Theory and Application of Dynamic Spatial Time Series Models

Andree, B.P.J.

2020

document version
Publisher's PDF, also known as Version of record

Chapter 2

Background Theory

Asymptotic theory is the cornerstone of inferential statistics. The limiting distribution of a basic quantity of interest delivers properties that are accurate in large samples and often reasonable when there is moderate data. In particular, limiting distributions can be used for approximate inference based on approximate confidence intervals and their associated test statistics. The benefit of the limiting distribution over exact distributional results is that it can often be derived following general rules that are valid even for complicated models that include heterogeneity, interaction and nonlinearity. The exact distributions are, however, often difficult to derive, and may not even apply in certain cases of interest. Asymptotic distribution theory is centered around the notion of an expected mean and an expected variance. The general steps to establish these quantities of interest are to establish convergence of the mean and convergence of the variance under a notion of growing data.

Because asymptotic theory is crucial for econometric analysis, it is useful to have general results with conditions that can be applied to as many estimators as possible to deliver standard and identical interpretation to a wide range of empirical results. The purpose of this chapter is to present such results in a brief and common format adapted to the setting of spatial time series. The basic exposition sets the table for the later chapters that establish and discuss properties of complex models,

including some that have not been used in existing literature. References to literature on specific results and proofs, but also to advanced textbooks that have wide coverage, will be provided in the relevant sections in later chapters.

2.1 Linear estimators

In introductory econometrics books, the properties of standard estimators have been extensively studied. However, basic theory only works in the simplistic setting of linear models and requires the very restrictive assumption that the model is an exact description of reality (i.e. that the model is correctly specified). Generally, as the dimensions of the data grow in time, space, and number of variables, it becomes increasingly unlikely that the same average description appropriately describes local processes across all dimensions and levels in the data. It is more likely that the derivatives that describe marginal effects between dependent and independent data vary from one local mean to another across regions or regimes. While flexibility to cope with these transitions may be a natural idea, it is not always possible to simply allow for more complex model dynamics without breaking assumptions that are made under standard theory. In particular, the linearity of the standard regression model was key to obtaining an analytical expression for a simple estimator and the assumption of correct specification of the model was used to express the estimator in terms of deviations around the *true* parameter. The linearity of the model also made it straightforward to derive stationary conditions and ensure that a Law of Large Numbers and Central Limit Theorems can be applied to obtain the consistency and asymptotic normality of the estimator. For example, the LSE of the linear autoregressive parameter β in the model given by $y_t = \beta y_{t-1} + \varepsilon_t$ takes the form:

$$\hat{\beta}_T = \frac{\sum_{t=2}^{T} y_t y_{t-1}}{\sum_{t=2}^{T} y_{t-1}^2}. \tag{2.1}$$

Deriving this expression was only possible because the model is linear, which drastically reduced the complexity of the calculus involved. Due to the simplicity, the properties of this estimator can also easily be analyzed if we assume that this linear description is correct, e.g. that our parametrization corresponds exactly with the *true* model that produced the observed data. This allows us to rewrite the estimator in terms of the *true* parameter β_0 and a remainder, β_r:

$$\hat{\beta}_T = \beta_0 + \beta_r, \qquad \beta_r = \frac{\sum_{t=2}^{T} \varepsilon_t y_{t-1}}{y_{t-1}^2}. \tag{2.2}$$

Furthermore, when dependence is linear, we can straightforwardly show that if $|\beta_0| < 1$, then the model is stationary. Stationarity then allows us to apply the LLN and CLT to β_r and as a result, following these simple steps, we can conclude that:

1. The remainder, β_r, vanishes to 0 as the time dimension T approaches infinity:

$$\frac{\sum_{t=2}^{T} \varepsilon_t y_{t-1}}{\sum_{t=2}^{T} y_{t-1}^2} \xrightarrow{p} 0 \quad \text{as} \quad T \to \infty,$$

   hence the estimator $\hat{\beta}_T$ is consistent toward β_0.

2. The remainder, β_r, is asymptotically normally distributed:

$$\frac{\sum_{t=2}^{T} \varepsilon_t y_{t-1}}{\sum_{t=2}^{T} y_{t-1}^2} \xrightarrow{d} N(0, \sigma^2) \quad \text{as} \quad T \to \infty,$$

   hence the estimator $\hat{\beta}_T$ is asymptotically normally distributed around β_0.

These simple results have proved extremely useful over time. For good reasons, the Law of Large Numbers, which took more than a staggering 300 years to complete, has been coined the *Golden Theorem*. In many cases, these simple results are more than just interesting, and remain the work horse of standard analysis approaches that are widely used to support policies and interventions across many domains. However, they are applicable only in the limited setting of linear models and under the very restrictive assumption that this linear relationship describes reality correctly.

2.1.1 The linear Least Squares Estimator

Many empirical problems dealing with repeated cross-sectional data can be analyzed by the linear regression model:

$$\mathbf{y}_t = \alpha + \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t \ \forall t \in \mathbb{N}, \tag{2.3}$$

where $\mathbf{y}_t$ is the dependent vector variable at time t containing $i \in \{1, ..., N\}$ values each observed at a different location, $\mathbf{X}_t$ is a d-dimensional matrix containing the independent or explanatory variables similarly observed at locations $i \in \{1, ..., N\}$ and time t, and $\boldsymbol{\varepsilon}_t$ are the unobserved residuals. The parameter α is a constant, and $\boldsymbol{\beta}$ is a vector of length d containing the marginal effects, or slope parameters, for each variable included in $\mathbf{X}_t$. The error term is assumed to satisfy $\mathbb{E}(\boldsymbol{\varepsilon}_t|\mathbf{X}_t) = 0$. Under this assumption, the linear regression model is a model of the conditional expectations of $\mathbf{y}_t$ given the observed $\mathbf{X}_t$. In particular, one can decompose the problem as follows:

$$\mathbb{E}(\mathbf{y}_t|\mathbf{X}_t) = \mathbb{E}(\alpha + \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t|\mathbf{X}_t). \tag{2.4}$$

Naturally, given that the expectation of a static parameter is simply the value of that parameter, the right hand side can be separated in

individual parts, α, $\mathbb{E}(\mathbf{X}_t|\mathbf{X}_t)\boldsymbol{\beta}$, and $\mathbb{E}(\boldsymbol{\varepsilon}_t|\mathbf{X}_t)$. Furthermore,

$$\mathbb{E}(\mathbf{X}_t|\mathbf{X}_t) = \mathbf{X}_t, \tag{2.5}$$

and by assumption,

$$\mathbb{E}(\boldsymbol{\varepsilon}_t|\mathbf{X}_t) = 0. \tag{2.6}$$

Hence, the expectation of $\mathbf{y}_t$ conditional on observables is simply:

$$\mathbb{E}(\mathbf{y}_t|\mathbf{X}_t) = \alpha + \mathbf{X}_t\boldsymbol{\beta}. \tag{2.7}$$

This interpretation will turn out to remain incredibly useful in the nonlinear case as well, as, no matter how complex the model gets, the modeled data can often be interpreted as local conditional expectations rather than global (average) expectations, which is still an intuitively accessible concept. The key exogeneity assumption used for this, can be summarized as follows:

ASSUMPTION. 1 (Exogeneity of the Regressors). $\mathbb{E}(\boldsymbol{\varepsilon}_t|\mathbf{X}_t) = 0 \ \forall \ t \in \mathbb{N}$.

REMARK. 1. *Note that by a Law of Total Expectation, the Exogeneity of Regressors assumption also implies*

$$\mathbb{E}(\boldsymbol{\varepsilon}_t\mathbf{x}_t) = \mathbb{E}(\mathbb{E}(\boldsymbol{\varepsilon}_t\mathbf{x}_t|\mathbf{x}_t)) = \mathbb{E}(\mathbf{x}_t\mathbb{E}(\boldsymbol{\varepsilon}_t|\mathbf{x}_t)) = \mathbb{E}(\mathbf{x}_t 0) = 0 \ \forall \ t \in \mathbb{N}.$$

Note that $\boldsymbol{\varepsilon}_t$ is a vector of residuals at time t for locations $i \in \{1, ..., N\}$. The conditional expectation condition is stated for vectors indexed by time intervals. Essentially, the parameters in the vector $\boldsymbol{\beta}$ measure the expected changes in the cross-section $\mathbf{y}_t$ given the changes in $\mathbf{X}_t$. While it may well be that $\mathbb{E}(\varepsilon_{it}|\mathbf{x}_{it}) = 0 \ \forall \ t \in \mathbb{N}$ for certain locations (or the cross-sectional mean), $\mathbb{E}(\boldsymbol{\varepsilon}_t|\mathbf{X}_t) = 0 \ \forall \ t \in \mathbb{N}$ may still break, if for example local errors have non-zero expectation $(\varepsilon_{it}|\mathbf{x}_{it}) \neq 0$, which for example occurs when there are expectations about missing components conditional on the data locally in the cross-section. One such example is clustering of residuals in regions in the cross-section, particularly if those clusters tend to remain in place over time. There are many reasons why

this assumption may be difficult to hold in practice. Advanced modeling techniques, including those discussed in later chapters, are in fact often aimed at mitigating these violations.

Let us now first consider the simple LSE that chooses the parameters that minimize the sum of squared residuals from a compact collection of potential solutions $(\mathcal{A}, \mathcal{B})$. Specifically:

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) \;=\; \arg \min_{(\alpha,\boldsymbol{\beta})\in(\mathcal{A},\mathcal{B})} \sum_{t=1}^{T} \boldsymbol{\varepsilon}_t^2 \;=\; \arg \min_{(\alpha,\boldsymbol{\beta})\in(\mathcal{A},\mathcal{B})} \sum_{t=1}^{T} (\mathbf{y}_t - \alpha + \mathbf{X}_t\boldsymbol{\beta})^2. \tag{2.8}$$

As always, the parameters can be found by simply taking the derivative of this Least Squares criterion with respect to it's parameters, and equating 0. Supposing we omit α for a moment, for example because we have demeaned the data such that the average is 0, and focus on the simple case of just one regressor, we can find $\hat{\beta}$ using the derivative:

$$\frac{\partial \sum_{t=1}^{T} (\mathbf{y}_t - \beta \mathbf{x}_t)^2}{\partial \beta} = \sum_{t=1}^{T} (\mathbf{y}_t - \beta \mathbf{x}_t)\mathbf{x}_t, \tag{2.9}$$

which can be rearranged to obtain our estimate explicitly:

$$\hat{\beta}_T = \frac{\sum_{t=1}^{T} \mathbf{y}_t \mathbf{x}_t}{\sum_{t=1}^{T} \mathbf{x}_t^2}. \tag{2.10}$$

Deriving estimators for multiple parameters, each being a marginal effect with respect to a different variable or a simple constant, only involves longer derivations. The linear LSE can always be derived analytically. This is incredibly useful. Even in the nonlinear case we often use flexible functionals that generate parameterizations that are locally linear, in which case the same strategies can be applied for the resulting locally linear expressions only at the cost of longer equations.

