A cure for misspecifying W: the Data Weighted Prior estimator

Esteban Fernández Vázquez (evazquez@uniovi.es)
Rosa Bernardini Papalia (rosella.bernardini@unibo.it)

Departamento: Economía Aplicada

Universidad: Universidad de Oviedo

Área Temática: Análisis económico espacial, métodos de análisis regional y econometría espacial

The traditional approach in spatial econometrics bases on the specification of a spatial weighting matrix **W**, which reflects the assumed spatial interactions between the nodes of a network. The elements of **W** are assumed to be perfectly known, but spatial models are often estimated by using an inaccurate **W** matrix with cells defined from a somewhat arbitrary choice. This paper explores how the Data Weighted Prior (DWP) estimator proposed by Golan (2001, Journal of Econometrics) can be adapted to be used in the context of discriminating between alternative potential **W** matrices to be used in spatial models. By means of numerical simulation, we evaluate the consequences of misspecifying the elements of **W** on the estimates of spatial models based on cross sectional datasets and how the use of a DWP estimator could alleviate these consequences. In particular, we pose two type of true spatial weighting matrices -dense versus sparse- and show that estimating spatial models by means of the DWP estimator alleviates the consequences of specifying a dense (sparse) **W** when the true **W** is sparse (dense).

The model

Our point of departure is a classical spatial lag model for N locations formulated as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \rho \boldsymbol{W}^{*}\boldsymbol{y} + \boldsymbol{u} \quad (1)$$

Where

$$\boldsymbol{W}^{*} = \alpha \boldsymbol{W}^{\boldsymbol{d}} + (1-\alpha)\boldsymbol{W}^{\boldsymbol{s}} \quad (2)$$

$\boldsymbol{W}^{*}$ is the spatial weighting matrix to be used to estimate our spatial lag model, where two possible options for its specification can be applied: a dense matrix as $\boldsymbol{W}^{\boldsymbol{d}}$ and a sparse matrix $\boldsymbol{W}^{\boldsymbol{s}}$. The final specification of this matrix that is going to be used in our model is given by a parameter α to be estimated. This parameter only takes two values: 0,1. $\boldsymbol{W}^{*}$ is set as $\boldsymbol{W}^{\boldsymbol{d}}$ if $\alpha = 1$, while $\boldsymbol{W}^{*}$ is set as $\boldsymbol{W}^{\boldsymbol{s}}$ when $\alpha = 0$.

Consequently, the formulation of the spatial lag model to be estimated is:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \rho \boldsymbol{W}^{*}\boldsymbol{y} + \boldsymbol{u} = \boldsymbol{X\beta} + \rho\left[\alpha \boldsymbol{W}^{\boldsymbol{d}} + (1-\alpha)\boldsymbol{W}^{\boldsymbol{s}}\right]\boldsymbol{y} + \boldsymbol{u} \quad (3)$$

Our objective is now twofold: (i) to estimate the relevant parameters of this model ($\boldsymbol{\beta}$ and ρ) and, simultaneously, (ii) to choice the bets specification of the weighting matrix given the two alternatives given (i.e., to estimate α). These goals can be achieved by applying a DWP estimator. The solution bases on minimizing the KL divergence with respects to our prior beliefs about all these parameters. For the structural parameters $\boldsymbol{\beta}$ and ρ we must set a vector that contains its $M \geq 2$ possible realizations. This information is included for the estimation by means of a supporting vector $\boldsymbol{b}_{\boldsymbol{\beta}}{}' = \left(b_{\beta 1}, \dots, b_{\beta M}\right)$ for each β_h with unknown probabilities $\boldsymbol{p}'_h = (p_{h1}, \dots, p_{hM})$. Similarly, it will be assumed that there are $L \geq 2$ possible realizations for the spatial parameter ρ in a support vector $\boldsymbol{b}_{\boldsymbol{\rho}}{}' = \left(b_{\rho 1}, \dots, b_{\rho L}\right)$, with corresponding probabilities $\boldsymbol{p}'_{\boldsymbol{\rho}} = \left(p_{\rho 1}, \dots, p_{\rho L}\right)$. Vectors $\boldsymbol{b}_{\boldsymbol{\beta}}$ and $\boldsymbol{b}_{\boldsymbol{\rho}}$ are based on the researcher's a priori belief about the likely values of the parameter. For the random term $\boldsymbol{u}$, a similar approach is followed. In contrast to other estimation techniques, DWP does not require rigid assumptions about a specific probability distribution function of the stochastic component, but it still is necessary to make some assumptions. $\boldsymbol{u}$ is assumed to have mean $E[\boldsymbol{u}] = 0$ and a finite covariance matrix. Basically, we represent our uncertainty about the realizations of vector $\boldsymbol{u}$ treating it as a discrete random variable with $J \geq 2$ possible outcomes contained in a convex set $\boldsymbol{v}' = \left\{v_1, \dots, v_J\right\}$, assuming that these possible realizations are symmetric around zero ($-v_1 = v_J$) with probabilities $\boldsymbol{p}'_{\boldsymbol{u}} = \left(p_{u1}, \dots, p_{uJ}\right)$. The traditional way of fixing the upper and lower limits of this set is to apply the three-sigma rule (see Pukelsheim, 1994).

