1 Optimization-based Inference

We will briefly examine, using the language and background previously outlined, a broad class of statistical inference procedures widely used in Econometrics. These are estimation and hypothesis testing procedures based on mathematical optimization problems. In these cases, the unknown value of the parameter θ_0 is characterized as the solution to an optimization problem arising from (part of) the structure of the underlying statistical model.

Many (semi-)parametric inference procedures fall under the aforementioned framework. For instance, likelihood theory and the resulting Maximum Likelihood Estimator (MLE), along with related Likelihood Ratio and Score (Lagrange Multipliers) tests in parametric models; the family of least squares estimators (OLSE, GLSE, FGLSE, etc.) in various versions of the semi-parametric linear model; the Generalized Method of Moments (GMM), the associated GMM Estimator (GMME), and corresponding hypothesis tests; and so on.

Here, we will describe a general framework for constructing estimators derived via optimization, along with properties related to those discussed in the previous chapter.

2 Objective Functions

In various cases, the unknown parameter value θ_0 can be characterized by a variational property, i.e., an optimality condition. The structure of the statistical model, determined by its specification and parameterization

properties, implies the existence of an objective function that is minimized, potentially uniquely, at θ_0.[1]

If this function were known, its minimization feasible, and its minimizer unique, it would be possible to determine θ_0 precisely, without necessarily using the information contained in the sample. Typically, however, this function depends (also) on θ_0, rendering the above reasoning invalid. It might, however, be approximated-in a suitable sense-by a sample-based function. Under certain assumptions, minimizing this sample-based approximation with respect to θ may yield a stochastic approximation of θ_0. This analogy leads to the concept of *Optimization-Based Estimators (OE)*.[2] Let us first examine some simple examples of the formulation of such objective functions.

Example 1. [LLS] For the model

$$\{\mathbb{E}(Y_n/\sigma(\mathbf{X}_n)) = \mathbf{X}_n\theta,\ \theta \in \mathbb{R}^p,\ \mathrm{Var}(Y_n/\sigma(\mathbf{X}_n)) = \mathbf{I}_n\}\,,$$

and given the assumption that $\mathrm{rank}(\mathbf{X}_n) = p$,[3][4] and assuming $Y_n = \mathbf{X}_n\theta_0 + \varepsilon_n$

[1] The choice of describing the process through minimization is made without loss of generality. Maximization is the dual concept: maximizing f is equivalent to minimizing $-f$.

[2] In the literature, these are also referred to as **M-Estimators**.

[3] Modifying the above terminology slightly, and given that this matrix is generally stochastic, we could more broadly assume that the set of sample values for which its columns become linearly dependent is negligible with respect to the underlying probability distribution, i.e., that the rank condition holds with probability 1. We do not do so here for simplicity of terminology.

[4] This means that the square matrix $\frac{1}{n}\mathbf{X}_n'\mathbf{X}_n$ is invertible (why?), and consequently, due to its construction, it is strictly positive definite (why?). It also implies that necessarily $n \geq p$ (why?).

with $\mathbb{E}(\varepsilon_n \mid \sigma(\mathbf{X}_n)) = \mathbf{0}_n$, we observe that

$$\theta_0 = \arg\min_{\theta \in \Theta} \mathbb{E}\left(\frac{1}{n}(Y_n - \mathbf{X}_n\theta)'(Y_n - \mathbf{X}_n\theta)/\sigma(\mathbf{X}_n)\right). \tag{1}$$

This is because

$$\mathbb{E}\left(\frac{1}{n}(Y_n - \mathbf{X}_n\theta)'(Y_n - \mathbf{X}_n\theta)/\sigma(\mathbf{X}_n)\right)$$

$$= (\theta - \theta_0)'\frac{\mathbb{E}(\mathbf{X}_n'\mathbf{X}_n/\sigma(\mathbf{X}_n))}{n}(\theta - \theta_0) - \frac{2}{n}(\theta - \theta_0)'\mathbb{E}(\mathbf{X}_n'\varepsilon_n/\sigma(\mathbf{X}_n)) + \frac{1}{n}\mathbb{E}(\varepsilon_n'\varepsilon_n/\sigma(\mathbf{X}_n)).$$

Using properties of the conditional expectation, we find that

$$\mathbb{E}(\mathbf{X}_n'\mathbf{X}_n/\sigma(\mathbf{X}_n)) = \mathbf{X}_n'\mathbf{X}_n, \quad \mathbb{E}(\mathbf{X}_n'\varepsilon_n/\sigma(\mathbf{X}_n)) = \mathbf{X}_n'\mathbb{E}(\varepsilon_n/\sigma(\mathbf{X}_n)) = \mathbf{X}_n'\mathbf{0}_n = \mathbf{0}_p,$$

and

$$\mathbb{E}(\varepsilon_n'\varepsilon_n/\sigma(\mathbf{X}_n)) = \mathbb{E}(\mathrm{tr}(\varepsilon_n'\varepsilon_n/\sigma(\mathbf{X}_n))) = \mathrm{tr}(\mathbb{E}(\varepsilon_n\varepsilon_n'/\sigma(\mathbf{X}_n))) = \mathrm{tr}(\mathbf{I}_n) = n.$$

5

Thus,

$$\mathbb{E}\left(\frac{1}{n}(Y_n - \mathbf{X}_n\theta)'(Y_n - \mathbf{X}_n\theta)/\sigma(\mathbf{X}_n)\right) = (\theta - \theta_0)'\frac{\mathbf{X}_n'\mathbf{X}_n}{n}(\theta - \theta_0) + 1,$$

and the latter, due to the strict positive definiteness of $\frac{\mathbf{X}_n'\mathbf{X}_n}{n}$, is strictly convex. It is therefore uniquely minimized if and only if $\theta - \theta_0 = \mathbf{0}_p$, and

[5] Recall that the trace tr of any square matrix is defined as the sum of its diagonal elements. This is a linear operator and thus commutes with the integral, which is also linear.

consequently, (1) holds because by the already adopted well-specification assumption $\theta_0 \in \Theta$.