The first important step now is to establish that the estimator is consistent toward the parameter of interest. That is, that it converges in probability toward te set of parameters, $(\alpha_0, \boldsymbol{\beta}_0)$, that deliver a correct description of

the data, as $T \to \infty$. This requires us to assume that this set of correct parameters is in fact included in the space of considered parameters $(\mathcal{A}, \mathcal{B})$. We will return to this assumption in later chapters and try to get an understanding of what this truly means, and more importantly, what it means if this assumption breaks. For now, let us summarize:

ASSUMPTION. 2 (Correct Specification of the Model). *The regression* $\mathbf{y}_t = \alpha + \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t \ \forall \ t \in \mathbb{N}$ *is correctly specified.*

As before, this allows us to write the estimator in terms of the *true* parameter and a remainder that involves the residuals, from where we can show that this remainder term converges to 0 as T grows, leaving us with an estimator that converges to the correct result. Let us now state the exact Theorem.

THEOREM. 1 (Bernoulli's Law of Large Numbers for Independent and Identically Distributed Data). *Let* z_1, z_2, z_T *be an Independent and Identically Distributed random variable with finite first moment,* $\mathbb{E}|z_t| < \infty$. *Then,*

$$\frac{1}{T}\sum_{t=1}^{T} z_t \xrightarrow{p} \mathbb{E}(z_t) \quad as \quad T \to \infty.$$

This Theorem tells us that disregard of the distribution of z, the sample average is a consistent estimator of the *true* mean. It is easy to see that this Theorem can also be applied to cross-sectional data, in which case we would index the observations cross-sectionally. The main issue that results is that observations are often not independent across space as by the definition of neighborhood relationships, independence is violated. This similarly applies to the endogenous time series case, in which we assume dependence of observations over time. For now, this Theorem is sufficient as we are interested in the relationship between $\mathbf{y}_t$ and exogenous variables $\mathbf{X}_t$ for which no process has been defined at this point. The application to the LSE follows by first noting that the criterion is a function of random variables, hence noting that it is itself is a random

variable, and then multiplying the numerator and the denominator of the remainder term by $\frac{1}{T}$, and applying the LLN to both components. In particular, again for the simple case,

$$\beta_r = \frac{\sum_{t=2}^{T} \boldsymbol{\varepsilon}_t \mathbf{x}_t}{\sum_{t=2}^{T} \mathbf{x}_t^2} = \frac{\frac{1}{T}\sum_{t=2}^{T} \boldsymbol{\varepsilon}_t \mathbf{x}_t}{\frac{1}{T}\sum_{t=2}^{T} \mathbf{x}_t^2}, \tag{2.11}$$

and if both $\{\boldsymbol{\varepsilon}_t \mathbf{x}_t\}$ and $\{\mathbf{x}_t^2\}$ are i.i.d. with finite first moment $|\boldsymbol{\varepsilon}_t \mathbf{x}_t| < \infty$ and $|\mathbf{x}_t^2| < \infty$, then

$$\frac{1}{T}\sum_{t=2}^{T} \boldsymbol{\varepsilon}_t \mathbf{x}_t \xrightarrow{p} \mathbb{E}(\boldsymbol{\varepsilon}_t \mathbf{x}_t) \quad \text{and} \quad \frac{1}{T}\sum_{t=2}^{T} \mathbf{x}_t^2 \xrightarrow{p} \mathbb{E}(\mathbf{x}_t^2) \quad \text{as} \quad T \to \infty.$$

Note that by our first assumption, $\mathbb{E}(\boldsymbol{\varepsilon}_t \mathbf{x}_t) = 0$, and because the Least Squares criterion is continuous, and functions are limit-preserving even if their arguments are sequences of random variables, the LLN thus delivers

$$\beta_r = \frac{\frac{1}{T}\sum_{t=2}^{T} \boldsymbol{\varepsilon}_t \mathbf{x}_t}{\frac{1}{T}\sum_{t=2}^{T} \mathbf{x}_t^2} \xrightarrow{p} \frac{0}{\mathbb{E}(\mathbf{x}_t^2)} = 0 \quad \text{as} \quad T \to \infty.$$

We have now proven that the estimator is consistent because the error in our estimation converges to zero as we collect more and more data over the time dimension. Note that the above derivations shows the criticality of assuming that the regressors are exogenous $\mathbb{E}(\boldsymbol{\varepsilon}_t | \mathbf{x}_t) = 0$, otherwise

$$\beta_r = \frac{\frac{1}{T}\sum_{t=2}^{T} \boldsymbol{\varepsilon}_t \mathbf{x}_t}{\frac{1}{T}\sum_{t=2}^{T} \mathbf{x}_t^2} \xrightarrow{p} \frac{\eta}{\mathbb{E}(\mathbf{x}_t^2)} = \epsilon \neq 0 \quad \text{as} \quad T \to \infty.$$

With η and ϵ being unknown non-zero components, hence β_r, and therefore $\hat{\beta}_T$, converge to unknown real-valued constants. In other words, we can't really tell what limit our criterion converges to, which renders the entire estimation result quite arbitrary.

Often, the finite moments of lower constituents of complex regression models are introduced as a separate assumption, and we shall see that

instead of assuming these conditions it is often possible to verify the assumptions by defining a process for the endogenous regressors and validating that certain stability conditions and moment-preserving properties hold within specified parameter ranges. For now, let us collect our simple assumption as follows:

ASSUMPTION. 3 (Finite First Moments). *Assume that*

1. $|\boldsymbol{\varepsilon}_t \mathbf{x}_t| < \infty$,
2. *and* $|\mathbf{x}_t^2| < \infty$.

for each $\mathbf{x}_t$ *contained in* $\mathbf{X}_t$.

We can collect the general consistency result of the LSE.

COROLLARY. 1 (Consistency of the Correctly Specified Least Squares Estimator). *Let* $\{\mathbf{y}_t\}_{t\in\mathbb{N}}$ *and* $\{\mathbf{X}_t\}_{t\in\mathbb{N}}$ *be observed sequences, and the model*

$$\mathbf{y}_t = \alpha + \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t \ \forall \ t \in \mathbb{N},$$

be correctly specified. Furthermore, let $\{\boldsymbol{\varepsilon}_t \mathbf{x}_t\}_{t\in\mathbb{N}}$ *and* $\{\mathbf{x}_t^2\}_{t\in\mathbb{N}}$ *be i.i.d. with* $\mathbb{E}(\boldsymbol{\varepsilon}_t|\mathbf{x}_t) = 0 \ \forall \ t \in \mathbb{N}$ *and* $|\boldsymbol{\varepsilon}_t \mathbf{x}_t| < \infty$ *and* $|\mathbf{x}_t^2| < \infty$ *for each* $\mathbf{x}_t$ *contained in* $\mathbf{X}_t$. *Then, the Least Squares estimator of* $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ *defined as*

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = \arg\min_{(\alpha,\boldsymbol{\beta})\in(\mathcal{A},\mathcal{B})} \sum_{t=1}^{T} (\mathbf{y}_t - \alpha + \mathbf{X}_t\boldsymbol{\beta})^2$$

is consistent

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) \xrightarrow{p} (\alpha_0, \boldsymbol{\beta}_0) \quad as \quad T \to \infty.$$

In practice, one is also interested in making statements about the probability that our estimates of individual components in $(\alpha_0, \boldsymbol{\beta}_0)$ are different from 0. That allows us to say that estimated economic effects are *significantly* different from 0, e.g. that an intervention had effect. This requires us to know the distribution of the estimator, which in practice is unknown. Luckily, we can approximate this distribution by appealing to the Central Limit Theorem and showing that the estimator is approximately normally distributed when T is large.

THEOREM. 2 (Lindeberg-Levy's Central Limit Theorem for Independent and Identically Distributed Data). *Let z_1, z_2, z_T be an Independent and Identically Distributed random variable with $\mathbb{E}|z_t| = \mu < \infty$ and $\mathbb{V}\text{ar}(z_t) = \sigma^2 < \infty$, then*

$$\sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T}(z_t - \mu)\right) \xrightarrow{d} N(0, \sigma^2) \quad as \quad T \to \infty.$$

We can now use the CLT to obtain the asymptotic normality of our correct LSE of any parameter by first writing $\sqrt{T}(\hat{\beta} - \beta_0)$ and then plugging in our estimator in terms of the *true* parameter and the remainder term:

$$\sqrt{T}(\hat{\beta}-\beta_0) = \sqrt{T}\big((\beta_0+\beta_r)-\beta_0\big) = \sqrt{T}(\beta_r) = \sqrt{T}\left(\frac{\frac{1}{T}\sum_{t=2}^{T}\boldsymbol{\varepsilon}_t\mathbf{x}_t - \mathbb{E}(\boldsymbol{\varepsilon}_t\mathbf{x}_t)}{\frac{1}{T}\sum_{t=2}^{T}\mathbf{x}_t^2}\right). \tag{2.12}$$

The term $\mathbb{E}(\boldsymbol{\varepsilon}_t\mathbf{x}_t)$ can be added, as by our first assumption, exogeneity of the regressors, this term equals 0. We can now apply the CLT to the numerator:

$$\sqrt{T}\left(\frac{1}{T}\sum_{t=2}^{T}\boldsymbol{\varepsilon}_t\mathbf{x}_t - \mathbb{E}(\boldsymbol{\varepsilon}_t\mathbf{x}_t)\right) \xrightarrow{d} N\left(0, \sigma^2\mathbb{E}\left(\mathbf{x}_t^2\right)\right) \quad \text{as} \quad T \to \infty,$$

and the LLN to the denominator:

$$\left(\frac{1}{T}\sum_{t=2}^{T}\mathbf{x}_t^2\right) \xrightarrow{p} \mathbb{E}\left(\mathbf{x}_t^2\right) \quad \text{as} \quad T \to \infty.$$

By Slutsky's Theorem, we now have

$$\sqrt{T}(\hat{\beta} - \beta_0) \xrightarrow{d} \frac{N\left(0, \sigma^2\mathbb{E}\left(\mathbf{x}_t^2\right)\right)}{\mathbb{E}\left(\mathbf{x}_t^2\right)} \; N\left(0, \sigma^2\mathbb{E}\left[\mathbf{x}_t^2\right]^{-1}\right).$$

This is the standard strategy to deliver asymptotic normality, which we can summarize in the following general result. First, note that the CLT imposes a stricter moment assumption. In particular:

ASSUMPTION. 4 (Finite Second Moments). *Assume that*

1. $\mathbb{V}\text{ar}(\boldsymbol{\varepsilon}_t\mathbf{x}_t) < \sigma^2 < \infty$,

for each $\mathbf{x}_t$ *contained in* $\mathbf{X}_t$.

