



QML estimation with non-summable weight matrices

Jakub Olejnik¹ · Alicja Olejnik²

Received: 15 October 2018 / Accepted: 8 June 2020 / Published online: 4 July 2020
© The Author(s) 2020

Abstract

This paper revisits the theory of asymptotic behaviour of the well-known Gaussian Quasi-Maximum Likelihood estimator of parameters in mixed regressive, high-order autoregressive spatial models. We generalise the approach previously published in the econometric literature by weakening the assumptions imposed on the spatial weight matrix. This allows consideration of interaction patterns with a potentially larger degree of spatial dependence. Moreover, we broaden the class of admissible distributions of model residuals. As an example application of our new asymptotic analysis we also consider the large sample behaviour of a general group effects design.

Keywords Spatial autoregression · Quasi-maximum likelihood estimation · High-order SAR model · Asymptotic analysis · Non-summable matrices

JEL Classification C21 · C23 · C51

1 Introduction

It is a broadly employed assumption in a wide range of theoretical studies on spatial econometrics that the spatial weight matrix is absolutely row and column summable. This restriction is mostly a result of the Central Limit Theorem (CLT) used in the derivation of the result on asymptotic behaviour. Historically, it can be traced to the works of Kelejian and Prucha, e.g. Kelejian and Prucha (2001), who were

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s10109-020-00326-2>) contains supplementary material, which is available to authorized users.

✉ Jakub Olejnik
jakub.olejnik@wmii.uni.lodz.pl

Alicja Olejnik
alicja.olejnik@uni.lodz.pl

¹ Faculty of Mathematics and Computer Science, University of Lodz, Lodz, Poland

² Faculty of Economics and Sociology, University of Lodz, Lodz, Poland

first to formulate their assumptions as explicit requirements regarding the spatial weight matrix. Their CLT, which turned out to be a milestone in the development of asymptotic theories for spatial econometric models, relies on the absolute summability of the weight matrix involved. In our study we attempt to reconsider this approach, and we focus specifically on the Quasi Maximum Likelihood (QML) estimator for the spatial autoregressive model. By revisiting the classical argument of Lee (2004) and, importantly, introducing a generalised CLT for linear-quadratic forms, we are able to provide a theory for consistency and asymptotic normality of QML estimates for high-order spatial autoregressive models under relaxed conditions. In particular, our approach allows for spatial weight matrices that, even if row-standardised, may not be absolutely column summable.

The standard approach, with the absolute summability requirement on the weight matrix, undoubtedly has the appeal of a simple, self-contained theory. Although there might be a perception that the constraint is necessary for showing the desired asymptotic behaviour of various estimation schemes, it has already been recognised that this is not the case, see, e.g. Gupta and Robinson (2018). Indeed, the boundedness of row and column sums can be replaced with the less restrictive requirement of boundedness of spectral norms.¹

Unfortunately, the standard analysis excludes some of the spatial interaction patterns in which the number of spatial units influenced by any given unit grows to infinity. In particular, under infill asymptotics, if spatial units are assumed to interact with other units within a given distance, then the number of nonzero elements in a row or column of the spatial weight matrix grows with the sample size, leaving the potential for either row or column non-summability. Similarly, under increasing domain asymptotics certain spatial weight matrices based on inverted distance also lead to non-summable interaction patterns.²

Theoretical considerations can also reveal another limitation of the standard analysis. For example, consider the case of an initial model specification which is subjected to a transformation (e.g. linear filtering or demeaning) to obtain its final estimated form. Under the standard analysis, it is necessary to ensure that the applied transformation preserves the summability of rows and columns of the spatial weight matrix. As a result, the requirements of the standard analysis narrow the class of possible transformations of the model. A somewhat similar problem occurs if the so-called linear structure representation for model innovations is assumed.³ Then, following the standard approach, the derivation of the asymptotic distribution requires additional restrictions on the class of possible linear relations involved.

This paper provides a positive solution to a problem left by Gupta and Robinson (2018), who have made the first attempt to replace the requirement

¹ In particular, Assumption 4 in this paper implies that the rows and columns of an admissible spatial weight matrix could be square summable rather than absolutely summable.

² A number of more specific examples, in which standard asymptotic analysis does not yield a limiting distribution, are given in Sect. 2.1.

³ See, e.g. Robinson (2011) or Delegano and Robinson (2015).

of uniform summability of the spatial weight matrix with boundedness in the spectral norm. In that earlier work, extending the scope of spatial weight matrices beyond the standard asymptotic analysis of Lee (2004) was found to be useful, and results analogous to our Theorem 1 on consistency were independently obtained. However, their derivation of the asymptotic distribution of the estimates still relies on the assumption of absolute summability. Whether a relevant asymptotic distribution theorem under relaxed conditions is possible has been left as an open question.

Therefore, the aim of this paper is to present a refinement to the asymptotic analysis of the Gaussian Quasi-Maximum Likelihood (QML) estimator for high-order, spatial autoregressive models, considering the assumptions imposed on the spatial weight matrix. We further present an example of a possible application of our theorems. To this end, we develop a general group effects, high-order Spatial Autoregressive (SAR) model. Our approach to eliminating the effects components from the spatial process generalises that of Lee and Yu (2010a, b) and Lee et al. (2010), as well as Olejnik and Olejnik (2017), to the high-order autoregressive case. Finally, we present statements on consistency and asymptotic normality of the resulting QML estimator, which would not be possible to obtain with the standard argument.

The paper is organized as follows. Section 2 describes the motivation for addressing the topic. Section 3 presents our statements on the consistency and asymptotic normality of the Gaussian QML estimator. Finally, Sect. 4 develops an estimator for a high-order, spatial autoregressive, general group effects model, together with an analysis of its large sample behaviour. Appendices contain some details of the proofs, as well as a set of Monte Carlo simulations that empirically demonstrate the validity of the theory under the relaxed conditions.

2 Motivation for the refined asymptotic analysis

This section presents some basic examples of the application of our asymptotic analysis. First we present a class of spatial interaction schemes which cannot be handled by the standard asymptotic analysis of Lee (2004). Then, we describe a class of theoretical arguments for which the new analysis demonstrates a clear advantage over the standard approach. We also discuss the assumptions made in relation to the error term. We conclude the section with a brief discussion on whether our results may be considered optimal.

First, however, let us introduce the notation used in this paper when referring to norms. Unless stated otherwise, vectors, i.e. elements of \mathbb{R}^m , are column vectors. For an arbitrary row or column vector x the symbol $\|x\|$ denotes the usual Euclidean vector norm, which will also be called the l_2 norm. The quantity $\sum_{i=1}^m |x_i|$ is referred to as the l_1 norm. The same symbol, when used for matrices, denotes the induced spectral norm. That is, for a matrix A , the quantity $\|A\|$ is the largest singular value of A . For square matrices A we will also use the Frobenius norm $\|A\|_{\mathcal{F}}$, the maximum absolute column sum norm $\|A\|_1$, and the absolute row sum norm $\|A\|_{\infty}$.

2.1 Elementary examples of non-summable interaction patterns

One essential feature of an asymptotic theory for spatial econometric models is the set of assumptions imposed on the spatial weight matrix. These assumptions limit the amount of spatial interactions to a manageable degree, such that the statements on the large sample behaviour of the estimation scheme under question remain valid. It is a widely adopted standard in contemporary spatial econometrics to require that the spatial weight matrix is row and column summable. More precisely, with the conventional notation that W_n is the spatial weight matrix for the sample size $n \in \mathbb{N}$, it is required that the quantities $\|W_n\|_1$ and $\|W_n\|_\infty$ are both uniformly bounded in n . These conditions turn out to be unnecessarily restrictive in limiting the scope of spatial interactions that can be incorporated in an econometric model.

Let us start with a theoretical remark. Notice that the row and column summability of W_n is equivalent to the rows and columns constituting a bounded set in l_1 . However, a reader familiar with the ubiquitous nature of the theory of square-summable functions and sequences in much of the mathematical econometrics and geostatistics literature⁴ might expect that the l_2 norm would play a major role in the asymptotic theory—at least for some simple cases. In fact we find that the connection between the sharp boundedness condition for an asymptotic theory and square-summability is more nuanced. Instead of examining the properties of rows and columns of the spatial weight matrices, it is necessary to consider the sequence of the respective spectral norms $\|W_n\|$. The requirement then turns out to be the boundedness of the sequence.⁵ This will be discussed in subsequent sections.

The following motivating examples are connected to the class of Inverse Distance Weighting (IDW) interaction schemes, that find a wide range of uses in spatial econometrics, and other quantitative methods of geography. Let us assume that the strength of spatial interaction, represented by the spatial weights, is of the form $w_{ij} = \frac{1}{\text{dist}(i,j)^\alpha}$, where $\alpha > 0$ is a parameter and $\text{dist}(i,j)$ is a measure of the distance between units i and j . A question then arises: what are the values of α for which an individual row or column of the matrix $W = [w_{ij}]$ satisfies the requirement of boundedness imposed by the standard analysis? That is, for which values of α is it absolutely summable?

The answer to this question will depend on the nature of the spatial domain and the type of asymptotic scheme employed. We focus our attention here on the increasing domain asymptotics, since that is the more natural asymptotic scheme in this context. Let us now consider a one-dimensional spatial domain in which the spatial units are more or less evenly spaced, or at least the distance between each pair of consecutive units does not exceed a constant distance $D > 0$. Then, for an

⁴ Square summability leads to ideas that are fundamental in those fields, such as orthogonality and projections.

⁵ That is to say the ultimate optimality of the boundedness condition for the asymptotic theory requires that the spatial weight matrix is regarded as a l_2 space operator rather than a numerical table.

arbitrary unit j the column sums (and, by symmetry, row sums as well) of such a matrix $W_n = [w_{ij}]_{i,j \leq n}$ satisfy the bound

$$\sup_{n \in \mathbb{N}} \|W_n\|_\infty \geq \sup_{n \in \mathbb{N}} \sum_{i=1}^n w_{ij} = \sup_{n \in \mathbb{N}} \sum_{i=1}^n \frac{1}{\text{dist}(i,j)^\alpha} \geq \lim_{n \rightarrow \infty} \frac{2}{D^\alpha} \sum_{i=1}^n \frac{1}{i^\alpha}. \quad (1)$$

The right-hand side in Eq. (1) converges if and only if $\alpha > 1$. Let us note that the condition on the j -th row/column square summability would, in turn, be satisfied if the series $\frac{2}{D^{2\alpha}} \sum_{i=1}^\infty \frac{1}{i^{2\alpha}}$ converges, i.e. if $\alpha > \frac{1}{2}$.

Non-summable distance-related matrices were also investigated by Lee (2002) in the context of the Ordinary Least Squares (OLS) estimator. In that work, the complementarity condition, that is $\alpha \leq \frac{1}{2}$, was derived as necessary and sufficient for consistency and asymptotic normality (at the rate \sqrt{n}) of the OLS estimator if the matrix of inverse distances is row-normalised prior to being used in the model. With $\alpha \leq \frac{1}{2}$ the amount of spatial interaction (in each row) is intractable for Maximum Likelihood (ML) estimation. However, dividing elements of the weight matrix by the increasing row sums leads to $\max_{i,j \leq n} w_{ij} \rightarrow 0$ and $\frac{1}{n} \text{tr}(G_n) \rightarrow 0$, with $G_n = W_n(I_n - \lambda_0 W_n)^{-1}$ and λ_0 being the autoregressive parameter.⁶ According to Lee (2002), the dominant component of the bias for the OLS estimator is proportional to $\frac{1}{n} \text{tr}(G_n)$. That is to say, the row-normalisation reduces the spatial dependence to the extent that the OLS estimator becomes consistent. It should be noted that the results of Lee (2002) still require the spatial weight matrix (after row-normalisation) to be summable in both rows and columns. Let us make clear that in our argument we do not assume that the procedure of row-normalisation is applied to the spatial weight matrix, for the reasons presented in Sect. 2.2.