The formulation of the DWP estimator that we present here will estimate the parameters and the errors by choosing as solution the unknown probabilities that minimize the KL divergence with respect to some a priori distributions ($\boldsymbol{q}_{\boldsymbol{\beta}}, \boldsymbol{q}_{\boldsymbol{\rho}}$ and $\boldsymbol{q}_{\boldsymbol{u}}$ respectively), being consistent at the same time with the in-sample information. The natural way of specifying these a priori distributions is to set them as uniform, which implies that all the values contained in the supporting vectors are, in principle, equally probable.

For the α parameter we follow a similar but not identical strategy in the specification of the supporting vectors and the a priori probability distributions. In particular, we set supporting vectors $\boldsymbol{b_\alpha}$ bounded by 0 and 1 as the lower and upper limits respectively. Even when we can consider more complex supporting vectors with K points, let assume, for the sake of simplicity, that we consider a supporting vector with only 2 values as $\boldsymbol{b_\alpha}' = (0,1)$. We consider two a priori probability distributions, which reflect the only two possible values that we define for α: one is an a priory probability distribution ($\boldsymbol{q_s}$) with mass probability at the lower bound of $\boldsymbol{b_\alpha}$; the second is another a priori spike distribution ($\boldsymbol{q_d}$) but now with mass probability at the upper limit of $\boldsymbol{b_\alpha}$. The DWP estimator will discriminate which one of these two alternative priors takes over. If the a priori distribution $\boldsymbol{q_s}$ is favored, we have reasons to think that the sparse weighting matrix $\boldsymbol{W^s}$ is preferable to $\boldsymbol{W^d}$; the opposite situations happens if the DWP estimators chooses $\boldsymbol{q_d}$ as the a priori probability distribution that should be taken as reference. In the first case, we estimate the value of the α parameter as 0, while the DWP estimate of α will be 1 in the second case.

Following the formulation of the DWP estimation, the objective proposed can be achieved by minimizing the following constrained problem:

$$\begin{aligned}\underset{\boldsymbol{P}}{\mathrm{Min}}\, D(\boldsymbol{P}/\boldsymbol{Q}) = \sum_{h=1}^{H} D\left(\boldsymbol{p_{\beta h}}/\boldsymbol{q_{\beta h}}\right) + D\left(\boldsymbol{p_\rho}/\boldsymbol{q_\rho}\right) + D(\boldsymbol{p_u}/\boldsymbol{q_u}) \\ +\gamma D(\boldsymbol{p_\alpha}/\boldsymbol{q_s}) + (1-\gamma) D(\boldsymbol{p_\alpha}/\boldsymbol{q_d}) \\ +D\left(\boldsymbol{p_\gamma}/\boldsymbol{q_\gamma}\right)\end{aligned} \tag{4}$$

subject to:

$$y_n = \sum_{h=1}^{H}\sum_{m=1}^{M} b_m p_{hm} x_{hnt} + \left(\sum_{l=1}^{L} p_{\rho l}\, b_{\rho l}\right) \sum_{i \neq n}^{N} w_{ni}^{*} y_i + \sum_{j=1}^{J} v_j\, u_{nj};\;\; n = 1, \dots, N \tag{5}$$

$$w_{ni}^{*} = \alpha w_{ni}^{d} + (1-\alpha) w_{ni}^{s} = \left(\sum_{k=1}^{K} p_{\alpha k}\, b_{\alpha k}\right) w_{ni}^{d} + \left(1 - \sum_{k=1}^{K} p_{\alpha k}\, b_{\alpha k}\right) w_{ni}^{s};\; \forall i \neq n \tag{6}$$

$$\gamma = \sum_{g=1}^{G} b_{\gamma g}\, p_{\gamma g} \tag{7}$$

$$\sum_{m=1}^{M} p_{hm} = 1\,;\ \forall h \tag{8}$$

$$\sum_{j=1}^{J} p_{uj} = 1\,;\ n = 1, \dots, N \tag{9}$$

$$\sum_{k=1}^{K} p_{\alpha k} = \sum_{l=1}^{L} p_{\rho l} = \sum_{g=1}^{G} p_{\gamma g} = 1 \tag{10}$$