While this assumption is stated in terms of the second moment, variance, of $\boldsymbol{\varepsilon}_t\mathbf{x}_t$, it is sometimes stated in terms of higher moments of the lower constituents $\boldsymbol{\varepsilon}_t$ and $\mathbf{x}_t$ individually. In particular, since the variance involves squared terms, it can be shown that this assumption involves the finiteness of the fourth moments of $\boldsymbol{\varepsilon}_t$ and each $\mathbf{x}_t$ contained in $\mathbf{X}_t$. Intuitively, if the fourth moments are finite, then the tails of the distributions are relatively short, so the probability that an unusually large observations occurs is small. In that regard, this is interpreted by many as an indication that Least Squares estimates are very sensitive to the presence of outliers. Similar assumptions are however made when establishing the properties of other estimators, including those that aim at outliers-robustness by assuming non-Gaussian distributions that can better accommodate tail events. It turns out that many proofs of multivariate nonlinear estimators require even higher moments to exist.

COROLLARY. 2 (Asymptotic Normality of the Correctly Specified Least Squares Estimator). *Let* $\{\mathbf{y}_t\}_{t\in\mathbb{N}}$ *and* $\{\mathbf{X}_t\}_{t\in\mathbb{N}}$ *be observed sequences, and the model*

$$\mathbf{y}_t = \alpha + \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t \;\forall\; t \in \mathbb{N},$$

be correctly specified. Let $\{\boldsymbol{\varepsilon}_t\mathbf{x}_t\}_{t\in\mathbb{N}}$ *and* $\{\mathbf{x}_t^2\}_{t\in\mathbb{N}}$ *be i.i.d. with* $\mathbb{E}(\boldsymbol{\varepsilon}_t|\mathbf{x}_t) = 0 \;\forall\; t \in \mathbb{N}$ *and* $|\boldsymbol{\varepsilon}_t\mathbf{x}_t| < \infty$ *and* $|\mathbf{x}_t^2| < \infty$ *for each* $\mathbf{x}_t$ *contained in* $\mathbf{X}_t$. *Suppose furthermore that the variances* $\mathbb{V}\text{ar}(\boldsymbol{\varepsilon}_t\mathbf{x}_t) < \sigma^2 < \infty$ *are finite for each* $\mathbf{x}_t$ *contained in* $\mathbf{X}_t$. *Then, the Least Squares estimator of* $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ *defined as*

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = \arg\min_{(\alpha,\boldsymbol{\beta})\in(\mathcal{A},\mathcal{B})} \sum_{t=1}^{T} (\mathbf{y}_t - \alpha + \mathbf{X}_t\boldsymbol{\beta})^2$$

is asymptotically normally distributed for each parameter $\theta \in (\alpha, \boldsymbol{\beta})$ *and variable* $\mathbf{x}_t$ *associated with that parameter*

$$\hat{\theta}_T \xrightarrow{approx} N\left(\theta_0, \sigma^2 \left[\sum_{t=1}^{T} \mathbf{x}_t^2\right]^{-1}\right).$$

Similar results can also be obtained when focusing on the case where $\mathbf{x}_t$ is replaced by a lag of the endogenous variable $\mathbf{y}_{t-1}$. In this case, the exogenous regressors assumption is stated $\mathbb{E}(\boldsymbol{\varepsilon}_t|\mathbf{y}_{t-1}) = 0 \ \forall \ t \in \mathbb{Z}$. This implies that conditional on the past, no further information about the residuals can be available. This essentially requires that the residual process must be free from further correlations after filtering the time-dependencies conditional on lags and observable components from the dependent variable. In many cases there may still be correlations in the innovations, for example because policies impact a process not only idiosyncratically but for prolonged periods. Models therefore often include lagged residuals as explanatory variables. Apart from the need to render an observed time series free from time correlations to fulfill the assumptions needed to apply the LLN and CLT, finite moments can also not simply be assumed when the model is correct. In fact, we know that for certain parameter values the process is explosive such that $\mathbf{y}_t$ is in fact expected to tend to infinity. To prevent this from occurring, we need an additional result that ensures that $\mathbf{y}_t$ is Stationary. The following result specifically, is useful in standard settings.

THEOREM. 3 (Strict Stationarity of a Linear Recursion). *Let* $\{\mathbf{y}_t\}_{t\in\mathbb{Z}}$ *be generated by:*

$$\mathbf{y}_t = \alpha + \phi\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t \ \forall \ t \in \mathbb{Z}.$$

If $|\phi| < 1$ *and* $\boldsymbol{\varepsilon}_t$ *are innovations drawn from* $NID(0, \sigma^2_{\boldsymbol{\varepsilon}})$*, then* $\{\mathbf{y}_t\}_{t\in\mathbb{Z}}$ *is Strictly Stationary, that is the distribution of every finite sub-vector is invariant in time*

$$F_Y(\mathbf{y}_1, ..., \mathbf{y}_\tau) = F_Y(\mathbf{y}_{t+1}, ..., \mathbf{y}_{t+\tau}) \ \forall \ (t, \tau) \in \mathbb{N} \times \mathbb{N}.$$

where $F_Y(\mathbf{y}_{t+1}, ..., \mathbf{y}_{t+\tau})$ *represents the cumulative distribution function of the unconditional joint distribution of* $\{\mathbf{y}_t\}_{t\in\mathbb{Z}}$ *at times* $t + 1, ..., t + \tau$*.*

This stationarity property is incredibly important to obtain properties of estimators because it allows us to make use of the Laws of Large Numbers

for Stationary and Ergodic data, and if the model is correctly specified the Central Limit Theorem for Stationary and Ergodic Martingale Difference Sequences, rather than appealing to the Theorems for *i.i.d.* data. This extension will be discussed in more detail in the next section. If the model remains linear, but multiple (cross-sectional) variables are included, or a single cross-sectional time series is modeled with multiple locational autoregressive parameters $\mathbf{\Phi}\mathbf{y}_{t-1}$ collected in the $N \times N$ matrix $\mathbf{\Phi}$, the linear Stationarity condition can be generalized as $\|\mathbf{\Phi}\| < 1$, using some norm or a spectral radius. However, when the process turns nonlinear, and we can no longer condition on static parameters, proofs for Stationarity become more complex. Particularly when analyzing cross-sectional time series we not only want observations to depend possibly on unique local histories, but also on those of neighbors and possibly even on the contemporaneous values of neighbors. In these cases, models begin to exhibit more complex feedback properties for which proving stability may turn out to be a nontrivial task. At this point, one may start to make explicit distinctions between various types of stability as sometimes weaker forms of stability, that are easier to verify, may already be sufficient to obtain useful properties of estimators.

We shall return extensively to both the stability conditions and the residual dependencies in later chapters. For now, let us explore what happens to our LSE if we would want to model contemporaneous dependencies on neighbors in addition to the exogenous covariates of interest. This will highlight what can already be done with the simple theory that we have developed so far and expose some of its limitations. Suppose we extend our regression model:

$$\mathbf{y}_t = \alpha + \rho W \mathbf{y}_t + \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t \; \forall \; t \in \mathbb{N}, \tag{2.13}$$

in which W is an N by N pre-defined parameter matrix with zero diagonal. We reserve discussion about this matrix, that defines contemporaneous relations with neighboring observations, for later chapters. For now it

is sufficient to see that $\mathbf{y}_t$ occurs on both sides of the equation and the exogenous regressors assumption is thus now stated $\mathbb{E}(\mathbf{y}_t - \alpha - \mathbf{X}_t\boldsymbol{\beta} - \rho W\mathbf{y}_t | W\mathbf{y}_t) = 0 \; \forall \; t \in \mathbb{N}$, which obviously makes little sense to impose since $W\mathbf{y}_t$ occurs on both sides. Only if $\rho = 0$, and the model is non-spatial, the expectation is zero by the fact that the residuals are *i.i.d.* In other words, $W\mathbf{y}_t$ is an endogenous regressor. Contrary to the time series case, where the lagged term of the dependent variable can be uncorrelated with the residual term if there is no serial residual correlation, e.g. if the model is correct, in the spatially lagged case, this correlation occurs regardless of the properties of the residual term. We had already seen that the Least Squares criterion converges to an unknown limit if the exogenous regressor assumption breaks, implying that standard application of the Least Squares criterion delivers arbitrary results.

One option is to invert the equation, and ensure that $\mathbf{y}_t$ only enters on the left side of the equation:

$$(I - \rho W)\mathbf{y}_t = \alpha + \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t \; \forall \; t \in \mathbb{N}, \tag{2.14}$$

with I being an identity matrix. At this point, our dependent variable contains unknown parameters. We can get rid of $(I - \rho W)$ on the left side by devision:

$$\mathbf{y}_t = (I - \rho W)^{-1}\alpha + (I - \rho W)^{-1}\mathbf{X}_t\boldsymbol{\beta} + (I - \rho W)^{-1}\boldsymbol{\varepsilon}_t \; \forall \; t \in \mathbb{N}. \tag{2.15}$$

This highlights that when $\mathbf{y}_t$ is in part a function of $W\mathbf{y}_t$, e.g. when $|\rho| > 0$, $\mathbf{y}_t$ is a nonlinear function of the data and residuals. The model cannot be parameterized and estimated in this form because the residuals result as a product of estimation, hence their values are not available *a priori* as regressors. Chapter 4 discusses that the nonlinearity can be approximated using an infinite power series approximation, which reveals that $\mathbf{y}_t$ not only depends on local observations and neighbors, but also

on the values of residuals and covariates of distant neighbors.

$$\mathbf{y}_t = (I + \rho W + \rho^2 W^2 + ...)\,(\alpha + \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t) \;\; \forall\; t \in \mathbb{N}. \tag{2.16}$$

The influence of distant neighbors will be small if ρ is not too high. This suggests that when spatial dependence is mild and residuals are small, a considerable share of the dependencies can be captured with a first order approximation of the spillover dynamics.

$$\begin{aligned}\mathbf{y}_t &\sim (I + \rho W)\,(\alpha + \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t) + \mu_t \\ &\sim (I + \rho W)\,(\alpha + \mathbf{X}_t\boldsymbol{\beta}) + \boldsymbol{\varepsilon}_t + \mu_t + \xi_t \;\forall\; t \in \mathbb{N},\end{aligned} \tag{2.17}$$

in which μ_t is an approximation error that results from restricting to dependence on first order neighbors, and ξ_t is an additional approximation error that results from neglecting the residual spillovers. The magnitude of both errors increases with $|\rho|$, and the magnitude of ξ_t increases with the magnitude of residuals $\boldsymbol{\varepsilon}_t$. The aim is then to specify as many lower-level constituents of the residuals by incorporating many covariates to ensure that residuals are small, and parameterize spatial dependence on covariates directly to capture the important first order spatial dependence dynamics. The resulting simplified model can consistently be estimated using Least Squares as it is simply equal to a standard regression introduced in the previous section:

$$\mathbf{y}_t = \alpha + \mathbf{X}_t\boldsymbol{\beta} + W\mathbf{X}_t\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_t \;\forall\; t \in \mathbb{N}. \tag{2.18}$$

In this equation, we made use of the fact that $(I + \rho W)\alpha$ simply remains a linear constant and introduced a new unknown set of parameters $\boldsymbol{\beta}_2$ to capture dependence on neighboring values of the exogenous covariates. Note that our simple estimation theorems at this point still require the correct specification assumption to be satisfied, which is unrealistic since we have already established sources of approximation error that stem from neglecting the spatial effects in residuals and dependence on distant

observations.