Although the condition of summability $\alpha > 1$ derived from Eq. (1) is not too restrictive, it is also fair to say that such a linear domain is rarely encountered in practice. If we now extend this example to higher dimensions, the restriction on α also changes. To see this, let us assume that the spatial domain is, a two-dimensional plane, as it is in majority of economic applications.⁷ Now, the crucial observation is the following. Let us imagine a circle of radius δ around a fixed spatial unit. The number of spatial units which are intersected by the circle is roughly proportional⁸ to its circumference, $2\pi\delta$. Similarly, for any given unit j on the plane, and any radius δ , the number $\eta(j, \delta)$ of units i for which $\delta - 1 < \text{dist}(i, j) \leq \delta$ is roughly proportional to δ . Thus, we have

⁶ Since $I_n + \lambda_0 G_n = (I_n - \lambda_0 W_n)^{-1}$, the value of $\frac{\lambda_0}{n} \text{tr}(G_n)$ may indicate the component of the average direct effect due to the feedback loop in spatial interactions, c.f. Le Sage and Pace (2009).

⁷ To maintain mathematical rigour, we may also have to assume that the distribution of the spatial units is neither pathologically dense nor sparse anywhere in the spatial domain. This is actually the case in all realistic settings. In particular, we might expect that the distances between nearby units are within a fixed range $D_1 \leq \text{dist}(i, j) \leq D_2$ or, alternatively, the units are regions with areas in a given range $A_1 \leq \text{area}(i) \leq A_2$. For example, tessellations obtained by generating a Voronoi diagram of a randomly distributed set of points constitute an adequate framework.

⁸ The actual constants would possibly involve D_1 , D_2 , A_1 , A_2 , see footnote 7.

$$\sup_{n \in \mathbb{N}} \|W_n\|_{\infty} \geq \sup_{n \in \mathbb{N}} \sum_{i=1}^n w_{n,ij} \geq \sum_{\delta=1}^{\infty} \frac{\eta(j, \delta)}{\delta^{\alpha}} \approx \sum_{\delta=1}^{\infty} \frac{1}{\delta^{\alpha-1}}. \quad (2)$$

Again, the series on the right hand side converges only if $\alpha > 2$. Thus, in most cases for IDW spatial models, unless the exponent α is well greater than two, the standard asymptotic analysis fails as the spatial weight matrix is simply not summable. This fact is apparently not widely known as it is still common to see inverse distance decay parameters, either set a priori or estimated, lie in $[0, 2]$, see Anselin (2002). In particular, this also applies to the case of $\alpha = 2$, which constitutes a popular econometric analogue of the Newtonian gravity model.⁹ In this planar case, an IDW row or column would be square summable if only $\alpha > 1$. Going further, it might be argued that in the less common, but still relevant¹⁰, three-dimensional domain the summability restriction on alpha becomes $\alpha > 3$, whereas row and column square-summability requires $\alpha > \frac{3}{2}$.¹¹

Limitations of the summability requirement of the standard asymptotic analysis are not only encountered in the case of increasing domain asymptotics. In fact, the situation is even worse, for infill asymptotics, as by definition, new neighbours are allowed to emerge within any radius about a given unit. Then, in an extreme case, even the simple common border spatial weight matrix may not be summable. Establishing asymptotics of a model based on such an interaction scheme is therefore highly problematic. Summability issues can also be found in interaction models of a non-geographical nature. As an example, consider a social networking model, where the relation of “friendship” in an online social networking service represents contiguity, and the distance is measured in terms of the contiguity path between given pair of individuals. Then, the average distance δ_n between members an n element group is expected to grow slowly with n . This results in a behaviour similar to infill asymptotics. In other words, the quantity $\eta(j, \delta)$ is expected to grow very rapidly (cf. Eq. (2)), similarly as it would be the case in a high dimensional space.

2.2 Applicability of the new asymptotic analysis

This section explains the circumstances under which our new asymptotic analysis applies to a non-summable interactions scheme. As we noted in the previous section, a sharp condition on boundedness requires a nuanced approach. Although it can be shown that a matrix is bounded in spectral norm only if its rows and columns are square summable, those two classes of matrices are not necessarily coincident. Unfortunately, there is no straightforward test which could be applied to matrix entries to decide whether its spectral norm is bounded. Nevertheless, we

⁹ It is also common for statistical software to provide the functionality of generating and using such spatial weight matrices without any warning.

¹⁰ One might consider autoregressive models of natural phenomena in environmental sciences, where an additional dimension may be present, e.g. depth, altitude etc.

¹¹ In general, in an m dimensional domain $\eta(j, \delta)$ is proportional to δ^{m-1} , thus the two condition are $\alpha > m$ and $\alpha > \frac{m}{2}$, respectively.

formulate some general suggestions on the use of non-summable matrices in model specifications.

For a square-summable matrix to be bounded in spectral norm it is sufficient that one of the two additional conditions is met. The first condition is when the number of rows and columns which are not absolutely summable is finite. This case is relevant to models that distinguish a set of units, called economic “centres of gravity”, for which the amount of spatial interaction is possibly non-summable. Similarly, a social networking model might distinguish a set of “leader/influencer” individuals with non-summable impact on other members of a group. In those cases the resulting matrices are bounded in spectral norm.

If the number of non-summable rows and columns is infinite or, in particular, all of them are non-summable, then the matrix still can be bounded in spectral norm. It is the case if those non-summable rows and columns are in a sense asymptotically not strongly correlated or, in other words, the corresponding weightings are not too similar. Unfortunately, this formulation is not at all precise, and thus for such interaction schemes, we suggest applying a rescaling factor as described below. Importantly, this operation preserves the structure of the spatial interdependence expressed by relative magnitudes between all weights in the matrix, e.g. Elhorst (2001) or Vega and Elhorst (2015).

Although rescaling is similar to the familiar procedure of row-normalisation, there are important differences that need to be highlighted. We note that row-normalisation is typically applied for the following three reasons. Firstly, almost by definition, it normalises the amount of spatial interaction received by each of the spatial units. This is believed to facilitate the interpretation of the autoregressive parameter. Secondly, together with non-negativity and a zero diagonal, the Greshgorin theorem¹² implies that the space for the autoregressive parameter can contain any compact¹³ subset of $(-1, 1)$. Lastly, the procedure trivially assures summability of rows, which is a part of the boundedness assumption of the standard analysis.

Although, normalisation of rows is indeed beneficial for interpretation in a variety of contexts, e.g. common border or k -nearest neighbours schemes, it may also be harmful in certain interaction patterns. In particular, as emphasised in (Vega and Elhorst, 2015, pp. 355) after (Anselin, 1988, pp. 23–24), and Kelejian and Prucha (2010), “row-normalising a weights matrix based on inverse distance causes its economic interpretation in terms of distance decay to no longer be valid”. Secondly, as the maximal modulus of an eigenvalue does not exceed spectral norm, a matrix rescaled by $\|W_n\|^{-1}$ allows any value in $(-1, 1)$ in its parameter space. Lastly, row-normalisation does not in general assure column summability required by the standard analysis, whereas rescaling by $\|W_n\|^{-1}$ produces a matrix with unit spectral norm.

¹² The Greshgorin circle theorem states that for any matrix $A = [a_{ij}]_{i,j \leq n}$ and any its eigenvalue v there is an $i \leq n$ such that $|v - a_{ii}| \leq \sum_{j \neq i} |a_{ij}|$. It follows that $\det(I - \lambda W_n) = 0$ only if $v = \frac{1}{\lambda}$ is an eigenvalue of $A = W_n$, thus only if $|\frac{1}{\lambda} - 0| \leq 1$, i.e. $\lambda \geq 1$.

¹³ It is prudent to recall that the parameter space for ML-type estimation should be compact.

The merits of the rescaling procedure have long been recognised and applied in practice. For example Elhorst (2001) and Vega and Elhorst (2015) rescale weight matrices by their largest characteristic roots. Notice that in the case of a symmetric weight matrix the root is equal to the spectral norm. However, it is not true in general, not even for arbitrary symmetric matrices, that such rescaling assures row and column summability. This makes our asymptotic theory necessary to justify inferences from such a model.

Although the described procedure may be used to rescale an arbitrary matrix, this does not imply that estimation is possible with any spatial weight matrix. A cautionary example has been given in Lee (2004) of a matrix $W_n = \frac{1}{n-1} (t_n t_n^T - I_n)$, where $t_n = (1, \dots, 1)^T$ is an $n \times 1$ vector of ones and I_n is the identity matrix. In this case, the ML estimator may be inconsistent, even though W_n is absolutely summable in both rows and columns. This is explained by Mynbaev and Ullah (2008), who analyse a class of weight matrices, of which W_n is a member. For matrices in the class, the identification condition fails. It is not clear to us whether the OLS estimator would be consistent in this case, and nor do the results of Lee (2002) seem to provide the answer. A related matrix $W_{n,m} = I_m \otimes W_n$ has also been considered in, e.g. Case (2015a, b), Lee (2004), and Lee (2007b). However, in the case of $W_{n,m}$, if m grows at a sufficient rate, then favourable asymptotic properties can be assured.

Mynbaev and Ullah (2008), and Mynbaev and Ullah (2010) study a class of weight matrices which approximate a kernel of an integral operator on the space of square-integrable functions. These may be related to the class of square-summable matrices. However, their assumptions, in particular the absolute summability of operator eigenvalues, preclude the use of ML estimation. In particular, such weight matrices contain an insufficient amount of information for identification of the autoregressive parameter. Instead, the asymptotic behaviour of the OLS estimator is investigated.

2.3 Applicability in theoretical arguments

In this section we argue that the consideration of a wider class of spatial weight matrices can also be beneficial in theoretical arguments when developing new model specifications. Let us start with a rather simplified description of one possible application. Let T_n be an operator on \mathbb{R}^n and let us assume that it is invertible (although in practice this is often not the case). Given a specification of a spatial model, for example the SAR specification $Y_n = \lambda W_n Y_n + X_n \beta + \varepsilon_n$, we might be interested in its transformed form $T_n Y_n = \lambda (T_n W_n T_n^{-1}) T_n Y_n + T_n X_n \beta + T_n \varepsilon_n$. Then it is natural to ask what transformations T_n are known to preserve the required properties of the spatial weight matrix, so that the asymptotics of the transformed model could follow easily from the properties of the original specification. In the case of the standard asymptotic analysis of Lee (2004), this means that we want the matrix $T_n W_n T_n^{-1}$ to be row and column summable whenever W_n is. Practically, this limits the class of possible matrices T_n to those which are themselves summable, i.e. whose entries we can and must finely control. In that respect, the standard analysis collapses, even in the simple case of T_n being an isometry, i.e. an orthogonal matrix.

Our refined asymptotic analysis, on the other hand, calls for the operator norm of T_n to be bounded. In many cases this is easily satisfied as T_n is often a projection, and thus an operator of norm one, for which we have $\|T_n W_n T_n^+\| \leq \|W_n\|$. We note that, although projections are generally not invertible, in practice we may still be able to exploit the fact that the Moore-Penrose inverse T_n^+ is an isometry on the range of T_n .

