Economics Letters 148 (2016) 87–90

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Composite marginal likelihood estimation of spatial autoregressive probit models feasible in very large samples

Pavlo Mozharovskyi [a], Jan Vogler [b,*]

[a] Agrocampus Ouest, 65 rue de Saint-Brieuc, 35000 Rennes, France

[b] Institute of Econometrics and Statistics, University of Cologne, Universitätsstr. 22a, D-50937 Cologne, Germany

HIGHLIGHTS

- We implement Composite Marginal Likelihood (CML) for spatial probit models.
- Existing CML implementations are infeasible in large samples.
- We achieve computational feasibility using sparse matrix techniques.
- We illustrate feasibility of our CML implementation through a Monte-Carlo study.

ARTICLE INFO

Article history:
Received 5 July 2016
Received in revised form
16 September 2016
Accepted 19 September 2016
Available online 1 October 2016

JEL classification:
C21
C25
C63
C87

Keywords:
Spatial probit models
Sparse matrices
Composite marginal likelihood
Partial maximum likelihood
Spatial econometrics

ABSTRACT

Composite Marginal Likelihood (CML) has become a popular approach for estimating spatial probit models. However, for spatial autoregressive specifications the existing brute-force implementations are infeasible in large samples as they rely on inverting the high-dimensional precision matrix of the latent state variable. The contribution of this paper is to provide a CML implementation that circumvents inversion of that matrix and therefore can also be applied to very large sample sizes.

1. Introduction

Composite Marginal Likelihood (CML) has become a popular technique for estimating spatial probit models. It has been introduced to the analysis of spatial binary data by Heagerty and Lele (1998) in the field of geostatistics while Bhat et al. (2010) were the first to apply CML in the area of spatial econometrics. The Partial Maximum Likelihood approach of Wang et al. (2013) is equivalent to CML. For a comprehensive overview on CML see Bhat (2014).

Relative to standard Maximum Likelihood (ML) the CML approach has the advantage of avoiding high-dimensional numerical integration. However, for spatial autoregressive specifications the existing brute-force implementations of CML rely on inverting the precision matrix of the latent state variable. Since that matrix has dimensions equal to the number of observations n brute-force CML is infeasible in large samples (Billé, 2013).

The contribution of this paper is to provide a CML implementation that avoids computation of the full inverse matrix. Instead we compute only those elements of the inverse matrix that are actually required by CML. We do so by exploiting a recursion formula derived by Takahashi et al. (1973). Sparse matrix techniques allow us to apply it even for large n with low computational costs so that

* Correspondence to: Institut für Ökonometrie und Statistik, Universität zu Köln, Universitätsstr. 22a, D-50937 Köln, Germany. Fax: +49 0 221 470 5074.
E-mail addresses: pavlo.mozharovskyi@univ-rennes1.fr (P. Mozharovskyi), vogler@statistik.uni-koeln.de (J. Vogler).

http://dx.doi.org/10.1016/j.econlet.2016.09.022
0165-1765/

CML becomes feasible in large samples. Originally the recursion has been introduced to the Bayesian spatial statistic context by Rue and Martino (2007). However, its potential for CML has not been recognized so far by the spatial econometrics literature.

2. Spatial probit models

Spatial probit models are defined as follows. Let y_i denote a binary dependent variable observed for spatial unit i:

$$y_i = \begin{cases} 1, & \text{if } y_i^* \geq 0, \\ 0, & \text{else} \end{cases} \quad i = 1, \ldots, n. \tag{1}$$

The factor y_i^* denotes a latent state variable following a linear spatial model of the form

$$y^* = m + u, \quad u \sim \mathrm{N}(0, H^{-1}), \tag{2}$$

where $y^* = (y_1^*, \ldots, y_n^*)'$ denotes the state vector with mean $m = (m_1, \ldots, m_n)'$ and $u = (u_1, \ldots, u_n)'$ is a vector of Gaussian errors with precision matrix H. For spatial autoregressive specifications

$$H = (I - \rho W)'(I - \rho W), \tag{3}$$

where I labels an $(n \times n)$ identity matrix, ρ denotes a correlation parameter and $W = (w_{ij})$ is an $(n \times n)$ matrix of spatial weights w_{ij}. In typical applications W consists of a large proportion of zero entries so that W and H are sparse matrices.

We consider two specifications for m. The first leads to the Spatial Autoregressive Lagged dependent variable (SAL) model, the second to the Spatial Autoregressive Error (SAE) model. Thus, the mean is given as

$$\text{SAL: } m = (I - \rho W)^{-1} X\beta \tag{4}$$

$$\text{SAE: } m = X\beta, \tag{5}$$

where the $(n \times \ell)$ matrix X denotes exogenous variables and β is an ℓ-dimensional vector of slope parameters.