While the validity of the correct specification assumption can be verified by diagnosing $\boldsymbol{\varepsilon}_t$, the approach may be seen as a dis-satisfactory as it provides no empirical strategy to dealing with residual spatial correlation or pure SAR processes in which exogenous covariates play no role. The question naturally arises if other, alternative, estimators can be thought of that are not prone to this problem and that can handle estimation of spatial disturbance terms directly. It turns out that the problem can be tackled with the framework of Maximum Likelihood.

2.1.2 The linear Maximum Likelihood Estimator

Given T observations $\mathbf{y}_1, ..., \mathbf{y}_T$ from the time series $\{\mathbf{y}_t\}_{t\in\mathbb{Z}}$, generated by the model

$$\mathbf{y}_t = \phi \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t \ \forall \ t \in \mathbb{Z}, \tag{2.19}$$

with $\boldsymbol{\varepsilon}_t$ being drawn from a standardized normal distribution with zero mean. Suppose we have a correctly specified regression. The likelihood function $\ell(\mathbf{y}_1, ..., \mathbf{y}_T; \boldsymbol{\theta})$ is simply the joint density function of the sequence $\mathbf{y}_1, ..., \mathbf{y}_T$ under the parameter vector $\boldsymbol{\theta} = (\phi, \sigma_\varepsilon^2)$ that defines the distribution of the data. Note that if our model would include more or other parameters, they would simply be part of this parameter vector (for example, if we would include a constant as we did earlier, it would be $\boldsymbol{\theta} = (\alpha, \phi, \sigma_\varepsilon^2)$). The MLE is the parameter vector that the maximizes the likelihood function:

$$\hat{\boldsymbol{\theta}}_T = \arg\max_{\boldsymbol{\theta}\in\Theta} \ell(\mathbf{y}_1, ..., \mathbf{y}_T; \boldsymbol{\theta}). \tag{2.20}$$

A useful property of joint density functions is that they can be factorized into the product of conditional and marginal densities:

$$\ell(\mathbf{y}_1, \mathbf{y}_2; \boldsymbol{\theta}) = \ell(\mathbf{y}_1; \boldsymbol{\theta}) \times \ell(\mathbf{y}_2; \boldsymbol{\theta}),$$

$$\ell(\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3; \boldsymbol{\theta}) = \ell(\mathbf{y}_1; \boldsymbol{\theta}) \times \ell(\mathbf{y}_2|\mathbf{y}_1; \boldsymbol{\theta}) \times \ell(\mathbf{y}_3|\mathbf{y}_2, \mathbf{y}_1; \boldsymbol{\theta}),$$

...

$$\ell(\mathbf{y}_1, ..., \mathbf{y}_T; \boldsymbol{\theta}) = \ell(\mathbf{y}_1; \boldsymbol{\theta}) \times \prod_{t=2}^{T} \ell(\mathbf{y}_t|\mathbf{y}_{t-1}, ..., \mathbf{y}_1; \boldsymbol{\theta}). \tag{2.21}$$

Writing the likelihood as a product of conditional densities is useful because we impose the distribution of $\mathbf{y}_t$ conditional on $\mathbf{y}_{t-1}$ through our parameterized model. For example, in the linear autoregressive case that we have assumed, with ϕ being the linear autoregressive parameter, it is

$$\mathbf{y}_t|\mathbf{y}_{t-1} \sim N(\phi \mathbf{y}_{t-1}, \sigma_\varepsilon^2). \tag{2.22}$$

It may also be possible to work with different distributions, for example distributions that can accommodate fatter tails. Different distributional assumptions or models will merely imply that the densities are of another form, which can be accounted for. Under the Gaussian assumption, it is given by the well-known formula:

$$\ell(\mathbf{y}_t|\mathbf{y}_{t-1}; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left[-\frac{(\mathbf{y}_t - \phi \mathbf{y}_{t-1})^2}{2\sigma_\varepsilon^2}\right]. \tag{2.23}$$

Taking logs allows us to express the products as sums, hence we have that the MLE can be written as

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} \sum_{t=2}^{T} -\log\sqrt{2\pi\sigma_\varepsilon^2} - \frac{(\mathbf{y}_t - \phi \mathbf{y}_{t-1})^2}{2\sigma_\varepsilon^2}. \tag{2.24}$$

Just as in the Least Squares case, we can find the estimator by calculating the derivative and setting it to zero. Since in this simple example we have assumed $\sigma = 1$, we will set it to unit. In practice, the variance is often estimated, in which case the derivations have to take into account that σ is itself a free parameter. For now, the estimator for ϕ is simply:

$$\frac{\partial \ell(\mathbf{y}_1, ..., \mathbf{y}_T; \phi)}{\phi} = \sum_{t=2}^{T} (\mathbf{y}_t - \phi \mathbf{y}_{t-1}) \mathbf{y}_{t-1}. \tag{2.25}$$

Equating to zero and rearranging gives us a familiar expression:

$$\hat{\phi} = \frac{\sum_{t=2}^{T} \mathbf{y}_t \mathbf{y}_{t-1}}{\sum_{t=2}^{T} \mathbf{y}_{t-1}^2}. \tag{2.26}$$

In this particular case, in which we have assumed the same model and distributional form of the residuals, the MLE is identical to the LSE that we explored earlier. Since we now have an analytical expression, and have again assumed correct specification, we might expect that that the proofs for consistency and normality will follow the exact same steps from here. This is almost correct. In the early Least Squares example, we worked with a model $\mathbf{y}_t = \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t,\ \forall\ t \in \mathbb{N}$ in which our dependent variable is generated by exogenous, independent, data $\mathbf{X}_t$. In the current example $\mathbf{y}_t = \phi\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t,\ \forall t \in \mathbb{Z}$, our dependent variable is generated only by innovations and temporal dependence. This implies that unlike in the exogenous regressor case, where we can assume that $\mathbf{X}_t$ is *i.i.d.*, we can now no longer assume that $\mathbf{y}_{t-1}$ is *i.i.d.* as we model its dependence explicitly. This implies that we can no longer make use of Bernoulli's LLN and the Lidneberg-Levy's CLT. Note that this is an issue that is not related to the MLE itself, or the LSE vice-versa, it is just because we are now formally considering a time series process. We already hinted earlier that stability of the time series was intuitively important to ensure that $\mathbf{y}_t$ does not wander off to infinity, in which case the expectations are infinite and moment assumptions would surely break. It turns out that the Stationarity property is also key to applying an LLN and CLT. Particularly, since the derivations are identical to the Least Squares example, we want to show that by application of an LLN that

$$\phi_r = \frac{\frac{1}{T}\sum_{t=2}^{T} \boldsymbol{\varepsilon}_t \mathbf{y}_{t-1}}{\frac{1}{T}\sum_{t=2}^{T} \mathbf{y}_{t-1}^2} \xrightarrow{p} \frac{0}{\mathbb{E}(\mathbf{y}_t^2)} = 0 \quad \text{as} \quad T \to \infty,$$

and by application of a CLT to the numerator and a LLN to the denominator, that

$$\sqrt{T}(\phi_r) = \sqrt{T}\left(\frac{\frac{1}{T}\sum_{t=2}^{T}\boldsymbol{\varepsilon}_t\mathbf{y}_{t-1} - \mathbb{E}(\boldsymbol{\varepsilon}_t\mathbf{y}_{t-1})}{\frac{1}{T}\sum_{t=2}^{T}\mathbf{y}_{t-1}^2}\right) \xrightarrow{d} \frac{N(0,\sigma_{\boldsymbol{\varepsilon}}^2\mathbb{E}(\mathbf{y}_t^2))}{\mathbb{E}(\mathbf{y}_{t-1}^2)} \sim$$
$$N\left(0,\sigma_{\boldsymbol{\varepsilon}}^2\mathbb{E}[\mathbf{y}_{t-1}^2]^{-1}\right),$$
$$\text{as} \quad T \to \infty.$$

We can do so by appealing to the following LLN for Strictly Stationary and Ergodic sequences and CLT for Martingale Difference Sequences.

THEOREM. 4 (Birkhoff-Khinchin's Law of Large Numbers for Strictly Stationary and Ergodic data). *Let the random sequence* $\{z_t\}_{t\in\mathbb{Z}}$ *be Strictly Stationary and Ergodic with finite first moment* $\mathbb{E}|z_t| < \infty$*, then we have*

$$\frac{1}{T}\sum_{t=2}^{T} z_t \xrightarrow{p} \mathbb{E}z_t \quad as \quad T \to \infty.$$

THEOREM. 5 (Billingsley's Central Limit Theorem for Stationary and Ergodic Martingale Difference Sequences). *Let the sequence* $\{z_t\}_{t\in\mathbb{Z}}$ *be Strictly Stationary and Ergodic with first moment* $\mathbb{E}(z_t) = \mu < \infty$ *and second moment* $\mathbb{V}\text{ar}(z_t) = \sigma^2 < \infty$*. Suppose furthermore that* $\{z_t\}_{t\in\mathbb{Z}}$ *is a Martingale Difference Sequence of random variables,* $\mathbb{E}(z_t|z_{t-1}, z_{t-2}, ...)\ \forall\ t \in \mathbb{Z}$*, then we have*

$$\sqrt{T}\left(\frac{1}{T}\sum_{t=2}^{T} z_t - \mu\right) \xrightarrow{d} N(0,\sigma^2) \quad as \quad T \to \infty.$$

Using these Theorems, together with the stationarity property, we come to the following results.

