cemmap

centre for microdata methods and practice

Properties of the maximum likelihood estimator in spatial autoregressive models

Grant Hillier
Federico Martellosio

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP44/13

An ESRC Research Centre

Institute for Fiscal Studies, 7 Ridgmount Street, London WC1E 7AE **Tel:** 44 (0)20 7291 4800 **Email:** info@cemmap.ac.uk **www.cemmap.ac.uk**

Properties of the Maximum Likelihood Estimator in Spatial Autoregressive Models

Grant Hillier[a] and Federico Martellosio[b,*]

[a]CeMMAP and Department of Economics, University of Southampton,
Highfield, Southampton, SO17 1BJ, UK
[b]School of Economics, University of Surrey,
Guildford, Surrey, GU2 7XH, UK

12 September 2013

Abstract

The (quasi-) maximum likelihood estimator (MLE) for the autoregressive parameter in a spatial autoregressive model cannot in general be written explicitly in terms of the data. The only known properties of the estimator have hitherto been its first-order asymptotic properties (Lee, 2004, Econometrica), derived under specific assumptions on the evolution of the spatial weights matrix involved. In this paper we show that the exact cumulative distribution function of the estimator can, under mild assumptions, be written down explicitly. A number of immediate consequences of the main result are discussed, and several examples of theoretical and practical interest are analyzed in detail. The examples are of interest in their own right, but also serve to illustrate some unexpected features of the distribution of the MLE. In particular, we show that the distribution of the MLE may not be supported on the entire parameter space, and may be nonanalytic at some points in its support.

Keywords: spatial autoregression, maximum likelihood estimation, group interaction, networks, complete bipartite graph.
JEL Classification: C12, C21.

*Corresponding author. Tel: +44 (0) 1483 683473
E-mail addresses: ghh@soton.ac.uk (G. Hillier), f.martellosio@surrey.ac.uk (F. Martellosio)

1 Introduction

Spatial autoregressive processes have enjoyed considerable recent popularity in modelling cross-sectional data in economics and in several other disciplines, among which are geography, regional science, and politics.[1] In most applications, such models are based on a fixed spatial weights matrix W whose elements reflect the modeler's assumptions about the pairwise interactions between the observational units. An autoregressive parameter λ measures the strength of this cross-sectional interaction. This paper is concerned with the *exact* properties of the (quasi-)maximum likelihood estimator (MLE) for this parameter that is implied by assuming a Gaussian likelihood.

The particular class of spatial autoregressive models we discuss have the form

$$y = \lambda W y + X\beta + \varepsilon, \tag{1.1}$$

where y is the $n \times 1$ vector of observed random variables, and X is a fixed $n \times k$ matrix of regressors with full column rank. Different choices for the weights matrix W produce a class of models that can have very different characteristics. We will refer to model (1.1) simply as the SAR (spatial autoregressive) model; it is also known as the spatial lag model, or as the mixed regressive, spatial autoregressive model. We refer to the model with the regression component ($X\beta$) missing as the pure SAR model. Initially we make no distributional assumptions on the error vector ε, but do assume that quasi-maximum likelihood estimation is conducted on the basis of the likelihood that would prevail if the assumption $\varepsilon \sim \mathrm{N}(0, \sigma^2 I_n)$ were added to equation (1.1). Many results we obtain do not require distribution assumptions, but we later add the Gaussianity assumption in order to obtain explicit formulae.

The parameter λ is usually of direct interest in applications. For example, in social interactions analysis measuring the strength of network effects is important to policy makers.[2] Although considerable progress has been made recently in establishing the first-order asymptotic properties of the MLE for λ in such models, there remain some compelling reasons for studying its exact properties - more so, perhaps, than usual. First, exact results reveal explicitly how the properties of the estimator depend on the characteristics of the underlying model. Second, exact results are

[1]For an introduction to spatial autoregressions see, e.g., Cliff and Ord (1973), Cressie (1993), and LeSage and Pace (2009). Empirical applications of spatial autoregressions in economics can be found in Case (1991), Besley and Case (1995), Audretsch and Feldmann (1996), Bell and Bockstael (2000), Bertrand, Luttmer and Mullainathan (2000), Topa (2001), Pinkse, Slade, and Brett (2002), Liu, Patacchini, Zenou, and Lee (2012), to name just a few.

[2]Of course, the parameter β is also of interest, but the exact properties of the MLE for β can be deduced from those of the MLE for λ, because the model reduces to a standard linear regression if λ is known.

useful for checking the accuracy of the available asymptotic results. This is important because the distribution of the estimator may (indeed, does) depend crucially on the spatial weights matrix, and on the assumptions made on how it evolves with the sample size. Until now, simulation studies have been virtually the only source of such information. Third, the exact distribution may possess important features that would be impossible to discover by asymptotic methods or Monte Carlo simulation - for example, non-differentiability, non-analyticity, or unboundedness of the density. Finally, exact results are informative when the assumptions needed to obtain asymptotic results are not plausible.

The first-order condition defining the MLE for λ is, in general, a polynomial of high degree from which no closed-form solution can be obtained. Hence, even the calculation of the MLE has been regarded as problematic in this model, let alone study of its exact properties. Ord (1975) presents a simplified procedure for maximum likelihood estimation of model (1.1). A rigorous (first-order) asymptotic analysis of the estimator was given only much later, in an influential paper by Lee (2004). Bao and Ullah (2007) provide analytical formulae for the second-order bias and mean squared error of the MLE for λ in the Gaussian pure SAR model. Bao (2013) and Yang (2013) extend such approximations to the case when exogenous regressors are included and when ε is not necessarily Gaussian. Several other papers have studied the performance of the MLE by simulation, particularly in relation to competing estimators such as the two-stage least squares (2SLS) estimator or more general GMM estimators.

The key observation that enables us to carry out an exact analysis of the MLE is that, when - as it always is in practice - the likelihood is defined only for an interval of values of λ containing the origin for which the matrix $I_n - \lambda W$ is positive definite, the profile (or concentrated) likelihood after maximizing with respect to (β, σ^2) is, under certain assumptions, *single-peaked*. This fact implies that an exact expression for the cdf of the MLE for λ can easily be written down, notwithstanding the unavailability of the MLE in closed form. This is the main result of the paper.

Starting from this fundamental result, we then present a number of exact results for the MLE that follow from it. In principle, knowledge of the cdf provides a starting point for a full exact analysis of the MLE, for arbitrary choices of W and X, and for an arbitrary distribution of ε. However, the distribution theory for the MLE is nonstandard, and, perhaps not unexpectedly, turns out to have key aspects in common with that for serial correlation coefficients (von Neumann (1941), Koopmans (1942)). In particular, the cdf can be non-analytic at certain points of its domain, and can have a different functional form in the intervals between those points. For this and other reasons, the distribution theory for the MLE that is implied by our main result is, for general (W, X), quite complicated. We give some general results of this nature in Section 4, including an explicit formula for the cdf in the pure Gaussian case that is valid for any symmetric W. But, we do not attempt a complete general analysis; that is almost certainly best accomplished on a case-by-case basis. We illustrate the usefulness of the main results by examining in detail some popular special cases of model (1.1).

It is intuitive that in model (1.1) the relationship between the matrices W and X must be important, and this will be evident at many points in the paper. The first of these is the observation that there can be (W, X) combinations that lead to an unbounded profile likelihood, and hence to non-existence of the MLE. These pathological cases, of course, we rule out. The interaction between W and X will also be seen to be fundamental in determining the properties of the MLE. A striking example of this is that the distribution of the MLE may not be supported on the entire parameter space. This result implies that the estimator cannot be uniformly consistent in such circumstances.

The rest of the paper is organized as follows. Section 2 describes the assumptions we make on the spatial weights matrix W and the parameter space for λ, and introduces some examples that will be used to illustrate the theoretical results. Section 3 discusses some key properties of the profile log-likelihood for λ. Section 4 gives the main results, along with a number of important consequences. The main results are then applied in Section 5 to the examples introduced earlier. The analysis up to this point is carried out under the assumption that the eigenvalues of W are real; the case of complex eigenvalues is discussed briefly in Section 6. Section 7 concludes by discussing generalizations and further work that our results suggest.

In many statements in the text the qualification "almost surely" (a.s.) should, strictly speaking, appear. To avoid tedious repetition we take this qualification to be understood and omit it, except in the formal statements of results. For an $n \times p$ matrix A, we denote the space spanned by the columns of A by $\mathrm{col}(A)$, and the null space of A by $\mathrm{null}(A)$. The Appendix contains proofs of the results that are not established directly in the main text.

2 Assumptions and Examples

2.1 Assumptions on the Weights Matrix

The following assumptions on W are maintained throughout the paper: (a) W is entrywise nonnegative; (b) W is non-nilpotent; (c) the diagonal entries of W are zero; (d) W is normalized so that its spectral radius is one.[3] Assumptions (a), (b), and (c) are virtually always satisfied in practical applications. Assumption (d) is automatically satisfied if W is row-stochastic; otherwise, the normalization

[3] Recall that the spectral radius of a matrix is the largest of the absolute values of its eigenvalues.

can be accomplished by rescaling, provided only that the spectral radius of W is nonzero, and this is guaranteed under Assumptions (a) and (b). We remark that Assumption (b) captures the "spatial" character of the models we wish to discuss. Given nonnegativity of W, assuming non-nilpotency is equivalent to requiring that there is no permutation of the observational units that would make W triangular, i.e., would make the autoregressive process unilateral (see Martellosio, 2011). Also, if W is nilpotent and nonnegative it can be shown that the ML and OLS estimators for λ coincide, in which case study of the MLE is straightforward.

The four assumptions above are not contentious, and will not be referred to in the statements of the formal results in the paper. Additional assumptions on the structure of W will be made from time-to-time; these will be explicitly stated in the statement of results. In particular, the main results in Section 4 are proved under the assumption that the eigenvalues of W are real. This assumption is very often satisfied in applications of the model, but some consequences of its removal will be discussed in Section 6.

Two assumptions that imply that all eigenvalues of W are real, and will be useful to simplify the results, are that W is similar to a symmetric matrix, or, more restrictively, that W is itself symmetric. The former assumption covers the common case in which W is the row-standardized version of a symmetric matrix,[4] and is equivalent to the assumption that W has real eigenvalues and is diagonalizable. An important context in which all eigenvalues of W are real is when W is the adjacency matrix of a simple graph, possibly row-standardized (a simple graph is an unweighted and undirected graph containing no loops or multiple edges).

2.2 The Parameter Space

In order for model (1.1) to uniquely determine the vector y (given $X\beta$ and ε) it is necessary and sufficient that the matrix $S_\lambda := I_n - \lambda W$ is nonsingular. Thus, the values of λ at which $\det(S_\lambda) = 0$ must be ruled out for the model to be complete, so the reciprocals of the nonzero real eigenvalues of W must be excluded as possible values for λ. This we assume throughout, but in practice the parameter space for λ is usually restricted much further, as explained next.

The normalization of the spectral radius to unity (Assumption (d) above) implies that the largest eigenvalue of W is 1.[5] We also assume that W has at least one real negative eigenvalue, and denote the smallest real eigenvalue of W by $\omega_{\min}$, the value

[4] If R is a diagonal matrix with the row sums of the symmetric matrix A on the diagonal, then the row-standardised matrix $W = R^{-1}A = R^{-1/2}(R^{-1/2}AR^{-1/2})R^{1/2}$ is similar to the symmetric matrix $R^{-1/2}AR^{-1/2}$.

[5] This follows by the Perron-Frobenius Theorem for nonnegative matrices (see, e.g., Horn and Johnson, 1985).

of which must be in $[-1, 0)$. Letting $\lambda_{\min} := \omega_{\min}^{-1}$, the interval $\Lambda := (\lambda_{\min}, 1)$, or a subset thereof, is, either implicitly or explicitly, virtually always regarded as the relevant parameter space for λ, because it is the largest interval containing the origin in which $\det(S_\lambda) \neq 0$ (see, e.g., Lee, 2004, and Kelejian and Prucha, 2010).[6] In this paper we take the parameter space to be Λ throughout. However, see footnote 8 below.

2.3 Examples

To illustrate our results the following examples will be used, chosen for their simplicity and their popularity in the literature.

Example 1 (Group Interaction Model)**.** The relationships between a group of m members, all of whom interact uniformly with each other, may be represented by a matrix whose elements are all unity except for a zero diagonal. When normalized so that its row sums are unity, such a matrix has the form

$$B_m := \frac{1}{m-1}\left(\iota_m \iota_m' - I_m\right),$$

where ι_m is the $m-$vector of ones. Suppose there are r groups, of possibly different sizes. Let u denote the number of different group sizes, and, for $i = 1, ..., u$, let r_i denote the number of groups of size m_i, with $m_1 \leq m_2 \leq .... \leq m_u$. Then $r = \sum_{i=1}^{u} r_i$, and the sample size is $n = \sum_{i=1}^{u} r_i m_i$. We refer to the SAR model with $n \times n$ spatial weights matrix

$$W = \operatorname{diag}(I_{r_i} \otimes B_{m_i}, i = 1, ..., u) \tag{2.1}$$

(a block-diagonal matrix whose diagonal blocks are the matrices B_{m_i}, each repeated r_i times) as the *Group Interaction model*. This model is popular in applications, and is often used to illustrate theoretical work by simulation (see, e.g., Baltagi, 2006, Kelejian et al., 2006, Lee, 2004, 2007). We say that the model is *balanced* if the groups are all of the same size (in which case $W = I_r \otimes B_m$), *unbalanced* otherwise. The eigenvalues of the weights matrix (2.1) are 1, with multiplicity r, and $-1/(m_i - 1)$, with multiplicity $r_i\,(m_i - 1)$, for $i = 1, ..., u$. The parameter space is therefore $\Lambda = (-(m_1 - 1), 1)$.

Example 2 (Complete Bipartite Model)**.** In a *complete bipartite graph* the n observational units are partitioned into two groups of sizes p and q, say, with all individuals

[6] If the assumption that W has at least one negative eigenvalue were not satisfied it would be appropriate to set $\lambda_{\min} = -\infty$. Note that if all eigenvalues of W are real the fact that W has at least one negative eigenvalue is implied by the assumption that $\operatorname{tr}(W) = 0$.

within a group interacting with all in the other group, but with none in their own group (e.g., Bramoullé et al., 2009, Lee et al., 2010). For $p = 1$ or $q = 1$ this corresponds to the graph known as a *star*, a particularly important case in network theory (see Estrada, 2011). The adjacency matrix of a complete bipartite graph is

$$A := \begin{bmatrix} 0_{pp} & \iota_p \iota_q' \\ \iota_q \iota_p' & 0_{qq} \end{bmatrix}.$$

The corresponding row-standardized weights matrix is

$$W = \begin{bmatrix} 0_{pp} & \frac{1}{q}\iota_p \iota_q' \\ \frac{1}{p}\iota_q \iota_p' & 0_{qq} \end{bmatrix}. \tag{2.2}$$

This is not symmetric unless $p = q$. Alternatively, A can be rescaled by its spectral radius, yielding the symmetric weights matrix

$$W = (pq)^{-\frac{1}{2}} A. \tag{2.3}$$

We refer to the SAR models with weights matrix (2.2) or (2.3), as, respectively, the *row-standardized Complete Bipartite model* and the *symmetric Complete Bipartite model*. In both cases, W has two nonzero eigenvalues (1 and -1, each with multiplicity 1), and $n-2$ zero eigenvalues, so that the parameter space is $\Lambda = (-1, 1)$.

These two examples will be used to illustrate theoretical results in Sections 3 and 4. Notice that for Group Interaction models W has full rank, while in the Complete Bipartite class it has rank 2 (the minimum possible, since we assume $\text{tr}(W) = 0$). In Section 5 we provide brief details of the properties of the MLE for λ in each case. More extensive treatment of the examples will be given elsewhere.