This objective function arises from the simple observation that, for a distribution in Euclidean space with finite second moments, the mean is the point minimizing the expected squared deviation around an arbitrary point in the space (**Exercise**: Try to demonstrate this in $\mathbb{R}$, using, for example, an argument based on dominated convergence for a related observation to facilitate the use of first-order conditions). It is thus based on variational properties of moments.

Note that in the case where $\text{rank}(\mathbf{X}_n) < p$, θ_0 minimizes-but not uniquely-the objective function. In any case, this function is not directly usable for determining or approximating θ_0-remember, this is the goal of inference-because it is evident that it directly depends on θ_0. Under certain conditions (to be examined later), it is suitably approximated by

$$c_n(\theta) := \frac{1}{n}(Y_n - \mathbf{X}_n\theta)'(Y_n - \mathbf{X}_n\theta),$$

whose minimization[6] over Θ yields the Ordinary Least Squares Estimator (OLSE)-see below.

Example 2 (IVE)**.** We consider the model of the form

$$\{\mathbb{E}(\mathbf{W}_n'(Y_n - \mathbf{X}_n\theta)) = \mathbf{0}_q,\ \theta \in \Theta\}.$$

Recall that the specification states that when integration is performed with respect to D_0, then $\theta = \theta_0 \Leftrightarrow \mathbb{E}(\mathbf{W}_n'(Y_n - \mathbf{X}_n\theta)) = \mathbf{0}_q$. We strengthen this with

[6] Or a monotonic transformation of it.

the assumption that, when integration is performed with respect to D_0, then

$$\theta = \theta_0 \Leftrightarrow \mathbb{E}(\mathbf{W}'_n(Y_n - \mathbf{X}_n\theta)) = \mathbf{0}_q.$$

Given that

$$\mathbb{E}(\mathbf{W}'_n(Y_n - \mathbf{X}_n\theta)) = \mathbb{E}(\mathbf{W}'_n\mathbf{X}_n)(\theta_0 - \theta) + \mathbb{E}(\mathbf{W}'_n\varepsilon_n),$$

this is equivalent to the matrix $\mathbb{E}(\mathbf{W}'_n\mathbf{X}_n)$ having rank p (why?), which directly implies that $q \geq p$ and $n \geq q$.

Let V be a strictly positive definite square matrix of dimension $q \times q$. It is easy to prove that the square root of the quadratic form with respect to V, i.e., the function

$$\mathbb{R}^q \ni \mathbf{z} \to \sqrt{\mathbf{z}'V\mathbf{z}}$$

is a norm on $\mathbb{R}^q$. (**Exercise**: prove this!)[7]

Therefore, based on the above, and because squaring and multiplication by a positive number are monotonic transformations, as well as the fact that $\theta_0 \in \Theta$, it follows that

$$\theta_0 = \arg\min_{\theta\in\Theta} \left(\frac{1}{n}\mathbb{E}(\mathbf{W}'_n(Y_n - \mathbf{X}_n\theta))\right)' V \left(\frac{1}{n}\mathbb{E}(\mathbf{W}'_n(Y_n - \mathbf{X}_n\theta))\right). \tag{2}$$

Due to that (see above) the objective function is

$$\frac{1}{n^2}(\theta_0 - \theta)'\mathbb{E}(\mathbf{X}'_n\mathbf{W}_n)V\mathbb{E}(\mathbf{W}'_n\mathbf{X}_n)(\theta_0 - \theta),$$

which directly depends on θ_0, it cannot be directly used to locate it. As in

[7] It is evident that when V is the identity matrix, the Euclidean norm is recovered.

the previous example, under certain conditions (to be examined later), it is suitably approximated by

$$c_n(\theta) := \frac{1}{n^2}(Y_n - \mathbf{X}_n\theta)'\mathbf{W}_n V \mathbf{W}_n'(Y_n - \mathbf{X}_n\theta),$$

whose minimization over Θ yields the Instrumental Variables Estimator (IVE) corresponding to the choice of V-see below.

Questions arise, such as: **(a)** Does the issue of optimal selection of V make sense? **(b)** Is it meaningful to use a norm that does not arise from a quadratic form with respect to a strictly positive definite matrix?

Clearly, the above indicate that the underlying structure can support the existence of a non-unique corresponding objective function.

Example 3 (GARCH(1,1))**.** In this example, it can be shown (see Chapter 5 of [10]) that when the model is enriched with the assumption that the support of the distribution of z_0 has cardinality greater than 2, then

$$\mathbb{P}\left(\frac{\sigma_0^2}{h_0(\theta)} = 1\right) = 1 \Leftrightarrow \theta = \theta_0.$$

Additionally, since the function $x - \ln(x), x > 0$, is uniquely minimized at 1 (Exercise: prove this!), and it can be shown that for every $\theta \in \Theta$, $0 < \mathbb{E}\left(\frac{\sigma_0^2}{h_0(\theta)}\right) < +\infty$, it follows that

$$\theta_0 = \arg\min_{\theta\in\Theta} \mathbb{E}\left(-\ln\left(\frac{\sigma_0^2}{h_0(\theta)}\right) + \frac{\sigma_0^2}{h_0(\theta)}\right).$$

Clearly, the above depends directly on θ_0, making its direct use for inference infeasible. Under certain conditions (to be discussed later), the above is

suitably approximated by the sample analogue

$$c_n(\theta) := \frac{1}{n}\sum_{t=1}^{n}\left(\ln h_t^*(\theta) + \frac{y_t^2}{h_t^*(\theta)}\right),$$

where h_0^* is an arbitrary positive constant, and

$$h_t^*(\theta) = \omega + \alpha y_{t-1}^2 + \beta h_{t-1}^*(\theta),\ t = 1, \ldots, n.$$

The above objective function is related to the (conditional) log-likelihood function of the sample-see again [10][8] in the case of the parametric version of the model where $z_0 \sim N(0,1)$. Even in the semi-parametric case, minimizing it leads to what is called the Maximum Gaussian Quasi Likelihood Estimator (Gaussian QMLE).