COROLLARY. 3 (Consistency of the MLE for the Correctly Specified Autoregressive Model). *Let the time series* $\{\mathbf{y}_t\}_{t\in\mathbb{Z}}$ *be generated by the Strictly Stationary autoregressive model* $\mathbf{y}_t = \phi_0\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t\ \forall\ t \in \mathbb{Z}$*,* $|\phi_0| < 1$*, with exogenous innovations* $\mathbb{E}(\boldsymbol{\varepsilon}_t|\mathbf{y}_{t-1}) = 0$ *that satisfy* $\{\boldsymbol{\varepsilon}_t\}_{t\in\mathbb{Z}} \sim NID(0,\sigma_{\boldsymbol{\varepsilon}}^2)$ *with finite variance* $\sigma_{\boldsymbol{\varepsilon}}^2 < \infty$*. Suppose furthermore that the regression model is correctly specified* $\phi_0 \in \Theta$*, then*

$$\hat{\phi}_T \to \phi_0 \quad as \quad T \to \infty.$$

COROLLARY. 4 (Asymptotic Normality of the MLE for the Correctly Specified Autoregressive Model). *Let the time series* $\{\mathbf{y}_t\}_{t\in\mathbb{Z}}$ *be generated by the Strictly Stationary autoregressive model* $\mathbf{y}_t = \phi_0\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t \ \forall\ t \in \mathbb{Z}$, $|\phi_0| < 1$, *with exogenous innovations* $\mathbb{E}(\boldsymbol{\varepsilon}_t|\mathbf{y}_{t-1}) = 0$ *that satisfy* $\{\boldsymbol{\varepsilon}_t\}_{t\in\mathbb{Z}} \sim NID(0, \sigma^2_{\boldsymbol{\varepsilon}})$ *with finite variance* $\sigma^2_{\boldsymbol{\varepsilon}} < \infty$. *Suppose furthermore that the regression model is correctly specified* $\phi_0 \in \Theta$, *then*

$$\sqrt{T}(\hat{\phi}_T - \phi_0) \xrightarrow{d} N\left(0, \sigma^2_{\boldsymbol{\varepsilon}}[\mathbb{E}\mathbf{y}^2_{t-1}]^{-1}\right) \quad as \quad T \to \infty.$$

Note that the consistency results from applying an LLN to $\frac{1}{T}\sum_{t=2}^{T}\boldsymbol{\varepsilon}_t\mathbf{y}_{t-1}$ and $\frac{1}{T}\sum_{t=2}^{T}\mathbf{y}^2_{t-1}$, which easily follows from the fact that when $\{\mathbf{y}_t\}_{t\in\mathbb{Z}}$ is Stationary and Ergodic, the sequences $\{\mathbf{y}^2_t\}_{t\in\mathbb{Z}}$ and $\{\boldsymbol{\varepsilon}_t\mathbf{y}_{t-1}\}_{t\in\mathbb{Z}}$ are trivially also Stationary and Ergodic. Furthermore, as long as $\sigma^2_{\boldsymbol{\varepsilon}} < \infty$ then $\mathbb{E}|\mathbf{y}^2_t| < \infty$ and $\mathbb{E}|\boldsymbol{\varepsilon}_t\mathbf{y}_{t-1}| < \infty$. Application of the the CLT to $\sqrt{T}\frac{1}{T}\sum_{t=2}^{T}\boldsymbol{\varepsilon}_t\mathbf{y}_{t-1} - \mathbb{E}(\boldsymbol{\varepsilon}_t\mathbf{y}_{t-1})$ requires first that $\mathbb{V}\text{ar}(\boldsymbol{\varepsilon}_t\mathbf{y}_{t-1})$ is finite and that $\boldsymbol{\varepsilon}_t\mathbf{y}_{t-1}$ is a Martingale Difference Sequence, $\mathbb{E}(\boldsymbol{\varepsilon}_t\mathbf{y}_{t-1}|\boldsymbol{\varepsilon}_{t-1}\mathbf{y}_{t-2}, \boldsymbol{\varepsilon}_{t-2}\mathbf{y}_{t-3}, ...) = 0$. The finiteness of the variance can naturally be stated in terms of a moment conditions on the innovations. In particular, if $|\boldsymbol{\varepsilon}^4_t| < \infty$, then $|\mathbf{y}^4_t| < \infty$ and $|(\boldsymbol{\varepsilon}_t\mathbf{y}_t)^2| < \infty$ are easily verified. The martingale difference property follows trivially by the fact that the *true* innovations are exogenous $\mathbb{E}(\boldsymbol{\varepsilon}_t|\mathbf{y}_{t-1}) = 0$ and the model is correctly specified and consistent. Hence, the residuals of the regression around the correct parameter are also exogenous and NID. To verify the martingale difference property, we only have to define $\mathcal{F}_{t-1} := (\boldsymbol{\varepsilon}_{t-1}\mathbf{y}_{t-2}, \boldsymbol{\varepsilon}_{t-2}\mathbf{y}_{t-3}, ...)$ and then need that $\mathbb{E}(\boldsymbol{\varepsilon}_t\mathbf{y}_{t-1}|\mathcal{F}_{t-1}) = 0$. This follows by application of a Law of Total Expectation,

$$\begin{aligned}\mathbb{E}(\boldsymbol{\varepsilon}_t\mathbf{y}_{t-1}|\mathcal{F}_{t-1}) &= \mathbb{E}\big(\mathbb{E}(\boldsymbol{\varepsilon}_t\mathbf{y}_{t-1}|\mathbf{y}_{t-1}, \mathcal{F}_{t-1})|\mathcal{F}_{t-1}\big) \\ &= \mathbb{E}\big(\mathbf{y}_{t-1}\mathbb{E}(\boldsymbol{\varepsilon}_t|\mathbf{y}_{t-1}, \mathcal{F}_{t-1})|\mathcal{F}_{t-1}\big) \\ &= \mathbb{E}\big(\mathbf{y}_{t-1}\mathbb{E}(\boldsymbol{\varepsilon}_t)|\mathcal{F}_{t-1}\big) = \mathbb{E}\big(\mathbf{y}_{t-1}0|\mathcal{F}_{t-1}\big) = 0. \qquad (2.27)\end{aligned}$$

Note that again, this result relies on the fact that we can substitute $\mathbb{E}(\boldsymbol{\varepsilon}_t) = 0$ which holds by the fact that our autoregressive parameter is

consistent with respect to the correct parameter. Hence, the result only follows due to the critical assumption that our simple model correctly reflects reality. Furthermore, Stationarity of the correctly specified model was crucial but simple to show because ϕ_0 is a linear parameter. As soon as we replace ϕ_0 with a nonlinear observation-driven function, the theory that we used to obtain stationarity no longer applies.

Maximum Likelihood is a very flexible framework, and a wide variety of models can be estimated as long as the conditional densities implied by the model can be expressed to derive the log likelihood function. In the case of the spatial autoregressive model, for which the Least Squares assumptions broke down, a likelihood function is also available. In this particular case, one can derive the joint distribution of the dependent variable from that of the residuals using the determinant of the first order derivatives of the functional relationship between the two. Doing the derivations, one will find that the log likelihood function contains new components that account for the feedback term $(I - \rho W)^{-1}$ that multiplies with the residuals. Several additional assumptions are now needed to show that the likelihood function with these additional terms is still continuous, such that it is limit-preserving. In addition, slightly more demanding stability conditions are needed to obtain Stationarity of the model. This has to factor in that stable feedback now has to account for dependence on both past observations and current neighbor values. The last difficulty is then that the added complexities to the log likelihood function result in difficult derivatives, that once set to zero, do not have analytical solutions. This prevents us from obtaining the analytical expression of the estimator, and in particular, showing that the remainder term vanishes as T grows. Analytical intractability is a key problem to solve before we can start to tackle the MLE's of the more complicate spatial autoregressive time series processes.

2.2 General Extremum Estimators

In practice, it is often the case that only simple econometric models lead to analytically tractable estimators. From a practical point of view, the estimation can be easily carried out using numerical methods that approximate the optima and derivatives of interest. However, from a theoretical perspective, the absence of an expression for the estimator implies that we can no longer analyze its properties in the manner that we have done earlier. As an effect, we might numerically find the parameters that maximize likelihood of the spatial autoregressive time series model, but without establishing consistency and normality, we can't really tell how the obtained results can be interpreted. This obviously calls for the need of a more general theory to establish the desired properties. In particular, we can classify most of the estimators of interest as *extremum estimators*, and state general conditions to verify their properties.

2.2.1 General Consistency

Given a probability space $(\Omega, \mathcal{F}, P)$, a random sample $\mathbf{y}_T$, shorthand for the entire sequence $(\mathbf{y}_1, ..., \mathbf{y}_T)$, and a parameter space Θ, we can define an extremum estimator as the measurable map $\hat{\boldsymbol{\theta}}_T : \Omega \to \Theta$

$$\hat{\boldsymbol{\theta}}_T \in \arg\max_{\boldsymbol{\theta} \in \Theta} Q_T(\mathbf{y}_T; \boldsymbol{\theta}). \tag{2.28}$$

The criterion function $Q_T : \mathbb{R}^T \times \Theta \to \mathbb{R}$ is real-valued and random because it is a function of the random sample $\mathbf{y}_T$, which is itself a measurable map $\mathbf{y}_T : \Omega \to \mathbb{R}^T$, hence Q_T is a map $Q_T(\mathbf{y}_T, \cdot) : \Omega \times \Theta \to \mathbb{R}$. When the random sample is realized and we observe $\mathbf{y}_T(\omega) \in \mathbb{R}^T$ for some event $\omega \in \Omega$, then $Q_T(\mathbf{y}_T(\omega), \cdot)$ is a real valued function $Q_T(\mathbf{y}_T(\omega), \cdot) : \Omega \times \Theta \to \mathbb{R}$. Hence, for every realization we get a new function to maximize, and we obtain a new maximizer that is our parameter estimate. Hence, the estimators we consider are random.

Note that the maximizer is a set in the arg max as at this point we have not yet said anything about the uniqueness of a maximum. Extremum estimators that take the form of a sum are called an M-estimator, while those with criterion functions $Q_T(\mathbf{y}_T, \cdot) : \Omega \times \Theta \to \mathbb{R}$ that are differentiable on the parameter space Θ can also be written as Z-estimators that directly set the derivative of the criterion to zero. If $Q_T(\mathbf{y}_T, \cdot) : \Omega \times \Theta \to \mathbb{R}$ is also strictly concave, then it ensures that the point where $\nabla Q_T(\mathbf{y}_T, \cdot) = 0$ is really the global, rather than a local, maximum of the function $Q_T(\mathbf{y}_T, \cdot)$. Strict concavity is not necessary, for a twice differentiable criterion one may also use the second derivative to infer which solution corresponds to the global maximum. In any case, one can define the estimate as an element

$$\hat{\boldsymbol{\theta}}_T \in \{\boldsymbol{\theta} \in \Theta : \nabla Q_T(\mathbf{y}_T, \cdot) = 0\}. \tag{2.29}$$

The first thing that one generally wants to ensure is that $\hat{\boldsymbol{\theta}}_T \notin \emptyset$, which can be shown by applying a Bolzano-Weierstrass Theorem. In particular, this Theorem tells us that every function that is continuous, has a maximum on a compact set. This leads us to the following standard assumption and the implied useful result.