3 Properties of the Profile Log-Likelihood

Quasi-maximum likelihood of the parameters in model (1.1) is, we assume, based on the Gaussian log-likelihood

$$l(\beta, \sigma^2, \lambda) := -\frac{n}{2}\ln(\sigma^2) + \ln\left(\det\left(S_\lambda\right)\right) - \frac{1}{2\sigma^2}(S_\lambda y - X\beta)'(S_\lambda y - X\beta), \tag{3.1}$$

where $(\beta, \sigma^2, \lambda) \in \mathbb{R}^k \times \mathbb{R}^+ \times \Lambda$ and additive constants are omitted. After maximizing $l(\beta, \sigma^2, \lambda)$ with respect to β and σ^2 we obtain the profile, or concentrated, log-likelihood

$$l_p(\lambda) := -\frac{n}{2}\ln\left(y' S_\lambda' M_X S_\lambda y\right) + \ln\left(\det\left(S_\lambda\right)\right), \tag{3.2}$$

where $M_X := I_n - X(X'X)^{-1}X'$. This is well defined as long as $y'S'_\lambda M_X S_\lambda y \neq 0$ for all $\lambda \in \Lambda$, an event with probability 1.[7] The estimator we consider in this paper is

$$\hat{\lambda}_{\mathrm{ML}} := \arg\max_{\lambda \in \Lambda} l_p(\lambda), \tag{3.3}$$

provided that the maximum exists and is unique. This is the MLE in most common use, but of course it might not be the MLE under a different specification of the parameter space for λ.[8]

Remark 3.1. If $M_X W = 0$ the first term in equation (3.2) does not involve λ, and $\hat{\lambda}_{\mathrm{ML}}$ in that case is not a function of the data y, being identically zero for all y. This happens if and only if $\mathrm{col}(W) \subseteq \mathrm{col}(X)$.[9] However, the condition $M_X W = 0$ does not imply that λ is unidentified in the conventional sense (see the discussion of Assumptions 8′ and 9 in Lee, 2004, or Rothenberg, 1971). An example in which $M_X W = 0$ is a symmetric or row-standardized Complete Bipartite model when X includes an intercept for each of the two groups.

3.1 Existence of the MLE

Before discussing the properties of $\hat{\lambda}_{\mathrm{ML}}$ it is prudent to check that it is well-defined, i.e., that $l_p(\lambda)$ is bounded above on Λ. For the pure SAR model it is straightforward to check that this is the case. In the general model, since $l_p(\lambda)$ is continuous on the interior of Λ, it is bounded on any closed subset of Λ, but might still be unbounded at the endpoints of Λ. The following lemma shows that there are in fact combinations of X and W such that, for every $y \in \mathbb{R}^n$, $l_p(\lambda) \to +\infty$ at the extremes of Λ.

Lemma 3.2. *For any $y \in \mathbb{R}^n$,*

$$\lim_{\lambda \to 1} l_p(\lambda) = \begin{cases} +\infty, & \text{if } M_X S_1 = 0 \\ -\infty, & \text{otherwise,} \end{cases} \tag{3.4}$$

and

$$\lim_{\lambda \to \lambda_{\min}} l_p(\lambda) = \begin{cases} +\infty, & \text{if } M_X S_{\lambda_{\min}} = 0 \\ -\infty, & \text{otherwise.} \end{cases} \tag{3.5}$$

[7] The event that $y'S'_\lambda M_X S_\lambda y = 0$ has zero probability because, for any $\lambda \in \Lambda$, $\mathrm{null}\,(S'_\lambda M_X S_\lambda)$ has dimension $n - \mathrm{rank}\,(S'_\lambda M_X S_\lambda) = n - \mathrm{rank}\,(M_X) = k < n$.

[8] The unrestricted maximizer of $l_p(\lambda)$ can, in general, be anywhere on the entire real line (with the points where $\det(S_\lambda) = 0$ excluded). Some authors suggest that λ should be restricted to $(-1, 1)$ (see, e.g., Kelejian and Prucha, 2010). When $\Lambda \neq (-1, 1)$, the estimator $\bar{\lambda}_{\mathrm{ML}} := \arg\max_{\lambda \in (-1,1)} l_p(\lambda)$ is a censored version of $\hat{\lambda}_{\mathrm{ML}}$. Since $\Pr(\bar{\lambda}_{\mathrm{ML}} = -1) = \Pr(\hat{\lambda}_{\mathrm{ML}} < -1)$ and $\Pr(\bar{\lambda}_{\mathrm{ML}} < z) = \Pr(\hat{\lambda}_{\mathrm{ML}} < z)$, for any $z \in (-1, 1)$, it is clear that the properties of $\bar{\lambda}_{\mathrm{ML}}$ follow from those of $\hat{\lambda}_{\mathrm{ML}}$.

[9] Because $M_X W = 0$ if and only if $M_X W \zeta = 0$ for any $\zeta \in \mathbb{R}^n$, which is so if and only if $W\zeta \in \mathrm{col}(X)$ for any $\zeta \in \mathbb{R}^n$.

According to Lemma 3.2 the estimator (3.3) *does not exist* for any $y \in \mathbb{R}^n$ if $M_X S_1 = 0$ or $M_X S_{\lambda_{\min}} = 0$. Fortunately, the exceptional cases are unlikely. A necessary (and sufficient if W is diagonalizable) condition for $M_X S_1 = 0$ (resp., $M_X S_{\lambda_{\min}} = 0$) is that all eigenvectors of W associated with eigenvalues other than 1 (resp., $\omega_{\min}$) are in $\mathrm{col}(X)$.[10] An important model where $\hat{\lambda}_{\mathrm{ML}}$ does not exist is the balanced Group Interaction model with group-specific fixed effects. The weights matrix $W = I_r \otimes B_m$ in that model is symmetric and has two eigenspaces: $\mathrm{col}(I_r \otimes \iota_m)$, associated to the eigenvalue 1, and its orthogonal complement, associated to the eigenvalue $-1/(m-1)$. Thus, in the balanced Group Interaction model, $\hat{\lambda}_{\mathrm{ML}}$ does not exist (whatever the value of y) if and only if the model contains group fixed effects (i.e. $I_r \otimes \iota_m$ is a submatrix of X).[11]

Remark 3.3. One may wonder whether the non-existence of the MLE is equivalent to the parameter λ being unidentified in the conventional sense. But, as in the case $M_X W = 0$ discussed in Remark 3.1, this is not so. For example, a balanced Group Interaction model with $\varepsilon \sim \mathrm{N}(0, \sigma^2 I_n)$ and fixed effects is globally identified,[12] even though, as we have just pointed out, $\hat{\lambda}_{\mathrm{ML}}$ does not exist in that model. Arnold (1979) considers a model similar to a balanced Group Interaction model with fixed effects, and discusses the non-existence of the MLE.

In the rest of the paper we assume that $M_X S_1 \neq 0$ and $M_X S_{\lambda_{\min}} \neq 0$, so that $l_p(\lambda) \to -\infty$ at the extremes of Λ. This simply amounts to ruling out the pathological cases in which the MLE does not exist.

3.2 The Profile Score

The profile log-likelihood $l_p(\lambda)$ is differentiable on Λ, with first derivative given by

$$\dot{l}_p(\lambda) = n \left[\frac{y' W' M_X S_\lambda y}{y' S_\lambda' M_X S_\lambda y} - \frac{1}{n} \mathrm{tr}(G_\lambda) \right], \tag{3.6}$$

where $G_\lambda := W S_\lambda^{-1}$. This matrix plays an important role in the sequel.

[10] To see this, note that $M_X S_1 = 0$ if and only if $M_X (I - W) v = 0$ for all $v \in \mathbb{R}^n$. For v in the eigenspace of W associated to the eigenvalue 1 this is nugatory, but for v an eigenvector associated to an eigenvalue $\omega \neq 1$ (i.e., $Wv = \omega v$), it is true if and only if $v \in \mathrm{col}(X)$. The condition $v \in \mathrm{col}(X)$ is therefore necessary, and is sufficient if there are n linearly independent eigenvectors of W, which is so if and only if W is diagonalizable (see Horn and Johnson (1985), Theorem 1.3.7)). The same argument applies for $M_X S_{\lambda_{\min}}$.

[11] See Lee (2007) for a different perspective on the inferential problems in a balanced Group Interaction model with fixed effects.

[12] This can be easily verified by solving the equations $\mathrm{E}_{\theta_1}(y) = \mathrm{E}_{\theta_2}(y)$ and $\mathrm{var}_{\theta_1}(y) = \mathrm{var}_{\theta_2}(y)$, where θ_1 and θ_2 denote two values of the parameter $\theta = (\lambda, \beta', \sigma^2)'$.

Differentiability of $l_p(\lambda)$ and the fact that Λ is an open set imply that the MLE must be a root of the equation $\dot{l}_p(\lambda) = 0$. The following result establishes an important property of $l_p(\lambda)$.

Lemma 3.4. *The first-order condition defining the* MLE, $\dot{l}_p(\lambda) = 0$, *is a.s. equivalent to a polynomial equation of degree equal to the number of distinct eigenvalues of* W.

Thus, the equation $\dot{l}_p(\lambda) = 0$ has, for any W, a number of complex roots (counting multiplicities) equal to the number of distinct eigenvalues of W. Any real roots lying in Λ are candidates for $\hat{\lambda}_{\mathrm{ML}}$. Since there is no explicit algebraic solution of polynomial equations of degree higher than four, Lemma 3.4 explains why $\hat{\lambda}_{\mathrm{ML}}$ cannot in general be obtained "in closed form". In spite of this, we shall see in the next section that the cdf of $\hat{\lambda}_{\mathrm{ML}}$ can be represented explicitly. The following result is the basis of the main theorem - Theorem 1 below.

Lemma 3.5. *If all eigenvalues of* W *are real, the function* $l_p(\lambda)$ *a.s. has a single critical point in* Λ, *and that point is a maximum.*

The key to this result is the observation that, when the pathological cases referred to in Lemma 3.2 are excluded, $l_p(\lambda) \to -\infty$ at both endpoints of Λ. Since $l_p(\lambda)$ is continuous on the interior of Λ, this implies that Λ must contain *at least one real zero* of $\dot{l}_p(\lambda)$. Under the assumption that all eigenvalues of W are real there is *exactly one* such critical point in Λ. The assumption that all eigenvalues of W are real is stronger than needed for the result in Lemma 3.5, but is convenient for expository purposes, and is satisfied in many applications. We defer a discussion of the possibility of extending the result to complex eigenvalues to Section 6.

Geometrically, Lemma 3.5 says that, when all eigenvalues of W are real, *the profile log-likelihood* $l_p(\lambda)$ *is single-peaked* on Λ, with no inflection points.

Remark 3.6. It might be expected that Lemma 3.5 would also apply to the *spatial error model* $y = X\beta + u$, $u = \rho W u + \varepsilon$, a popular alternative to model (1.1). This is not the case; see Ross-Parker (1975) for a counter-example.

Remark 3.7. In many applications, W is the adjacency matrix of a (unweighted and undirected) graph. It is well known in graph theory that the number of distinct eigenvalues of an adjacency matrix is related to the degree of symmetry of the graph (see Biggs, 1993). On the other hand, in algebraic statistics the degree of the score equation is regarded as an index of algebraic complexity of ML estimation (see Drton et al., 2009). Thus Lemma 3.4 establishes a connection between the algebraic complexity of $\hat{\lambda}_{\mathrm{ML}}$ and the degree of symmetry satisfied by the graph underlying W.

3.3 Invariance Properties

Before presenting the main results we mention some general properties of the MLE for λ that can be deduced directly from the model itself, and the profile score (3.6). The reduced form equation for y,

$$y = S_\lambda^{-1} X\beta + S_\lambda^{-1}\varepsilon, \tag{3.7}$$

implies certain invariance properties for the distribution of y that have significant consequences for the properties of $\hat{\lambda}_{\mathrm{ML}}$. First, assuming ε has a distribution free of β and λ, and that $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2 I_n$, it is clear that scale transformations of y, $y \to \kappa y$, $\kappa > 0$, leave the family of densities for y invariant, inducing only transformations $(\beta, \sigma^2, \lambda) \to (\kappa\beta, \kappa^2\sigma^2, \lambda)$ on the parameters. A maximal invariant under this group of scale changes is $v := y(y'y)^{-\frac{1}{2}}$, a point on the unit sphere in n dimensions, $\mathcal{S}^{n-1}$. Since the profile score (3.6) is a function of y only through v, we conclude at once that *the distribution of* $\hat{\lambda}_{\mathrm{ML}}$ *depends on* $(\beta, \sigma^2, \lambda)$ *only through a maximal invariant in the parameter space, namely* $(\beta/\sigma, \lambda)$.

There is a second consequence of scale invariance that is even more important for the properties of $\hat{\lambda}_{\mathrm{ML}}$, namely

Proposition 3.8. *The distribution of* $\hat{\lambda}_{\mathrm{ML}}$ *induced by a particular distribution of* y *is constant on the family of distributions generated by forming scale-mixtures of the initial distribution of* y.

In particular, all results obtained under Gaussian assumptions continue to hold under scale mixtures of the Gaussian distribution for y. Thus, assuming (as we will later) a Gaussian distribution for ε is far less restrictive on the generality of the results obtained that it would usually be.

As stated above, the distribution of v, and therefore $\hat{\lambda}_{\mathrm{ML}}$, generally depends on β/σ as well as λ. In the class of pure SAR models (with β absent) the group of scale changes acts transitively on the variance σ^2, which implies that *in pure SAR models the distribution of* $\hat{\lambda}_{\mathrm{ML}}$ *is free of* σ^2, *and so depends only on* λ. In the model with regressors, under certain restrictions on (X, W), the distribution of y induced by the model is invariant under a larger group of transformations, and in that case we can prove the following

Proposition 3.9. *If* $\mathrm{col}(X)$ *is spanned by* k *eigenvectors of* W, *the distribution of* $\hat{\lambda}_{\mathrm{ML}}$ *depends only upon* λ.

The importance of this result is that the distribution of $\hat{\lambda}_{\mathrm{ML}}$ is free of nuisance parameters, making the statistic an "ideal" basis for inference on λ. The assumption that the column space of X is spanned by k eigenvectors of W is, of course, restrictive, but it certainly holds in some applications. For models in which $X = \iota_n$, for example, the condition required is simply that W is row-stochastic. We will see later (in Theorem 4) that, under Gaussian assumptions and with W symmetric, there is an almost complete analogy with the pure SAR model when $\mathrm{col}(X)$ is spanned by k eigenvectors of W.

4 Main Results

4.1 The Main Theorem and Some Immediate Consequences

Before stating the main result we introduce some further notation (essentially a modification of that used in Lee (2004)). Let $C_\lambda := G_\lambda - (\mathrm{tr}(G_\lambda)/n)I_n$. The log-likelihood derivative $\dot{l}_p(\lambda)$ in equation (3.6) can be rewritten as

$$\dot{l}_p(\lambda) = \frac{n}{2}\frac{y'S'_\lambda Q_\lambda S_\lambda y}{y'S'_\lambda M_X S_\lambda y}, \tag{4.1}$$

where

$$Q_\lambda := M_X C_\lambda + C'_\lambda M_X. \tag{4.2}$$

The key to the main result is the simple observation that single-peakedness of $l_p(\lambda)$ established in Lemma 3.5 implies that, for any $z \in \Lambda$,

$$\Pr(\hat{\lambda}_{\mathrm{ML}} \leq z) = \Pr(\dot{l}_p(z) \leq 0),$$

because the single peak of $l_p(\lambda)$ is to the left of a point $z \in \Lambda$ if and only if the slope at z is negative. Thus, we have the following explicit representation for the cdf of $\hat{\lambda}_{\mathrm{ML}}$.

Theorem 1. *If all eigenvalues of W are real, the* cdf *of $\hat{\lambda}_{\mathrm{ML}}$ at each point $z \in \Lambda$ is given by*

$$\Pr(\hat{\lambda}_{\mathrm{ML}} \leq z) = \Pr(y'S'_z Q_z S_z y \leq 0). \tag{4.3}$$

This result has both analytical and computational utility. On the computational side, equation (4.3) provides a simple way of obtaining the cdf, and hence the pdf, of $\hat{\lambda}_{\mathrm{ML}}$ numerically. Importantly, this can be accomplished *without the need to directly maximize the likelihood.* Indeed, the right hand side of equation (4.3) can be approximated very efficiently by Monte Carlo simulation, for any $z \in \Lambda$, for *any distribution of* ε, and for any choices of (W, X). Here we shall concentrate on the analytical consequences of Theorem 1. We begin by pointing out some simple but important general results that can be seen immediately from (4.3).