The objective functions c_n can be viewed as stochastic processes as long as they satisfy the relatively weak requirement that $c_n(\theta)$ is a well-defined random variable for every possible value of θ in Θ. The interested reader is referred to Section 1.4 of [11] to examine how the Daniell-Kolmogorov theorem's validity follows from the fact that Θ has the topological property of separability as a subset of Euclidean space. If Θ is uncountable, its countable subset consisting of elements of Θ with rational components is necessarily dense in Θ, and between any two elements of Θ, such a vector exists. Consequently, the following essentially pertain to optimization issues in stochastic processes, and the corresponding estimation procedures as their extrema.

[8] Specifically, its negative version.

3 Extremum Estimators

Let us assume-as in the previous examples-that the statistical model is suitably structured so that there exists a function $c_n : \mathbb{R}^{k \times n} \times \Theta \to \mathbb{R}$, i.e., a function evaluated at the sample values and the parameter,[9] and which returns real numbers. This function may, as seen in the examples above, represent some kind of sample analogue of another, not fully accessible, function supported by the structure of the model. Optimizing the former over Θ for every possible sample value will yield a well-defined function of the sample with values in Θ, which we hope will be a sample-based approximation of θ_0. Thus, we arrive at an estimation procedure defined by

$$\theta_n \in \arg\min_{\theta \in \Theta} c_n(\theta),$$

the properties of which are of interest to us.

For this to be well-defined, it would be useful if the set of sample values for which $\arg\min_{\theta \in \Theta} c_n(\theta) = \emptyset$ constitutes a negligible subset of $\mathbb{R}^{n \times k}$ with respect to the underlying distribution D_0-i.e. a set of zero D_0-probability. This is not necessary, however, if it is simply ensured that $c_n(\theta)$ is bounded below on Θ for almost every possible sample value,[10] and if some kind of optimization error is tolerated:

Definition 1. Given c_n and a random variable u_n, with $\mathbb{P}(u_n \geq 0) = 1$, θ_n will

[9] For the sake of symbolic simplicity, the dependence on the sample will be represented only by the index n in the notation of the function.

[10] That is, on a set of full probability under D_0.

be called an extremum estimator (OE) for θ_0 if and only if it satisfies

$$c_n(\theta_n) \leq \inf_{\theta \in \Theta} c_n(\theta) + u_n. \tag{3}$$

Regarding the existence of a solution to the above defining inequality, we note that for any sample value where c_n is bounded below, $\inf_{\theta \in \Theta} c_n \in \mathbb{R}$. If $\inf_{\theta \in \Theta} c_n = \min_{\theta \in \Theta} c_n$,[11] then $\arg\min$ is non-empty, and u_n is allowed to take the value zero. When the function has no minimum for a given sample value, a simple argument referring to the properties of $\inf$ as the greatest lower bound (with respect to θ) of the function implies that, in this case, $u_n > 0$, and there will necessarily exist an element of Θ for which (3) is satisfied. Thus, the existence of the random variable u_n, when it takes appropriate values, ensures the existence of a solution to the above defining inequality in any case where the infimum is well-defined. Hence, u_n can be viewed as a random element that almost always allows the above inequality to be satisfied when the infimum is almost always well-defined.

However, u_n also has another role. Even when $\arg\min$ is non-empty, the form of c_n, and/or the characteristics of Θ, may imply that the solution to the optimization problem-although well-defined-is not analytically known due to its complexity. In such cases, the estimator is obtained via computational methods, and optimization is carried out numerically, at least for the sample values where it is not analytically feasible. Computational methods are usually approximate and, therefore, may not yield the exact minimizing point of the function but something that approximately min-

[11] This may not hold; the infimum can be well-defined without the function being minimized. For a simple example, consider $f(\theta) = \theta^2$ for $\theta \in (0, 1)$.

imizes it. In these cases, u_n represents the approximation error and can thus be characterized as the optimization error. Its inclusion is therefore useful in covering estimators that, with positive probability, arise from computational optimization procedures.

The second issue that needs to be ensured is the measurability of the estimator. This will guarantee that it is a well-defined random vector and thus-among other things-can be categorized based on the properties of its distribution. We will not delve into this issue in detail. We note that verifying measurability requires concepts that are beyond the scope of the present notes. For example, even in the case where $\arg\min$ is almost surely non-empty but also non-unique with positive probability, the estimator arises from some sort of selection from the set of minimizing points for each sample value where non-uniqueness occurs. Therefore, this is outside the scope of the current book. It is noted, however, that verifying measurability is greatly facilitated by the fact that, by construction, Θ is a subset of Euclidean space. Due to properties of the space,[12] it can be shown that when $\inf_{\theta\in\Theta} c_n(\theta)$ is almost always well-defined, it is a well-defined random variable. The interested reader may consult the Measurable Projection Theorem in Section 1.7 of [11].

> It can therefore be shown that, for example, when:
>
> - Θ is compact, and c_n is continuous with respect to θ for almost every sample value, or

[12] Specifically, its separability; this means that the space has a dense countable subset-vectors with rational components. Between any two arbitrary elements of $\mathbb{R}^p$, there is at least one such vector-see, for instance, Section 1.4 in [11].

- Θ is closed and convex, and c_n is convex with respect to θ for almost every sample value,

then θ_n, which arises from the above definition, is a well-defined estimator.