A more concrete example can be derived from the work of Lee and Yu (2010a, b), where the incidental parameter problem is addressed in the context of a spatial autoregressive panel model and fixed spatial and temporal effects. In this case, the natural candidate for T_n is the demeaning operator, i.e. the projection on the space of zero-mean vectors within units and time periods. A similar idea is applied in Lee et al. (2010) for group effects¹⁴ in social interaction models. With this technique, the expected multiplicative bias correction is derived for consistent estimates. Those results strongly rely on the summability of the demeaning operator T_n . However, incorporation of arbitrary group effects, where groups are not necessarily disjoint and the number of groups may increase with n , leads to a demeaning operator which does not have to be summable. We show in Sect. 4 that such generalisation is possible with the refined asymptotic analysis.

2.4 On the distribution of innovations

A significant merit of the original paper of Lee (2004) is the consideration of the QML estimation scheme rather than the classical maximum likelihood estimator with the assumption of Gaussian errors. That is to say, the error term in the spatial autoregressive model is allowed to be an arbitrary random vector of independent and identically distributed components, as long as the shared distribution allows moments of order higher than four. This seemingly technical improvement has substantially changed how the validity of inference may be perceived, as it no longer has to rely on a belief in the joint normality of errors.

Similarly, with heterogeneous processes governing the underlying phenomena across spatial units, the expectation of identical distribution of disturbances does not appear to be well-founded. In our analysis, the components of the error term are, for purely technical reasons, assumed to be homogeneous in terms of variance. However, they are not required to follow the same probability law.

Lastly, we have found that the universally employed requirement that the distribution of the residuals should have finite moments higher than four is not necessary. In fact, integrability with the fourth power is sufficient to obtain the Gaussian asymptotics for QML estimates. This result opens the possibility of considering heavier tailed probability laws. For example, one might consider distribution functions with tails of order $x^{-5} \log^{-2} x$, as x grows to infinity.

¹⁴ Notice that group and fixed effects are both algebraically equivalent to discriminating certain subsets of observations with respect to the constant term in the model.

2.5 Ultimate optimality

It can be argued that our boundedness condition with the spectral norm (Assumption 4) is already optimal. That is to say, the class of spatial weight matrices considered, in general, cannot be broadened. Indeed, if we consider a matrix W_n for which $\lim_{n \rightarrow \infty} \|W_n\| = \infty$, invalidating our assumptions, then we arrive at the rather uninteresting case of the parameter space of Λ not containing any positive elements. To see this, consider the example of a symmetric spatial weight matrix W_n with non-negative entries. Then, $\|W_n\|$ is the maximum eigenvalue and, by a well-known argument, any interval $[0, t) \subset \Lambda$ satisfies $t \leq \frac{1}{\|W_n\|} \rightarrow 0$.

Obtaining a Gaussian asymptotic distribution for the estimates requires the use of a CLT. Specifically, it is applied for a random variable Q_n which is a linear-quadratic form of the residual. If the fourth moments of ε_n were not finite, then the normalising factor for Q_n , namely $\sqrt{\text{Var } Q_n}$, would also be infinite. This seriously compromises any effort to derive the limiting distribution. As a result, we believe that any substantial generalisation is unlikely.

3 Revisiting asymptotic analysis of high-order SAR models

Let $W_{n,1}, \dots, W_{n,d}$ be arbitrary $n \times n$ matrices.¹⁵ We consider the high-order SAR model described by the following equation

$$Y_n = \sum_{r=1}^d \lambda_r W_{n,r} Y_n + X_n \beta + \varepsilon_n, \quad (3)$$

where $\lambda = (\lambda_r)_{r=1}^d \in \Lambda \subset \mathbb{R}^d$ and $\beta \in \mathbb{R}^k$. Furthermore, Y_n is a vector of n observations on the dependent variable, X_n is the matrix of k non-stochastic explanatory variables and ε_n is the error term, for which the assumptions given below hold.

Assumption 1 The matrix $X_n^T X_n$ is invertible, and both $\left\| \frac{1}{n} X_n^T X_n \right\|$ and $\left\| \left(\frac{1}{n} X_n^T X_n \right)^{-1} \right\|$ are uniformly bounded in $n \in \mathbb{N}$.

Let us note that Assumption 1, used for the consistency argument, does not require the sequence $\frac{1}{n} X_n^T X_n$ to be convergent. Instead, our reasoning stipulates that this sequence is merely bounded.¹⁶ The necessity of non-singularity of $X_n^T X_n$ is straightforward, as regressors should not be linearly dependent for the slope parameter to be identifiable. Note that our assumption that $\|(X_n^T X_n)^{-1}\| = O(\frac{1}{n})$ is not

¹⁵ In practice, the spatial weight matrices typically have non-negative elements and zero diagonals to facilitate interpretation of the autoregressive parameter. However, for the purpose of the argument of Sect. 4, we deliberately do not make this assumption.

¹⁶ This does not imply that elements of X_n should be bounded in absolute value, as is explicitly assumed in e.g. Lee (2004) or Lee and Yu (2010a).

far from the well-known condition necessary for consistency of the OLS estimator for non-spatial regression, i.e. $\|(X_n^T X_n)^{-1}\| = o(1)$.¹⁷

Assumption 2 One of the following holds¹⁸

- (a) $\varepsilon_n = (\varepsilon_{n,i})_{i \leq n}$ is a vector of independent random variables with zero mean, variance σ^2 and uniformly bounded fourth moments,
- (b) ε_n is of the form $\varepsilon_n = F \bar{\varepsilon}_m$ where F is an $n \times m$ matrix with orthogonal rows of norm one; the underlying $\bar{\varepsilon}_m$ is a vector of independent random variables with zero mean, variance σ^2 and uniformly bounded fourth moments.

Assumption 3 For every λ in respective parameter space $\Lambda \subset \mathbb{R}^d$ the matrix $\Delta_n(\lambda) = I_n - \sum_{r \leq d} \lambda_r W_{n,r}$ is invertible.

We investigate the asymptotic behaviour of the widely applied Gaussian QML estimator, which maximises the log-likelihood of the observed sample as if the model innovations were Gaussian, namely the function

$$\ln L_n(\theta) = -\frac{n}{2} \ln(2\pi\sigma^2) + \ln |\det \Delta_n(\lambda)| - \frac{1}{2\sigma^2} \|\Delta_n(\lambda)Y_n - X_n\beta\|^2, \quad (4)$$

where $\theta = (\lambda^T, \beta^T, \sigma^2)^T$ is the model parameter. It turns out that, under certain regularity conditions, this estimation scheme is consistent, even if the model residuals do not follow the normal distribution (c.f. Assumption 2). The estimator will be denoted $(\hat{\lambda}_n^T, \hat{\beta}_n^T, \hat{\sigma}_n^2)^T$ or $\hat{\theta}_n$. Our result on the asymptotic behaviour of $\hat{\theta}_n$ requires the following boundedness assumption, which gives the essential condition imposed on a spatial weight matrix.

Assumption 4 The set Λ is compact in \mathbb{R}^d . There exists a universal constant K_Λ such that for all $n \in \mathbb{N}$, $\lambda \in \Lambda$, $r = 1, \dots, d$ the matrix norms $\|W_{n,r}\|$ and $\|\Delta_n(\lambda)^{-1}\|$ do not exceed K_Λ .

Notice that any matrix with absolutely summable rows and columns is also bounded in the spectral norm.¹⁹ That is to say, the asymptotic theory presented in this paper is indeed a generalisation of the theory initiated in Lee (2004). Moreover,

¹⁷ For any non-negative sequences (a_n) , (b_n) and (c_n) such that $a_n \leq b_n c_n$, we write $a_n = O(b_n)$ when $\limsup_{n \rightarrow \infty} c_n < \infty$, and $a_n = o(b_n)$ whenever $\limsup_{n \rightarrow \infty} c_n = 0$.

¹⁸ In particular, elements of the error term do not need to be identically distributed. Also notice that trivially (a) implies (b) for $F = I_n$, $m = n$. Although condition (a) is simpler and sufficient in a standard setting, the relaxed condition (b) will be crucial in the argument of Sect. 4 when an independent vector of residuals will be transformed by such a matrix F . Note that (b) implies $\mathbb{E}\varepsilon_n = 0$ and $\mathbb{E}\varepsilon_n \varepsilon_n^T = \sigma^2 I_n$, thus its components may be merely uncorrelated. We emphasise the distinction as the innovations are not assumed to be Gaussian.

¹⁹ This follows the fact that $\|A\|^2 \leq \|A\|_1 \|A\|_\infty$, for an arbitrary matrix A .

it is also a proper generalisation, as there are non-summable interaction schemes bounded in the spectral norm. Remark 1 in the appendix describes an example of such a weight matrix, which is additionally row-standardised. This also shows that row-normalisation, in general, does not ensure absolute summability of columns.

The following identification assumption guarantees that the Gaussian QML estimator is able to asymptotically identify the true value of the spatial autoregressive parameter.

Assumption 5 For every $\lambda_1, \lambda_2 \in \Lambda$, such that $\lambda_1 \neq \lambda_2$, at least one of the statements (a) or (b) is satisfied:

- (a) $\liminf_{n \rightarrow \infty} \frac{1}{\sqrt{n}D_n(\lambda)} \|\Delta_n(\lambda_1)\Delta_n(\lambda_2)^{-1}\|_{\mathcal{F}} > 1$,
- (b) $\liminf_{n \rightarrow \infty} \frac{1}{\sqrt{n}} \|M_{X_n} \Delta_n(\lambda_1)\Delta_n(\lambda_2)^{-1}X_n\beta\| > 0$, for every $\beta \in \mathbb{R}^k$,

where $D_n(\lambda) = \left| \det(\Delta_n(\lambda_1)\Delta_n(\lambda_2)^{-1}) \right|^{1/n}$, $M_{X_n} = I_n - X_n(X_n^T X_n)^{-1}X_n^T$.

Assumption 5 is typically called the identification condition. It ensures that there is enough information in the observed process to decrease uncertainty of $\hat{\lambda}_n$, with increasing n . The distinction between the statements (a) and (b) reflects the fact that this information can come from either the spatial autocorrelation of Y_n or via the accumulated spatial lag of regressors.

Theorem 1 Under Assumptions 1–5 the Gaussian QML estimator $\hat{\theta}_n$ is consistent.

In order to establish the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ we need to adopt a number of additional assumptions. Let $\Xi \subset \prod_{n=1}^{\infty} \mathbb{R}^n$ be the linear space²⁰ of all sequences $(x_n)_{n \in \mathbb{N}}$, with $x_n = (x_{n,i})_{i \leq n} \in \mathbb{R}^n$, $n \in \mathbb{N}$, for which $\max_{i \leq n} x_{n,i}^2 = o(n)$. Additionally, let us set $G_{n,r} = W_{n,r}\Delta_n(\lambda_0)^{-1}$ for $r \leq d$.

Assumption 1' The earlier Assumption 1 on X_n is satisfied. Moreover, each column of the matrices X_n and $G_{n,r}X_n\beta_0$, $r \leq d$, is a member of the linear space Ξ .

The above is a technical assumption necessary for obtaining asymptotic normality of the deviation $\sqrt{n}(\hat{\theta}_n - \theta_0)$. Intuitively, the limiting distribution can be normal regardless of the original distribution of ε_n only when none of the observations within the regressor matrices makes an overwhelming contribution to the estimate of the corresponding slope coefficient. Let us emphasise that this assumption is also necessary in the simple case of non-spatial least squares regression. Although implicitly, it is also present in the standard analysis as a consequence of other assumptions adopted therein, in particular, the boundedness of elements of X_n .²¹

²⁰ Naturally, the set $\prod_{n=1}^{\infty} \mathbb{R}^n = \mathbb{R} \times \mathbb{R}^2 \times \dots$ is a vector space when endowed with element-wise addition. Then, Ξ is its linear subspace.