To understand the logic of this DWP estimator an explanation on the objective function of the previous minimization program is required. The objective function contained in equation (4) is an information theory divergence measure based on the Kullback-Leibler (1951) divergences between several prior and posterior probability distributions generally denoted as $D(\boldsymbol{P}/\boldsymbol{Q})$ and it is divided into several terms. The first term $\sum_{h=1}^{H} D\left(\boldsymbol{p_{\beta h}}/\boldsymbol{q_{\beta h}}\right)$ quantifies the Kullback-Leibler divergence between the estimated and the a priori probabilities for the parameters in $\boldsymbol{\beta}$. In a similar fashion, $D\left(\boldsymbol{p_\rho}/\boldsymbol{q_\rho}\right)$ and $D(\boldsymbol{p_u}/\boldsymbol{q_u})$ are the KL divergences for the spatial parameter ρ and the noise term $\boldsymbol{u}$.

Additionally, we find the KL divergence corresponding to the α parameter in the terms $D(\boldsymbol{p_\alpha}/\boldsymbol{q_s})$ and $D(\boldsymbol{p_\alpha}/\boldsymbol{q_d})$. Note that the divergence with respect to the prior $\boldsymbol{q_s}$ that assumes a sparse $\boldsymbol{W}$ matrix is weighted by the parameter γ, while the KL divergence to the prior $\boldsymbol{q_d}$ that assumes a dense $\boldsymbol{W}$ matrix is weighted $1-\gamma$. This weighting parameter γ is bounded between 0 and 1 and is estimated simultaneously to the other elements in our spatial lag model. γ measures the weight given to the prior $\boldsymbol{q_s}$. Its estimation requires the definition of a supporting vector $\boldsymbol{b'_\gamma} = \left\{b_{\gamma 1}, \dots, b_{\gamma G}\right\}$ containing its G possible realizations. Defined as $\gamma = \sum_{g=1}^{G} b_{\gamma g}\, p_{\gamma g}$, $b_{\gamma 1} = 0$ and $b_{\gamma G} = 1$ are respectively the lower and upper bound set in $\boldsymbol{b'_\gamma}$. It is necessary, additionally, to specify an a priori probability distribution $\boldsymbol{q_\gamma}$ for the values in $\boldsymbol{b'_\gamma}$, which given the uncertainty about this parameter are usually specified as uniform. The last element in (4), finally, contains the KL divergence corresponding to parameter γ.

The constraint in (5) sets that the estimates must be consistent with the data observed in our sample of N spatial units. Equation (6) defines the elements of the spatial matrix $\boldsymbol{W}^*$ from the estimates of the parameter α, while (7) defines the estimates of the weighting parameter γ. Finally (8), (9) and (10) are regularity constraints. By solving this DWP formulation, we will get estimates for the parameters $\boldsymbol{\beta}$ and ρ of the spatial lag model and simultaneously "picks up" the preferable specification of $\boldsymbol{W}$ given the data.

Some preliminary simulation results

Table 1. Results of the numerical experiment
(N=47 Spanish NUTS3 regions; 1,000 replications)

		ML									
		Assumed W matrix: W^d (dense)			**Assumed W matrix: W^{knn} (sparse)**			**DWP**			
	True W matrix	β_0	β_1	ρ	β_0	β_1	ρ	β_0	β_1	ρ	α
$\rho = 0.25$	**W^d (dense)**	0.168 [0.082] (0.086)	0.200 [0.001] (0.001)	0.205 [0.032] (0.035)	0.399 [0.007] (0.096)	0.200 [0.001] (0.001)	0.058 [0.003] (0.039)	0.190 [0.001] (0.009)	0.207 [0.001] (0.001)	0.163 [0.001] (0.008)	0.950
	W^{knn} (sparse)	-0.726 [0.052] (0.735)	0.200 [0.001] (0.001)	0.780 [0.022] (0.302)	0.105 [0.005] (0.005)	0.200 [0.001] (0.001)	0.246 [0.002] (0.002)	0.117 [0.001] (0.001)	0.204 [0.001] (0.001)	0.220 [0.001] (0.001)	1.000
$\rho = 0.50$	**W^d (dense)**	0.182 [0.158] (0.165)	0.200 [0.001] (0.001)	0.464 [0.029] (0.030)	0.970 [0.014] (0.771)	0.200 [0.001] (0.001)	0.127 [0.002] (0.141)	0.197 [0.001] (0.001)	0.207 [0.001] (0.001)	0.436 [0.001] (0.001)	0.001
	W^{knn} (sparse)	-2.382 [0.050] (6.213)	0.200 [0.001] (0.001)	1.569 [0.010] (1.153)	0.101 [0.006] (0.006)	0.200 [0.001] (0.001)	0.499 [0.001] (0.001)	0.163 [0.001] (0.001)	0.208 [0.001] (0.001)	0.447 [0.001] (0.001)	1.000
$\rho = 0.75$	**W^d (dense)**	0.322 [0.554] (0.604)	0.200 [0.001] (0.001)	0.702 [0.026] (0.028)	2.665 [0.052] (6.629)	0.200 [0.001] (0.001)	0.201 [0.002] (0.304)	0.197 [0.001] (0.001)	0.207 [0.001] (0.001)	0.717 [0.001] (0.001)	0.001
	W^{knn} (sparse)	-6868 [0.061] (48.613)	0.200 [0.001] (0.001)	2.263 [0.003] (2.292)	0.098 [0.008] (0.008)	0.200 [0.001] (0.001)	0.750 [0.001] (0.001)	0.176 [0.001] (0.006)	0.217 [0.001] (0.001)	0.707 [0.001] (0.002)	1.000