Note that in the aforementioned cases, it is not necessary for the optimization error to be identically zero. Existence results can be derived in much more general cases; we will not pursue these as the above cover the cases we will deal with later. To orient the interested reader towards the relevant literature, we provide a skeleton of the proof for the case of compactness and continuity:

Theorem 1. *If Θ is compact and c_n is continuous with respect to θ for almost every sample value, then there exists an estimator θ_n that satisfies 3.*

Proof. The subset of Θ consisting of its elements that satisfy 3 is non-empty and compact for almost every sample value due to the continuity of the criterion and the compactness of the space. Continuity ensures that $c_n^{-1}(\{\theta \in \Theta : \text{ satisfies 3}\})$ is closed for almost every sample value, while compactness and continuity ensure it is non-empty and, therefore, compact and non-empty. Lemma 1.4.1 in [11] and Proposition 3.10.(iii) in Chapter 5 of [7] guarantee the measurability of $\inf_{\theta \in \Theta} c_n(\theta) + u_n$, and consequently the measurability of the aforementioned set. The result follows from Theorem 2.13 in [7], which guarantees the existence of a measurable selection function. □

Let us examine the above within the context of our examples:

Example 4 (LLS)**.** For the model

$$\{\mathbb{E}(Y_n/\sigma(\mathbf{X}_n)) = \mathbf{X}_n\theta, \theta \in \mathbb{R}^p, \mathrm{Var}(Y_n/\sigma(\mathbf{X}_n)) = \mathbf{I}_n\},$$

supplemented with the assumption that $\mathrm{rank}(\mathbf{X}_n) = p$, we have that

$$c_n(\theta) := \frac{1}{n}(Y_n - \mathbf{X}_n\theta)'(Y_n - \mathbf{X}_n\theta)$$

is strictly convex. When $\Theta = \mathbb{R}^p$, the first-order conditions

$$\frac{\partial c_n(\theta)}{\partial \theta} = \mathbf{0}_p \Leftrightarrow -\frac{2}{n}\mathbf{X}_n'(Y_n - \mathbf{X}_n\theta) = \mathbf{0}_p$$

are both necessary and sufficient for optimization. Based on the previous observation regarding the strictly positive definiteness-and thus invertibility-of $\frac{\mathbf{X}_n'\mathbf{X}_n}{n}$, the resulting unique solution, and consequently the estimator corresponding to the choice $u_n = 0$ in the above definition, is

$$\theta_n := (\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{X}_n'Y_n,$$

which is also the standard form of the OLSE.

Regarding optimization for general $\Theta \subseteq \mathbb{R}^p$, we observe the following: Using the second-order Taylor expansion of the criterion around $(\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{X}_n'Y_n$, we obtain the expression

$$c_n(\theta) = \frac{1}{n}e_n'e_n - \frac{2}{n}(\theta - (\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{X}_n'Y_n)'\mathbf{X}_n'e_n + (\theta - (\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{X}_n'Y_n)'\frac{\mathbf{X}_n'\mathbf{X}_n}{n}(\theta - (\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{X}_n'Y_n),$$

where $e_n := (\mathbf{I}_n - \mathbf{X}_n(\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{X}_n')Y_n$.

Observing that $\mathbf{X}_n' e_n = \mathbf{0}_p$ (**Exercise**: derive this!), and that the term $\frac{1}{n} e_n' e_n$ does not depend on the parameter and thus does not affect the optimization, we conclude that in the general case, the estimator equivalently satisfies

$$\theta_n \in \arg\min_{\theta \in \Theta} \left\| (\mathbf{X}_n' \mathbf{X}_n)^{-1} \mathbf{X}_n' Y_n - \theta \right\|_{\frac{\mathbf{X}_n' \mathbf{X}_n}{n}} .$$

Thus, due to the strict convexity and quadratic nature of the least squares criterion, it minimizes the distance between the parameter space and the unrestricted estimator derived from the quadratic form with respect to the positive definite matrix $\frac{\mathbf{X}_n' \mathbf{X}_n}{n}$. This brings us back to the expression $(\mathbf{X}_n' \mathbf{X}_n)^{-1} \mathbf{X}_n' Y_n$ when $\Theta = \mathbb{R}^p$. When Θ is compact or closed and convex, the above indicates that the estimator exists. In the case of convexity of the parameter space, it is also the unique minimizing element.

Example 5 (IVE)**.** Consider the model

$$\{\mathbb{E}(\mathbf{W}_n'(Y_n - \mathbf{X}_n \theta)) = \mathbf{0}_q, \theta \in \Theta\}$$

supplemented with the assumptions that $\text{rank}(\mathbf{X}_n) = p$ and $\text{rank}(\mathbf{W}_n) = q$. Due to this supplementation, the function

$$c_n(\theta) := \frac{1}{n^2} (Y_n - \mathbf{X}_n \theta)' \mathbf{W}_n V \mathbf{W}_n' (Y_n - \mathbf{X}_n \theta)$$

is strictly convex since V is strictly positive definite.[13] When $\Theta = \mathbb{R}^p$, the

[13] Exercise: Try to prove this; you can use the so-called Cholesky factorization of a positive definite matrix or consider the matrix as a composition of appropriate linear operators, which under these conditions are one-to-one.

first-order conditions

$$\frac{\partial c_n(\theta)}{\partial \theta} = \mathbf{0}_p \Leftrightarrow -\frac{2}{n^2}\mathbf{X}_n'\mathbf{W}_n V \mathbf{W}_n'(Y_n - \mathbf{X}_n\theta) = \mathbf{0}_p,$$

are both necessary and sufficient for optimization. Based on the previous observation regarding the strictly positive definiteness-and therefore invertibility-of $\mathbf{X}_n'\mathbf{W}_n V \mathbf{W}_n'\mathbf{X}_n$, the resulting unique solution, and consequently the estimator corresponding to the choice $u_n = 0$ in the above definition, is

$$\theta_n := (\mathbf{X}_n'\mathbf{W}_n V \mathbf{W}_n'\mathbf{X}_n)^{-1}\mathbf{X}_n'\mathbf{W}_n V \mathbf{W}_n' Y_n,$$

which is also the standard form of the Instrumental Variables Estimator (IVE). We observe that when $p = q$, the above becomes