²¹ Under the assumption of boundedness of elements of X_n , as in Lee (2004) and Lee and Yu (2010a), Assumption 1' is implied by the relation $\|G_{n,r}\|_1 = o(n)$ or, if $W_{n,r}$ is row-normalized, $\|\Delta_n(\lambda_0)^{-1}\|_1 = o(n)$.

Assumption 2' The error term satisfies (a) in Assumption 2. Moreover, the family of random variables $\varepsilon_{n,i}^4$, $n \in \mathbb{N}$, $i \leq n$, is uniformly integrable.

Derivation of the asymptotic distribution requires strengthening of the Assumption 2. However, in Assumption 2' the elements of the error term still do not need to follow the same distribution. Instead, we impose the requirement that those distributions have uniformly integrable tails in terms of the fourth moment.

Assumption 6 The true value of parameter λ lies in the interior of the space Λ , that is $\lambda_0 \in \text{Int } \Lambda$.

Assumption 7 For the matrices $\mathfrak{F}_n = -\mathbb{E}_{\theta_0} \frac{1}{n} \frac{\partial^2 \ln L_n}{\partial \theta \partial \theta^T}(\theta_0)$ and $\Sigma_{\mathcal{S},n} = \mathbb{E}_{\theta_0} \frac{1}{n} \mathcal{S}_n^T \mathcal{S}_n$, where $\mathcal{S}_n = \frac{\partial \ln L_n}{\partial \theta}(\theta_0)$, $n \in \mathbb{N}$, the following limits exist: $\mathfrak{F} = \lim_{n \rightarrow \infty} \mathfrak{F}_n$ and $\Sigma_{\mathcal{S}} = \lim_{n \rightarrow \infty} \Sigma_{\mathcal{S},n}$.²² Moreover, the matrix \mathfrak{F} is non-singular.

Assumption 7 spells out the necessary conditions for the existence of the limiting distribution variance. Note that for consistency of $\hat{\theta}_n$ the sequences $(\mathfrak{F}_n)_{n \in \mathbb{N}}$, $(\Sigma_{\mathcal{S},n})_{n \in \mathbb{N}}$ do not need to converge. A limiting distribution theorem could also be easily obtained under the mere assumption that the norms $\|\mathfrak{F}_n\|$, $\|\mathfrak{F}_n^{-1}\|$, $\|\Sigma_{\mathcal{S},n}\|$ exist and are uniformly bounded for sufficiently large n . However, its statement would be expressed in terms of a transformation of $\sqrt{n}(\hat{\theta}_n - \theta_0)$, rather than the deviation itself, as is the case in, e.g. Gupta and Robinson (2018). The requirement of invertibility of the matrix \mathfrak{F} could also be relaxed. However, using the present argument, it would only be possible to obtain partial results. An approach to the problem of the singularity of \mathfrak{F} which considers various convergence rates has been described in Lee (2004). Finally, we obtain the following generalisation of Theorem 3.2 in Lee (2004).

Theorem 2 Under Assumptions 1', 2' and 3–6, the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is multivariate normal with zero mean and variance $\mathfrak{F}^{-1} \Sigma_{\mathcal{S}} \mathfrak{F}^{-1}$. Furthermore, in the special case when the error term is normally distributed, the limiting distribution is $\mathcal{N}(0, \mathfrak{F}^{-1})$.

4 Application to a higher-order general group effects model

This section provides an illustration of an application of our refined asymptotic analysis to a theoretical argument. We describe a new group effect elimination scheme for the high-order SAR model, and using arguments of Sect. 3, we derive the asymptotics of the corresponding QML estimator. In the simple case of a constant number of group-specific effect dummy variables a consistent, asymptotically normal QML estimator can be obtained quite straightforwardly. In general, however,

²² We employ the Jacobian formulation of vector derivatives, as a result, \mathcal{S}_n is treated as a row vector.

a more careful approach is necessary. Firstly, the incidental parameter problem must be accounted for to assure consistency of estimates. Secondly, a certain degree of compatibility with the spatial weight matrix has to be assured.

Let us consider a modified version of the SAR model specification (c.f. Eq. (3)) in which an additional term associated with group-specific effects is introduced. The model specification then becomes

$$Y_n = \sum_{r=1}^d \lambda_r W_{n,r} Y_n + X_n \beta + \Phi_n \mu + \varepsilon_n, \quad (5)$$

where $\mu \in \mathbb{R}^\kappa$ is the vector of group-specific effects, with $\kappa = \kappa(n)$ possibly increasing with n , and the columns of the corresponding matrix Φ_n are typically dummy variables distinguishing non-overlapping groups of observations.²³ Importantly, we note that, as the number of columns in Φ_n may change with sample size, the theorems of Sect. 3 cannot be applied directly.

In applied spatial econometrics it is common to eliminate fixed effects by means of the demeaning procedure, see e.g. Elhorst (2014), which can be understood as a simple projection on the space orthogonal to the columns of Φ_n . This is therefore closely related to the well-known Frisch-Waugh theorem, see Baltagi (2005). However, with an increasing number of groups, we must be concerned about singularity of the resulting variance, as expressed in e.g. Anselin et al. (2008) and the incidental parameter problem that can arise in such a setting. An effective method of dealing with those issues is developed in Lee and Yu (2010a), where a projection onto a lower dimensional space is applied to properly derive the asymptotic distribution of the resulting QML estimator. The technique presented in this paper extends this idea. Our approach consists in handling the group effect term together with its spatial lags, that is $W_{n,r} \Phi_n, W_{n,r}^2 \Phi_n, \dots$. At the same time, the transformed model is projected onto a lower dimensional space, following the idea of Lee and Yu (2010a).

Let $\mathcal{K}_n \subset \mathbb{R}^n$ be the linear subspace generated by iterating $W_{n,r}$ on the columns of the matrix Φ_n . In other words, \mathcal{K}_n is the smallest $W_{n,r}$ -invariant subspace containing the columns of Φ_n . Indeed, any spatial lag of Φ_n (member of \mathcal{K}_n), when multiplied by W_n , is yet another spatial lag, thus it is a member of \mathcal{K}_n . The same is true for all linear combinations of such spatial lags. Our idea is to filter out those vector components of both Y_n and X_n which lie in \mathcal{K}_n . Under the assumption that the orthogonal complement \mathcal{K}_n^\perp is sufficiently rich, we can obtain a consistent QML estimator of $\theta = (\lambda^T, \beta^T, \sigma^2)^T$.

Let $n_* = n - \dim \mathcal{K}_n$ and fix an $n_* \times n$ matrix F whose rows form an orthonormal basis of \mathcal{K}_n^\perp . It is easy to observe that $M_{\mathcal{K}} = F^T F$ is the projection onto \mathcal{K}_n^\perp and $FF^T = I_{n_*}$.

²³ Groups are understood as subsets of observations within the sample to which the specific effect is attributed. As an example, this contains the individual fixed effects model as a special case. That is, for balanced panel data when a longitudinal sample of size n is indexed in a way that distinguishes N spatial units and T time periods, $n = NT$, groups contain observations relevant to respective spatial units. Similarly, when each group contains observations ascribed to the respective time periods, we arrive at the time-specific fixed effects specification. Let us also note that, formally, columns of Φ_n are allowed to be arbitrary vectors, as long as the relevant assumptions of this section hold.

Denote $Y_n^* = FY_n$, $X_n^* = FX_n$, $\varepsilon_n^* = F\varepsilon_n$ and $W_{n,r}^* = FW_{n,r}F^T$. As $I_n - F^TF$ projects onto \mathcal{K}_n we have $FW_{n,r}(I_n - F^TF) = 0$. Thus, transforming the specification of Eq. (5) with F , we obtain

$$\begin{aligned} Y_n^* &= FY_n = \sum_{r \leq d} \lambda_r FW_{n,r} Y_n + FX_n \beta + F\Phi_n \mu + F\varepsilon_n \\ &= \sum_{r \leq d} \lambda_r FW_{n,r} F^T FY_n + FX_n \beta + F\varepsilon_n \\ &= \sum_{r \leq d} \lambda_r W_{n,r}^* Y_n^* + X_n^* \beta + \varepsilon_n^*, \end{aligned} \quad (6)$$

since $FW_{n,r} = FW_{n,r}F^TF$ and, by definition, $F\Phi_n = 0$.

Let us observe that ε_n^* satisfies Assumption 2 (b) if the original ε_n satisfies Assumption 2 (a) or (b). The crucial observation, however, is that Assumptions 3 and 4 are satisfied when $W_{n,r}^*$ is substituted for $W_{n,r}$. Indeed, with $\Delta_n^*(\lambda) = I_{n_*} - \sum_{r \leq d} \lambda_r W_{n,r}^* = F(I_n - \sum_{r \leq d} \lambda_r W_{n,r})F^T$ and $\Delta_n^*(\lambda)^{-1} = F(I_n - \sum_{r \leq d} \lambda_r W_{n,r})^{-1}F^T$ it is sufficient to note that $\|W_{n,r}^*\| \leq \|W_{n,r}\|$ and $\|\Delta_n^*(\lambda)^{-1}\| \leq \|\Delta_n(\lambda)^{-1}\|$, as $\|F\| = 1$, whenever $n_* > 0$.

Assumption A The earlier Assumptions 1 and 5 hold for the transformed specification of Eq. (6), that is with X_n^* , $W_{n,r}^*$, n^* substituted in place of X_n , $W_{n,r}$, n , respectively.

Assumption B We have $\lim_{n \rightarrow \infty} n_* = \infty$.

The following result is an immediate consequence of Theorem 1.

Theorem 3 Under Assumptions 2 (a), 3, 4 as well as the above Assumptions A and B the Gaussian QML estimator $\hat{\theta}_n^* = (\hat{\lambda}_n^{*\top}, \hat{\beta}_n^{*\top}, \hat{\sigma}_n^{2*})^T$ for Eq. (6) is consistent.

Let us note that the mere application of Theorem 2 is not sufficient to establish asymptotic normality of $\sqrt{n_*}(\hat{\theta}_n^* - \theta_0)$. The main difficulty is that the components of $F\varepsilon_n$ do not have to be independent, even if the original ε_n is.²⁴ However, with our asymptotic analysis, a valid argument is still possible.

Assumption A' The earlier Assumption A holds and Assumptions 1' is satisfied for $F^TX_n^*$, $F^TG_{n,r}^*X_n^*\beta_0$ and n_* . Moreover, Assumption 7 holds with Y_n^* , X_n^* , n_* , \mathfrak{F}_n^* etc. in the capacity of Y_n , X_n , n , \mathfrak{F}_n

Theorem 4 Under Assumptions 2', 3, 4, 6, A' and B the rescaled deviation $\sqrt{n_*}(\hat{\theta}_n^* - \theta_0)$ is asymptotically normally distributed with zero mean and variance $(\mathfrak{F}^*)^{-1}\Sigma_S(\mathfrak{F}^*)^{-1}$.

²⁴ Unless normality of ε_n is assumed, c.f. the proof of Theorem 2 in Lee and Yu (2010a).