It is convenient to rewrite (4.3) as

$$\Pr(\hat{\lambda}_{\mathrm{ML}} \leq z) = \Pr\left(\tilde{y}' A(z,\lambda)\tilde{y} \leq 0\right), \tag{4.4}$$

where $\tilde{y} := S_\lambda y = X\beta + \varepsilon$, and

$$A(z,\lambda) := (S_z S_\lambda^{-1})' Q_z (S_z S_\lambda^{-1}).$$

The structure of the matrix $A(z,\lambda)$ is evidently crucial in determining the properties of the MLE. In particular, if $\varepsilon \sim \mathrm{N}(0, \sigma^2 I_n)$, a spectral decomposition of $A(z,\lambda)$ shows that $\tilde{y}' A(z,\lambda)\tilde{y}$ is distributed as a linear combination of independent (possibly non-central) χ^2 variates, with coefficients the distinct eigenvalues of $A(z,\lambda)$. This would be the "crudest" use of Theorem 1. However, by exploiting the special structure of $A(z,\lambda)$, and imposing some conditions on the relationship between W and X, it is possible to be much more precise. This will become clearer as we proceed.

The particular case $z = \lambda$, corresponding to $\Pr(\hat{\lambda}_{\mathrm{ML}} \leq \lambda)$, is especially important. In that case $A(\lambda,\lambda) = Q_\lambda$, so $\Pr(\hat{\lambda}_{\mathrm{ML}} \leq \lambda) = \Pr(\tilde{y}' Q_\lambda \tilde{y} \leq 0)$. Apart from providing a simple device for computing the probability of underestimating λ, it is also clear that the asymptotic behavior of $\hat{\lambda}_{\mathrm{ML}}$ is governed by that of the quadratic form $\tilde{y}' Q_\lambda \tilde{y}$.

Next, observe that, because only the sign of the quadratic form in (4.4) matters, we can divide the statistic $\tilde{y}' A(z,\lambda)\tilde{y}$ by any positive quantity, without altering the probability. Dividing by $\tilde{y}'\tilde{y}$, we immediately obtain:

Corollary 4.1. *If all eigenvalues of W are real,*

$$\Pr(\hat{\lambda}_{\mathrm{ML}} \leq z) = \Pr\left(h' A(z,\lambda) h \leq 0\right), \tag{4.5}$$

where $h := \tilde{y}/(\tilde{y}'\tilde{y})^{1/2}$ is a random vector distributed on the surface of the unit sphere in n dimensions, $\mathcal{S}^{n-1}$.

The representation (4.5) allows one to appeal to known results for quadratic forms on the sphere. In particular, with the added assumption that the distribution of ε is spherically symmetric, h is uniformly distributed on $\mathcal{S}^{n-1}$ in the pure SAR model, but in general non-uniformly distributed on $\mathcal{S}^{n-1}$ in the presence of regressors. An expression for the cdf suitable for the latter case is given in Forchini (2005), while the uniformly distributed case was dealt with in Hillier (2001). As the following result shows, the distribution of $\hat{\lambda}_{\mathrm{ML}}$ is certainly non-standard. The result follows directly from Mulholland (1965), see also Saldanha and Tomei (1996). These authors show that the distribution of $h' A(z,\lambda) h$ in (4.5) is non-analytic at the eigenvalues of $A(z;\lambda)$. The important point for us is the point 0 (on the right in (4.5)), and, since $\mathrm{rank}(A(z,\lambda)) = \mathrm{rank}(Q_z)$ for $\lambda \in \Lambda$, we have

Corollary 4.2. *If all eigenvalues of W are real, the cdf of $\hat{\lambda}_{\mathrm{ML}}$ is non-analytic at every point $z \in \Lambda$ at which* $\mathrm{rank}(Q_z) < n$.

As we will see, this property of the distribution of $\hat{\lambda}_{\mathrm{ML}}$ is not a mere curiosity: for any (W, X) there will usually be a number of points at which the cdf is non-analytic. Note too that this result does not depend on the distribution assumptions made (see Forchini (2002)). It implies that the functional form of the cdf varies with z, and we will see several examples of this later. And, in some cases these properties of the distribution persist asymptotically, the Complete Bipartite model being one example.

Before continuing we remark that the argument used to obtain Theorem 1 has implications for the relationship between $\hat{\lambda}_{\mathrm{ML}}$ and the ordinary least squares estimator, $\hat{\lambda}_{\mathrm{OLS}}$.

Proposition 4.3. *When all eigenvalues of W are real the distribution function of $\hat{\lambda}_{\mathrm{OLS}}$ is above that of $\hat{\lambda}_{\mathrm{ML}}$ for $\hat{\lambda}_{\mathrm{OLS}} < 0$, crosses it at $\hat{\lambda}_{\mathrm{OLS}} = \hat{\lambda}_{\mathrm{ML}} = 0$, and is below it for $\hat{\lambda}_{\mathrm{OLS}} > 0$.*

The proof is immediate from the fact that, when defined, $\hat{\lambda}_{\mathrm{OLS}}$ is the solution to $y'W'M_X S_\lambda y = 0$, so that $\dot{l}_p(\hat{\lambda}_{\mathrm{OLS}}) = -\mathrm{tr}(G_{\hat{\lambda}_{\mathrm{OLS}}})$, and the easily established fact that, if all the eigenvalues of W are real, $\mathrm{tr}(G_\lambda)$ has the same sign as λ.[13] The single-peaked property of $l_p(\lambda)$ means that $\hat{\lambda}_{\mathrm{OLS}} < 0$ implies $\dot{l}_p(\hat{\lambda}_{\mathrm{OLS}}) > 0$ so that $\hat{\lambda}_{\mathrm{OLS}} < \hat{\lambda}_{\mathrm{ML}}$, $\hat{\lambda}_{\mathrm{OLS}} = 0$ implies $\hat{\lambda}_{\mathrm{OLS}} = \hat{\lambda}_{\mathrm{ML}}$, and $\hat{\lambda}_{\mathrm{OLS}} > 0$ implies $\dot{l}_p(\hat{\lambda}_{\mathrm{OLS}}) < 0$ so that $\hat{\lambda}_{\mathrm{OLS}} > \hat{\lambda}_{\mathrm{ML}}$. Note particularly that Proposition 4.3 holds for any X, and any distribution of ε.[14]

Thus, for instance, $\Pr(\hat{\lambda}_{\mathrm{OLS}} < \lambda)$ is greater than (less than) $\Pr(\hat{\lambda}_{\mathrm{ML}} < \lambda)$ for any negative (positive) value of λ, and the two coincide when $\lambda = 0$. Also, the density of $\hat{\lambda}_{\mathrm{ML}}$ is necessarily above that of $\hat{\lambda}_{\mathrm{ML}}$ at the origin. We do not investigate the properties of the OLS estimator further in the present paper.

Remark 4.4. The quadratic form $y'S_z'Q_z S_z y$ can always be written as a linear combination of the three quadratic forms $y'W'M_X W y$, $y'M_X y$, and $y'W'M_X y$, and it is true that the distribution of $\hat{\lambda}_{\mathrm{ML}}$ is determined by the joint distribution of these three statistics. However, it is considerably simpler to combine them, as we have done in the statement of Theorem 1, because we can then focus attention on the properties of a single quadratic form.

[13] When all eigenvalues of W are real $d\mathrm{tr}(G_\lambda)/d\lambda = \mathrm{tr}(G_\lambda^2) > 0$, so that $\mathrm{tr}(G_\lambda)$ is monotonic increasing in λ, and $\mathrm{tr}(G_0) = 0$.

[14] The support of $\hat{\lambda}_{\mathrm{OLS}}$ can be larger than Λ, but this single-crossing property also applies for $\hat{\lambda}_{\mathrm{OLS}}$ outside Λ, where the cdf of $\hat{\lambda}_{\mathrm{ML}}$ must necessarily be either 0 or 1.

4.2 Canonical Forms

Theorem 1 permits, in principle at least, an exact analysis of the properties of $\hat{\lambda}_{\mathrm{ML}}$ for any given W and X. However, as the general results mentioned so far suggest, that research agenda would certainly be non-trivial, and it is not our major objective in the present paper. Instead, in the remainder of the paper we first discuss some further general results that are reasonably straightforward consequences of Theorem 1, and then, in Section 5, explore the detailed consequences of Theorem 1 for the examples described earlier. Before doing so we shall show that, *under the assumption that W is similar to a symmetric matrix*, we can express the quadratic form in equation (4.3) in a canonical form which helps to simplify analysis of its consequences. Recall that the condition that W is similar to a symmetric matrix is satisfied whenever W is a row-standardized version of a symmetric matrix.

To begin with, let us fix some notation. We denote by T be the number of *distinct* eigenvalues of W. If the distinct eigenvalues of W are real we denote them by, in ascending order, $\omega_1, \omega_2, ..., \omega_T$, the eigenvalue ω_t occurring with algebraic multiplicity n_t (so that $\sum_{t=1}^{T} n_t = n$). Thus, $\omega_1 = \omega_{\min}$ and $\omega_T = 1$.

If W is similar to a symmetric matrix, its eigenvalues are real, and W is diagonalizable by a real nonsingular matrix. Hence, we can write $W = HDH^{-1}$, with H a nonsingular matrix (orthogonal if W is symmetric) whose columns are the eigenvectors of W, and $D := \operatorname{diag}(\omega_t I_{n_t}, t = 1, ..., T)$. Using this decomposition, we have $C_z = HD_1H^{-1}$, and $S_z S_\lambda^{-1} = HD_2H^{-1}$, with

$$D_1 := \operatorname{diag}(\gamma_t(z) I_{n_t}, t = 1, ..., T),$$

where $\gamma_t(\lambda) := \omega_t/(1 - \lambda\omega_t) - \operatorname{tr}(G_\lambda)/n$, $t = 1, ..., T$, are the distinct eigenvalues of C_λ, and

$$D_2 := \operatorname{diag}\left(\frac{1 - z\omega_t}{1 - \lambda\omega_t} I_{n_t}, t = 1, ..., T\right).$$

We can now write the matrix of the quadratic form in (4.4) as

$$A(z, \lambda) = (H')^{-1} D_2 (D_1 M + M D_1) D_2 H^{-1}, \tag{4.6}$$

where

$$M := H' M_X H. \tag{4.7}$$

The key matrix M depends on both W and X, and it is through this matrix that the interaction between the spatial weights matrix W and the regressor matrix X manifests itself.

Next, let $M = (M_{st}; s, t = 1, ..., T)$ be the partition of M conformable with D_1 and D_2, so that the blocks $M_{st} = (M_{ts})'$ are of dimension $n_s \times n_t$. We have

$$D_2 (D_1 M + M D_1) D_2 = (d_{st} M_{st}; s, t = 1, ..., T),$$

where

$$d_{st} := \frac{(1 - z\omega_s)(1 - z\omega_t)}{(1 - \lambda\omega_s)(1 - \lambda\omega_t)} \left[\gamma_s(z) + \gamma_t(z)\right] = d_{ts}. \tag{4.8}$$

Note that the coefficients d_{st} are functions of z, λ, and W, but do not depend on X, and that $d_{tt} = -2\mathrm{tr}(G_z)/n$ for all $z \in \Lambda$ if $\omega_t = 0$. The diagonal terms d_{tt} are labelled in the same order as are the eigenvalues of W.

Writing $x := H^{-1}\tilde{y}$, and partitioning x conformably with the partition of M (so that x_t is $n_t \times 1$, for $t = 1, ..., T$), we obtain the following theorem.

Theorem 2. *(i) If W is similar to a symmetric matrix,*

$$\Pr(\hat{\lambda}_{\mathrm{ML}} \leq z) = \Pr\left(\sum_{t=1}^{T} d_{tt}(x_t' M_{tt} x_t) + 2 \sum_{s,t=1,s>t}^{T} d_{st}(x_s' M_{st} x_t) \leq 0\right). \tag{4.9}$$

(ii) If W is similar to a symmetric matrix, the bilinear terms in (4.9) all vanish if and only if the matrix $M_X W$ is symmetric. In that case,

$$\Pr(\hat{\lambda}_{\mathrm{ML}} \leq z) = \Pr\left(\sum_{t=1}^{T} d_{tt}(x_t' M_{tt} x_t) \leq 0\right). \tag{4.10}$$

(iii) If W and $M_X W$ are both symmetric (4.10) simplifies further to

$$\Pr(\hat{\lambda}_{\mathrm{ML}} \leq z) = \Pr\left(\sum_{t=1}^{T} d_{tt}(\tilde{x}_t' \tilde{x}_t) \leq 0\right), \tag{4.11}$$

where $\tilde{x}_t$ is a subvector of x_t of dimension $n_t - n_t(X)$, where $n_t(X)$ is the number of columns of X in the eigenspace associated to ω_t. The vector $\tilde{x}_t$ contains those elements of x_t that correspond to eigenvectors not in $\mathrm{col}(X)$.

Equation (4.9) provides a general canonical representation of the cdf of $\hat{\lambda}_{\mathrm{ML}}$ in terms of a linear combination of quadratic and bilinear forms in the vectors x_t. Under the additional conditions in Theorem 2 (ii) and (iii) the representation contains only quadratic forms in the x_t, and subvectors of them.

Remark 4.5. The conditions of Theorem 2 (iii) are equivalent to the condition that $\mathrm{col}(X)$ is spanned by k eigenvectors of W, so Proposition 3.9 applies to that case.

Three examples where $M_X W$ is symmetric will be met in Section 5: the Group Interaction model with constant mean, the unbalanced Group Interaction model with $X = \bigoplus_{i=1}^{r} \iota_{m_i}$ (i.e., X contains an intercept for each of the r groups, and no other regressors),[15] and the Complete Bipartite model with row-standardized W and constant mean.

Before proceeding we mention the following properties of the coefficients d_{tt}. These will play an important role in some subsequent results.

Proposition 4.6. *If all eigenvalues of W are real, we have:*

(i) Regarded as functions of z, the eigenvalues $\gamma_t(z)$ of C_z for $t = 1, ..., T$ are in increasing order for all $z \in \Lambda$ (i.e., $s > t$ implies $\gamma_s(z) > \gamma_t(z)$ for all $z \in \Lambda$). For all $z \in \Lambda, \gamma_1(z) < 0, \gamma_T(z) > 0$, and, for $t = 2, ..., T-1$, each $\gamma_t(z)$ changes sign exactly once on Λ;

(ii) For $T \geq 2$, $d_{11} < 0$ and $d_{TT} > 0$ for all $z \in \Lambda$. If $T > 2$, the coefficients d_{tt}, $t = 2, ..., T-1$, each change sign exactly once on Λ, with $d_{tt} > 0$ if $z < z_t$, $d_{tt} < 0$ if $z > z_t$, where z_t denotes the unique value of $z \in \Lambda$ at which $\gamma_t(z) = 0$.

4.3 Support of the MLE

We are now in a position to discuss another important consequence of Theorem 1: *the support of $\hat{\lambda}_{\mathrm{ML}}$ is not necessarily the entire interval* Λ (by support of $\hat{\lambda}_{\mathrm{ML}}$ we mean the set on which the density of $\hat{\lambda}_{\mathrm{ML}}$ is positive). To see this, note that the first-order condition implies that the only possible candidates as the MLE are the values of λ for which the matrix Q_λ is indefinite (see equation (4.1)). More decisively, Theorem 1 shows that if there are values of $z \in \Lambda$ for which Q_z is either positive or negative definite, those will either be impossible (i.e., $\Pr(\hat{\lambda}_{\mathrm{ML}} \leq z) = 0$), or certain (i.e., $\Pr(\hat{\lambda}_{\mathrm{ML}} \leq z) = 1$). In such cases the support of the MLE will be a (proper) subset of Λ. This cannot happen for the pure SAR model, because in that case $Q_z = (G_z + G'_z) - n^{-1}\mathrm{tr}(G_z + G'_z)I_n$, which is necessarily indefinite (since $n^{-1}\mathrm{tr}(G_z + G'_z)$ is the average of the eigenvalues of $G_z + G'_z$). But, when regressors are introduced, there can be choices for (W, X) for which the support for $\hat{\lambda}_{\mathrm{ML}}$ is restricted.