$$\theta_n := (\mathbf{W}_n'\mathbf{X}_n)^{-1}\mathbf{W}_n' Y_n,$$

which does not depend on the choice of V. In the special case where $\mathbf{W}_n = \mathbf{X}_n$, the standard form of the OLSE is recovered. When the optimization is performed generally for $\Theta \subseteq \mathbb{R}^p$, it is shown (*Exercise*!) that the estimator equivalently satisfies

$$\theta_n \in \arg\min_{\theta \in \Theta} \left\| (\mathbf{X}_n'\mathbf{W}_n V \mathbf{W}_n'\mathbf{X}_n)^{-1}\mathbf{X}_n'\mathbf{W}_n V \mathbf{W}_n' Y_n - \theta \right\|_{\frac{\mathbf{X}_n'\mathbf{W}_n}{n} V \frac{\mathbf{W}_n'\mathbf{X}_n}{n}},$$

i.e., similar to the above example, it minimizes the Euclidean distance between the parameter space and the unrestricted estimator. This collapses to the standard form when $\Theta = \mathbb{R}^p$. When Θ is compact or closed and convex, the above tells us that the estimator exists. In the case where the parameter

space is convex, it is also the unique minimizing element.

Example 6 (GARCH(1,1))**.** When Θ is compact, it can be shown-via investigating the solutions of the recursive relations involved in the likelihood function that the existence of the estimator falls under Theorem 1. However, the analytical solution to the optimization problem is not feasible. Thus, we have a situation-frequently encountered in econometrics, though not among the most complex ones-where the form of the estimator as a function of the sample is unknown, even though it is well-defined. The estimation is realized computationally, and the properties of the estimator are influenced by the characteristics of the corresponding optimization algorithm used, making the optimization error relevant in this context.

Our examples, though relatively simple, provide sufficient insight. OE estimators often do not-and generally do not-have a known analytical form due to the complexity of the underlying optimization process. In such cases, they are obtained through computational procedures, meaning their properties are generally influenced by the specific characteristics of the corresponding optimization algorithm and the parameters determining its operation. Generally, even their unknown analytical form is not a linear function of the sample, so properties like unbiasedness are not expected to hold. For example, within the context of Example 7.4, when $\Theta = \mathbb{R}^p$, it is easy to show-using the Law of Iterated Expectations (LIE)-Exercise!-that the OLSE is unbiased, due to the strong assumption that $\mathbb{E}(\varepsilon_n/\sigma(\mathbf{X}_n)) = \mathbf{0}_p$ and the linear form of the estimator with respect to Y_n, a property that can be lost if the aforementioned strong assumption is replaced with the weaker orthogonality assumption for instrumental variables with $\mathbf{W}_n = \mathbf{X}_n$, or if

the parameter space is altered, which may distort the aforementioned linear form.[14] Similarly, the IVE, although linear with respect to Y_n, is generally not unbiased due to the insufficiency of the orthogonality condition that constitutes the model. The QMLE in the GARCH case informs us that exploring properties corresponding to a fixed and arbitrary n in even slightly more complex models than linear ones is particularly challenging.

The following table describes in pseudo code a general algorithm for the evaluation of such an estimator. The algorithms accepts as inputs the sample, the criterion, an initial parameter value, as well as several stopping criteria, evaluates the criterion at the initial value, changes the parameter according to some computational optimization methodology, and continues until at least one of the stopping criteria is fulfilled.

[14] Unbiasedness is a property not preserved under nonlinear transformations; integrals are linear creatures!

Algorithm 1 Optimization-Based Estimator (OE)

Input: Data: X (observed data sample) Objective function: $c_n(\theta, X)$ (criterion depending on parameters θ and data X) Parameter space: Θ (feasible set for θ) Error tolerance: ϵ (convergence threshold for optimization) Maximum iterations: `max_iter` (to prevent infinite loops in iterative optimization)

Output:

- Estimator: θ_n (minimizer of c_n)

Initialize:

- Choose initial guess: θ_{init} (e.g., random point in Θ or domain-specific guess)
- Set iteration counter: `iter` $\leftarrow 0$
- Set convergence flag: `converged` $\leftarrow$ `False`

Define stopping criteria:

- `Condition 1:` $|c_n(\theta_k, X) - c_n(\theta_{k+1}, X)| < \epsilon$
- `Condition 2:` $\|\theta_{k+1} - \theta_k\| < \epsilon$
- `Condition 3: iter` $\geq$ `max_iter`

Optimization Loop: `not converged` (a) Compute gradient (if available) or approximate it:

- $\nabla \leftarrow \texttt{gradient}(c_n, \theta_k, X)$
- If gradient is unavailable, use a gradient-free optimization method (e.g., Nelder-Mead).

(b) Update parameters:

- For gradient descent: $\theta_{k+1} \leftarrow \theta_k - \texttt{learning\_rate} \cdot \nabla$
- For second-order methods (e.g., Newton-Raphson):

$$\theta_{k+1} \leftarrow \theta_k - \nabla^2 c_n(\theta_k, X)^{-1} \nabla$$

- For gradient-free methods:

$$\theta_{k+1} \leftarrow \texttt{next\_iterate}(\theta_k, X, c_n)$$

(c) Check feasibility of θ_{k+1} in Θ:

- If $\theta_{k+1} \notin \Theta$, project onto Θ (e.g., $\texttt{project}(\theta_{k+1}, \Theta)$).

(d) Update iteration counter:

$$\texttt{iter} \leftarrow \texttt{iter} + 1$$

(e) Evaluate stopping criteria:

- If any stopping condition is satisfied:

$$\texttt{converged} \leftarrow \texttt{True}$$

Output Results:

- If `converged:` return $\theta_n \leftarrow \theta_{k+1}$
- Else: Print warning: "`Optimization did not converge.`" and return $\theta_n \leftarrow \theta_{k+1}$.