To capture the relationship between our group effect elimination technique and the classical demeaning operator, let us observe that, if specification of Eq. (6) is further transformed with F^T , then we arrive at the proper Gaussian log-likelihood of θ , given $Y_n^\dagger = F^T F Y_n = M_K Y_n$ and $X_n^\dagger = M_K X_n$. Indeed, we obtain

$$\log L_n^\dagger(\theta) = -\frac{n_*}{2} \ln(2\pi\sigma^2) + \ln \left| \text{pdet}(\Delta_n^\dagger(\lambda)) \right| - \frac{1}{2\sigma^2} \left\| \Delta_n^\dagger(\lambda) Y_n^\dagger - X_n^\dagger \beta \right\|^2, \quad (7)$$

where $\Delta_n^\dagger(\lambda) = M_K \Delta_n(\lambda) M_K$ and pdet denotes the pseudo-determinant, i.e. the product of all non-zero singular values.

One advantage of the log-likelihood in Eq. (7) is that it does not depend on a particular choice of matrix F . Moreover, given $\det \Delta_n(\lambda)$ the pseudo-determinant might be computed using the determinant formula for block matrices. If E is a matrix such that $[F^T, E^T]$ is an orthogonal matrix, then we have the relation

$$\det \Delta_n(\lambda) = \det \left(\begin{bmatrix} F \\ E \end{bmatrix} \Delta_n(\lambda) [F^T, E^T] \right) = \text{pdet}(\Delta_n^\dagger(\lambda)) \det(E \Delta_n(\lambda) E^T).$$

For some specifications of the group or fixed effects, the determinant of $E \Delta_n(\lambda) E^T$ can be found analytically. For example, in a panel setting, with $n = mt$, time-invariant matrices $W_{n,r} = I_t \otimes \bar{W}_{m,r}$ and a usual matrix Φ_n of spatial unit dummy variables, it can be shown that $\det(E \Delta_n(\lambda) E^T)$ equals to $\det(I_m - \sum_{r \leq d} \lambda_r \bar{W}_{m,r})$. If the matrices in $W_{m,r}$ are additionally row-normalised and the matrix Φ_n incorporates both temporal and spatial fixed effects, we have $\det(E \Delta_n(\lambda) E^T) = (1 - \sum_{r \leq d} \lambda_r)^{t-1} \det(I_m - \sum_{r \leq d} \lambda_r \bar{W}_{m,r})$. Lastly, in the case of temporal fixed effects specification, it can be shown that $\det(E \Delta_n(\lambda) E^T) = (1 - \sum_{r \leq d} \lambda_r)^t$.

5 Closing remarks

In this paper we have revisited the analysis of the asymptotic behaviour of the well-known Gaussian QML estimator for higher-order SAR models. Our findings indicate that the standard assumptions on row and column summability of the spatial weight matrix can be weakened to cover econometric models with a greater degree of spatial dependence. Additionally, it is possible to apply a broader class of model transformations in theoretical arguments, without violating the essential boundedness requirement. Secondly, weaker conditions on the existence of moments of the error term can be imposed and its elements do not need to be identically distributed as long as their kurtosis is uniformly bounded. We expect that our results can be used to reconsider the asymptotic behaviour of QML estimation in more general specifications. Moreover, large sample theories for other estimators, in particular, General Method of Moments or Two-Stage Least Squares, can benefit from reapplication of our arguments, especially our Theorem 5 in ‘‘Appendix C’’. We should also mention that we have made the effort to avoid certain mathematical imprecision that can be found in the arguments of the standard analysis. For example, we properly derive the asymptotic distribution based on the Cramér-Wald theorem. Moreover,

our proofs rely neither on the existence of the Lagrange remainder in the Taylor expansion nor on the mean value theorem.²⁵

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendix A: Monte Carlo experiments

We have conducted computer simulations to show that, under the relaxed boundedness condition, the asymptotic theory is valid. We considered four different spatial interaction schemes in a linear setting. The first matrix considered, W_n^1 , is a common nearest neighbour matrix, with one distinguished central unit, whose interaction with other units is defined by the IDW scheme with the power parameter $\alpha = 2$. This is a summable matrix and the results of Monte Carlo experiments may serve as a point of reference for non-summable settings.

The matrix W_n^2 is analogous to the matrix W_n^1 , with the crucial difference that the power parameter $\alpha = 1$. This leads to an interaction scheme which is no longer summable. The third matrix, W_n^3 , is yet another variation on the same idea. However, instead of using the IDW scheme, the non-summable interaction pattern is now uniform, with all its weights equal to $1/\sqrt{n}$. It is the largest possible square-summable, uniform pattern, with respect to size of the weights. Lastly, the matrix W_n^4 is obtained from the symmetric non-summable IDW matrix, all elements of which are proportional to $\frac{1}{|i-j|}$, with i, j being its indices (here we have $\alpha = 1$). This matrix has been rescaled by its norm.

The Monte Carlo simulations for all of the matrices W_n^i , $i = 1, 2, 3, 4$, were conducted with a mixed regressive autoregressive specification, see Eq. (3), with a single autoregressive parameter $\lambda_0 = 0.3$. The regressor matrix X_n contained a constant term $c_0 = 2$ and one regressor, uniformly distributed in an interval symmetric around zero, with the corresponding slope $\beta_0 = 3$. In all simulated models the innovations were Gaussian with variance $\sigma_0^2 = 1$.

A constant number of Monte Carlo samples, $m = 5000$, was used for all trials. Tables 1, 2, 3, 4 show the expectation estimate, the standard error for the estimator, and the Kolmogorov-Smirnov assessment of normality of the individual components of the scaled difference $\sqrt{n}(\hat{\theta}_n - \theta_0)$. For all matrices a clear tendency can be seen for the values of the bias and its standard deviation to diminish with increasing n . Moreover, in most cases the quotient of the absolute bias divided by the standard deviation does not

²⁵ Recall that the Lagrange remainder in the Taylor series expansion of a vector valued function is not available. The function $f : [0, 1] \rightarrow \mathbb{C}$, $f(t) = e^{2\pi i t}$, $t \in [0, 1]$, can serve as a counter-example, see also Feng et al. (2014). Instead, our technique makes use of an original bound (inequalities of Eq. (C.5) and Eq. (C.6) in "Appendix C").

Table 1 Simulation results for the summable matrix W_n^1 of the Inverse Distance Weighting scheme with the power parameter $\alpha = 2$

	$n = 125$	$n = 250$	$n = 500$	$n = 750$	$n = 1000$	$n = 1500$
$\lambda_0 = 0.3$						
Estimate	0.2800	0.2905	0.2950	0.2965	0.2973	0.2982
Standard error	0.0811	0.0575	0.0413	0.0339	0.0292	0.0241
KS test figure	0.0152	0.0081	0.0142	0.0154	0.0112	0.0115
Sign. level	**	* * *	**	**	* * *	* * *
$c_0 = 2$						
Estimate	2.0555	2.0268	2.0134	2.0101	2.0080	2.0055
Standard error	0.2436	0.1764	0.1252	0.1038	0.0881	0.0736
KS test figure	0.0170	0.0103	0.0129	0.0139	0.0116	0.0140
Sign. level	**	* * *	**	**	* * *	**
$\beta_0 = 3$						
Estimate	3.0855	3.0407	3.0213	3.0152	3.0115	3.0078
Standard error	0.3472	0.2467	0.1771	0.1456	0.1255	0.1036
KS test figure	0.0167	0.0078	0.0128	0.0153	0.0107	0.0094
Sign. level	**	* * *	**	**	* * *	* * *
$\sigma_0^2 = 1$						
Estimate	0.9787	0.9879	0.9954	0.9970	0.9975	0.9982
Standard error	0.1262	0.0917	0.0652	0.0528	0.0462	0.0373
KS test figure	0.0228	0.0092	0.0130	0.0083	0.0185	0.0135
Sign. level	—	* * *	**	* * *	*	**

The Kolmogorov-Smirnov normality test significance levels are: *** for $p > 0.5$, ** for $p > 0.1$ and * for the usual $p > 0.05$

Table 2 Simulation results for the non-summable matrix W_n^2 of the Inverse Distance Weighting scheme with the power parameter $\alpha = 1$

	$n = 125$	$n = 250$	$n = 500$	$n = 750$	$n = 1000$	$n = 1500$
$\lambda_0 = 0.3$						
Estimate	0.2805	0.2895	0.2955	0.2964	0.2975	0.2986
Standard error	0.0823	0.0574	0.0412	0.0329	0.0289	0.0237
KS test figure	0.0164	0.0195	0.0124	0.0098	0.0081	0.0139
Sign. level	**	—	**	* * *	* * *	**
$c_0 = 2$						
Estimate	2.0589	2.0314	2.0136	2.0105	2.0077	2.0041
Standard error	0.2580	0.1781	0.1269	0.1011	0.0893	0.0726
KS test figure	0.0234	0.0143	0.0120	0.0095	0.0083	0.0117
Sign. level	—	**	**	* * *	* * *	* * *
$\beta_0 = 3$						
Estimate	3.0834	3.0450	3.0193	3.0152	3.0109	3.0061
Standard error	0.3528	0.2463	0.1767	0.1411	0.1242	0.1015
KS test figure	0.0156	0.0185	0.0103	0.0093	0.0075	0.0134
Sign. level	**	*	* * *	* * *	* * *	**

Table 2 (continued)

	$n = 125$	$n = 250$	$n = 500$	$n = 750$	$n = 1000$	$n = 1500$
$\lambda_0 = 0.3$						
$\sigma_0^2 = 1$						
Estimate	0.9771	0.9894	0.9953	0.9961	0.9978	0.9983
Standard error	0.1263	0.0909	0.0632	0.0522	0.0452	0.0373
KS test figure	0.0207	0.0153	0.0073	0.0111	0.0149	0.0093
Sign. level	—	**	***	***	**	***

The Kolmogorov-Smirnov normality test significance levels are: *** for $p > 0.5$, ** for $p > 0.1$ and * for the usual $p > 0.05$

Table 3 Simulation results for the non-summable matrix W_n^3 of uniform interaction scheme at the rate $\frac{1}{\sqrt{n}}$

	$n = 125$	$n = 250$	$n = 500$	$n = 750$	$n = 1000$	$n = 1500$
$\lambda_0 = 0.3$						
Estimate	0.2967	0.2988	0.2992	0.2997	0.2996	0.2998
Standard error	0.0245	0.0170	0.0121	0.0099	0.0086	0.0070
KS test figure	0.0081	0.0114	0.0101	0.0073	0.0101	0.0071
Sign. level	***	***	***	***	***	***
$c_0 = 2$						
Estimate	2.0133	2.0041	2.0027	2.0011	2.0019	2.0003
Standard error	0.1543	0.1055	0.0723	0.0591	0.0507	0.0412
KS test figure	0.0105	0.0124	0.0060	0.0074	0.0078	0.0098
Sign. level	***	**	***	***	***	***
$\beta_0 = 3$						
Estimate	3.0144	3.0054	3.0033	3.0012	3.0018	3.0010
Standard error	0.1059	0.0741	0.0523	0.0429	0.0371	0.0303
KS test figure	0.0076	0.0125	0.0100	0.0104	0.0135	0.0090
Sign. level	***	**	***	***	**	***
$\sigma_0^2 = 1$						
Estimate	0.9784	0.9903	0.9938	0.9956	0.9977	0.9986
Standard error	0.1278	0.0885	0.0626	0.0520	0.0433	0.0362
KS test figure	0.0227	0.0208	0.0143	0.0114	0.0159	0.0137
Sign. level	—	—	**	***	**	**

The Kolmogorov-Smirnov normality test significance levels are: *** for $p > 0.5$, ** for $p > 0.1$ and * for the usual $p > 0.05$