Notes: figures in the table show, respectively, the average estimates, the empirical variances (in brackets) and the Mean Squared Errors (in parentheses) through the simulation. Values of parameters β_0 and β_1 are 0.1 and 0.2 respectively. The spatial parameter $\boldsymbol{\rho}$ takes three different values: 0.25, 0.5 and 0.75. The supporting vectors for the parameters β_0 and β_1 in the DWP estimator have been set as common in $\boldsymbol{b}' = [-\mathbf{1}, \mathbf{0}, \mathbf{1}]$; for the spatial autoregressive parameter this vector is again $\boldsymbol{b}' = [-\mathbf{1}, \mathbf{0}, \mathbf{1}]$.

Table 2. Results of the numerical experiment

(N=164 EU NUTS2 regions; 1,000 replications)

		ML									
		Assumed W matrix: W^d (dense)			Assumed W matrix: W^{knn} (sparse)			DWP			
	True W matrix	β_0	β_1	ρ	β_0	β_1	ρ	β_0	β_1	ρ	α
$\rho = 0.25$	W^d (dense)	0.136 [0.049] (0.051)	0.200 [0.001] (0.001)	0.226 [0.023] (0.023)	0.430 [0.003] (0.112)	0.199 [0.001] (0.001)	0.025 [0.001] (0.052)	0.146 [0.001] (0.002)	0.201 [0.001] (0.001)	0.213 [0.001] (0.001)	0.001
	W^{knn} (sparse)	-0.563 [0.061] (0.501)	0.203 [0.001] (0.001)	0.700 [0.028] (0.231)	0.095 [0.002] (0.002)	0.200 [0.001] (0.001)	0.254 [0.023] (0.023)	0.116 [0.001] (0.001)	0.201 [0.001] (0.001)	0.235 [0.001] (0.001)	1.000
$\rho = 0.50$	W^d (dense)	0.136 [0.114] (0.115)	0.200 [0.001] (0.001)	0.484 [0.024] (0.024)	1.068 [0.006] (0.942)	0.199 [0.001] (0.001)	0.058 [0.001] (0.196)	0.212 [0.001] (0.013)	0.202 [0.001] (0.001)	0.443 [0.001] (0.003)	0.001
	W^{knn} (sparse)	-2.321 [0.115] (5.978)	0.207 [0.001] (0.001)	1.596 [0.024] (1.226)	0.086 [0.003] (0.003)	0.200 [0.001] (0.001)	0.506 [0.001] (0.001)	0.153 [0.001] (0.003)	0.201 [0.001] (0.001)	0.471 [0.001] (0.001)	1.000
$\rho = 0.75$	W^d (dense)	0.130 [0.441] (0.442)	0.200 [0.001] (0.001)	0.743 [0.023] (0.023)	2.953 [0.025] (8.167)	0.198 [0.001] (0.001)	0.096 [0.001] (0.429)	0.211 [0.001] (0.012)	0.202 [0.001] (0.001)	0.722 [0.001] (0.001)	0.001
	W^{knn} (sparse)	-7.419 [0.311] (5.978)	0.214 [0.001] (0.001)	2.456 [0.016] (2.925)	0.069 [0.005] (0.006)	0.200 [0.001] (0.001)	0.750 [0.001] (0.001)	0.051 [0.001] (0.003)	0.216 [0.001] (0.001)	0.751 [0.001] (0.001)	0.010

Notes: figures in the table show, respectively, the average estimates, the empirical variances (in brackets) and the Mean Squared Errors (in parentheses) through the simulation. Values of parameters β_0 and β_1 are 0.1 and 0.2 respectively. The spatial parameter $\boldsymbol{\rho}$ takes three different values: 0.25, 0.5 and 0.75. The supporting vectors for the parameters β_0 and β_1 in the DWP estimator have been set as common in $\boldsymbol{b}' = [-\mathbf{1}, \mathbf{0}, \mathbf{1}]$; for the spatial autoregressive parameter this vector is again $\boldsymbol{b}' = [-\mathbf{1}, \mathbf{0}, \mathbf{1}]$.