4 Asymptotic Properties

The previous section essentially motivates us to investigate the asymptotic properties of these estimators.[15] We will limit ourselves to the properties outlined in the previous chapter. The fact that we generally do not have an analytical form of the estimator-as a function of the sample-compels us to derive asymptotic properties from corresponding properties of the criteria involved in their construction. We will see that it is crucial for these properties to remain unaffected by the optimization procedures. In the following, limits pertain to $n \to \infty$. Whenever this is not the case, it will be explicitly mentioned.

4.1 Weak Consistency

We begin with the issue of weak consistency of the estimator defined in (3). We expect this to result from a kind of asymptotic behavior of c_n that also influences its extrema. The methodology we will follow can be summarized as follows:

1. The sequence of objective functions (c_n) converges appropriately in probability to a limiting function. The latter, denoted c, is non-stochastic and maps Θ to $\mathbb{R}$.[16]. The form of convergence ensures a kind of approximation of the optimizers of the sequence with those of c.

2. The function c has a unique minimum at θ_0, and $c(\theta_0)$ is not accessible via diverging sequences of parameter values (*"well distinguishability"* of the minimizing θ_0).

[15] This could allow us to relax the measurability requirements in our definitions of estimators. For example, a property that ensures measurability asymptotically could suffice. Clearly, such an investigation lies well beyond the scope of this text. Readers interested in such details are referred to [11]. In any case, such considerations are more relevant in nonparametric statistical models.

[16] Although it could be stochastic and map to extended real values

Let us start with the first component of our methodology. Our goal is to describe a mode of convergence of c_n to c that ensures the aforementioned extrema approximation. We will not delve into the most general-and thus weakest-convergence notion. Instead, we describe a relatively strong one, which suffices for the examples we are developing.

Given the notion of convergence in probability for sequences of random variables, a natural extension to sequences of stochastic processes taking real values is pointwise convergence in probability. These processes can be considered as collections of random variables, and the corresponding sequences as collections of sequences of random variables. From this perspective, pointwise convergence in probability simply involves the convergence in probability of the random variable (more precisely, the n-th member of the sequence[17]) $c_n(\theta)$ to the real number $c(\theta)$ for each $\theta \in \Theta$:

Definition 2. c_n converges pointwise in probability to c if and only if $\forall \theta \in \Theta$, $c_n(\theta) \xrightarrow{\text{p}} c(\theta)$.

In some cases, the above may be facilitated by the validity of appropriate Laws of Large Numbers, potentially combined with tools like the Continuous Mapping Theorem. Unfortunately-as it can be shown-pointwise convergence does not behave as we would wish with respect to optimization.[18] The reason is that it does not control how c_n approaches c jointly across elements within appropriate neighborhoods of the involved extrema. The following mode of convergence is strong enough to preclude such "anomalies" and essentially pertains to the characteristics of (c_n) as a sequence of stochastic processes:

[17] We allow ourselves this slight abuse of terminology for brevity!

[18] A simple example, in a much simpler non-stochastic context-so here convergence in probability is trivial-but indicative of how pointwise convergence fails to ensure the convergence of optimization characteristics is as follows: let $f_n(\theta) := \begin{cases} \exp(-n\theta), & \theta \in (0, n) \\ n, & \theta \in [n, +\infty) \end{cases}$. This converges pointwise to the constant function $f(\theta) = 1$ on $(0, +\infty)$. However, $\max_{\theta \in (0,+\infty)} f_n(\theta) = n \nrightarrow 1 = \max_{\theta \in (0,+\infty)} f$.

Definition 3. The sequence c_n converges continuously in probability to c, denoted as $c_n \xrightarrow{\text{cp}} c$, if and only if for all $\theta \in \Theta$ and for all $\theta_n \in \Theta$ such that $\theta_n \to \theta$, we have $c_n(\theta_n) \xrightarrow{\text{p}} c(\theta)$.[19]

Continuous convergence in probability concerns the asymptotic behavior of c_n when evaluated at members of convergent sequences of elements in Θ. It is evidently stronger than pointwise convergence, as the latter pertains only to one type of convergent sequences inside the parameter space-the constant ones. Before showing that it suffices for our methodology, we can ask how pointwise convergence in probability could be augmented to imply continuous convergence. We will demonstrate that if c_n is Lipschitz continuous with respect to θ, with a Lipschitz constant that is suitably bounded with probability converging to 1, then pointwise convergence implies continuous convergence:

Lemma 1. *Let c_n converge pointwise in probability to c, and let there exist a positive random variable k_n and a positive constant C, such that $\lim_{n\to\infty} \mathbb{P}(k_n > C) = 0$, and for all $\theta, \theta_\star \in \Theta$,*

$$|c_n(\theta) - c_n(\theta_\star)| \leq k_n \|\theta - \theta_\star\|.$$

Then $c_n \xrightarrow{cp} c$.

Proof. Let $\theta \in \Theta$, $\theta_n \to \theta$ with $\theta_n \in \Theta$, and let $\delta > 0$. If we show that $\lim_{n\to\infty} \mathbb{P}(|c_n(\theta_n) - c(\theta)| > \delta) = 0$, we have proved the claim (why?).

Using the triangle inequality, we have:

$$|c_n(\theta_n) - c(\theta)| \leq |c_n(\theta_n) - c_n(\theta)| + |c_n(\theta) - c(\theta)|.$$

Due to the Lipschitz property of c_n, the right-hand side is bounded by:

$$k_n \|\theta_n - \theta\| + |c_n(\theta) - c(\theta)|.$$

[19] Both definitions 2 and 3 are readily extendable to the mode of almost sure convergence. Provide the details!