ASSUMPTION. 5 (Compactness of the Parameter Space). *Let Θ be a compact space in $\mathbb{R}^{n \in \mathbb{N}}$.*

The compactness assumption is standard, and apart from it's critical role in establishing existence and measurability, it will again play a crucial role in the uniform convergence of the estimator.

THEOREM. 6 (Existence and Measurability of the Estimator). *Let Θ be a compact space in $\mathbb{R}^{n \in \mathbb{N}}$ and $Q_T(\cdot; \cdot)$ be continuous in its arguments, then there exists a measurable map $\hat{\boldsymbol{\theta}}_T : \Omega \to \Theta$ satisfying*

$$\hat{\boldsymbol{\theta}}_T \in \arg\max_{\boldsymbol{\theta} \in \Theta} Q_T(\mathbf{y}_T; \boldsymbol{\theta}).$$

Apart from Existence and Measurability, one typically wants the maximizing point $\boldsymbol{\theta}_0$ to be identifiable and unique.

ASSUMPTION. 6 (Identifiable Uniqueness of the Maximizer of the Limit Criterion). *Let* $\boldsymbol{\theta}_0 \in \Theta$ *be the identifiable unique maximizer of the limit criterion* $Q_\infty : \Theta \to \mathbb{R}$.

There are different definitions with varying mathematical detail. Typically, we mean that $\boldsymbol{\theta}_0$ not only maximizes the limit criterion Q_∞ , i.e. that $Q_\infty(\boldsymbol{\theta}_0) \geq Q_\infty(\boldsymbol{\theta}) \; \forall \; \boldsymbol{\theta} \in \Theta$, but that this point is well separated from other points. If $S(\boldsymbol{\theta}_0, r)$ is the set of points contained in a ball with fixed radius $r > 0$ and center-point $\boldsymbol{\theta}_0$ and $S^c(\boldsymbol{\theta}_0, r)$ denotes its complement set in Θ, i.e

$$S(\boldsymbol{\theta}_0, r) := \{\boldsymbol{\theta} \in \Theta : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < r\} \text{ and } S^c(\boldsymbol{\theta}_0, r) := \{\boldsymbol{\theta} \in \Theta : \boldsymbol{\theta} \notin S(\boldsymbol{\theta}_0, r)\},$$

then $\boldsymbol{\theta}_0$ is not only the maximizer of Q_∞, but also the identifiable unique maximizer if

$$\sup_{\boldsymbol{\theta} \in S^c(\boldsymbol{\theta}_0, r)} Q_\infty(\boldsymbol{\theta}_0) > Q_\infty(\boldsymbol{\theta}). \tag{2.30}$$

Essentially, this says that if we draw a sphere around the correct parameter $\boldsymbol{\theta}_0$ with any positive real valued radius, then the criterion always judges that any parameter outside of that sphere is not optimal, even as the radius of that sphere becomes arbitrarily small. Note that the identifiability is thus a property of the criterion, and characterizes it's ability to differentiate between possible likely solutions.

We now have the right conditions in place to establish consistency of the extremum estimator. In particular, an estimator is (weakly) consistent, *if and only if* it convergences in probability $\hat{\boldsymbol{\theta}}_T \xrightarrow{p} \boldsymbol{\theta}_0$ as $T \to \infty$, and strongly consistent, *if and only if* it convergences *almost surely* $\hat{\boldsymbol{\theta}}_T \xrightarrow{a.s.} \boldsymbol{\theta}_0$ as $T \to \infty$. Weak consistency states that for a specified large T, the estimator $\hat{\boldsymbol{\theta}}_T$ is likely to be near it's correct value $\boldsymbol{\theta}_0$, leaving open the possibility that one can find some arbitrary $\epsilon > 0$ for which $|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0| > \epsilon$ still happens an infinite number of times, although at infrequent intervals. Strong consistency instead states that this will in fact *almost surely* not occur. In particular, it implies that with probability 1, we have that

for any $\epsilon > 0$ the inequality $|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0| < \epsilon$ holds when T has become large enough. Either result can be obtained following a similar strategy, though strong consistency requires a stricter condition that may in some cases not hold while the conditions for (weak) consistency may still be verified.

The general consistency theorem for the criterion of an extremum estimator requires uniform convergence of the criterion function to a limit deterministic function. We say that the criterion function Q_T converges point-wise in probability over Θ to a limit function Q_∞ if it holds true that $|Q_T(\mathbf{y}_T; \boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta})| \xrightarrow{p} 0 \; \forall \; \boldsymbol{\theta} \in \Theta$ as $T \to \infty$. Moreover, we say that the criterion function Q_T converges uniformly in probability over Θ to a limit function Q_∞ if it holds true that $\sup_{\boldsymbol{\theta} \in \Theta} |Q_T(\mathbf{y}_T; \boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta})| \xrightarrow{p} 0$ as $T \to \infty$. The difference lies in the fact that the latter is expressed for the *supremum*, which can loosely be interpreted as the "worst case"-convergence across all elements in Θ. The point-wise convergence is thus a much weaker condition than the uniform convergence, since for point-wise convergence the rate of convergence can be different for each element in Θ. More so, while uniform convergence implies point-wise convergence, point-wise convergence does not imply uniform convergence. Unfortunately, directly establishing uniform convergence is often not easy. However, due to a remarkable result known as the stochastic Arzelà-Ascoli Theorem, it is known that point-wise convergence of the criterion function over a compact parameter space implies uniform convergence if the estimator is stochastically equicontinuous. A family of functions is equicontinuous if all the functions are continuous and they have equal variation over a given neighborhood. By itself, stochastic equicontinuity is not an easy to use concept, but a Lipschitz condition implies stochastic uniform equicontinuity. This gives us the very easy to use condition that if $\sup_T \mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} \|\partial Q_T(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}\| < \infty$, then the sequence of random criterion functions at sample size T generated under $\omega \in \Omega$ is stochastically equicontinuous. Furthermore, in order to obtain strong

uniform convergence of Q_T to Q_∞ we need it to be strongly stochastically equicontinuous, which requires that the derivative be uniformly bounded rather than bounded in expectation $\sup_T \sup_{\boldsymbol{\theta}\in\Theta} \|\partial Q_T(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\| < \infty$ *almost surely*. As a result, it is quite straightforward to verify that the criterion function of an extremum estimator is (strongly) consistent. In particular by applying a suitable LLN to obtain point-wise convergence and using the bounded expectation of the derivative of the criterion to obtain the uniform convergence. Strong consistency of the criterion can be obtained by applying an LLN to obtain point-wise convergence and using the uniform boundedness of the derivative of the criterion to obtain strong uniform convergence. This can be summarized as follows.

THEOREM. 7 (General Consistency for M-estimators). *Let* $(\Omega, \mathcal{F}, P)$ *be a probability space, and let the criterion function* $Q_T : \Omega \times \Theta \to \mathbb{R}$ *be a sequence of random continuous functions that take the form*

$$Q_T(\mathbf{y}_T; \boldsymbol{\theta}) = \textstyle\sum_{t=2}^{T} q(\mathbf{y}_t, \mathbf{y}_{t-1}; \boldsymbol{\theta}),$$

and q *be differentiable on a convex compact parameter space* Θ*. Assume that also that* $q(\mathbf{y}_t, \mathbf{y}_{t-1}; \boldsymbol{\theta})$ *is Stationary and Ergodic and has bounded first moment* $\mathbb{E}|q(\mathbf{y}_t, \mathbf{y}_{t-1}; \boldsymbol{\theta})| < \infty$*. Then the criterion satisfies a Law of Large Numbers*

$$\frac{1}{T}\textstyle\sum_{t=2}^{T} q(\mathbf{y}_t, \mathbf{y}_{t-1}; \boldsymbol{\theta}) \xrightarrow{p} \mathbb{E}(q(\mathbf{y}_t, \mathbf{y}_{t-1}; \boldsymbol{\theta})) \quad as \quad T \to \infty \quad \forall\ \boldsymbol{\theta} \in \Theta,$$

hence the sequence $\{Q_T\}_{T\in\mathbb{N}}$ *converges point-wise in probability to a limit function* $Q_\infty = \mathbb{E}(q(\mathbf{y}_t, \mathbf{y}_{t-1}; \boldsymbol{\theta}))\ \forall\ \boldsymbol{\theta} \in \Theta$*. If furthermore,* $q(\mathbf{y}_t, \mathbf{y}_{t-1}; \boldsymbol{\theta})$ *has a derivative with bounded expectation,*

$$\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} \|\partial q(\mathbf{y}_t, \mathbf{y}_{t-1}; \boldsymbol{\theta})/\partial\boldsymbol{\theta}\| < \infty,$$

then $\{Q_T\}_{T\in\mathbb{N}}$ *is stochastically equicontinuous. Together with the point-wise convergence this implies that* $\{Q_T\}_{T\in\mathbb{N}}$ *then converges uniformly in probability to the limit function* Q_∞

$$\sup_{\boldsymbol{\theta}\in\Theta} |Q_T(\mathbf{y}_T; \boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta})| \xrightarrow{p} 0 \quad as \quad T \to \infty.$$

The compactness of Θ *together with the continuity of* q *implies that the measurable map* $\hat{\boldsymbol{\theta}}_T : \Omega \to \Theta$ *exists, satisfying*

$$\hat{\boldsymbol{\theta}}_T \in \arg\max_{\boldsymbol{\theta}\in\Theta} Q_T(\mathbf{y}_T; \boldsymbol{\theta}).$$

If furthermore the parameter $\boldsymbol{\theta}_0$ *is the identifiably unique maximizer of the limit criterion function* Q_∞

$$\sup_{\boldsymbol{\theta}\in S^c(\boldsymbol{\theta}_0, r)} Q_\infty(\boldsymbol{\theta}_0) > Q_\infty(\boldsymbol{\theta}),$$

then the uniform convergence implies that $\hat{\boldsymbol{\theta}}_T$ *is consistent for* $\boldsymbol{\theta}_0$ *since*

$$|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0| \xrightarrow{p} 0 \quad as \quad T \to \infty.$$

If finally, the derivative is also uniformly bounded,

$$\sup_{\boldsymbol{\theta}\in\Theta} \|\partial q(\mathbf{y}_t, \mathbf{y}_{t-1}; \boldsymbol{\theta})/\partial\boldsymbol{\theta}\| < \infty \ a.s.$$

then $\{Q_T\}_{T\in\mathbb{N}}$ *is strongly stochastically equicontinuous. The strong stochastic equicontinuity together with the established point-wise convergence implies that* $\{Q_T\}_{T\in\mathbb{N}}$ *converges uniformly almost surely to the limit function* Q_∞

$$\sup_{\boldsymbol{\theta}\in\Theta} |Q_T(\mathbf{y}_T; \boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta})| \xrightarrow{a.s.} 0 \quad as \quad T \to \infty,$$

hence that $\hat{\boldsymbol{\theta}}_T$ *is also strongly consistent for the identifiably unique maximizer of the limit criterion function* $\boldsymbol{\theta}_0$ *since*

$$|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0| \xrightarrow{a.s.} 0 \quad as \quad T \to \infty.$$

2.2.2 General asymptotic Normality

The general consistency theorem can be applied in a wide range of settings. The asymptotic normality follows by a very similar argument. In particular, we can show uniform convergence of the second derivative in turn obtained from the point-wise convergence of the second derivative together with boundedness of the third derivative that implies the stochastic equicontinuity of the second derivative.