It is difficult to specify general conditions on (W, X) that lead to restricted support for $\hat{\lambda}_{\mathrm{ML}}$, but in the context of Theorem 2 (ii) the conditions that do so are straightforward, and we confine ourselves here to that case. The situation arises when some of the matrices M_{tt} in equation (4.10) vanish, and the coefficients d_{tt} of the remaining terms have a common sign for z in some subset of Λ. As in Proposition 4.6, for $t = 2, ..., T-1$, z_t is the unique point $z \in \Lambda$ at which $\gamma_t(z) = 0$.

Proposition 4.7. *Assume that W is similar to a symmetric matrix and $M_X W$ is symmetric. The support of $\hat{\lambda}_{\mathrm{ML}}$ is:*

[15] Note that here it is essential that the model is unbalanced: as we have seen in Section 3.1, the MLE does not exist in the balanced case if X includes group fixed effects.

(i) $(\lambda_{\min}, z_t)$ *if* $\operatorname{col}(X)$ *contains all eigenvectors of* W *associated to eigenvalues* ω_s *for* $s > t$*, and*

(ii) $(z_t, 1)$ *if* $\operatorname{col}(X)$ *contains all eigenvectors of* W *associated to eigenvalues* ω_s *for* $s < t$*.*

In particular, under the conditions of Proposition 4.7, $\hat{\lambda}_{\mathrm{ML}}$ cannot be positive if $\operatorname{col}(X)$ contains all eigenvectors of W associated to positive eigenvalues (and a similar statement applies with negative replaced by positive in both places). This is because, in this case, z_t in Proposition 4.7 (i) must be nonpositive, by Proposition 4.6 and the fact that $\gamma_t(0) = \omega_t$. Intuitively, the eigenvectors of W associated to positive eigenvalues capture all positive (resp. negative) spatial autocorrelation, so that the remaining autocorrelation, measured by $\hat{\lambda}_{\mathrm{ML}}$, can only be negative (resp. positive). An example of this effect arises with the row-standardized Complete Bipartite model when $X = \iota_n$, because in that case ι_n is an eigenvector of W corresponding to the eigenvalue 1. This has multiplicity 1, and is the only positive eigenvalue of W. In this model $\Lambda = (-1, 1)$, but the MLE cannot be positive, even if the true value of λ is positive. This will be discussed in Section 5.3.2.

Another interesting example for which the support of $\hat{\lambda}_{\mathrm{ML}}$ is a subset of Λ is the unbalanced Group Interaction model with group fixed effects and no other regressors, i.e., $X = \bigoplus_{i=1}^{r} \iota_{m_i}$. The columns of X span the eigenspace of W associated to $\omega_{u+1} = 1$. Hence, by Proposition 4.7, the support of $\hat{\lambda}_{\mathrm{ML}}$ is $(\lambda_{\min}, z_u)$ (see also Section 5.2.4 below).

Although beyond the scope of the present paper, the restricted support phenomenon certainly seems to demand further investigation. The conditions of Theorem 2 (ii) do not seem to be necessary for its occurrence, but we do not explore this further here. It is clear, however, that if the support of $\hat{\lambda}_{\mathrm{ML}}$ is restricted then asymptotic approximations to its distribution that are supported on the entire interval Λ are unlikely to be satisfactory. It is worth remarking, too, that this phenomenon is not confined to the MLE, but is also true of the OLS estimator, for example, under certain circumstances.

4.4 Gaussian Pure SAR Models with Symmetric W

In this section we show that the results above simplify considerably when (i) there are no regressors, (ii) W is symmetric, and (iii) ε is a scale mixture of the $\mathrm{N}(0, \sigma^2 I_n)$ distribution. The resulting model provides a fairly simple context in which to discuss the general properties of the distribution of the MLE. Bao and Ullah (2007) have given finite sample approximations to the moments of the MLE in a Gaussian pure SAR model. Our focus here is on the exact distribution of the MLE.

According to Proposition 3.8 any property of the distribution of $\hat{\lambda}_{\mathrm{ML}}$ that holds under the assumption $\varepsilon \sim \mathrm{N}(0, \sigma^2 I_n)$ continues to hold under the assumption that ε belongs to the family of scale mixtures of this density, which we denote by $\varepsilon \sim \mathrm{SMN}(0, \sigma^2 I_n)$. Note that these are spherically symmetric distributions for ε, which need not be i.i.d. Letting (here and elsewhere) χ^2_ν denote a (central) χ^2 random variable with ν degrees of freedom, Theorem 2 (iii) yields the following result

Theorem 3. *If W is symmetric and $\varepsilon \sim \mathrm{SMN}(0, \sigma^2 I_n)$, then for any pure SAR model the* cdf *of the* MLE *is given by*

$$\Pr(\hat{\lambda}_{\mathrm{ML}} \leq z) = \Pr\left(\sum_{t=1}^{T} d_{tt}\chi^2_{n_t} \leq 0\right), \tag{4.12}$$

where the $\chi^2_{n_t}$ variates are independent.

The *spectrum* of an $n \times n$ matrix is defined to be the multiset of its n eigenvalues, each eigenvalue appearing with its algebraic multiplicity. Matrices with the same spectrum are called *cospectral*. According to (4.12), the distribution of $\hat{\lambda}_{\mathrm{ML}}$, and hence all of its properties, depends on W only through its spectrum.

Corollary 4.8. *In a pure SAR model with $\varepsilon \sim \mathrm{SMN}(0, \sigma^2 I_n)$, the distribution of $\hat{\lambda}_{\mathrm{ML}}$ is constant on the set of cospectral symmetric matrices.*

One simple application of Corollary 4.8 is as follows: since the spectrum of the weights matrix (2.3) depends on p and q only through their sum n, *the distribution of $\hat{\lambda}_{\mathrm{ML}}$ is the same for any pure Gaussian symmetric Complete Bipartite model on n observational units*, regardless of the partition of n into p and q. In case p or q is 1 (i.e., the graph is a star graph), we may also consider the class of all symmetric weights matrices that are "compatible" with a star graph on n vertices (i.e., matrices having positive (i, j)-th entry if and only if (i, j) is an edge of the star graph).[16] It is a simple exercise to show that all such weights matrices have (after normalization by the spectral radius) eigenvalues 0, with multiplicity $n - 2$, and -1, 1, and hence are cospectral with the adjacency matrix of the graph. We conclude that *the distribution of $\hat{\lambda}_{\mathrm{ML}}$ is the same for any Gaussian pure SAR model with symmetric weights matrix compatible with a star graph.*

Another application of Corollary 4.8 is to (non-isomorphic, to avoid trivial cases) cospectral graphs, which are well-studied in graph theory; see, e.g., Biggs (1993). Corollary 4.8 implies that the distribution of $\hat{\lambda}_{\mathrm{ML}}$ is constant on the family of pure Gaussian SAR models with weights matrices that are the adjacency matrices of cospectral graphs.

[16] That is, W is not restricted to be the $(0, 1)$ adjacency matrix associated to the star graph, but is allowed to be any symmetric matrix compatible with that graph.

A second corollary to Theorem 3 can be deduced for matrices W with *symmetric spectrum*. The spectrum of a matrix is said to be symmetric if, whenever ω is eigenvalue, $-\omega$ is also an eigenvalue, with the same algebraic multiplicity.[17] The weights matrix of a balanced Group Interaction model with $m = 2$ is an example of this type, as is that of the Complete Bipartite model, when symmetrically normalized.[18]

Corollary 4.9. *In a pure SAR model with $\varepsilon \sim \mathrm{SMN}(0, \sigma^2 I_n)$, W symmetric, and the spectrum of W symmetric about the origin, the density of $\hat{\lambda}_{\mathrm{ML}}$ satisfies the symmetry property* $\mathrm{pdf}_{\hat{\lambda}_{\mathrm{ML}}}(z;\lambda) = \mathrm{pdf}_{\hat{\lambda}_{\mathrm{ML}}}(-z;-\lambda)$.

That is, under the stated assumptions, the density of $\hat{\lambda}_{\mathrm{ML}}$ when the value of the autoregressive parameter is λ is the reflection about the vertical axis of the density when the value of the autoregressive parameter is $-\lambda$. Note that this implies that (subject to its existence) the mean of $\hat{\lambda}_{\mathrm{ML}}$ satisfies $\mathrm{E}(\hat{\lambda}_{\mathrm{ML}};\lambda) = -\mathrm{E}(\hat{\lambda}_{\mathrm{ML}};-\lambda)$.

Theorem 3 shows that in pure SAR models with symmetric W the cdf of $\hat{\lambda}_{\mathrm{ML}}$ is induced by that of a linear combination of independent χ^2 random variables with coefficients d_{tt}. Proposition 4.6 shows that, in this representation, both positive and negative coefficients *must* occur, and that, as long as $T > 2$, the number of positive and negative coefficients varies with z – in fact changing $T - 2$ times as z increases through Λ. We may now use this fact to obtain an explicit form for the cdf of $\hat{\lambda}_{\mathrm{ML}}$ in such models.[19] Before doing so we note the special case of Corollary 4.2 that applies in this context: *In a pure SAR model with W symmetric and $T > 2$, the cdf of $\hat{\lambda}_{\mathrm{ML}}$ is non-analytic at the $T - 2$ points z_t where the $\gamma_t(z)$ change sign, and has a different functional form on each interval between those points.* This result follows immediately because in this context the d_{tt} are themselves the eigenvalues of the matrix $A(z,\lambda)$.

Now, for a fixed $z \in \Lambda$ at which none of the d_{tt} vanishes, let $T_1 = T_1(z)$ and $T_2 = T_2(z)$ denote the numbers of positive and negative terms d_{tt}, respectively, in (4.12), with the T_1 positive terms first. Let $v_1 := \sum_{t=1}^{T_1} n_t$ and $v_2 := \sum_{t=T_1+1}^{T} n_t$, with $v_1 + v_2 = n$. The numbers T_1 and T_2 vary with z, as do v_1 and v_2. Next, partition x into (x_1', x_2'), with x_i of dimension $v_i \times 1$, for $i = 1, 2$, and let A_1 be the $v_1 \times v_1$ matrix $\mathrm{diag}(d_{tt} I_{n_t}; t = 1, .., T_1)$, and A_2 the $v_2 \times v_2$ matrix $\mathrm{diag}(-d_{tt} I_{n_t}; t = T_1{+}1, .., T)$. Both matrices are diagonal with positive diagonal elements, and as z varies the dimensions of the two square matrices A_1 and A_2 necessarily vary (subject to $v_1 + v_2 = n$).

[17] Note that if W is non-negative and normalised to have largest eigenvalue 1, then $\Lambda = (-1, 1)$ when W has a symmetric spectrum.

[18] In fact, for any matrix W that is the adjacency matrix of a graph $\mathcal{G}$, it is known that the spectrum is symmetric if and only if $\mathcal{G}$ is bipartite.

[19] The cdf of the OLS estimator has exactly the same form as equation (4.12), under the same assumptions, but with the d_{tt} replaced by $\omega_t(1 - z\omega_t)/(1 - \lambda\omega_t)^2$. Again, some of these must be positive, some negative, for $z \in \Lambda$. The results to follow also hold for the OLS estimator with this modification.

Let $Q_i := x_i' A_i x_i$, for $i = 1, 2$. The statistics Q_1 and Q_2 are independent linear combinations of central χ^2 random variables with positive coefficients. From (4.12),

$$\Pr(\hat{\lambda}_{\mathrm{ML}} \le z) = \Pr(Q_1 \le Q_2) = \Pr(R \le 1), \tag{4.13}$$

where $R := Q_1/Q_2$. That is, the distribution of $\hat{\lambda}_{\mathrm{ML}}$ in symmetric Gaussian pure SAR models is determined by that of a ratio of positive linear combinations of independent χ^2 random variables *at the fixed point* $r = 1$.

Before giving the general result, notice that if $T = 2$ (i.e., W has only two distinct eigenvalues), then $T_1 = T_2 = 1$, $v_1 = n_1, v_2 = n_2, Q_1 = d_{11}\chi^2_{n_1}$, $Q_2 = d_{22}\chi^2_{n_2}$, and so from (4.13) we obtain

$$\Pr(\hat{\lambda}_{\mathrm{ML}} \le z) = \Pr\left(\mathrm{F}_{n_1,n_2} \le -\frac{n_2 d_{22}}{n_1 d_{11}}\right). \tag{4.14}$$

where F_{ν_1,ν_2} denotes a random variable with an F-distribution on (ν_1, ν_2) degrees of freedom. Thus, when $T = 2$ the cdf is remarkably simple; we will shortly see that the balanced Group Interaction model has this form. However, it is clear that *only* in the case $T = 2$ is there no point of non-analyticity.

To state the general result, let $C_j(A)$ denote the top-order zonal polynomial of order j in the eigenvalues of the matrix A (Muirhead, 1982, Chapter 7), i.e., the coefficient of θ^j in the expansion of $(\det(I_n - \theta A))^{-1/2}$. Then, the result for general T is the following consequence of Theorem 3.[20]

Corollary 4.10. *If W is symmetric and $\varepsilon \sim \mathrm{SMN}(0, \sigma^2 I_n)$, then for any pure SAR model, for z in the interior of any one of the $T-1$ intervals in Λ determined by the points of non-analyticity, z_t,*

$$\begin{aligned}\Pr(\hat{\lambda}_{\mathrm{ML}} \le z) &= \left[\det(\tau_1 A_1)\det(\tau_2 A_2)\right]^{-\frac{1}{2}} \\ &\quad\times \sum_{j,k=0}^{\infty} \frac{\left(\frac{1}{2}\right)_j \left(\frac{1}{2}\right)_k}{j!k!} C_j(\tilde{A}_1) C_k(\tilde{A}_2) \Pr\left(\mathrm{F}_{v_1+2j, v_2+2k} \le \frac{(v_2+2k)\,\tau_1}{(v_1+2j)\,\tau_2}\right),\end{aligned} \tag{4.15}$$

where $\tau_i := \mathrm{tr}(A_i^{-1})$ and $\tilde{A}_i := I_{v_i} - (\tau_i A_i)^{-1}$, for $i = 1, 2$.

Because the matrices A_1 and A_2 vary as z varies over Λ, it is probably impossible to obtain the density function of $\hat{\lambda}_{\mathrm{ML}}$ directly from equation (4.15), but we shall see in Section 5 that this problem can often be avoided by a conditioning argument.[21]

[20] It is easily confirmed that the cdf (4.15) is a bivariate *mixture* of the distributions of random variables that are conditionally, given the values of two independent non-negative integer-valued random variables J and K, say, distributed as $\mathrm{F}_{v_1+2j,v_2+2k}$. The probability $\Pr(J = j)$ is the coefficient of t^j in the expansion of $\det[(1-t)\tau_1 A_1 + t I_{v_1}]$, with a similar expression for $\Pr(K = k)$.

[21] The top-order zonal polynomials in (4.15) can be computed very efficiently by methods described recently in Hillier, Kan, and Wang (2009). This is further simplified in the present case because the distinct terms in A_1 and A_2 typically occur with multiplicity greater than one.

The introduction of regressors, or the removal of the assumption that W is symmetric, does not change the general nature of these results, see Corollary 4.2 above. A generalized version of equation (4.15) for the SAR model with arbitrary X can certainly be obtained, but would require lengthy explanation. To end this section we provide, instead, a generalization of Theorem 3 to the model with W symmetric and regressors present, but subject to a restriction on the relationship between W and X.

4.5 General Gaussian Models with Symmetric W

When W is symmetric the relatively simple representation of the cdf of $\hat{\lambda}_{\mathrm{ML}}$ in terms of a linear combination of χ^2 variates is not confined to pure SAR models. When the assumption $\varepsilon \sim \mathrm{SMN}(0, \sigma^2 I_n)$ is added, Theorem 2 (iii) becomes:

Theorem 4. *Assume that W is symmetric, $\varepsilon \sim \mathrm{SMN}(0, \sigma^2 I_n)$, and $\mathrm{col}(X)$ is spanned by k eigenvectors of W. Then the cdf of $\hat{\lambda}_{\mathrm{ML}}$ is given by*

$$\Pr(\hat{\lambda}_{\mathrm{ML}} \leq z) = \Pr\left(\sum_{t=1}^{T} d_{tt}\chi^2_{n_t - n_t(X)} \leq 0\right), \tag{4.16}$$

where the χ^2 variates involved are central, and independent, and $\chi^2_0 \equiv 0$.