By the monotonicity of $\mathbb{P}$ and the above inequality, we obtain:

$$\mathbb{P}(|c_n(\theta_n) - c(\theta)| > \delta) \leq \mathbb{P}(k_n\|\theta_n - \theta\| + |c_n(\theta) - c(\theta)| > \delta).$$

The probability on the right-hand side is further bounded by:

$$\mathbb{P}(k_n\|\theta_n - \theta\| > \frac{\delta}{2}) + \mathbb{P}(|c_n(\theta) - c(\theta)| > \frac{\delta}{2}).^{20}$$

Using the bound on k_n, we have:

$$\mathbb{P}(k_n\|\theta_n - \theta\| > \frac{\delta}{2}) \leq \mathbb{P}(\|\theta_n - \theta\| > \frac{\delta}{2C}),$$

so summarizing, for each n, we get:

$$0 \leq \mathbb{P}(|c_n(\theta_n) - c(\theta)| > \delta) \leq \mathbb{P}(\|\theta_n - \theta\| > \frac{\delta}{2C}) + \mathbb{P}(|c_n(\theta) - c(\theta)| > \frac{\delta}{2}).$$

The result follows by taking limits on both sides of the above inequality and observing that since $\theta_n \to \theta$, $\mathbb{P}(\|\theta_n - \theta\| > \frac{\delta}{2C})$ equals zero for sufficiently large n, and that due to pointwise convergence in probability, $\lim_{n\to\infty} \mathbb{P}(|c_n(\theta) - c(\theta)| > \frac{\delta}{2}) = 0$. □

The aforementioned property addresses the Lipschitz continuity of c_n as a function of θ for every possible sample value and for every n. It requires the corresponding set of functions $\Theta \to \mathbb{R}$ to exhibit a uniform behavior regarding this strong continuity property-the collection of involved Lipschitz constants (one for each possible sample value and for each n) must be bounded above by a common constant, denoted C, with a probability that converges to one.[21] This constitutes a fairly strong requirement, which, however-as we shall see below-is sufficient for our examples.

[20] Note that $\mathbb{P}(a + b > \epsilon) \leq \mathbb{P}(a > \frac{\epsilon}{2}) + \mathbb{P}(b > \frac{\epsilon}{2})$.

[21] The boundedness in probability condition for the Lipschitz coefficient, can be for example substituted by a uniform integrability condition of the form $\sup_n \mathbb{E}(k_n) < +\infty$ without affecting the result. Provide the details!

We will show that when the parameter space is compact and the involved functions are continuous with respect to θ, then this type of convergence suffices for $\inf_{\theta\in\Theta} c_n(\theta) \xrightarrow{\text{p}} \inf_{\theta\in\Theta} c(\theta)$. The limit is well-defined because c, by necessity, must also be a continuous function due to the related property of c_n and the specific mode of convergence. The compactness of Θ then ensures that c will have optimizers. However, this alone is not sufficient to guarantee that the approximate minimizer of c_n, and therefore the estimator, will converge to a minimizer of c. It must first be ensured that the optimization error asymptotically vanishes-this necessarily requires a strong assumption that must be verified on a case-by-case basis. Yet, even this is not sufficient. Without further restrictions on where the limiting function c achieves its minimum, the approximate minimizer of c_n might asymptotically approach various elements of $\arg\min_{\theta\in\Theta} c(\theta)$ with a probability that tends to one, which might not stabilize and could depend on the specific sample realization.[22]

To exclude such behavior, we employ the second part of our methodology, namely that the limiting function has a unique and "well distinguishable" minimum at θ_0, thus eliminating the aforementioned complex asymptotic behavior. This, which forms the second pillar of our methodology, is a condition of *asymptotic identification*. In some sense, it ensures that the structure of the model and the specific estimation process are sufficient to allow the sample information to increasingly pinpoint θ_0 as $n \to \infty$.

The above considerations lead us to the theorem of weak consistency for θ_n, which we present below:

Theorem 2. *Let (a) Θ be compact, (b) c_n be continuous with respect to θ almost surely, (c) there exist a function $c : \Theta \to \mathbb{R}$ such that $c_n \xrightarrow{cp} c$, (d) $u_n \xrightarrow{p} 0$, and (e) $\theta_0 = \arg\min_{\theta\in\Theta} c(\theta)$. Then, given (a)-(d), $\inf_{\theta\in\Theta} c_n(\theta) + u_n \xrightarrow{p} \inf_{\theta\in\Theta} c(\theta)$. Furthermore, given (e), $\theta_n \xrightarrow{p} \theta_0$.*

The proof will be provided without full details,[23] but it will also direct

[22] This implies that the sequence might have stochastic accumulation points, but this is a complex behavior outside the scope of the notes.

[23] In fact, we provide the framework of a proof that is more complex than necessary.

the interested reader towards a relaxation of continuous convergence to a form of convergence that is both more general and sufficient for the result in question.

Proof. First, using a construction involving Skorokhod representations-see the relevant footnote in the notes regarding convergence in distribution-it is possible to reduce the examination of convergence in probability to the examination of almost sure convergence. This is achieved by extending the definitions of the involved random elements to a common, richer probability space, which is permitted by Theorem 3.7.25 in [3].[24]

Let $\mathbf{epi}_n$ denote the stochastic epigraph of (the modified) c_n, that is, the set $\{(\theta, x) \in \Theta \times \mathbb{R} : x \geq c_n(\theta)\}$. Due to the background of Theorem 1, it is shown through Proposition A.2 in [7] that this is a suitably measurable multivalued function with values in the subset of 2^{Θ} consisting of the non-empty closed subsets of the parameter space.[25] Assumptions (a), (b), and (c) essentially ensure (and in fact, are much stronger than) the almost sure epiconvergence of c_n to (the modified version of) c with respect to the modified probability space (see Definition D.1 and Proposition D.2 in [7]).

Additionally, using the triangle inequality, for any $\theta \in \Theta$ and any $\theta_n^\star \to \theta$ in Θ, we have:

$$|c(\theta_n^\star) - c(\theta)| \leq |c(\theta_n^\star) - c_n(\theta_n^\star)| + |c(\theta) - c_n(\theta_n^\star)|.$$

The second term on the right-hand side of the inequality above converges almost surely to 0, due to (c). It can also be shown that, owing to (c) and (a), the first term behaves similarly. Consequently, the (modified) version is continuous, and its infimum is well-defined due to (a). Furthermore, the epigraph of c, denoted $\mathbf{epi}$, is closed. From Section 1.1 of [7] and Definition 4.5.1 in [6], this almost sure epiconvergence is equivalent to the following:

[24] Details of this construction are omitted. We are fortunate that continuous functions on compact domains are bounded. Interested readers may consult Theorem 1.10.3 in [11].