The reason behind the central role of the second derivative can be easily understood. First, focus on the fact that we are interested in obtaining an approximate limit distribution for $\sqrt{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0}$. Remember that by construction of our estimate as the optimum of the criterion, it holds that, at the estimate, the derivative of the criterion is zero $\nabla Q_T(\mathbf{y}_T, \hat{\boldsymbol{\theta}}) = 0$. Suppose we introduce a new point $\boldsymbol{\theta}_T^*$ that lies between $\hat{\boldsymbol{\theta}}_T$ and $\boldsymbol{\theta}_0$, then we can use the *Mean Value Theorem* to write the derivative as

$$\nabla Q_T(\mathbf{y}_T, \hat{\boldsymbol{\theta}}_T) = \nabla Q_T(\mathbf{y}_T, \boldsymbol{\theta}_0) + \nabla^2 Q_T(\mathbf{y}_T, \boldsymbol{\theta}_T^*)(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) = 0. \quad (2.31)$$

We can now obtain an expression for $\sqrt{(}\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0)$ by rewriting the second equality and multiplying both sides by the square root of T.

$$\sqrt{(}\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) = \left(\nabla^2 Q_T(\mathbf{y}_T, \boldsymbol{\theta}_T^*)\right)^{-1} \times \sqrt{T}\nabla Q_T(\mathbf{y}_T, \boldsymbol{\theta}_0). \quad (2.32)$$

This immediately suggests that obtaining the asymptotic normality of $\hat{\boldsymbol{\theta}}_T$ at $\boldsymbol{\theta}_0$ follows in three steps. First, by showing that $\sqrt{T}\nabla Q_T(\mathbf{y}_T, \boldsymbol{\theta}_0)$ converges in distribution to $N(0, \Sigma)$. Second, by showing that $\left(\nabla^2 Q_T(\mathbf{y}_T, \boldsymbol{\theta}_T^*)\right)^{-1}$ converges in probability to $\left(\nabla^2 Q_T(\boldsymbol{\theta}_0)\right)^{-1}$ as $T \to \infty$. Since $\boldsymbol{\theta}_T^*$ is evaluated between $\hat{\boldsymbol{\theta}}_T$ and $\boldsymbol{\theta}_0$, the consistency of $\hat{\boldsymbol{\theta}}_T$ implies that $\boldsymbol{\theta}_T^*$ approaches $\boldsymbol{\theta}_0$. Hence, the convergence of $\left(\nabla^2 Q_T(\mathbf{y}_T, \boldsymbol{\theta}_T^*)\right)^{-1}$ to $\left(\nabla^2 Q_T(\boldsymbol{\theta}_0)\right)^{-1}$ in probability, follows by showing that $\nabla^2 Q_T$ converges uniformly over Θ to $\nabla^2 Q_\infty$. Finally, to obtain the convergence to $\left(\nabla^2 Q_T(\boldsymbol{\theta}_0)\right)^{-1}$, we must establish that the limit is invertible.

The first condition, that the scaled criterion derivative $\sqrt{(}T)\nabla Q_T(\mathbf{y}_T, \boldsymbol{\theta}_0)$ converges in distribution to $N(0, \Sigma)$ can be obtained by applying a CLT to the derivative $\nabla q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta})$. This is straightforward because

$$\sqrt{T}\nabla Q_T(\mathbf{y}_T, \boldsymbol{\theta}_0) = \sqrt{T}\frac{1}{T}\sum_{t=2}^{T} \nabla q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}_0)$$

$$= \sqrt{T}\left(\frac{1}{T}\sum_{t=2}^{T}\nabla q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}_0) - \mathbb{E}\big(\nabla q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}_0)\big)\right) \tag{2.33}$$

where, just as before, $\mathbb{E}\big(\nabla q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}_0)$ can be added as it equals zero by construction. Recall that the CLT will require the derivative of the criterion $\nabla q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}_0)$ to be Stationary and Ergodic Martingale Difference Sequence and be bounded in second moment $\mathbb{E}\|\nabla q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}_0)\|^2 < \infty$.

The uniform convergence of the second derivative of the criterion function $\sup_{\boldsymbol{\theta}\in\Theta}\|\nabla^2 Q_T(\mathbf{y}_T, \boldsymbol{\theta}) - \nabla^2 Q_\infty(\boldsymbol{\theta})\| \xrightarrow{p} 0 \;\text{ as }\; T \to \infty$ can be obtained by the same strategy as followed in the consistency Theorem. In particular, the stochastic Arzelà-Ascoli Theorem tells us we can focus the argument on the point-wise convergence of

$$\|\nabla^2 Q_T(\mathbf{y}_T, \boldsymbol{\theta}) - \nabla^2 Q_\infty(\boldsymbol{\theta})\| \xrightarrow{p} 0 \;\;\forall\;\; \boldsymbol{\theta} \in \Theta \;\text{ as }\; T \to \infty,$$

and the stochastic equicontinuity of $\{\nabla^2 Q_T\}$, in turn implied by a Lipschitz condition ensured by a bounds on it's derivative

$$\sup_T \mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\|\nabla^3 Q_T(\mathbf{y}_T, \boldsymbol{\theta})\| < \infty.$$

The final invertibility requirement is strongly related to the identification of $\boldsymbol{\theta}_0$. In particular, when $\boldsymbol{\theta}_0$ is well-separated, then the limit criterion Q_∞ must have strong curvature around $\boldsymbol{\theta}_0$. This strong curvature implies that Q_∞ accelerates moving from $\boldsymbol{\theta}_0$ to any point around it, hence that the second derivative $\nabla^2 Q_\infty$ is non-singular and invertible. If, on the other hand, Q_∞ is flat around $\boldsymbol{\theta}_0$, then $\nabla^2 Q_\infty$ is singular and hence not invertible.

We can thus summarize the general normality theorem for extremum estimators as follows.

THEOREM. 8 (General Asymptotic Normality for M-estimators). *Let Θ be a compact parameter space and $\hat{\boldsymbol{\theta}}_T$ be a consistent M-estimator for an identifiable unique point $\boldsymbol{\theta}_0 \in \Theta$.*

Suppose that the derivative of the criterion $\nabla q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}_0)$ *is a Stationary and Ergodic Martingale Difference Sequence and bounded in second moment* $\mathbb{E}\|\nabla q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}_0)\|^2 < \infty$*, then it is asymptotically normal at*

$$\sqrt{T}\left(\frac{1}{T}\sum_{t=2}^{T}\nabla q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}_0)\right) \xrightarrow{d} N(0, \Sigma) \quad as \quad T \to \infty.$$

Suppose also that the second derivative is Stationary and Ergodic and bounded

$$\sup_T \mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\|\nabla^2 q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}_0)\| < \infty,$$

then application of a Law of Large Numbers yields that that the second derivative converges point-wise

$$\|\frac{1}{T}\sum_{t=2}^{T}\nabla^2 q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}) - \mathbb{E}\nabla^2 q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta})\| \xrightarrow{p} 0 \quad \forall \quad \boldsymbol{\theta} \in \Theta \quad as \quad T \to \infty.$$

Suppose furthermore that the third derivative is bounded

$$\sup_T \mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}\|\nabla^3 q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}_0)\| < \infty,$$

then the point-wise convergence and the stochastic equicontinuity of the second derivative imply that it also converges uniformly

$$\sup_{\boldsymbol{\theta}\in\Theta}\|\frac{1}{T}\sum_{t=2}^{T}\nabla^2 q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}) - \mathbb{E}\nabla^2 q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta})\| \xrightarrow{p} 0.$$

Finally, by the invertibility of the limit $\mathbb{E}\nabla^2 q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta})$*, implied by a strong curvature of the criterion around* $\boldsymbol{\theta}_0$ *in turn ensured by the identifiability of* $\boldsymbol{\theta}_0$*, together with the established uniform convergence and asymptotic normality at* $\boldsymbol{\theta}_0$*, implies that*

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, E\Sigma E') \quad as \quad T \to \infty,$$

with $E = \left(\mathbb{E}\nabla^2 q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}_0)\right)^{-1}$.

We now have a general theory that can be applied to show the consistency and normality of possibly complex models like the spatial autoregressive

time series model. All the exact derivations of the conditions and steps will not be provided here. Instead, Chapter 4 provides a more general proof that covers the linear spatial autoregressive model but also allows for the possible failure of several simplifying assumptions that have been made here. In particular, the result also covers cases in which the spatial autoregressive parameter is nonlinear, possibly observation-driven, and under a more general distributional assumption than the Gaussian one. The theory there shall also detail what happens in the case of multiple optima.

2.3 Further complications when modeling dynamic spatial time series

To conclude this chapter, we will look at the steps of the General Consistency and Normality Theorems more closely and discuss them further with regard to the MLE of a general, unparameterized, nonlinear autoregressive model. This broad setting covers, among possible other dynamics, the spatial autoregressive ones that we were particularly interested in. We then discuss each of the assumptions that were made and aim to provide relevant meaning to them for the cases in which one would consider certain parameterizations. We then wish to find out what might still be easily violated or difficult to show, and set several theoretical objectives to remedy those situations.