Table 4 Simulation results for the non-summable matrix $W_n^4 = [|i - j|^{-\alpha}]_{i \neq j}$, with $\alpha = 1$, rescaled by its spectral norm

	$n = 125$	$n = 250$	$n = 500$	$n = 750$	$n = 1000$	$n = 1500$
$\lambda_0 = 0.3$						
Estimate	0.2972	0.2983	0.2988	0.2993	0.2990	0.2992
Standard error	0.0325	0.0250	0.0191	0.0162	0.0145	0.0126
KS test figure	0.0128	0.0075	0.0075	0.0141	0.0061	0.0108
Sign. level	**	***	***	**	***	***
$c_0 = 2$						
Estimate	2.0078	2.0036	2.0029	2.0016	2.0033	2.0023
Standard error	0.1278	0.0954	0.0699	0.0577	0.0516	0.0444
KS test figure	0.0134	0.0098	0.0122	0.0126	0.0091	0.0085
Sign. level	**	***	**	**	***	***
$\beta_0 = 3$						
Estimate	3.0073	3.0050	3.0034	3.0021	3.0032	3.0024
Standard error	0.0868	0.0712	0.0574	0.0497	0.0451	0.0399
KS test figure	0.0133	0.0083	0.0093	0.0136	0.0048	0.0091
Sign. level	**	***	***	**	***	***
$\sigma_0^2 = 1$						
Estimate	0.9768	0.9869	0.9934	0.9960	0.9974	0.9983
Standard error	0.1235	0.0895	0.0632	0.0518	0.0436	0.0362
KS test figure	0.0205	0.0133	0.0124	0.0134	0.0151	0.0101
Sign. level	—	**	**	**	**	***

The Kolmogorov-Smirnov significance levels are: *** for $p > 0.5$, ** for $p > 0.1$ and * for the usual $p > 0.05$

exceed 0.1256, which implies that true values of parameters lie within 0.05 one-sided confidence intervals for the centre of the distribution of estimates. We have also examined the differences between the theoretical variance of the estimator (implied by the matrix \mathfrak{F}^{-1}) and the values derived from the samples. The relative differences ranged from zero to four percent, with an average of roughly two percent, which is consistent with the relative standard deviation $\sqrt{2/m}$ of the $\chi^2(m)$ distribution.

Appendix B: Additional facts

Remark 1 There is a row-normalised matrix W_n which is bounded in spectral norm, yet $\|W_n\|_\infty$ is unbounded.

Proof Set $D_n = [d_{ij}]_{i,j \leq n}$ with non-zero entries $d_{1,j} = 1/j$, $d_{i,1} = 1/i$ and $d_{i,i+1} = d_{j+1,j} = 1$ if $i, j > 1$. As an illustration, we have

$$D_6 = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{2} & 0 & 1 & 0 & 0 & 0 \\ \frac{1}{3} & 1 & 0 & 1 & 0 & 0 \\ \frac{1}{4} & 0 & 1 & 0 & 1 & 0 \\ \frac{1}{5} & 0 & 0 & 1 & 0 & 1 \\ \frac{1}{6} & 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Let W_n be the row-normalisation of D_n . Then $\|W_n\|_\infty \geq \sum_{j=1}^n w_{j,1} \geq \frac{1}{1+\frac{1}{2}} \sum_{j=2}^n \frac{1}{j}$ which escapes to infinity with $n \rightarrow \infty$. Let \tilde{W}_n be equal to the matrix W_n with the values in the first column set to zero. Then $\|W_n\| \leq \|\tilde{W}_n\| + \|W_n - \tilde{W}_n\|$. Lastly, $\|W_n - \tilde{W}_n\| \leq \sum_{j=2}^n \frac{3}{2j^2} < \infty$ and \tilde{W}_n is outright summable.²⁶ \square

Remark 2 If $W_{n,r}$, $r \leq d$, and Λ satisfy Assumptions 3 and 4, then there exists a bounded open set $U_\Lambda \subset \mathbb{R}^d$, $U_\Lambda \supset \Lambda$, independent of n , such that $\Delta_n(\lambda)$ is invertible for each $\lambda \in U$ and the norms $\|\Delta_n(\lambda)^{-1}\|$, $\|\Delta_n(\lambda)\|$ do not exceed $K_\Lambda + 1$.

Proof of this non-trivial fact can be found in the supplementary material.

Lemma 1 Let $U \subset \mathbb{R}^d$ be an open set and $A \subset U$ its compact subset. If $F : U \rightarrow \mathbb{R}^m$ is differentiable and $\|F\|$, $\|F'\| \leq M < \infty$ for a constant M , then F is Lipschitz continuous on A with a constant $K_L = K_L(M, A)$.

Lemma 2 Let $\varepsilon_n = (\varepsilon_{n,i})_{i \leq n}$ be an $n \times 1$ random vector satisfying Assumption 2, and let $(A_n)_{n \in \mathbb{N}}$ and $(P_n)_{n \in \mathbb{N}}$ be sequences of $n \times n$ matrices satisfying

$$\sup_{n \in \mathbb{N}} \|A_n\| < \infty, \quad \sup_{n \in \mathbb{N}} \|P_n\| \leq 1 \quad \text{and} \quad P_n = P_n^T P_n.$$

If $x_n \in \mathbb{R}^n$, for $n \in \mathbb{N}$, is a non-random vector satisfying $\|x_n\|^2 = O(n)$, then

- (a) for $Z_n^a = \frac{1}{n} x_n^T A_n \varepsilon_n$ we have $\text{Var } Z_n^a \leq \frac{\sigma^2}{n} \|A_n\|^2 \frac{1}{n} \|x_n\|^2$,
- (b) for $Z_n^b = \frac{1}{n} \varepsilon_n^T A_n \varepsilon_n$ we have $\text{Var } Z_n^b \leq \frac{3}{n} \|A_n\|^2 \sup_{n,i} \mathbb{E} \varepsilon_{n,i}^4$.

Proof of Lemma 2 is given in the supplementary material.

Appendix C: Theorems

Proof of Theorem 1 Let $\theta_0 = (\lambda_0^T, \beta_0^T, \sigma_0^2)^T$ be the true value of parameter θ . Let $S_n(\lambda) = \Delta_n(\lambda) \Delta_n(\lambda_0)^{-1}$, $\lambda \in \Lambda$. It is a standard approach to use the first-order optimality conditions for β and σ^2 , that is

²⁶ A considerably more intricate example is also possible with infinitely many non-summable columns.

$$\hat{\beta}_n(\lambda) = (X_n^T X_n)^{-1} X_n^T \Delta_n(\lambda) Y_n, \quad \hat{\sigma}_n^2(\lambda) = \frac{1}{n} \left\| M_{X_n} \Delta_n(\lambda) Y_n \right\|^2, \quad (\text{C.1})$$

to obtain the concentrated log-likelihood $\ln L_n^c(\lambda) = -\frac{n}{2} \ln(2\pi \hat{\sigma}_n^2(\lambda)) - \frac{n}{2} + \ln |\det \Delta_n(\lambda)|$, which is maximised by $\hat{\lambda}_n$. Let us set $R_n(\lambda) = \frac{1}{n} \ln \hat{L}_n^c(\lambda)$ and note that each R_n is a random function.

The proof proceeds as follows. First, we obtain consistency of $\hat{\lambda}_n$ by the generic argument presented in Lemma 3.1 of Pötscher and Prucha (1997). Thus, we introduce a new, non-random function

$$\bar{R}_n(\lambda) = \frac{1}{2} \ln \mathbb{E} \exp(2R_n) = \ln \sqrt{2\pi \mathbb{E} \hat{\sigma}_n^2(\lambda)} + \frac{1}{2} + \frac{\ln |\det \Delta_n(\lambda)|}{n}, \quad (\text{C.2})$$

which uniformly approximates R_n for large n .²⁷ Then we show that λ_0 is an identifiably unique maximiser of each \bar{R}_n , as in Definition 3.1 of Pötscher and Prucha (1997) and, as a result, we obtain consistency of $\hat{\lambda}_n$. Finally, we deduce consistency of $\hat{\beta}_n$ and $\hat{\sigma}_n^2$.

Using Eq. (C.1) we expand $\hat{\sigma}_n^2(\lambda)$ and obtain, with $S_n(\lambda) = \Delta_n(\lambda) \Delta_n(\lambda_0)^{-1}$,

$$\hat{\sigma}_n^2(\lambda) = \frac{1}{n} \left\| M_{X_n} S_n(\lambda) X_n \beta_0 + M_{X_n} S_n(\lambda) \varepsilon_n \right\|^2 = \mathbb{E} \hat{\sigma}_n^2(\lambda) + \xi_n(\lambda),$$

where $\mathbb{E} \hat{\sigma}_n^2(\lambda) = \frac{1}{n} \left\| M_{X_n} S_n(\lambda) X_n \beta_0 \right\|^2 + \frac{\sigma_0^2}{n} \left\| M_{X_n} S_n(\lambda) \right\|_F^2$ and the remaining random deviation $\xi_n(\lambda) = \frac{2}{n} (S_n(\lambda) X_n \beta_0)^T M_{X_n} S_n(\lambda) \varepsilon_n + \frac{1}{n} \varepsilon_n^T S_n(\lambda)^T M_{X_n} S_n(\lambda) \varepsilon_n - \frac{1}{n} \sigma_0^2 \left\| M_{X_n} S_n(\lambda) \right\|_F^2$ converges to 0 in probability, uniformly in $\lambda \in \Lambda$, as a consequence of Lemma 2, with the use of Assumptions 1 and 4.

Now, we will show that $\sup_{\lambda \in \Lambda} |R_n(\lambda) - \bar{R}_n(\lambda)|$ converges to zero in probability. With K_Λ defined in Assumption 4 and $k = \text{rank}(X_n)$ we have

$$\begin{aligned} \|M_{X_n} S_n(\lambda)\|_F^2 &= \|S_n(\lambda)\|_F^2 + (\|M_{X_n} S_n(\lambda)\|_F - \|S_n(\lambda)\|_F) (\|M_{X_n} S_n(\lambda)\|_F + \|S_n(\lambda)\|_F) \\ &\geq \|S_n(\lambda)\|_F^2 - 2\|S_n(\lambda)\|_F \|P_{X_n} S_n(\lambda)\|_F \geq \|S_n(\lambda)\|_F^2 - 2kK_\Lambda^2 \sqrt{n}. \end{aligned}$$

Since $\|S_n(\lambda)\|_F^2 \geq \|S_n(\lambda)^{-1}\|^{-2}$, the value of $\mathbb{E} \hat{\sigma}_n^2(\lambda)$ is uniformly separated from zero. Finally, we obtain the uniform convergence in probability of

$$\bar{R}_n(\lambda) - R_n(\lambda) = \frac{1}{2} \ln \frac{\hat{\sigma}_n^2(\lambda)}{\mathbb{E} \hat{\sigma}_n^2(\lambda)} = \frac{1}{2} \ln \left(1 + \frac{\xi_n(\lambda)}{\mathbb{E} \hat{\sigma}_n^2(\lambda)} \right).$$

Since $\bar{R}_n(\lambda_0) = \frac{1}{n} \ln |\det \Delta_n(\lambda_0)| - \frac{1}{2} \ln \frac{n-k}{n} \sigma_0^2$ (cf. Eq. (C.2)), we have

²⁷ Note that the value of \bar{R}_n represents a log-root-mean-square of $\sqrt{L_n^c(\lambda)}$ with $L_n^c(\lambda)$ being the concentrated likelihood of λ . Although simply choosing $\bar{R}'_n = \mathbb{E} R_n$ instead of current $\bar{R}_n = \frac{1}{2} \ln \mathbb{E} e^{2R_n}$ might seem more natural, the use of \bar{R}_n results in simpler computations and the difference between \bar{R}'_n and \bar{R}_n diminishes as the randomness of R_n decreases with $n \rightarrow \infty$.

$$\begin{aligned}
2(\bar{R}_n(\lambda_0) - \bar{R}_n(\lambda)) &= \ln \frac{\frac{1}{(n-k)\sigma_0^2} \|M_{X_n} S_n(\lambda) X_n \beta_0\|^2 + \frac{1}{n-k} \|M_{X_n} S_n(\lambda)\|_{\mathcal{F}}^2}{|\det S_n(\lambda)|^{2/n}} \\
&\geq \ln \left(\frac{C}{n\sigma_0^2} \|M_{X_n} S_n(\lambda) X_n \beta_0\|^2 + \frac{\frac{1}{n} \|S_n(\lambda)\|_{\mathcal{F}}^2}{|\det S_n(\lambda)|^{2/n}} - \frac{2CkK_{\Lambda}^2 \sqrt{n}}{n-k} \right),
\end{aligned}$$

with a constant C such that $C \geq |\det S_n(\lambda)|^{-2/n}$. Furthermore, Assumption 5 implies that for any $\lambda \in \Lambda$ we have $\liminf_{n \rightarrow \infty} \bar{R}_n(\lambda_0) - \bar{R}_n(\lambda) > 0$.