It is clear here that the cdf of $\hat{\lambda}_{\mathrm{ML}}$ in equation (4.16) depends only on λ (i.e., is free of (β, σ^2)), as anticipated in Proposition 3.9. An explicit expression for the cdf analogous to that in Corollary 4.10 obviously holds, as do the other corollaries of Theorem 3 discussed above, with only minor modifications.

Remark 4.11. The convention $\chi^2_0 \equiv 0$ means that any term for which $n_t(X) = n_t$ does not appear in the sum on the right in (4.16). For example, in the Complete Bipartite model the eigenspaces associated with the eigenvalues ± 1 are both one-dimensional, so if either of these is in $\mathrm{col}(X)$ that term does not appear. Subject to the other conditions of Theorem 4 holding, the cdf is then particularly simple, involving only two independent χ^2 variates.[22]

In fact, in some models a special case of the condition used in Theorem 4 holds, in that $\mathrm{col}(X)$ is contained in a single eigenspace of W. In that case the columns of X itself are eigenvectors of W, and the condition needed automatically holds. In that case we have the following simpler form of equation (4.16): if $\mathrm{col}(X)$ is a subspace of the eigenspace associated to the eigenvalue ω_t, then

$$\Pr(\hat{\lambda}_{\mathrm{ML}} \leq z) = \Pr\left(d_{tt}\chi^2_{n_t - k} + \sum_{s=1; s\neq t}^{T} d_{ss}\chi^2_{n_s} \leq 0\right). \tag{4.17}$$

[22]If both one dimensional eigenspaces were in $\mathrm{col}(X)$ we would have $M_X W = 0$ (see Remark 3.1).

For example, in the unbalanced Group Interaction model with $X = \bigoplus_{i=1}^{r} \iota_{m_i}$ the columns of X are eigenvectors associated with the unit eigenvalue. Hence, equation (4.17) holds with $k = r$.

5 Applications

In this section we apply the general results to the examples introduced earlier. Our main purpose here is to illustrate the various aspects of the distribution of $\hat{\lambda}_{\text{ML}}$ that are unusual, but we also provide some completely new exact results for these examples, and some new asymptotic results for cases not covered by Lee's (2004) assumptions.

In Section 5.1 we consider the balanced Group Interaction Model, and then generalize the results to the possibly unbalanced case in Section 5.2. Section 5.3 is devoted to the Complete Bipartite model. For all models we consider both the pure case and the constant mean case. In the case of an unbalanced Group Interaction model we also consider the case when the mean is constant across groups, and in the case of the Complete Bipartite model we also briefly consider the case of arbitrary regressors. In stating the results we assume $\varepsilon \sim \text{SMN}(0, \sigma^2 I_n)$.[23]

5.1 The Balanced Group Interaction Model

In a balanced Group Interaction model there are r groups of size m. The weights matrix is $W = I_r \otimes B_m$, which has eigenvalues 1, with multiplicity r, and $-1/(m-1)$, with multiplicity $r(m-1)$. The parameter space is $\Lambda = (-(m-1), 1)$.

5.1.1 Zero Mean

Because the matrix (2.1) has only two distinct eigenvalues, equation (4.14) applies, giving the following strikingly simple result.

Proposition 5.1. *In the pure balanced Group Interaction model with* $\varepsilon \sim \text{SMN}(0, \sigma^2 I_n)$, *the* cdf *of* $\hat{\lambda}_{\text{ML}}$ *is, for* $z \in \Lambda$,

$$\Pr(\hat{\lambda}_{\text{ML}} \leq z) = \Pr(\text{F}_{r,r(m-1)} \leq c(z, \lambda)), \tag{5.1}$$

where

$$c(z, \lambda) := \frac{(1-\lambda)^2 (z+m-1)^2}{(1-z)^2 (\lambda+m-1)^2}.$$

[23] For the balanced Group interaction model, and the Complete Bipartite model, $\hat{\lambda}_{\text{ML}}$ is the unique root in Λ of either a quadratic or a cubic (by Lemma 3.4), and is therefore available in closed form. However, obtaining the exact distribution from such a closed form seems exceedingly difficult. Theorem 1 provides a much more convenient approach.

Taking $z = \lambda$, equation (5.1) gives $\Pr(\hat{\lambda}_{\text{ML}} \leq \lambda) = \Pr(\text{F}_{r,r(m-1)} \leq 1)$. Thus, in this model the probability of underestimating λ is independent of the true value of λ. A necessary condition for the consistency of $\hat{\lambda}_{\text{ML}}$ is clearly that $\text{F}_{r,r(m-1)} \to_p 1$, which suggests that $r \to \infty$ will be sufficient, but $m \to \infty$ may not.[24] More on the asymptotics for this model below.

Differentiating the cdf produces:

Proposition 5.2. *In the pure balanced Group Interaction model with* $\varepsilon \sim \text{SMN}(0, \sigma^2 I_n)$, *the density of* $\hat{\lambda}_{\text{ML}}$ *is, for* $z \in \Lambda$,

$$\text{pdf}_{\hat{\lambda}_{\text{ML}}}(z; \lambda) = \frac{2m\delta^{\frac{r}{2}}}{B(\frac{r}{2}, \frac{r(m-1)}{2})} \frac{(1-z)^{r(m-1)-1}(z+m-1)^{r-1}}{[(1-z)^2 + \delta(z+m-1)^2]^{\frac{rm}{2}}}, \tag{5.2}$$

where $\delta := (1-\lambda)^2/[(m-1)(\lambda+m-1)^2]$.

Figure 1 displays the density (5.2) for $\lambda = 0.5$, and for $m = 10$ and various values of r (left panel), and for $r = 10$ and various values of m (right panel). For convenience the densities are plotted for $z \in (-1, 1) \subseteq \Lambda$. It is apparent that the density is much more sensitive to r (the number of groups) than to m (the group size). Analogs of these plots for other values of λ exhibit similar characteristics.

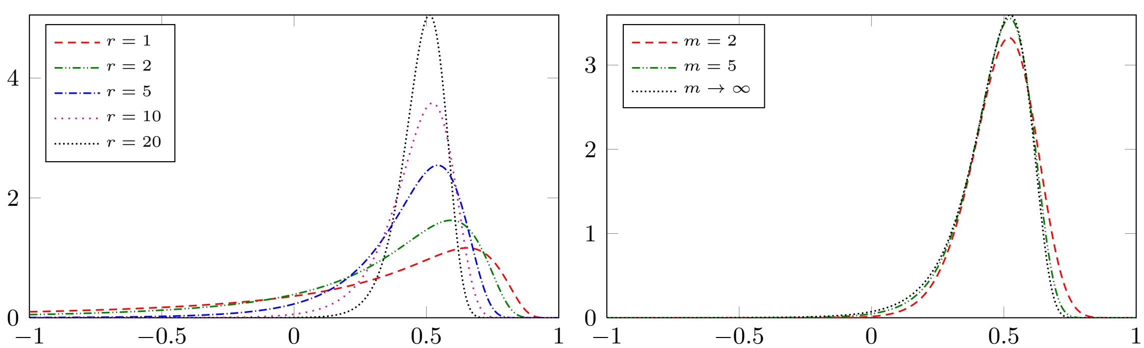

Figure 1: Density of $\hat{\lambda}_{\text{ML}}$ for the Gaussian pure balanced Group Interaction model with $\lambda = 0.5$, and with $m = 10$ (left panel), $r = 10$ (right panel).

In this model, if $r \to \infty$ is assumed, Lee's (2004) Assumptions 3 and 8' are satisfied, as is his condition (4.3), so $\hat{\lambda}_{\text{ML}}$ is consistent and asymptotically normal by Lee's Theorems 4.1 and 4.2. On the other hand, if $n \to \infty$ because $m \to \infty$ Lee's Assumption 3 is *not* satisfied, and his results leave open that $\hat{\lambda}_{\text{ML}}$ may be

[24] $\text{E}(\text{F}_{r,r(m-1)}) \to 1$ as either r or $m \to \infty$, but $\text{var}(\text{F}_{r,r(m-1)}) \to 0$ when $r \to \infty$, but not when $m \to \infty$.

inconsistent in this case. This is an example of so-called infill asymptotics. In fact, it may easily be shown (using equation (5.1) and the known result $v_1 \mathrm{F}_{v_1,v_2} \to_d \chi^2_{v_1}$ as $v_2 \to \infty$) that, for fixed r,

$$\Pr(\hat{\lambda}_{\mathrm{ML}} \le z) \overset{m\to\infty}{\longrightarrow} \Pr\left(\chi^2_r \le r\left(\frac{1-\lambda}{1-z}\right)^2\right), \quad -\infty < z < 1.$$

Thus, in fact $\hat{\lambda}_{\mathrm{ML}}$ is *inconsistent* under infill asymptotics. The associated limiting density as $m \to \infty$ with r fixed is

$$\mathrm{pdf}_{\hat{\lambda}_{\mathrm{ML}}}(z;\lambda) \overset{m\to\infty}{\longrightarrow} \frac{r^{\frac{r}{2}}(1-\lambda)^r}{2^{\frac{r}{2}-1}\Gamma(\frac{r}{2})(1-z)^{r+1}} e^{-\frac{r}{2}\left(\frac{1-\lambda}{1-z}\right)^2},$$

so $\hat{\lambda}_{\mathrm{ML}}$ converges to a random variable supported on $(-\infty, 1)$. It is clear from Figure 1 that increasing m but not r provides very little extra information on λ, at least as embodied in the MLE, and that the effective sample size under this asymptotic regime is r, and *not* $n = rm$. However, with the exact result now available, and simple, under mixed-Gaussian assumptions there is no need to invoke either form of asymptotic approximation.[25]

The exact results given in Propositions 5.1 and 5.2 enable a complete analysis of the exact properties of $\hat{\lambda}_{\mathrm{ML}}$ in this model, and the results needed for inference based upon it. For example, exact expressions for the moments and the median of $\hat{\lambda}_{\mathrm{ML}}$, and exact confidence intervals for λ based on $\hat{\lambda}_{\mathrm{ML}}$ can be obtained quite directly; these details are given in a separate paper (Hillier and Martellosio, 2013).

5.1.2 Constant Mean

The results given above for the pure balanced Group Interaction model can be extended immediately to the case of an unknown constant mean (i.e., $X = \iota_n$) by using the result in Theorem 4 (in fact the stronger version in equation (4.17)), because ι_n is in the eigenspace associated to the unit eigenvalue.

Proposition 5.3. *For the balanced Group Interaction model with $X = \iota_n$ and $\varepsilon \sim \mathrm{SMN}(0, \sigma^2 I_n)$, the* cdf *of $\hat{\lambda}_{\mathrm{ML}}$ is, for $z \in \Lambda$,*

$$\Pr(\hat{\lambda}_{\mathrm{ML}} \le z) = \Pr\left(\mathrm{F}_{r-1,r(m-1)} \le \frac{r}{r-1} c(z,\lambda)\right).$$

Because this is only a trivial modification of the result in Proposition 5.1, we omit further details for this case.

[25] Interestingly, notwithstanding its inconsistency under infill asymptotics, it seems clear that the MLE is a perfectly reasonable estimator in this model as long as r is at least of moderate size.

5.2 The Group Interaction Model

The results obtained above for the balanced Group Interaction model can be extended to the unbalanced case. Recall that in the balanced case the number of distinct eigenvalues is $T = 2$, but in the unbalanced case $T > 2$. For the pure model, the cdf is as given in Corollary 4.10, and, as discussed earlier, there are points of non-analyticity in the distribution of $\hat{\lambda}_{\mathrm{ML}}$. The density in each interval between the z_t is difficult to derive from equation (4.15), but can be obtained readily by a conditioning argument.

When there are just two group sizes (i.e., $u = 2$), $T = 3$, so there is a single point of non-analyticity, and we discuss only this case here. It is known that, in the general case (u arbitrary), $\hat{\lambda}_{\mathrm{ML}}$ is consistent if $r_t \to \infty$ for at least one t. In the Supplementary Material we briefly discuss the asymptotic behavior of $\hat{\lambda}_{\mathrm{ML}}$ that can be deduced from the exact result. In particular, we see that when the r_t are fixed, $\hat{\lambda}_{\mathrm{ML}}$ converges in distribution to a random variable if some or all of the $m_t \to \infty$, generalizing the results above for the balanced case. We first illustrate the conditioning argument for the case $T = 3$.

5.2.1 Conditioning Argument with Three Distinct Eigenvalues

In the case $T = 3$ Theorem 3 says that the cdf of $\hat{\lambda}_{\mathrm{ML}}$ can, under the assumptions of the theorem, be represented in the form

$$\Pr(\hat{\lambda}_{\mathrm{ML}} \leq z) = \Pr(d_{11}q_{n_1} + d_{22}q_{n_2} + d_{33}q_{n_3} \leq 0),$$

where q_{n_i}, $i = 1, 2, 3$ are independent $\chi^2_{n_i}$ random variables. Proposition 4.6 also says that $d_{11} < 0$ and $d_{33} > 0$ for all $z \in \Lambda$, while d_{22} changes sign at the point $z = z_2$. The density has a different functional form on each side of z_2.

For $z > z_2$ we may immediately write down the conditional cdf, given q_{n_1} and q_{n_2},

$$\Pr(\hat{\lambda}_{\mathrm{ML}} \leq z | q_{n_1}, q_{n_2}) = \mathcal{G}_{n_3}(\psi_{11}q_{n_1} + \psi_{12}q_{n_2}),$$

where $\mathcal{G}_v$ denotes the cdf of a χ^2_v variate, and the coefficient functions $\psi_{1t} := -d_{tt}/d_{33}, t = 1, 2$, are both positive for $z_2 < z < 1$. Likewise, for $z < z_2$ we have the conditional cdf q_{n_2} and q_{n_3},

$$\Pr(\hat{\lambda}_{\mathrm{ML}} \leq z | q_{n_2}, q_{n_3}) = 1 - \mathcal{G}_{n_1}(\psi_{22}q_{n_2} + \psi_{23}q_{n_3}).$$

with coefficients $\psi_{2t} := -d_{tt}/d_{11}, t = 2, 3$, that are again both positive. Differentiating these expressions with respect to z gives the two conditional densities, which can then be averaged with respect to the conditioning variates to produce the unconditional densities. Lemma 8.1 in the Supplementary Material provides formulae for the density that this process produces.

5.2.2 Two Group Sizes; Zero Mean

When $u = 2$ the r groups have size m_1 or m_2, with $m_1 < m_2$. The coefficients d_{tt} are as defined in equation (4.8) (with $\omega_1 = -1/(m_1 - 1), \omega_2 = -1/(m_2 - 1), \omega_3 = 1$). The coefficient d_{22} changes sign (once) at the point[26]

$$z_2 := -\frac{n(m_1 - 1)}{n + (m_2 - m_1)\, r_1 m_1} < 0. \tag{5.3}$$

The result in Lemma 8.1 in the Supplementary Material applies, and formulae for the density function on the two intervals $-(m_1 - 1) < z < z_2$ and $z_2 < z < 1$ can be found in the Supplementary Material.

In Figure 2 we display the exact density for the case when $r = 2$ and one of the two groups has fixed size $m_1 = 2$, varying the size of the other group, and hence varying $n = m_1 + m_2$. The density is plotted for three different values of λ. When $m_2 = 2$ (so that $n = 4$), the model is balanced, and the density is analytic on $\Lambda = (-1, 1)$ by Corollary 4.10. For any $m_2 > 2$, there is a point z_2 of non-analyticity in Λ. This point is $-.429$ for $n = 10$, and it approaches $-1/3$ from the left as $n \to \infty$.

