
ALASSO	7.657	7.623	15.768	(5.30; 11.68)	(10.70; 22.76)
SCAD	7.785	7.528	15.691	(5.26; 11.67)	(10.81; 22.80)
ENET	7.827	7.598	15.963	(5.31; 11.81)	(10.92; 23.52)
$l = 50$, $k = 70$					
TRUE		15	30		
OLS	15.081	14.990	31.597	(10.92; 2.31)	(22.69; 44.87)
LASSO	15.255	15.274	31.273	(12.43; 1.88)	(27.17; 38.27)
ALASSO	15.026	15.135	31.781	(11.03; 2.33)	(22.76; 45.34)
SCAD	15.129	15.057	31.617	(10.98; 2.32)	(22.66; 45.08)
ENET	15.169	15.128	32.058	(10.99; 2.33)	(22.88; 45.69)

Results obtained after 1000 replications. See page 5 for details.

Table 4: Variogram Parameter Estimation Using Non Linear Least Squares from the Residuals of each Penalized Regression Techniques With Nugget Effect

	mse	$\theta_0 + \theta_1$	θ_2	θ_1: 95% C.I.	θ_2: 95% C.I.
$l = 10,\ k = 20$					
TRUE		4.5	6		
OLS	1.599	1.353	15.350	(0.81; 2.05)	(1.11; 34.95)
LASSO	1.592	1.439	12.058	(0.84; 2.33)	(1.11; 35.36)
ALASSO	1.534	1.444	12.597	(0.86; 2.22)	(1.10; 35.82)
SCAD	1.638	1.392	13.347	(0.83; 2.09)	(1.08; 35.47)
ENET	1.690	1.449	11.094	(0.86; 2.26)	(1.11; 34.75)
$l = 25,\ k = 35$					
TRUE		11.25	15		
OLS	3.867	3.735	15.568	(2.61; 5.75)	(10.67; 24.12)
LASSO	3.865	3.775	15.832	(2.622; 5.83)	(10.92; 24.41)
ALASSO	3.831	3.803	15.781	(2.651; 5.88)	(10.82; 24.38)
SCAD	3.892	3.760	15.708	(2.626; 5.80)	(10.91; 24.10)
ENET	3.911	3.797	15.974	(2.653; 5.90)	(10.95; 24.49)
$l = 50,\ k = 70$					
TRUE		22.5	30		
OLS	7.684	7.702	32.028	(5.57; 11.23)	(23.92; 43.82)
LASSO	7.681	7.708	32.310	(5.57; 11.20)	(23.85; 43.97)
ALASSO	7.655	7.700	32.153	(5.62; 11.38)	(23.99; 44.02)
SCAD	7.706	7.703	32.122	(5.57; 11.24)	(24.08; 43.57)
ENET	7.729	7.773	32.457	(5.59; 11.24)	(23.84; 44.13)

Results obtained after 1000 replications. See page 5 for details.

References

Anderson, E. S., J. A. Thompson, and R. E. Austin (2005) "LIDAR density and linear interpolator effects on elevation estimates," *International Journal of Remote Sensing*, Vol. 26, No. 18, pp. 3889–3900.

Antoniadis, Anestis and Jianqing Fan (2001) "Regularization of Wavelet Approximations," *Journal of the American Statistical Association*, Vol. 96, pp. 939–967.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012) "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain," *Econometrica*, Vol. 80, No. 6, pp. 2369–2429.

Chu, Tingjin, Jun Zhu, and Haonan Wang (2011) "Penalized maximum likelihood estimation and variable selection in geostatistics," *The Annals of Statistics*, Vol. 39, No. 5, pp. 2607–2625.

Cressie, Noel (1985) "Fitting variogram models by weighted least squares," *Journal of the International Association for Mathematical Geology*, Vol. 17, No. 5, pp. 563–586.

Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani (2004) "Least angle regression," *Annals of Statistics*, Vol. 32, pp. 407–499.

Fan, Jianqing and R. Li (2001) "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, Vol. 96, pp. 1348–1360.

Genton, MarcG. (1998) "Variogram Fitting by Generalized Least Squares Using an Explicit Formula for the Covariance Structure," *Mathematical Geology*, Vol. 30, No. 4, pp. 323–345.

George, Edward I. and Robert E. McCulloch (1993) "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, Vol. 88, No. 423, pp. 881–889.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*: Springer, corrected edition.

Hoerl, Arthur E. and Robert W. Kennard (1970) "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, Vol. 12, No. 1, pp. 55–67.

Hoeting, Jennifer A., Richard A. Davis, Andrew A. Merton, and Sandra E. Thompson (2006) "Model selection for geostatistical models," *Ecological Applications*, Vol. 16, pp. 87–98.

Huang, Hsin-Cheng and Chun-Shu Chen (2007) "Optimal Geostatistical Model Selection," *Journal of the American Statistical Association*, Vol. 102, pp. 1009–1024.

Huang, Hsin-Cheng, Nan-Jung Hsu, David M. Theobald, and F. Jay Breidt (2010) "Spatial Lasso With Applications to GIS Model Selection," *Journal of Computational and Graphical Statistics*, Vol. 19, No. 4, pp. 963–983.

Lark, R. M. (2000) "Estimating variograms of soil properties by the method-of-moments and maximum likelihood," *European Journal of Soil Science*, Vol. 51, No. 4, pp. 717–728.

Marchant, B.P. and R.M. Lark (2007) "Robust estimation of the variogram by residual maximum likelihood," *Geoderma*, Vol. 140, No. 1,2, pp. 62–72.

Tibshirani, R. (1996) "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society (Series B)*, Vol. 58, pp. 267–288.

Wang, Haonan and Jun Zhu (2009) "Variable selection in spatial regression via penalized least squares," *Canadian Journal of Statistics*, Vol. 37, No. 4, pp. 607–624.

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

Zou, Hui and Trevor Hastie (2005) "Regularization and variable selection via the Elastic Net," *Journal of the Royal Statistical Society, Series B*, Vol. 67, pp. 301–320.