We will show that $(\lambda_0)_{n \in \mathbb{N}}$ is an identifiably unique sequence of maximisers of \bar{R}_n . To this end, let us begin by assuming the contrary. Then, there is a number $\epsilon > 0$ for which, for some increasing sequence $\{k(n)\}_{n \in \mathbb{N}} \subset \mathbb{N}$ and some sequence $\{\tilde{\lambda}_n\}_{n \in \mathbb{N}} \subset C_\epsilon = \{\lambda \in \Lambda : \|\lambda - \lambda_0\| \geq \epsilon\}$, we have $\lim_{n \rightarrow \infty} \bar{R}_{k(n)}(\lambda_0) - \bar{R}_{k(n)}(\tilde{\lambda}_n) \leq 0$. Since C_ϵ is closed in compact Λ , the sequences $\{\tilde{\lambda}_n\}_{n \in \mathbb{N}}$ and $\{k(n)\}_{n \in \mathbb{N}} \subset \mathbb{N}$ can be chosen in such a way that $\tilde{\lambda}_n \rightarrow \tilde{\lambda}$, for some $\tilde{\lambda} \neq \lambda_0$. Let $\delta = \liminf_{n \rightarrow \infty} \bar{R}_n(\lambda_0) - \bar{R}_n(\tilde{\lambda}) > 0$. Using Assumption 4 it can be verified²⁸ that both R_n and its derivative are bounded, thus Lemma 1 implies that all \bar{R}_n are Lipschitz continuous on Λ with a uniform constant K_L . We can choose $n_0 \in \mathbb{N}$ such that $\|\tilde{\lambda}_m - \tilde{\lambda}\| < \frac{\delta}{3K_L}$ for all $m \geq n_0$. The contradiction then follows from the inequality

$$\delta \leq \liminf_{n \rightarrow \infty} (\bar{R}_{k(n)}(\lambda_0) - \bar{R}_{k(n)}(\tilde{\lambda}_n) + |\bar{R}_{k(n)}(\tilde{\lambda}_n) - \bar{R}_{k(n)}(\tilde{\lambda})|) \leq \frac{\delta}{3}.$$

Lastly, the convergence $\|\hat{\lambda}_n - \lambda_0\| = o_{\mathbb{P}}(1)$, follows from Lemma 3.1 in Pötscher and Prucha (1997) as, by definition, for each $n \in \mathbb{N}$, $\hat{\lambda}_n$ is a maximiser of R_n .

Notice that, by Eq. (C.1), we have $\hat{\beta}_n(\hat{\lambda}_n) = \beta_0 - \zeta_n^{(1)} + \zeta_n^{(2)} - \zeta_n^{(3)}$, where

$$\begin{aligned}
\zeta_n^{(1)} &= (X_n^T X_n)^{-1} X_n^T \sum_{r \leq d} (\hat{\lambda}_{n,r} - \lambda_r^0) W_{n,r} \Delta_n(\lambda_0)^{-1} X_n \beta_0, \\
\zeta_n^{(2)} &= (X_n^T X_n)^{-1} X_n^T \epsilon_n, \\
\zeta_n^{(3)} &= (X_n^T X_n)^{-1} X_n^T \sum_{r \leq d} (\hat{\lambda}_{n,r} - \lambda_r^0) W_{n,r} \Delta_n(\lambda_0)^{-1} \epsilon_n.
\end{aligned}$$

From Assumptions 1 and 4 we have $\|\zeta_n^{(1)}\| = O(\|\hat{\lambda}_n - \lambda_0\|) = o_{\mathbb{P}}(1)$.

The convergence $\|\zeta_n^{(3)}\| = o_{\mathbb{P}}(1)$ can be deduced from the Chebyshev inequality, as, for a constant C' , we have $(\mathbb{E}_{\theta_0} \|\zeta_n^{(3)}\|)^2 \leq C' \sigma_0^2 \mathbb{E}_{\theta_0} \|\hat{\lambda}_n - \lambda_0\|^2$, by the Schwartz inequality. Lastly, by Assumption 1 we have $\|\text{Var}_{\theta} \zeta_n^{(2)}\| = O(1/n)$. Thus, $\hat{\beta}_n$ is consistent.

Again, using Eq. (C.1) and the consistency of $\hat{\lambda}_n$ and $\hat{\beta}_n$, we have

²⁸ See Remark S.5 in the supplementary material.

$$\begin{aligned}\hat{\sigma}_n^2(\hat{\lambda}_n) &= \frac{1}{n} \left\| \Delta_n(\lambda_0 + o_{\mathbb{P}}^{d \times 1}(1)) Y_n - X_n(\beta_0 + o_{\mathbb{P}}^{k \times 1}(1)) \right\|^2 \\ &= \frac{1}{n} \left\| \Delta_n(\lambda_0) Y_n - X_n \beta_0 - \sum_{r \leq d} o_{\mathbb{P}}(1) W_{n,r} Y_n - X_n o_{\mathbb{P}}^{k \times 1}(1) \right\|^2.\end{aligned}$$

Similar arguments imply that $\frac{1}{n} \left\| \sum_{r \leq d} o_{\mathbb{P}}(1) W_{n,r} Y_n + X_n o_{\mathbb{P}}^{k \times 1}(1) \right\|^2 = o_{\mathbb{P}}(1)$. Finally, as $\sigma_0^2 = \text{plim}_{n \rightarrow \infty} \frac{1}{n} \|\varepsilon_n\|^2$, by statement **(b)** of Lemma 2, we also have $\text{plim}_{n \rightarrow \infty} \hat{\sigma}_n^2 = \sigma_0^2$. \square

Theorem 5 Let $(\varepsilon_n)_{n \in \mathbb{N}}$ satisfy Assumption 2' and let $x_n = (x_{n,i})_{i \leq n}$, $n \in \mathbb{N}$, be column vectors. Denote $Q_n = \varepsilon_n^T (x_n + A_n \varepsilon_n)$ and assume that $\text{Var } Q_n > 0$ for sufficiently large $n \in \mathbb{N}$. If $\|x_n\|^2 + \|A_n\|_F^2 = O(\text{Var } Q_n)$, $\|A_n\|^2 = o(\text{Var } Q_n)$ and $\max_{i \leq n} x_{n,i}^2 = o(\text{Var } Q_n)$, then $\frac{Q_n - \mathbb{E} Q_n}{\sqrt{\text{Var } Q_n}}$ converges in distribution to a standardised normal variable $\mathcal{N}(0, 1)$.

Proof of Theorem 5 is given in the supplementary material. The argument is based on bounds originally developed in Bhansali et al. (2007), where a CLT for quadratic forms of i.i.d. vectors is shown.

Corollary 1 Let $(\varepsilon_n)_{n \in \mathbb{N}}$ satisfy Assumption 2' and let $x_n = (x_{n,i})_{i \leq n}$, $n \in \mathbb{N}$, be column vectors. Denote $Q_n = \varepsilon_n^T (x_n + A_n \varepsilon_n)$. If $\lim_{n \rightarrow \infty} \text{Var } Q_n$ exists and is positive, $\|x_n\|^2 + \|A_n\|_F^2 = O(1)$ and $\|A_n\|^2 + \max_{i \leq n} x_{n,i}^2 = o(1)$, then $\frac{Q_n - \mathbb{E} Q_n}{\sqrt{\text{Var } Q_n}}$ converges in distribution to a standard normal variable $\mathcal{N}(0, 1)$.

Proof of Theorem 2 With S_n defined in Assumption 7 and $G_{n,r} = W_{n,r} \Delta_n(\lambda_0)^{-1}$, a straightforward calculation²⁹ shows that the consecutive entries of $\frac{1}{\sqrt{n}} S_n$ are

$$\begin{aligned}\frac{1}{\sqrt{n}} \frac{\partial \ln L_n}{\partial \beta}(\theta_0) &= \frac{1}{\sqrt{n} \sigma_0^2} \varepsilon_n^T X_n, \\ \frac{1}{\sqrt{n}} \frac{\partial \ln L_n}{\partial \lambda}(\theta_0) &= \frac{1}{\sqrt{n} \sigma_0^2} \left([\varepsilon_n^T G_{n,r}]_{r \leq d}^T + [\varepsilon_n^T G_{n,r} \varepsilon_n - \sigma_0^2 \text{tr}(G_{n,r})]_{r \leq d}^T \right), \\ \frac{1}{\sqrt{n}} \frac{\partial \ln L_n}{\partial \sigma^2}(\theta_0) &= \frac{1}{2\sqrt{n} \sigma_0^4} (\varepsilon_n^T \varepsilon_n - n \sigma_0^2).\end{aligned}$$

We will show that $\frac{1}{\sqrt{n}} S_n^T$ converges in distribution to $\mathcal{N}(0, \Sigma_S)$.³⁰

²⁹ c.f. Remark S.1 in the supplementary material.

³⁰ Naturally, it is not sufficient to establish asymptotic normality of the above formulae, c.f. Lee (2004). Our argument follows by considering two cases and makes use of the standard Cramér-Wald theorem (see e.g. Billingsley (1995)).

Let $\alpha = (a^T, b^T, c)$, where $(a_r) \in \mathbb{R}^d$, $b \in \mathbb{R}^k$ and $c \in \mathbb{R}$. First, assume that $\alpha^T \Sigma_S \alpha \neq 0$. Then, we can observe that $\frac{1}{\sqrt{n}} \alpha^T \mathcal{S}_n^T$ is a centred linear-quadratic form of the residual ε_n . That is, $\frac{1}{\sqrt{n}} \alpha^T \mathcal{S}_n^T = Q_n - \mathbb{E} Q_n$, with $Q_n = x_n^T \varepsilon_n + \varepsilon_n^T A_n \varepsilon_n$, where

$$x_n = \frac{1}{\sqrt{n} \sigma_0^2} \left(X_n b + \sum_{r \leq d} a_r G_{n,r} X_n \beta_0 \right) \quad \text{and} \quad A_n = \frac{1}{\sqrt{n} \sigma_0^2} \left(\frac{c}{2 \sigma_0^2} I_n + \sum_{r \leq d} a_r G_{n,r} \right).$$

Note that by Assumptions 1' and 7 we have $\max_{i \leq n} x_{n,i}^2 = o(1)$, $\|A_n\|^2 = O(1/n)$, $\|x_n\|^2 + \|A_n\|_F^2 = O(1)$; and, using Assumption 7, we have $\lim_{n \rightarrow \infty} \text{Var} \frac{1}{\sqrt{n}} \alpha^T \mathcal{S}_n^T > 0$.