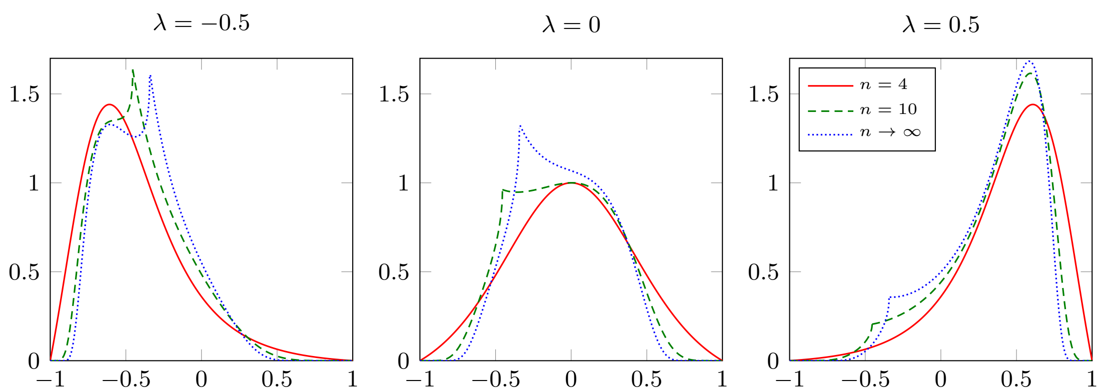

Figure 2: Density of $\hat{\lambda}_{\mathrm{ML}}$ for the Gaussian pure Group Interaction model with two groups, one of which has size $m_1 = 2$.

The plots show clearly that the density has a single component only when $m_1 = m_2$. As the difference between m_1 and m_2 increases, the difference between that two components becomes more apparent, and the density becomes less smooth at the point z_2. Note that this may be regarded as a consequence of imposing the same parameter λ on the two different groups.

[26]Note that this becomes $-(m_1 - 1)$ when $m_1 = m_2$ (the balanced case), so that only the interval $-(m_1 - 1) < z < 1$ is relevant, and the density has a single functional form over all of Λ.

The analogs of these figures for values of $m_1 > 2$ are given in the Supplementary Material. All of these figures show that the properties of $\hat{\lambda}_{\text{ML}}$ are, in this model with just two groups, almost invariant to the sample size, a property related to, but not implied by the asymptotic properties for a fixed number of groups mentioned earlier. However, even though the estimator is not consistent under some asymptotic regimes, there is certainly no evidence here that suggests not using maximum likelihood in this model.

5.2.3 Constant Mean

For the Group Interaction model with constant mean (i.e., $X = \iota_n$) it is easy to check that ι_n is in the eigenspace associated to the eigenvalue 1, so that equation 4.17 applies (again, the stronger version). Thus, the cdf in this case is again only a slightly modified version of that for the pure unbalanced Group Interaction case discussed above, and we omit further details.

5.2.4 Fixed Effects

We have already seen in Section 4.3 that in the case of an unbalanced Group Interaction model with group fixed effects and no other regressors (i.e., $X = \bigoplus_{i=1}^{r} \iota_{m_i}$), the support of $\hat{\lambda}_{\text{ML}}$ is a subset of Λ. In this case the columns of X span the eigenspace of W associated to $\omega_{u+1} = 1$, so that, by Proposition 4.7, the support of $\hat{\lambda}_{\text{ML}}$ is $(-(m_1 - 1), z_u)$. In the notation of Theorem 4 we have $n_{u+1} = n_{u+1}(X)$, so the term for $t = u + 1$ does not appear. The remaining coefficients are all negative for $z_u < z < 1$, so we have:

Proposition 5.4. *In the unbalanced Group Interaction model with $X = \bigoplus_{i=1}^{r} \iota_{m_i}$ and $\varepsilon \sim \text{SMN}(0, \sigma^2 I_n)$, the* cdf *of $\hat{\lambda}_{\text{ML}}$ is given by*

$$\Pr(\hat{\lambda}_{\text{ML}} \leq z) = \begin{cases} \Pr\left(\sum_{t=1}^{u} d_{tt}\chi^2_{n_t} \leq 0\right), & \textit{if} - (m_1 - 1) < z < z_u \\ 1, & \textit{if } z_u \leq z < 1, \end{cases} \tag{5.4}$$

where the χ^2 random variables involved are independent.

The density can be obtained from (5.4), but the main point of interest here is that $\hat{\lambda}_{\text{ML}}$ is restricted to a subset of Λ, whatever the true value of λ. We will meet another example of the same type shortly.[27]

[27] In the case of two group sizes ($u = 2$), z_2 is given in (5.3), and the cdf in (5.4) reduces to $\Pr\left(\text{F}_{n_2,n_1} \leq -n_1 d_{11}/(n_2 d_{22})\right)$ for $-(m_1 - 1) < z < z_u$.

5.3 The Complete Bipartite Model

We now apply the general results to the Complete Bipartite model introduced in Section 2.3. In Section 5.3.1 we discuss the simple case of a pure symmetric Complete Bipartite model. Then, in Section 5.3.2, we discuss the case of the row-standardized Complete Bipartite model with unknown constant mean (i.e., $X = \iota_n$). This provides a second important illustration of the restricted support phenomenon described in Section 4.3.

5.3.1 Symmetric W, Zero Mean

In the symmetric Complete Bipartite model, W again has $T = 3$ distinct eigenvalues: -1, 0, 1. According to Corollary 4.10, the pdf of $\hat{\lambda}_{\mathrm{ML}}$ in the pure Gaussian case is analytic everywhere on $\Lambda = (-1, 1)$ except at the point z_2, and it is readily verified that $z_2 = 0$. Moreover, since the spectrum of W is symmetric, the symmetry established in Corollary 4.9 may be used to obtain the density for $z \in (-1, 0)$ from that for $z \in (0, 1)$.

Proposition 5.5. *In the pure symmetric Complete Bipartite model with* $\varepsilon \sim \mathrm{SMN}(0, \sigma^2 I_n)$*,*

$$\Pr(\hat{\lambda}_{\mathrm{ML}} \leq z) = \Pr(\phi_1 \chi_1^2 \leq \phi_2 \chi_1^2 + 2z\chi_{n-2}^2), \tag{5.5}$$

for $-1 < z < 1$*, where*

$$\phi_1 := \frac{(1-z)^2 \left[n + (n-2)\, z\right]}{(1-\lambda)^2}, \phi_2 := \frac{(1+z)^2 \left[n - (n-2)\, z\right]}{(1+\lambda)^2},$$

and the three χ^2 *random variables involved are independent.*

Proposition 5.5 confirms the fact remarked upon in the discussion of Corollary 4.8, that the distribution, and hence all the properties of $\hat{\lambda}_{\mathrm{ML}}$, depends on p and q only through their sum n.[28] The coefficients ϕ_1, ϕ_2 in (5.5) are both positive for all $z \in \Lambda = (-1, 1)$, but z changes sign of course. Applying results reported in the Supplementary Material, we obtain:[29]

[28]Note that taking $z = 0$ in (5.5) gives the following very simple formula for the probability that $\hat{\lambda}_{\mathrm{ML}}$ is negative:

$$\Pr(\hat{\lambda}_{\mathrm{ML}} \leq 0) = \Pr\left(|\xi| \leq \frac{1-\lambda}{1+\lambda}\right),$$

where ξ has a Cauchy distribution. Since this does not depend on n, this holds for all sample sizes.

[29]The function ${}_2F_1(\cdot)$ here is the Gaussian Hypergeometric function, see Muirhead (1982), Chapter 1.

Proposition 5.6. *In the pure symmetric Complete Bipartite model with* $\varepsilon \sim \mathrm{SMN}(0, \sigma^2 I_n)$ *the density of* $\hat{\lambda}_{\mathrm{ML}}$ *for* $z \in (0,1)$ *is*

$$\mathrm{pdf}_{\hat{\lambda}_{\mathrm{ML}}}(z;\lambda) = \frac{B\left(\frac{1}{2},\frac{n}{2}\right)c}{2\pi a^{\frac{1}{2}}(1+c)^{\frac{n}{2}}}\left[\frac{\alpha\dot{a}}{a}\,{}_2F_1\left(\frac{n}{2},\frac{3}{2},\frac{n+1}{2};\eta\right) + \frac{\beta\dot{c}}{c}\,{}_2F_1\left(\frac{n}{2},\frac{1}{2},\frac{n+1}{2};\eta\right)\right], \quad (5.6)$$

where $a := \phi_2/\phi_1$*,* $c := 2z/\phi_1$*, and* $\eta := \phi_1(\phi_2 - 2z)/\phi_2(\phi_1 + 2z)$*. For* $z \in (-1,0)$ *the density is defined by* $\mathrm{pdf}_{\hat{\lambda}_{\mathrm{ML}}}(z;\lambda) = \mathrm{pdf}_{\hat{\lambda}_{\mathrm{ML}}}(-z;-\lambda)$*.*

The asymptotic distribution as $n \to \infty$ can be obtained easily, as follows. For every fixed $z \in \Lambda$, the characteristic function of the random variable $V_n := (\phi_1\chi_1^2 - \phi_2\chi_1^2 - 2z\chi_{n-2}^2)/(n-2)$ is easily seen to converge to that of

$$\bar{V}_n := \bar{\phi}_1\chi_1^2 - \bar{\phi}_2\chi_1^2 - 2z,$$

where $\bar{\phi}_1 := \lim_{n\to\infty}(\phi_1/(n-2)) = (1-z)^2(1+z)/(1-\lambda)^2$ and $\bar{\phi}_2 := \lim_{n\to\infty}(\phi_2/(n-2)) = (1+z)^2(1-z)/(1+\lambda)^2$. Therefore, $V_n \to_d \bar{V}_n$, and so (from Proposition 5.5), $\Pr(\hat{\lambda}_{\mathrm{ML}} \le z) \to \Pr\left(\chi_1^2 \le \bar{\psi}_1\chi_1^2 + \bar{\psi}_2\right)$, with

$$\bar{\psi}_1 := \left(\frac{1+z}{1-z}\right)\left(\frac{1-\lambda}{1+\lambda}\right)^2, \bar{\psi}_2 := \frac{2z(1-\lambda)^2}{(1+z)(1-z)^2},$$

for $z \in (0,1)$, and the two χ_1^2 variates are independent. For $z \in (0,1)$, therefore, the usual conditioning argument yields

$$\Pr(\hat{\lambda}_{\mathrm{ML}} \le z) \to \mathrm{E}_{q_1}\left[\mathcal{G}_1(\bar{\psi}_1 q_1 + \bar{\psi}_2)\right], \quad (5.7)$$

where $q_1 \equiv \chi_1^2$. Thus, as in the case when $m \to \infty$ in a Balanced Group Interaction model, $\hat{\lambda}_{\mathrm{ML}}$ is not consistent, but converges in distribution to a random variable as $n \to \infty$. The limiting pdf can be obtained from (5.7), but is omitted for brevity.

The density (5.6) is plotted in Figure 3 for $\lambda = -0.5, 0, 0.5$, for $n = 5, 10$, and for $n \to \infty$. It is clear from the plots that the density is again very insensitive to the sample size, so in this model increasing the sample size yields little extra information about λ. As a consequence, the non-standard asymptotic density is an excellent approximation to the actual distribution under mixed-normal assumptions. The expected non-analyticity at $z = 0$ is evident, and in fact for this model the density of $\hat{\lambda}_{\mathrm{ML}}$ is unbounded at $z = 0$.

Given the cdf and pdf, other exact properties of $\hat{\lambda}_{\mathrm{ML}}$ can be derived following techniques similar to those used in Hillier and Martellosio (2013) for the balanced Group Interaction model, but this is not pursued here.

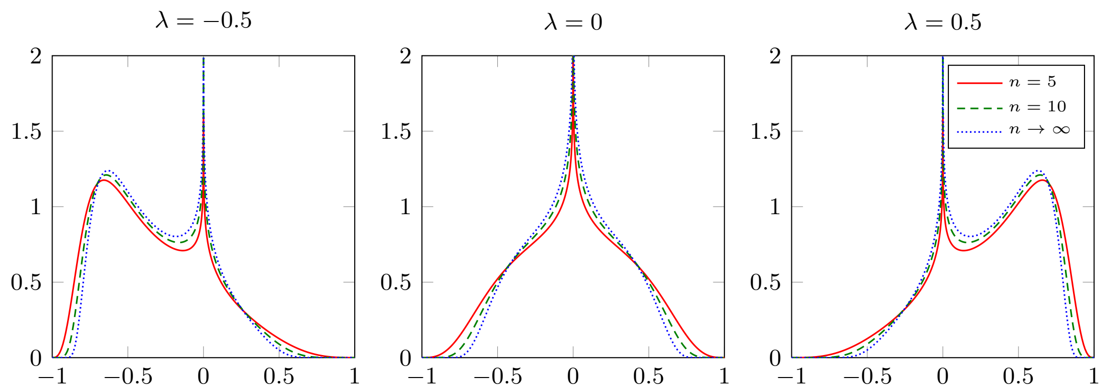

Figure 3: Density of $\hat{\lambda}_{\mathrm{ML}}$ for the Gaussian pure symmetric Complete Bipartite model.

5.3.2 Row-Standardized W, Constant Mean

As already anticipated in the discussion of Proposition 4.7, the support of $\hat{\lambda}_{\mathrm{ML}}$ in the row-standardized Complete Bipartite model with constant mean is not the entire interval $\Lambda = (-1, 1)$, but the subset $(-1, 0)$ (regardless of whether the true value of λ is positive or negative).

Proposition 5.7. *For the row-standardized Complete Bipartite model with $X = \iota_n$ and $\varepsilon \sim \mathrm{SMN}(0, \sigma^2 I_n)$,*

$$\Pr(\hat{\lambda}_{\mathrm{ML}} \leq z) = \begin{cases} \Pr\left(\mathrm{F}_{1,n-2} > -(n-2)g(z;\lambda)\right), & \text{if } -1 < z < 0 \\ 1, & \text{if } 0 \leq z < 1, \end{cases}$$

where

$$g(z;\lambda) := \frac{2z(1+\lambda)^2}{(1+z)^2[n-(n-2)\,z]}.$$

Differentiating the cdf we obtain the following expression for the density.

Proposition 5.8. *For the row-standardised Complete Bipartite model with $\varepsilon \sim \mathrm{SMN}(0, \sigma^2 I_n)$, and with $X = \iota_n$,*

$$\mathrm{pdf}_{\hat{\lambda}_{\mathrm{ML}}}(z;\lambda) = \frac{1}{B\left(\frac{1}{2}, \frac{n-2}{2}\right)} \frac{\dot{g}(z;\lambda)}{g(z;\lambda)^{\frac{1}{2}}\left[1-g(z;\lambda)\right]^{\frac{n-1}{2}}}, \tag{5.8}$$

for $z \in (-1, 0)$. For $z \in (0, 1)$, $\mathrm{pdf}_{\hat{\lambda}_{\mathrm{ML}}}(z;\lambda) = 0$.

The limiting cdf and pdf as $n \to \infty$ can be obtained immediately from the results above. Letting

$$h(z;\lambda) := \lim_{n\to\infty}\left[-(n-2)g(z;\lambda)\right] = -\frac{2z(1+\lambda)^2}{(1+z)^2(1-z)},$$

we obtain that, as $n \to \infty$, and for $z \in (-1, 0)$,

$$\Pr(\hat{\lambda}_{\text{ML}} \leq z) \to \Pr\left(\chi_1^2 > h(z;\lambda)\right),$$

and

$$\text{pdf}_{\hat{\lambda}_{\text{ML}}}(z;\lambda) \to -\frac{\dot{h}(z;\lambda)}{\sqrt{2\pi h(z;\lambda)}} e^{-\frac{h(z;\lambda)}{2}}.$$

Again, $\hat{\lambda}_{\text{ML}}$ is not consistent, but converges in distribution to a random variable supported on the non-positive real line as $n \to \infty$. Note that row-standardization of W is critical here: the symmetric Complete Bipartite model with constant mean does satisfy the assumptions for consistency and asymptotic normality in Lee (2004).

The density (5.8) is plotted in Figure 4 for $\lambda = -0.5, 0, 0.5$, for $n = 5, 10$, and for $n \to \infty$. Note that the shape of the density for $z < 0$ is similar to the case of the pure symmetric Complete Bipartite model (Figure 3).