3. Composite marginal likelihood

Evaluating the likelihood of spatial probit models amounts to numerical integration over the n-dimensional interdependent state vector u, which is computationally demanding for large n. The idea underlying CML is to avoid high-dimensional integration by dividing the spatial units into groups and then to account for dependence within each group but to ignore dependence between groups. Typically these groups are pairs of units. The objective function to be maximized by CML then results as a product of bivariate Gaussian cdfs which are amenable to standard numerical integration. Larger group sizes of three or four spatial units increase statistical efficiency but come at the cost of increased computational burden. In fact the pairwise approach has proven to be a good balance between computational and statistical efficiency (see also Bhat, 2014).

In total there are $n(n-1)/2$ possibilities to create pairs from n observations. Since spatial dependence decreases rapidly with rising distance most correlation can be captured by accounting only for pairs including nearby observations. Two spatial units i and j are classified as nearby observations if the spatial weight w_{ij} exceeds a threshold value. For further discussions including computation of standard errors as well as efficiency and robustness of CML compared to ML see Bhat (2014).

The pairwise CML implementation divides the spatial units into pairs indexed by $g = 1, \ldots, G$ with the members of each pair g denoted as $g1$ and $g2$. Furthermore define $z = Qu$ and $\nu = -Qm$, where Q is a diagonal matrix with entries $1 - 2y_i$ for $i = 1, \ldots, n$. The likelihood function is obtained as the joint probability $P(z \leq \nu)$. The CML approach maximizes the product over the marginal likelihoods associated to each pair:

$$L_{CML} = \prod_{g=1}^{G} P(z_{(g)} \leq \nu_{(g)}) = \prod_{g=1}^{G} \Phi_2(\nu_{(g)}; 0, \Sigma_g), \tag{6}$$

where the (2×1) vector $z_{(g)} = (z_{g1}, z_{g2})'$ collects the elements from z associated to pair g and $\nu_{(g)} = (\nu_{g1}, \nu_{g2})'$ denotes the corresponding upper bound. Thus $z_{(g)}$ is bivariate Gaussian with cdf $\Phi_2(\cdot)$ and covariance matrix Σ_g.

4. Computational issues

In principle Σ_g can be computed by extracting the corresponding elements from $\Sigma = Cov(z) = QH^{-1}Q'$. However, inversion of the $(n \times n)$ precision matrix H becomes prohibitively costly for large samples ($n = 50{,}000+$) so that the existing brute-force implementations of CML are infeasible. In order to overcome this problem Billé (2013) proposes the use of sparse matrix methods to implement a computationally efficient power series expansion of H^{-1}. Nevertheless this approach is just an approximation and furthermore is still costly for large n. Alternatively the sparse conjugate gradient method Smirnov (2010) suggested for pseudo-maximum likelihood estimation of a spatial random utility choice model[1] may be considered.

Fortunately, inverting H is actually not necessary since only a subset of the elements in Σ is required. According to Takahashi et al. (1973) single elements of the inverse of a positive-definite matrix can be computed by using the following recursion. Let L denote a lower triangular Cholesky factor such that $LL' = \Sigma^{-1} = Q'HQ$. Then the element in row i and column j of Σ is given by:

$$\Sigma_{ij} = \frac{\delta_{ij}}{L_{ii}^2} - \frac{1}{L_{ii}} \sum_{k=i+1}^{n} L_{ki}\Sigma_{kj}, \quad \text{where } j \geq i,\ i = n \to 1 \tag{7}$$

and $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise.[2]

For computational efficiency it is critical that L is obtained as a sparse matrix. In this case most of the terms in the sum are zero and can therefore be ignored which greatly accelerates computation time making the recursion feasible even in very large samples. Although the precision matrix $Q'HQ$ is sparse this does not translate directly into sparsity of its Cholesky factor L. However, reordering the spatial units $i = 1, \ldots, n$ according to a symmetric approximate minimum degree permutation of the rows and columns of $Q'HQ$ reduces the 'fill-in' occurring during computations resulting in a sparse L (see Rue and Martino, 2007 and Ch. 4 in LeSage and Pace, 2009).

5. Monte-Carlo study

In the following, we illustrate our computational efficient CML implementation by conducting a Monte-Carlo study. We generate data sets with $n = 50{,}000$ spatial units. The spatial weight matrix W is constructed by simulating a pair of coordinates from a uniform (0, 1) distribution for each spatial unit. We construct a row-standardized spatial weight matrix by means of the MATLAB function *fasymneighbors2* from the spatial statistics toolbox by Robert Kelley Pace (http://www.spatial-statistics.com), which – exploiting Delaunay triangulation – assigns exactly 6 neighbors to each spatial unit.

[1] In this model the error term of each state variable follows a type 1 extreme value distribution.

[2] The study by Takahashi et al. (1973) may be difficult to access. For a detailed discussion of recursion (7) we refer to Rue and Martino (2007).