Suppose we have possibly nonlinear model that depends on both past and current values

$$\mathbf{y}_t = \psi(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\varepsilon}_t; \boldsymbol{\theta}) \ \forall \ t \in \mathbb{Z}. \tag{2.34}$$

Note that this includes also popular linear spatial time series, for example of the form

$$\mathbf{y}_t = \alpha + \rho W \mathbf{y}_t + \phi \mathbf{y}_{t-1} + \phi_2 W \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t \ \forall \ t \in \mathbb{Z}. \tag{2.35}$$

In this model, the values of $\mathbf{y}_t$ depend on past local and neighbor values, and spillover contemporaneously across regions. The extension to dependence on past residuals, and possible spatial lags thereof, will be made in Chapter 4. In 6, extensions will also be made that allow dependence on past residuals of a different spatial variable. We already noted that models currently discussed can be written in the following form

$$\mathbf{y}_t = f(\mathbf{y}_{t-1}, \boldsymbol{\theta}) + v(\boldsymbol{\varepsilon}_t, \boldsymbol{\theta}) \; \forall \; t \in \mathbb{Z}. \tag{2.36}$$

This highlights that the current notation is general enough to also account for an additional spatial error process. In contrast to the earlier, more simplistic, example in which the residuals were not nonlinearly transformed such that we could obtained the density directly from the dependencies implied by the regression model, we now obtain the density $\boldsymbol{\varepsilon}_t \sim p_{\boldsymbol{\varepsilon}}(\boldsymbol{\theta})$ by inverting v. In general, for a regression model that has instantaneous dependence

$$\mathbf{y}_t = h(\boldsymbol{\theta})\mathbf{y}_t + g(\mathbf{y}_{t-1}, \boldsymbol{\theta}) + \boldsymbol{\varepsilon}_t \; \forall \; t \in \mathbb{Z}, \tag{2.37}$$

one can define the contribution to the log likelihood at time t as

$$\ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1}) = \log\det|I - h(\boldsymbol{\theta})| + \log p_{\boldsymbol{\varepsilon}}\Big(\big(I - h(\boldsymbol{\theta})\big)\mathbf{y}_t - f(\mathbf{y}_{t-1}, \boldsymbol{\theta}), \boldsymbol{\theta}\Big). \tag{2.38}$$

The log likelihood function is again defined as $\ell_T(\mathbf{y}_T, \boldsymbol{\theta}) = \sum_{t=2}^{T} \ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1})$. The component $\log\det|I - h(\boldsymbol{\theta})|$ stems from the fact that inverting the model leads to a residual dynamic $(I - h(\boldsymbol{\theta}))^{-1}\boldsymbol{\varepsilon}_t$. Hence, if no autoregressive dynamics would be modeled but only a spatial error process, a similar component would enter the log likelihood function. In the standard nonlinear case, without feedback, $I - h(\boldsymbol{\theta}) = I - 0 = I$, hence $\log\det|I - h(\boldsymbol{\theta})| = \log\det|I| = 0$. We thus obtain the standard nonlinear log likelihood contribution $\log p_{\boldsymbol{\varepsilon}}(\mathbf{y}_t - f(\mathbf{y}_{t-1}, \boldsymbol{\theta}), \boldsymbol{\theta})$ in the standard case in which $h(\boldsymbol{\theta})$ does not transform the data. When $h(\boldsymbol{\theta})$ does transform the data, then immediately, the established continuity prop-

erties of many density functions that are used in empirical applications are complicated by the additional component $\log\det|I - h(\boldsymbol{\theta})|$. We must thus ensure the non-singularity and boundedness of this additional term before we can obtain any result that relies on the continuity of $\ell_T(\mathbf{y}_T, \boldsymbol{\theta})$. Furthermore, calculating the derivatives of the log likelihood function can get particularly complicated. To apply an LLN, we need $\ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1})$ to be differentiable, Stationary and Ergodic, and be bounded in first moment. In contrast to the standard case, the verification of these properties has to take into account the additional component $h(\boldsymbol{\theta})$ that would normally not complicate any result. In particular, proving a suitable stationarity result under $\boldsymbol{\theta}_0$, that cannot be assumed as this is now a property of the model, can be significantly more complex when we need to control for both temporal dependence and instantaneous feedback. The stationarity results known in time series literature that only focus on stability of $f(\mathbf{y}_{t-1}, \boldsymbol{\theta})$ are not sufficient, neither are the stability results from spatial literature that only focus on $h(\boldsymbol{\theta})$. Suppose, however, that suitable forms of stability across the space-time dimension have been verified, then we could obtain the point-wise convergence

$$\frac{1}{T}\sum_{t=2}^{T}(\ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1})) \xrightarrow{p} \mathbb{E}(\ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1})) \text{ as } T \to \infty \ \forall\ \boldsymbol{\theta} \in \Theta,$$

hence we would establish that the sequence $\{\ell_T(\mathbf{y}_T, \boldsymbol{\theta})\}_{T\in\mathbb{N}}$ converges point-wise to a limit deterministic function $\ell_\infty(\boldsymbol{\theta}) = \mathbb{E}(\ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1}))$. Next we need the derivative of the log likelihood, the score, at each step $\ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1})'$ to be bounded and the parameter space, that also includes any components part of $h(\boldsymbol{\theta})$, to be compact, to obtain strong stochastic equicontinuity and thus the uniform convergence of the log likelihood function

$$\sup_{\boldsymbol{\theta}\in\Theta}|\ell_T(\mathbf{y}_T, \boldsymbol{\theta}) - \ell_\infty(\boldsymbol{\theta})| \xrightarrow{a.s.} 0 \text{ as } T \to \infty.$$

For example, if the score is uniformly bounded we have *almost sure* uniform convergence by the strong stochastic equicontinuity and point-

wise convergence. If $\boldsymbol{\theta}_0$ is the identifiable unique maximizer in a compact space Θ we then obtain strong consistency of the estimator $\hat{\boldsymbol{\theta}}_T$ for $\boldsymbol{\theta}_0$. Remember that if the weaker stochastic equicontinuity is obtained instead, we can still show the point-wise convergence of the log likelihood function and obtain the weak consistency result. It is not uncommon for nonlinear models to introduce identification complications, particularly when the data is in fact linear. For example, for a function $\delta/(\gamma(\mathbf{y}_t))$, the parameter δ can be any value if $\gamma(\mathbf{y}_t) = 0$, or, if instead $\delta = 0$, the quantity $\gamma(\mathbf{y}_t)$ can take on any value without affecting the value of the criterion function. We will see in Chapter 4 that the estimator can still be set-consistent, but for normality, however, identification is critical.

In particular, to establish normality we need first that the score is a Stationary and Ergodic Martingale Difference Sequence and bounded in second moment.

$$\mathbb{E}\|\ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1})'\|^2 < \infty.$$

The stationarity and ergodicity can be obtained by continuous differentiability of ℓ and the stationarity and ergodicity of $\{\ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1})\}_{t\in\mathbb{Z}}$. The second moment can be similarly obtained from $\{\ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1})\}_{t\in\mathbb{Z}}$ by deriving moment preserving properties that again have to factor in the properties of $h(\boldsymbol{\theta})$. The score is a Martingale Difference Sequence if the model is correctly specified and the criterion is consistent. Naturally, $h(\boldsymbol{\theta})$ can tremendously improve the fit and help ensure the appropriateness of the Martingale Difference Sequence assumption. Application of a CLT then delivers

$$\sqrt{T}\frac{1}{T}\textstyle\sum_{t=2}^{T} \ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1})' \xrightarrow{d} N(0, \Sigma) \;\forall\; \boldsymbol{\theta} \in \Theta.$$

Twice continuous differentiability of ℓ delivers stationarity and ergodicity of $\{\ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1})''\}_{t\in\mathbb{Z}}$, and together with the moment bound

$$\mathbb{E}\|\ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1})''\| < \infty,$$

this allows us to apply an LLN for every $\boldsymbol{\theta} \in \Theta$ to obtain

$$\frac{1}{T}\sum_{t=2}^{T} \ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1})'' \xrightarrow{p} \mathbb{E}\ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1})'' \text{ as } T \to \infty \ \forall\ \boldsymbol{\theta} \in \Theta.$$

Three times continuous differentiability of ℓ together with a uniform bound on the third derivative ensures the stochastic equicontinuity of the second derivative and thus the uniform convergence. Together with the invertibility of the limit second derivative, this deliver the asymptotic normality of the estimator. Again, the invertibility relies on parameter identification that can be complicated by identification problems of nonlinear functions. This is especially problematic since one would normally use the approximate distribution to infer whether a parameter is significantly different from zero. In many cases, nonlinear models are only asymptotically normal under the alternative assumption that the data is nonlinear. If one needs to assume nonlinearity to test for nonlinearity, then that test statistic is deeply flawed in some sense.

The proof for normality relies heavily on continuity and bounds of derivatives of the log likelihood function. In some sense this can be understood as smoothness, or, well-behavedness properties. The additional component $\log\det|I - h(\boldsymbol{\theta})|$ can complicate the derivations significantly which may make verifying these properties quite unpleasant. At a high level, one can easily understand that the smoothness of the log likelihood function depends on the type of nonlinearities generated by $h(\boldsymbol{\theta})$ and $g(\mathbf{y}_{t-1}, \boldsymbol{\theta})$. In these cases one may note that the stochastic equicontinuity, in turn implied by Lipschitz conditions, is only used as an optional tool that allows one to exploit the often easier to obtain point-wise convergence. It may naturally be possible to show uniform convergence directly. For example, instead of stochastically bounding the derivatives, such computations can be avoided when the Ergodic Theorem for random elements with values in a separable Banach space is applied. This is the strategy that we will take in Chapter 4. In other situations, one may try to show that the Lipschitz conditions are themselves ensured by higher-level smoothness

properties such as higher moment bounds of a function. For a sufficient degree of smoothness, the Lipschitz conditions may stretch out across a sufficient number of derivatives to immediately ensure that the second derivative of the log likelihood function is stochastically equicontinuous without the need of taking the derivations.

Finally, the discussions here all circled around compact parameter spaces. It was mentioned that in Chapter 4, set-consistency would be developed for the case of multiple solutions to the criterion. Nevertheless, stochastic equicontinuity relied on compactness, hence it was a crucial ingredient for normality too. We already analyzed that intuitively, there must be strong curvature around the solution to the criterion function to obtain an approximate distribution around a parameter estimate. In Chapter 6 we will consider a simple penalty modification to the criterion function that remedies a known form of non-identifiability and whose effect becomes negligible in the limit. However, non-parametric models may need penalization that does not vanish in the limit. In Chapter 5 we will analyze this. It turns out that penalties force the criterion function to favor simple solutions over complex ones. This, similar to the strategy of obtaining boundedness of derivatives through high-level smoothness conditions, essentially limits possible solutions of the criterion function to only those that are available in lower frequency domains, which again emphasizes that understanding the properties of $h(\boldsymbol{\theta})$ and $g(\mathbf{y}_{t-1}, \boldsymbol{\theta})$ is critical to establish the desired theoretical results needed to apply them to analyze spatial time series data.