Thus, Corollary 1 can be used to deduce that $\frac{1}{\sqrt{n}} \alpha^T \mathcal{S}_n^T$ converges in distribution to $\mathcal{N}(0, \alpha^T \Sigma_S \alpha)$. In the case of $\alpha^T \Sigma_S \alpha = 0$ the convergence holds trivially.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space on which ε_n , $n \in \mathbb{N}$, are defined. Let τ be an open bounded subset of $\Lambda \times \mathbb{R}^k \times (0, +\infty)$ and let B^{λ_0} be an open ball centred at λ_0 contained entirely in Λ . Set $U_{\theta_0} = \tau \cap (B^{\lambda_0} \times \mathbb{R}^{k+1})$. Also denote $\tilde{\mathfrak{T}}_n = -\frac{1}{n} \frac{\partial^2 \ln L_n}{\partial \theta \partial \theta^T}(\theta_0)$ and $M_n^\tau = \sup_{\tilde{\theta} \in \tau} \left\| \frac{1}{n} \frac{\partial^3 \ln L_n}{\partial \theta \partial \theta \partial \theta}(\tilde{\theta}) \right\|$. Evaluation of the third derivative reveals that $\mathbb{E} M_n^\tau$ is uniformly bounded.³¹ It can be verified³² that $\tilde{\mathfrak{T}}_n$ converges in probability to \mathfrak{T} as $n \rightarrow \infty$, hence $\mathbb{P}(\{\det \tilde{\mathfrak{T}}_n = 0\}) = o(1)$ and $\|\tilde{\mathfrak{T}}_n^{-1}\| = O_{\mathbb{P}}(1)$. By **Theorem 1** we have $\|\hat{\theta}_n - \theta_0\| = o_{\mathbb{P}}(1)$ and Remark S.3 yields $\sup_{n \in \mathbb{N}} \mathbb{E}_{\theta_0} M_n^\tau < \infty$. Thus, it follows that for the sets $\Omega_n = \{\hat{\theta}_n \in U_{\theta_0}\} \cap \{\det \tilde{\mathfrak{T}}_n \neq 0\} \cap \{M_n^\tau \|\tilde{\mathfrak{T}}_n^{-1}\| \|\hat{\theta}_n - \theta_0\| < 1\}$ we have $\lim_{n \rightarrow \infty} \mathbb{P}(\Omega_n) = 1$.

For any $\omega \in \Omega_n$, by the Taylor expansion theorem, see e.g. Theorem 107 in Hájek and Johannis (2014), applied for the function $f_{\omega,n}(\theta) = \frac{1}{\sqrt{n}} \frac{\partial \ln L_n}{\partial \theta}(\theta, \omega)^T$ in $\theta = \theta_0$ we have $f_{\omega,n}(\theta) = \frac{1}{\sqrt{n}} \mathcal{S}_n^T - \tilde{\mathfrak{T}}_n \left(\sqrt{n}(\theta - \theta_0) \right) + \mathcal{R}_n(\theta)$, $\theta \in U_{\theta_0}$, where \mathcal{R}_n is the expansion remainder satisfying $\|\mathcal{R}_n(\theta)\| \leq \frac{1}{2} \sup_{\tilde{\theta} \in U_{\theta_0}} \|f_{\omega,n}''(\tilde{\theta})\| \|\theta - \theta_0\|^2$. Substituting $\theta = \hat{\theta}_n(\omega)$ we obtain

$$\tilde{\mathfrak{T}}_n \left(\sqrt{n}(\hat{\theta}_n - \theta_0) \right) = \frac{1}{\sqrt{n}} \mathcal{S}_n^T + \mathcal{R}_n(\hat{\theta}_n) \quad (\text{C.3})$$

and

$$\|\mathcal{R}_n(\hat{\theta}_n)\| \leq \frac{\sqrt{n}}{2} M_n^\tau \|\hat{\theta}_n - \theta_0\|^2. \quad (\text{C.4})$$

With $a_n = \tilde{\mathfrak{T}}_n \sqrt{n}(\hat{\theta}_n - \theta_0)$ and $b_n = \frac{1}{\sqrt{n}} \mathcal{S}_n^T$, the crucial observation is that $\|a_n\| < 2\|b_n\|$ on the sets Ω_n . Indeed, otherwise we would have

$$\|a_n\| \leq 2\|a_n - b_n\| = 2\|\mathcal{R}_n(\hat{\theta}_n)\| \leq M_n^\tau \|\tilde{\mathfrak{T}}_n^{-1}\| \|\hat{\theta}_n - \theta_0\| \|a_n\| < \|a_n\|. \quad (\text{C.5})$$

Using Eq. (C.4) and the fact that $\sup_{\theta \in \tau} \left\| \frac{1}{\sqrt{n}} \frac{\partial \ln L_n}{\partial \theta} \right\| = O_{\mathbb{P}}(1)$ we conclude that

³¹ See also Remark S.3 in the supplementary material.

³² See Remark S.2 in the supplementary material.

$$\|\mathcal{R}_n(\hat{\theta}_n)\| \leq M_n^r \|\tilde{\mathfrak{F}}_n^{-1}\| \left\| \frac{1}{\sqrt{n}} \mathcal{S}_n^T - f_{\omega,n}(\hat{\theta}_n) \right\| \|\hat{\theta}_n - \theta_0\| = o_{\mathbb{P}}(1). \quad (\text{C.6})$$

Finally, by combining Eq. (C.3) with Eq. (C.6) the desired convergence in distribution follows. \square

Proof of Theorem 4 The proof relies on the same argument as the proof of Theorem 2, up to the point where our CLT is used to deduce asymptotic normality of the linear-quadratic form $\frac{1}{\sqrt{n_*}} \mathcal{S}^* \alpha$, with α as previously. Then it can be seen that, for arbitrary x_n and A_n we have $x_n^T \varepsilon_n^* + (\varepsilon_n^*)^T A_n \varepsilon_n^* = x_n^T F \varepsilon_n + (\varepsilon_n)^T F^T A_n F \varepsilon_n$, hence a linear quadratic form of ε_n^* is, in fact, a linear-quadratic form of ε_n . Finally, it is sufficient to note that, in the case of $\frac{1}{\sqrt{n_*}} \mathcal{S}^* \alpha$, by Assumptions A' and 4 we have $(\sqrt{n_*} F^T x_n)_{n \in \mathbb{N}} \in \Xi^*$, $\|F^T A_n F\|^2 = O(1/n_*)$, $\|F^T x_n\|^2 = O(1)$, $\|F^T A_n F\|_{\mathcal{F}}^2 = O(1)$ and $\lim_{n \rightarrow \infty} \text{Var} \frac{1}{\sqrt{n_*}} \mathcal{S}^* \alpha = \alpha^T \Sigma_{\mathcal{S}^*} \alpha$. Thus, Corollary 1 can be used to deduce that $\frac{1}{\sqrt{n_*}} \mathcal{S}^* \alpha$ converges in distribution to $\mathcal{N}(0, \alpha^T \Sigma_{\mathcal{S}^*} \alpha)$.

Again, the remainder of the proof proceeds accordingly to the proof of Theorem 2. \square

References

- Anselin L (1988) Spatial econometrics: methods and models. Kluwer Academic, Dordrecht
- Anselin L (2002) Under the hood. Issues in the specification and interpretation of spatial regression models. *Agric Econ* 27(3):247–267
- Anselin L, Le Gallo J, Jayet H (2008) Spatial panel econometrics. In: Matyas L, Sevestre P (eds) The econometrics of panel data, fundamentals and recent developments in theory and practice, 3rd edn. Springer, Berlin Heidelberg, pp 625–660
- Baltagi BH (2005) Econometric analysis of panel data. Chichester West Sussex, England
- Bhansali RJ, Giraitis L, Kokoszka PS (2007) Convergence of quadratic forms with non-vanishing diagonal. *Stat Probab Lett* 77(7):726–734
- Billingsley P (1995) Probability and measure. Wiley, New York
- Case AC (2015) Spatial patterns in household demand. *Econometrica* 59(4):953–965
- Case AC (2015) Neighborhood influence and technological change. *Reg Sci Urban Econ* 22(3):491–508
- Delegano MA, Robinson PM (2015) Non-nested testing of spatial correlation. *J Econom* 187(1):385–401
- Elhorst JP (2014) Spatial econometrics: from cross-sectional data to spatial panels. Springer Briefs in Regional Science, Springer, Heidelberg
- Elhorst JP (2001) Dynamic models in space and time. *Geogr Anal* 33(2):119–140
- Feng C, Wang H, Han Y, Xia Y, Tu XM (2014) The mean value theorem and Taylor's expansion in statistics. *Am Stat* 67(4):245–248
- Gupta A, Robinson PM (2015) Inference on higher-order spatial autoregressive models with increasingly many parameters. *J Econom* 186(1):19–31
- Gupta A, Robinson P (2018) Pseudo maximum likelihood estimation of spatial autoregressive models with increasing dimension. *J Econom* 202(1):92–107
- Hájek P, Johannis M (2014) Smooth analysis in banach spaces. De Gruyter, Berlin
- Hall P, Hyde CC (1980) Martingale limit theory and its application. Academic Press Inc, New York
- Kelejian HH, Prucha IR (2001) On the asymptotic distribution of the Moran I test statistic with applications. *J Econom* 104(2):219–257
- Kelejian HH, Prucha IR (2010) Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *J Econom* 157(1):53–67

- Lee L-F (2002) Consistency and efficiency of least-squares estimation for mixed regressive, spatial autoregressive models. *Econom Theory* 18(2):252–277
- Lee L-F (2004) Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica* 72(6):1899–1925
- Lee L-F (2007b) Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. *J Econom* 140(2):333–374
- Lee L-F, Liu X, Lin X (2010) Specification and estimation of social interaction models with network structures. *Econom J* 13(2):145–176
- Lee L-F, Yu J (2010a) Estimation of spatial autoregressive panel data models with fixed effects. *J Econom* 154(2):165–185
- Lee L-F, Yu J (2010b) A spatial dynamic panel data model with both time and individual fixed effects. *Econom Theory* 26(2):564–597
- Le Sage J, Pace RK (2009) Introduction to spatial econometrics. CRC Press, Boca Raton
- Li K (2017) Fixed-effects dynamic spatial panel data models and impulse response analysis. *J Econom* 198(1):102–121
- Mynbaev KT, Ullah A (2010) Asymptotic distribution of the OLS estimator for a mixed spatial model. *J Multivar Anal* 101(3):733–748
- Mynbaev KT, Ullah A (2008) Asymptotic distribution of the OLS estimator for a purely autoregressive spatial model. *J Multivar Anal* 99(2):245–277
- Olejnik A, Olejnik J (2017) An alternative to partial regression in maximum likelihood estimation of spatial autoregressive panel data model. *Stat Rev* 64(3):323–337
- Pötscher MB, Prucha IR (1997) Dynamic nonlinear econometric models: asymptotic theory. Springer, Berlin
- Robinson PM (2011) Asymptotic theory for nonparametric regression with spatial data. *J Econom* 165(1):5–19
- Vega SH, Elhorst JP (2015) The SLX model. *J Reg Sci* 55(3):339–363

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.