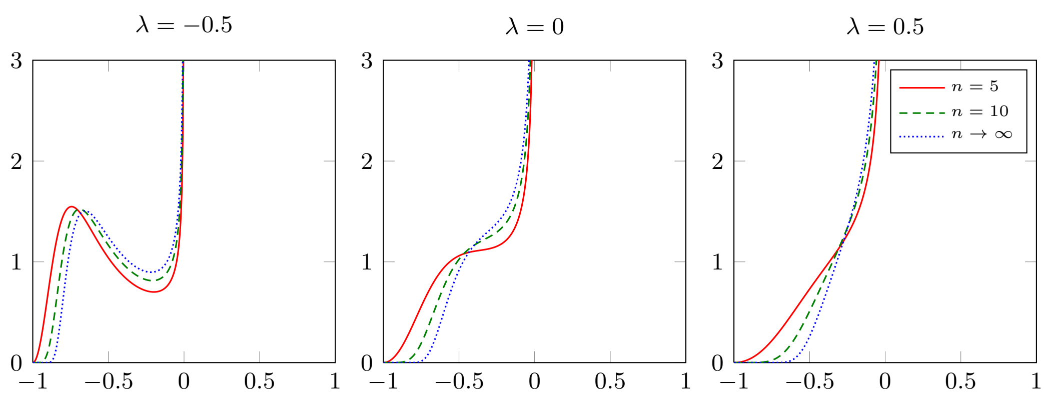

Figure 4: Density of $\hat{\lambda}_{\text{ML}}$ for the Gaussian row-standardized Complete Bipartite model with constant mean.

6 The Single-Peaked Property Generally

The exact expression for the cdf of $\hat{\lambda}_{\text{ML}}$ given in Theorem 1 depends only upon the fact that the profile log-likelihood $l_p(\lambda)$ is single-peaked on Λ, which was established in Lemma 3.5 under the condition that all eigenvalues of W are real. That condition makes the single-peaked property easy to prove, but it is certainly not necessary. It is obviously desirable to investigate the issue of single/multi-peakedeness of the log-likelihood further.[30] Let

$$\delta(\lambda) := [\text{tr}(G_\lambda)]^2 - n\text{tr}(G_\lambda^2).$$

The proof of Lemma 3.5 shows that whenever W has the property that $\delta(\lambda) < 0$ for all $\lambda \in \Lambda$, every critical point of $l_p(\lambda)$ is a local maximum, implying that $l_p(\lambda)$ is again single-peaked on Λ. Thus, we have the following more general version of Theorem 1.

Theorem 5. *For any W such that $\delta(\lambda) < 0$ for all $\lambda \in \Lambda$, the* cdf *of $\hat{\lambda}_{\text{ML}}$ is as given in Theorem 1.*

Theorem 5 generalizes Theorem 1 to cases in which some eigenvalues of W may be complex. It seems difficult to characterize the class of matrices W for which $\delta(\lambda) < 0$ for all $\lambda \in \Lambda$, but for any particular choice of W it is straightforward to check, by simply graphing $\delta(\lambda)$, whether this sufficient condition holds. Note that the condition depends only on W, not on X. The following example provides some evidence that the condition $\delta(\lambda) < 0$ for all $\lambda \in \Lambda$ is much more general than requiring real eigenvalues.

Example 3. Consider the weights matrix W obtained by row-standardizing the band matrix

$$A = \begin{bmatrix} 0 & a_3 & a_4 & 0 & \cdots \\ a_1 & 0 & a_3 & a_4 & \\ a_2 & a_1 & 0 & a_3 & \\ 0 & a_2 & a_1 & 0 & \\ \vdots & & & & \ddots \end{bmatrix},$$

for fixed a_1, a_2, a_3, a_4. If $a_1 = a_3$ and $a_2 = a_4$, all the eigenvalues of W are real and therefore $l_p(\lambda)$ is a.s. single-peaked by Lemma 3.5. Other configurations of the a_i can induce multi-peakedeness of $l_p(\lambda)$. To see this, fix $n = 20$, $a_1 = a_2 = a_3 = 1$, and consider values of a_4 in $[0, 1]$. For any value of a_4 larger than about 0.55, $\delta(\lambda) < 0$ for all $\lambda \in \Lambda$, so, even though not all eigenvalues of W are real, $l_p(\lambda)$ is a.s. single-peaked by Theorem 5. For smaller values of a_4 $\delta(\lambda)$ is not negative for all $\lambda \in \Lambda$, and there is a positive probability that $l_p(\lambda)$ is multi-peaked. Figure 6 displays $\delta(\lambda)$ when $a_4 = 0.9$ (left panel) and $a_4 = 0$ (right panel). Note that Λ depends on a_4. One can check by simulation that, whatever the value of X, $a_4 = 0$ entails a high probability of multi-peakedeness as y ranges over $\mathbb{R}^n$.

[30] In general, when $l_p(\lambda)$ is not a.s. single-peaked on Λ, there will nevertheless be a non-zero probability that the property holds. If that probability is large, the results in this paper could perhaps provide an approximation to the cdf of the MLE, rather than the exact cdf.

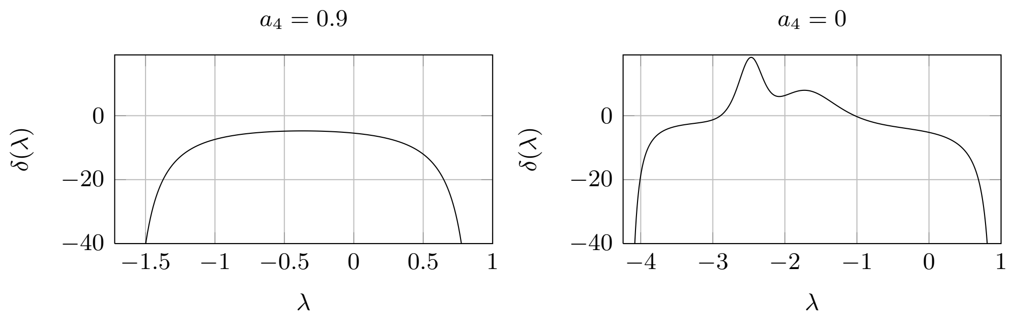

Figure 5: $\delta(\lambda)$, $\lambda \in \Lambda$, for the weights matrix W in Example 3.

A complete understanding of the cases in which the single-peaked property fails to hold is beyond the scope of this paper, but the next result is a first step in that direction. It says multi-peakedeness must always involve peaks at negative values of λ, for any W and X.

Proposition 6.1. *$l_p(\lambda)$ has at most one maximum in the interval* $[0, 1)$.

7 Discussion

The main result in this paper - Theorem 1 - provides a starting point for an examination of the properties of the maximum likelihood estimator for λ. Whatever the matrices W and X involved in model (1.1), and whatever the distribution assumptions entertained for ε, equation (4.3) provides a simple basis for simulation study of the properties of $\hat{\lambda}_{\mathrm{ML}}$. The result is also a useful starting point for the study of the higher-order asymptotic properties of $\hat{\lambda}_{\mathrm{ML}}$, a subject not embarked upon here. Finally, we have seen that in reasonably simple models with a high degree of structure (when W has only a few distinct eigenvalues), it can provide both exact results directly useful for inference, and new asymptotic results for cases not covered by the known results in Lee (2004). The present paper is just a beginning.

The study of quadratic forms of the type involved in equation (4.5) was begun by John von Neumann and Tjalling Koopmans in the 1940's when studying the distribution of serial correlation coefficients. The papers by von Neumann (1941) and Koopmans (1942), both discuss the unusual aspects of the distribution of serial correlation coefficients. Interestingly, Corollary 4.1 in this paper shows that the distributional properties of the MLE in spatial autoregressive models have closely related characteristics, at least in the Gaussian pure SAR case, a result that perhaps might have been anticipated but was, a priori, certainly not obvious. However, two aspects of our results for this model did not occur in that earlier work: the possibility that the MLE might not exist, and the possibility that the support of the estimator might not be the entire parameter space. These are subjects that clearly demand further work. In particular, the question of how these properties of the model relate to the identification of λ needs careful analysis.

Finally, the results discussed in the paper highlight the key role played by the design matrices W and X in determining the properties of the maximum likelihood estimator, and inference generally in this class of models. The fact that different assumptions about the asymptotic evolution of W can produce different outcomes is well-understood (Lee, 2004), and Theorem 1 provides a convenient new framework for further study of this issue.

Appendix A Proofs of Main Results

Proof of Lemma 3.2. We only prove (3.4); the proof of (3.5) is essentially identical. Let us look at the two terms, $l_{p1} := -n/2 \ln (y'S'_\lambda M_X S_\lambda y)$ and $l_{p2} := \ln(\det(S_\lambda))$, that make up the profile log-likelihood $l_p(\lambda)$. Note that $y'S'_1 M_X S_1 y$ is a.s. non-negative for any W and X, and is zero for all $y \in \mathbb{R}^n$ if and only if $M_X S_1 = 0$. Thus, $\lim_{\lambda\to 1} l_{p1} \in (-\infty, 0]$ if $M_X S_1 \neq 0$, and $\lim_{\lambda\to 1} l_{p1} = 0$ if $M_X S_1 = 0$. On the other hand l_{p2} always goes to $-\infty$ as $\lambda \to 1$, since $\det(S_1) = 0$. This proves that $\lim_{\lambda\to 1} l_p(\lambda) = -\infty$ when $M_X S_1 \neq 0$, but a further step is required to evaluate the limit when $M_X S_1 = 0$, because l_{p1} and l_{p2} diverge to ∞ in different directions in that case. The further step relies on the fact that if $M_X S_1 = 0$ then $M_X W = M_X$ and hence $M_X S_\lambda = (1-\lambda) M_X$. Letting Ω denote the spectrum of W, and n_ω the algebraic multiplicity of $\omega \in \Omega$, we obtain, under the condition $M_X S_1 = 0$,

$$\begin{aligned} l_p(\lambda) &= \ln\left(\frac{\det(S_\lambda)}{\left(y'S'_\lambda M_X S_\lambda y\right)^{\frac{n}{2}}}\right) \\ &= \ln\left(\frac{\prod_{\omega\in\Omega/\{1\}}(1-\lambda\omega)^{n_\omega}}{(y'M_X y)^{\frac{n}{2}}}\right) - (n-n_1)\ln(1-\lambda), \end{aligned}$$

The first term is a.s. finite because no term in the product vanishes, but since $n_1 < n$ (because $W \neq I_n$ by the assumption that $\operatorname{tr}(W) = 0$), the second term $\to +\infty$ as $\lambda \to 1$. Thus, if $M_X S_1 = 0$, $\lim_{\lambda\to 1} l_p(\lambda) = +\infty$. □

Proof of Lemma 3.4. The equation $\dot{l}_p(\lambda) = 0$ appears to give rise to a polynomial of degree $T+1$. However, the coefficient of the highest-order term vanishes, but that of the term of order T does not. The full proof can be found in the Supplementary Material. □

Proof of Lemma 3.5. Recall that we are assuming that $M_X S_1 \neq 0$ and $M_X S_{\lambda_{\min}} \neq 0$, so that $l_p(\lambda) \to -\infty$ at the extremes of Λ. Then, because it is a.s. differentiable on Λ, $l_p(\lambda)$ must a.s. have at least one maximum on Λ. We now show that it has a.s. exactly one maximum, and no other stationary points, on Λ. The second derivative of $l_p(\lambda)$ can be written as

$$\ddot{l}_p(\lambda) = \frac{-n(ac-b^2)}{(a\lambda^2-2b\lambda+c)^2} + \frac{n(b-a\lambda)^2}{(a\lambda^2-2b\lambda+c)^2} - \operatorname{tr}(G_\lambda^2),$$

where

$$a := y'W'M_X W y,\ b := y'W'M_X y,\ c := y'M_X y.$$

But at any point where $\dot{l}_p(\lambda) = 0$,

$$\frac{n(b-a\lambda)^2}{(a\lambda^2-2b\lambda+c)^2} = \frac{1}{n}\left[\operatorname{tr}\left(G_\lambda\right)\right]^2,$$

so that, at any critical point,

$$\ddot{l}_p(\lambda) = \left\{\frac{-n(ac-b^2)}{(a\lambda^2-2b\lambda+c)^2}\right\} + \frac{1}{n}\left\{[\operatorname{tr}(G_\lambda)]^2 - n\operatorname{tr}(G_\lambda^2)\right\}. \tag{A.1}$$

By the Cauchy-Schwarz inequality the first term on the right hand side of (A.1) is nonpositive. When the eigenvalues of W are real, the second term in curly brackets is also nonpositive, again by the Cauchy-Schwarz inequality, and cannot be zero because G_λ cannot be a scalar multiple of I_n. That is, at every point where $\dot{l}_p(\lambda)$ vanishes, $\ddot{l}_p(\lambda) < 0$. Thus, $l_p(\lambda)$ has exactly one maximum in Λ, and no other stationary points. □

Proof of Proposition 3.8. See the Supplementary Material.[31] □

Proof of Proposition 3.9. If the column space of X is spanned by k linearly independent eigenvectors of W, $X = V_k C$, say, where the columns of V_k are k linearly independent eigenvectors of W, and C is non-singular. Then $WV_k = V_k D$ for some $k \times k$ diagonal matrix of eigenvalues of W. It is easily checked that $S_\lambda X = XA$, say, with $A = C^{-1}[I_k - \lambda D]C$ a non-singular matrix for $\lambda \in \Lambda$. Now consider the transformations

$$y \to \kappa y + X\delta,\ \kappa > 0,\ \delta \in \mathbb{R}^k.$$

[31] For a formal treatment of the argument used to establish Proposition 3.8 - averaging over the group - see Eaton (1989), particularly Chapters 4 and 5. A direct proof of Proposition 3.8 can be found in the Supplementary Material.

These leave the family of densities for y invariant, inducing only the group of transformations

$$(\beta, \sigma^2, \lambda) \to (\kappa\beta + A\delta, \kappa^2\sigma^2, \lambda)$$

on the parameter space. The distribution of any invariant statistic will depend on $(\beta, \sigma^2, \lambda)$ only through the maximal invariant in the parameter space. But, the induced group acts transitively on the parameter space for (β, σ^2), and leaves λ invariant. It follows that the distribution of the maximal invariant cannot depend on (β, σ^2), and depends only on λ. □

Proof of Theorem 2. (i) Proved in the text. (ii) Under the assumption that W is similar to a symmetric matrix, the off-diagonal blocks in M vanish if and only if $MD = DM$, where D contains the eigenvalues of W and $M = H'M_XH$ is as in the text, because the eigenvalues in the decomposition of D are distinct. One can then easily check that this is so if and only if $M_XW = W'M_X$. (iii) With the added assumption of symmetry of W, H is orthogonal, and

$$M_{ij} = h_i'M_Xh_j = \begin{cases} 0 & \text{if } i \neq j \\ 0 & \text{if } i = j \text{ and } h_i \in \text{col}(X) \\ 1 & \text{if } i = j \text{ and } h_i \notin \text{col}(X). \end{cases}$$

The first line relies on the fact that $h_i'M_Xh_j = 0$ if either h_i or h_j is in col(X), and also that when $h_i \notin \text{col}(X)$, $M_Xh_i = h_i$, and $h_i'h_j = 0$. □

Proof of Theorem 3. The assumption $\varepsilon \sim \text{SMN}(0, \sigma^2 I_n)$ gives $\tilde{y} \sim \text{SMN}(0, \sigma^2 I_n)$ and $x \sim \text{SMN}(0, (H'H)^{-1})$. But $H'H = I_n$ if W is symmetric, and hence the stated result follows from (4.11). □

Proof of Proposition 4.6. (i) Obviously, $\omega_r > \omega_s$ implies $\omega_r/(1 - z\omega_r) > \omega_s/(1 - z\omega_s)$ for all $z \in \Lambda$, which in turn implies $\gamma_r(z) > \gamma_s(z)$. Consider the functions $\gamma_{1t}(z) := \omega_t/(1 - z\omega_t)$, labelled in the same order as the ω_t. If $\omega_t = 0$, $\gamma_{1t}(z) = 0$ is constant. For the non-zero eigenvalues, since $d\gamma_{1t}(z)/dz = \gamma_{1t}^2(z) > 0$, each of these functions is strictly increasing on Λ. The function $\gamma_{11}(z) = \omega_{\min}/(1 - z\omega_{\min}) \to -\infty$ as $z \downarrow \omega_{\min}^{-1}$, and is finite $(= \omega_{\min}/(1 - \omega_{\min}))$ at $z = 1$. Likewise, the function $\gamma_{1T}(z) = 1/(1 - z)$ is finite at $z = \omega_{\min}^{-1}$ $(= \omega_{\min}/(\omega_{\min} - 1))$ and $\gamma_{1T}(z) \to +\infty$ as $z \uparrow 1$. The remaining functions $\gamma_{1t}(z)$ are all finite at both endpoints of the interval Λ. The average of the γ_{1t} is

$$\frac{1}{n}\text{tr}(G_z) = \frac{1}{n}\sum_{t=1}^{T} \frac{n_t\omega_t}{1 - z\omega_t} = \sum_{t=1}^{T} \alpha_t\gamma_{1t}(z)$$

(with $\alpha_t := n_t/n$). Since this is a convex combination of the $\gamma_{1t}(z)$, it is between the smallest and largest of them, for all $z \in \Lambda$, i.e.,

$$\gamma_{11}(z) < \frac{1}{n}\operatorname{tr}(G_z) < \gamma_{1T}(z).$$

Thus, for all $z \in \Lambda$, $\gamma_1(z) < 0$, and $\gamma_T(z) > 0$, so these two functions do not change sign on Λ. Next, the properties of the γ_{1t} imply that $\operatorname{tr}(G_z)/n$ is monotonic increasing on Λ, going to $-\infty$ as $z \downarrow \omega_{\min}$, and to $+\infty$ as $z \uparrow 1$. It follows that $\operatorname{tr}(G_z)/n$ crosses all $T-2$ of the functions $\gamma_{1t}(z)$, $t \neq 1, T$, at least once, somewhere in Λ. To show that the two functions can only cross once, simply observe that, at a point z where $\gamma_t(z) = 0$,

$$\dot{\gamma}_{1t}(z) = \gamma_{1t}^2(z) = \left(\sum_{t=1}^{T} \alpha_t \gamma_{1t}(z)\right)^2 < \sum_{t=1}^{T} \alpha_t \gamma_{1t}^2(z) = \frac{d}{dz}\left(\frac{1}{n}\operatorname{tr}(G_z)\right).$$

(the inequality is strict because the $\gamma_{1t}(z)$ cannot all be equal). That is, at every point of intersection, $\operatorname{tr}(G_z)/n$ intersects $\gamma_{1t}(z)$ from below, which implies that there can be only one such point. (ii) This follows from part (i) and the fact that the signs of the d_{tt} are those of the γ_t. □

Proof of Proposition 4.7. Under the first stated condition all diagonal blocks M_{ss} for $s > t$ vanish, and for $s \leq t$ the $\gamma_s(z)$ are all negative for $z > z_t$, so that, by Theorem 2 (iii), $\Pr(\hat{\lambda}_{\mathrm{ML}} \leq z) = 1$ for $z \geq z_t$. In the second case the $\gamma_s(z)$ are all positive for $z < z_t$ by Proposition 4.6, so that $\Pr(\hat{\lambda}_{\mathrm{ML}} \leq z) = 0$ for $z \leq z_t$. □

Proof of Corollary 4.9. We prove this result under Gaussian assumptions. For notational convenience, let us rewrite expression (4.12) as

$$\Pr(\hat{\lambda}_{\mathrm{ML}} \leq z; \lambda) = \Pr\left(\sum_{s=1}^{S} d_{\omega_s}(\lambda, z)\, \chi^2_{n_s} \leq 0\right).$$

Since $d_{\omega_s}(\lambda, z) = -d_{-\omega_s}(-\lambda, -z)$, we have

$$\Pr(\hat{\lambda}_{\mathrm{ML}} \leq z; \lambda) = \Pr\left(\sum_{s=1}^{S} -d_{-\omega_s}(-\lambda, -z)\, \chi^2_{n_s} \leq 0\right),$$

which is equal to $\Pr(\hat{\lambda}_{\mathrm{ML}} \geq -z; -\lambda) = 1 - \Pr(\hat{\lambda}_{\mathrm{ML}} \leq -z; -\lambda)$ if the spectrum of W is symmetric. The stated result follows on differentiating $\Pr(\hat{\lambda}_{\mathrm{ML}} \leq z; \lambda) = 1 - \Pr(\hat{\lambda}_{\mathrm{ML}} \leq -z; -\lambda)$ with respect to z. □

Proof of Corollary 4.10. We first note the following slight modification of a result due to James (1964) for the density of a positive definite quadratic form in standard normal variables: If $Q := \sum_{i=1}^{S} a_i \chi^2_{n_i}$ is a linear combination of independent $\chi^2_{n_i}$ random variables with positive coefficients a_i, the density of Q is given by

$$\mathrm{pdf}_Q(q; A) = \frac{\exp\left(-\frac{1}{2}\tau q\right) q^{\frac{n}{2}-1}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})\left(\det(A)\right)^{\frac{1}{2}}} {}_1F_1\left(\frac{1}{2}; \frac{n}{2}; \frac{q}{2}A^*\right) \tag{A.2}$$

where $n = \sum_{i=1}^{S} n_i$, $A := \mathrm{diag}(a_i I_{n_i}, i = 1, ..., S)$, $\tau := \mathrm{tr}(A^{-1})$, and $A^* := \tau I_n - A^{-1}$. The confluent hypergeometric function here is of matrix argument (see Muirhead, 1982), but, importantly, only top-order zonal polynomials are involved. Using this result for both Q_1 and Q_2, transforming to (R, Q_2), and integrating out the redundant variable termwise gives an expression involving only r (it is straightforward to check that the term-by-term integration involved is justified). Integrating this over $0 < r < 1$ gives the result. □

Proof of Proposition 5.1. For the pure balanced Group Interaction model, W is symmetric, $T = 2$, $n_1 = r$, $n_2 = r(m-1)$, $\omega_1 = 1$, $\omega_2 = -1/(m-1)$. Also, by direct computation, $\mathrm{tr}(G_z)/n = (rm)^{-1}\left[r/(1-z) - r(m-1)/\left(z+m-1\right)\right] = z/[(1-z)(z+m-1)]$, and hence $d_{11} = 2(m-1)(1-z)/[(1-\lambda)^2(z+m-1)]$ and $d_{22} = -2(z+m-1)/[(\lambda+m-1)^2(1-z)]$. The stated result follows from equation (4.14). □

Proof of Proposition 5.5. For a symmetric Complete Bipartite model $\mathrm{tr}(W S_z^{-1}) = -1/(1+z) + 1/(1-z) = 2z/\left(1-z^2\right)$, and hence $\gamma_1(z) = -[n-(n-2)z]/[n\left(1-z^2\right)]$, $\gamma_2(z) = -2z/[n\left(1-z^2\right)]$, and $\gamma_3(z) = [n+(n-2)z]/[n\left(1-z^2\right)]$. The stated result follows by using expression (4.12). □

Proof of Proposition 5.6. For $z \in (0,1)$ the density is obtained as an application of Lemma 8.2 in the Supplementary Material, with $\gamma = 1$, $\alpha = n-2$, $\beta = 1$, $a(z) = 2z/\phi_1$, and $c(z) = \phi_2/\phi_1$. The proof is completed on using Corollary 4.9. □

Proof of Proposition 5.7. For the row-standardised Complete Bipartite model the matrix H is

$$H = \begin{bmatrix} \iota_p/\sqrt{n} & \iota_p/\sqrt{n} & L_{p,p-1} & 0 \\ \iota_q/\sqrt{n} & -\iota_q/\sqrt{n} & 0 & L_{q,q-1} \end{bmatrix},$$

where $L_{p,p-1}$ $(p \times (p-1))$ satisfies $L'_{p,p-1}\iota_p = 0$ and $L'_{p,p-1}L_{p,p-1} = I_{p-1}$. Thus,

$$M = H' M_{\iota_n} H = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{4pq}{n^2} & 0 \\ 0 & 0 & I_{n-2} \end{bmatrix}.$$

This is certainly block-diagonal, as expected, and in addition the $(1,1)$ block also vanishes. The mean of $x = H^{-1}\tilde{y}$ is $\mathrm{E}(x) = \beta\sqrt{n}(n,0,0)'$. Therefore, from equation (4.9), we have

$$\Pr(\hat{\lambda}_{\mathrm{ML}} \leq z) = \Pr\left(d_{22}\chi_1^2 + d_{33}\chi_{n-2}^2 \leq 0\right),$$

i.e.,

$$\Pr\left(-2\left(\phi_2\chi_1^2 + 2z\chi_{n-2}^2\right) \leq 0\right).$$

But, if $z \geq 0$, both coefficients here are non-negative, so for $z \geq 0$, $\Pr(\hat{\lambda}_{\mathrm{ML}} \leq z) = 1$. This yields the result stated. □

Proof of Proposition 6.1. When $\lambda \in [0,1)$, G_λ can be expanded as $\sum_{r=0}^{\infty} \lambda^r W^{r+1}$, showing that nonnegativity of W implies nonnegativity of G_λ. Letting g_{ij} denote (i,j)-th entry of G_λ, we have

$$n\mathrm{tr}\left(G_\lambda^2\right) = n\sum_{i,j=1}^{n} g_{ij}g_{ji} = n\sum_{i=1}^{n} g_{ii}^2 + n\sum_{i,j=1,i\neq j}^{n} g_{ij}g_{ji}.$$

By the sum of squares inequality $n\sum_{i=1}^{n} g_{ii}^2 \geq \left(\sum_{i=1}^{n} g_{ii}\right)^2 = [\mathrm{tr}(G_\lambda)]^2$. Also, note that $n\sum_{i,j=1,i\neq j}^{n} g_{ij}g_{ji} > 0$ when $\lambda \in [0,1)$, by the nonnegativity of G_λ. It follows that, for any $\lambda \in [0,1)$, $n\mathrm{tr}\left(G_\lambda^2\right) > [\mathrm{tr}(G_\lambda)]^2$ or, equivalently, $\delta(\lambda) < 0$. As stated in the proof of Lemma 3.5, the first term on the right hand side of expression (A.1) is nonpositive by the Cauchy-Schwarz inequality. The second term is equal to $\delta(\lambda)$. Hence all stationary points of $l_p(\lambda)$ in $[0,1)$ must be maxima, which completes the proof. □

References

Arnold, S.F. (1979) Linear models with exchangeably distributed errors, *Journal of the American Statistical Association* 74, 194-199.

Audretsch, D.B. and Feldmann, M.P. (1996) R&D Spillovers and the Geography of Innovation and Production. *American Economic Review* 86, 630-640.

Baltagi, B.H. (2006) Random effects and spatial autocorrelation with equal weights, *Econometric Theory* 22, 973-984.

Bao, Y. (2013) Finite sample bias of the QMLE in spatial autoregressive models, *Econometric Theory*, 29, 68-88.

Bao, Y. and Ullah, A. (2007) Finite sample moments of maximum likelihood estimator in spatial models, *Journal of Econometrics* 137, 396–413.

Bell, K.P. and Bockstael, N.E. (2000) Applying the Generalized Method of Moments Approach to Spatial Problems Involving Micro-Level Data, *Review of Economic and Statistics* 82, 72-82.

Bertrand, M., Luttmer, E.F.P., and Mullainathan S. (2000) Network effects and welfare cultures, *Quarterly Journal of Economics* 115, 1019–1055.

Besley, T. and Case, A. (1995) Incumbent Behavior: Vote-Seeking, Tax-Setting, and Yardstick Competition, *American Economic Review* 85, 25-45.

Biggs, N.L. (1993) *Algebraic Graph Theory*, second edition. Cambridge University Press.

Bramoullé, Y., Djebbari, H. and Fortin, B. (2009) Identification of peer effects through social networks, *Journal of Econometrics* 150, 41-55.

Case, A. (1991) Spatial Patterns in Household Demand, *Econometrica* 59, 953-966.

Cliff, A.D. and Ord, J.K. (1973) *Spatial autocorrelation.* London, Pion

Cressie, N. (1993) *Statistics for spatial data.* Wiley,

Drton, M., Sturmfels, B. and Sullivant, S. (2009). *Lectures on Algebraic Statistics*, Birkhauser.

Eaton, M.L. (1989) *Group Invariance Applications in Statistics*, Institute of Mathematical Statistics and American Statistical Association.

Estrada, E., (2011) *The Structure of Complex Networks, Theory and Practice.* Oxford University Press.

Forchini, G. (2002) The exact cumulative distribution function of a ratio of quadratic forms in normal variables, with application to the AR(1) model, *Econometric Theory* 18, 823-852.

Forchini, G. (2005) The distribution of a ratio of quadratic forms in noncentral normal variables, *Communications in Statistics - Theory and Methods* 34, 999-1008.

Hillier, G.H. (2001) The density of a quadratic form in a vector uniformly distributed on the n-sphere, *Econometric Theory* 17, 1-28.

Hillier, G.H., Kan, R., and Wang, X. (2009) Computationally efficient recursions for top-order invariant polynomials, with applications, *Econometric Theory* 25, 211–242.

Hillier, G.H. and Martellosio, F. (2013). Exact Likelihood inference in the spatial group-interaction model. Manuscript.

Horn, R. and Johnson, C.R. (1985) *Matrix Analysis.* Cambridge University Press, Cambridge.

James, A.T. (1964) Distributions of matrix variates and latent roots derived from normal samples. *Annals of Mathematical Statistics* 35, 475–501.

Kelejian, H.H., I.R. Prucha, & Y. Yuzefovich (2006) Estimation problems in models with spatial weighting matrices which have blocks of equal elements, *Journal of Regional Science* 46, 507-515.

Kelejian, H.H., Prucha, I.R. (2010) Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Econometrics* 157, 53-67.

Koopmans, T.C. (1942) Serial correlation and quadratic forms in normal variables. *Annals of Mathematical Statistics*, 12, 14–33.

Lee, L.F. (2004) Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models, *Econometrica* 72, 1899–1925.

Lee, L.F. (2007) Identification and estimation of econometric models with group interactions, contextual factors and fixed effects, *Journal of Econometrics* 140, 333–374.

Lee, L.F., Liu, X., Lin, X. (2010) Specification and estimation of social interaction models with network structures, *The Econometrics Journal* 13, 145-176.

LeSage, J.P. and Kelley Pace, R. (2009) *Introduction to Spatial Econometrics*, Boca Raton, CRC Press/Taylor & Francis.

Liu, X., Patacchini, E., Zenou, Y. and Lee, L.F. (2012) Criminal networks: who is the key player?, working paper.

Martellosio, F. (2011) Efficiency of the OLS estimator in the vicinity of a spatial unit root. *Statistics & Probability Letters* 81, 1285–1291.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*, Wiley, New York.

Mulholland, H.P. (1965) On the degree of smoothness and on singularities in distributions of statistical functions, *Proceedings of the Cambridge Philosophical Society* 61, 721–739.

Ord, J.K. (1975) Estimation methods for models of spatial interaction, *Journal of the American Statistical Association* 70, 120–126.

Pinkse, J., Slade, M.E. and Brett, C. (2002) Spatial Price Competition: A Semiparametric Approach, *Econometrica* 70, 1111-1153.

Ross-Parker, H. (1975) Inter-plant competition models and their sampling distributions, *Advances in Applied Probability* 7, 453-4.

Rothenberg, T.J. (1971) Identification in Parametric Models, *Econometrica* 39, 577-591.

Saldanha, N.C. & C. Tomei (1996) The accumulated distribution of quadratic forms on the sphere, *Linear Algebra and Its Applications,* 245, 335–351.

Topa, G. (2001) Social interactions, local spillovers and unemployment, *The Review of Economic Studies* 68, 261-295.

Yang, Z. (2013) A general method for third-order bias and variance corrections on a nonlinear estimator, manuscript, Singapore Management University.

von Neumann, J. (1941) Distribution of the ratio of the mean square successive difference to the variance, *Annals of Mathematical Statistics* 12, 367–395.