[25] The intricate details regarding the sigma-algebra on its domain are beyond the scope of this book. Interested readers may consult [7].

1. For sufficiently large n, in a set of full probability with respect to the modified distribution, $\mathbf{epi}_n \cap (\Theta \times (\inf_{\theta\in\Theta} c(\theta), +\infty)) \neq \varnothing$, because $(\inf_{\theta\in\Theta} c(\theta), +\infty)$ is open and $\mathbf{epi} \cap (\Theta \times (\inf_{\theta\in\Theta} c(\theta), +\infty)) \neq \varnothing$. Hence, $\inf_{\theta\in\Theta} c_n(\theta) \geq \inf_{\theta\in\Theta} c(\theta)$ almost surely for sufficiently large n. Therefore, $\liminf_n \inf_{\theta\in\Theta} c_n(\theta) \leq \inf_{\theta\in\Theta} c(\theta)$ almost surely.

2. For any $\delta > 0$, there exists a sufficiently large $n^\star$ such that for every $n \geq n^\star$, $\mathbf{epi}_n \cap (\Theta \times [\inf_{\theta\in\Theta} c(\theta) - 2\delta, \inf_{\theta\in\Theta} c(\theta) - \delta]) = \varnothing$ almost surely, because $\Theta \times [\inf_{\theta\in\Theta} c(\theta) - 2\delta, \inf_{\theta\in\Theta} c(\theta) - \delta]$ is compact and $\mathbf{epi} \cap (\Theta \times [\inf_{\theta\in\Theta} c(\theta) - 2\delta, \inf_{\theta\in\Theta} c(\theta) - \delta]) = \varnothing$. Thus, $\limsup_n \inf_{\theta\in\Theta} c_n(\theta) \leq \inf_{\theta\in\Theta} c(\theta)$ almost surely.

Combining the above with (d) and returning to the original probability space, we obtain the first conclusion. Now, if x_n arises from a measurable selection from:

$$\{\theta \in \Theta : c_n(\theta) \leq \inf_{\theta\in\Theta} c_n(\theta) + u_n\},$$

such that x is a limit point almost surely, its existence follows from Theorem 2.13 in [7] and (a). Working in the modified probability space, we have almost surely:

$$c(x) \leq \liminf_n c_n(x_n) \leq \limsup_n c_n(x_n) \leq \limsup_n (\inf_{\theta\in\Theta} c_n(\theta) + u_n) \leq \inf_{\theta\in\Theta} c(\theta),$$

which, given (e), ensures the second conclusion due to (a), once we return to the original probability space. The continuous convergence in probability to the limiting c along with the continuity of c_n, imply that the limiting criterion is continuous. Continuity and the compactness of Θ imply that the unique optimizer is "well distinguishable"; minimizing c outside any neighborhood of θ_0 results into something strictly greater than $c(\theta_0)$. Then, if θ_n does not converge in probability to θ_0, there must exist some $\varepsilon, \delta > 0$ such that $\mathbb{P}(\|\theta_n - \theta_0\| > \varepsilon) \geq \delta$ for an infinite number of n's. But this and "well distinguishability" imply that there exists some $\epsilon > 0$ such that for

those events $\mathbb{P}(|c(\theta_n) - c(\theta_0)| > \epsilon) > \delta$, and thus for those events

$$\mathbb{P}(|c(\theta_n) - c(\theta_0)| > \epsilon) > \delta \Rightarrow$$

$$\mathbb{P}(|c(\theta_n) - c_n(\theta_n)| > \frac{\epsilon}{2}) + \mathbb{P}(|c_n(\theta_n) - c(\theta_0)| > \frac{\epsilon}{2}) > \delta \Rightarrow$$

$$\mathbb{P}(\sup_{\theta\in\Theta} |c(\theta) - c_n(\theta)| > \frac{\epsilon}{2}) + \mathbb{P}(|\inf_{\theta\in\Theta} c_n - \inf_{\theta\in\Theta} c| > \frac{\epsilon}{4}) + \mathbb{P}(u_n > \frac{\epsilon}{4}) > \delta.$$

But the each of the three probabilities converges to zero due to the assumptions and the previous result, and thereby their sum cannot remain above δ for an infinite number of n's. We arrived at a contradiction due to the hypothesis that θ_n does not converge in probability to θ_0. □

The above proof is relatively complex; however, as previously mentioned, it implicitly suggests that:

(i) The aforementioned form of continuous convergence is stronger than necessary. To ensure the convergence of the infima when Θ is compact, weaker forms of convergence (yet stronger than the pointwise) are sufficient, e.g. epi-convergence.

(ii) The "well distinguishability" of the unique minimizer θ_0 essentially follows from its uniqueness, the continuity of c and the compactness of Θ.

(iii) Changing the assumptions involving convergence in probability to almost sure convergence leads to almost sure convergence of the estimator to θ_0, yielding a stronger form of consistency (the estimator is then termed strongly consistent).

The above can be generalized to cases where Θ is not compact. One way to generalize the theorem is to consider cases where the estimator can be shown to belong to a compact subset of Θ containing θ_0 with probability converging to 1. If this is ensured, an application of the arguments leading to the above proof would also yield the weak consistency of the estimator. Sufficient conditions facilitating the asymptotic containment of the estimator in such a compact set with probability converging to 1 could arise from

additional structure in the involved functions.

For example, it can be shown that if c_n is also strictly convex almost surely for each sample realization, and c is also strictly convex, then the aforementioned containment holds. Moreover, it can be shown (see, for instance, Corollary 2.C in [9]) that in this case, pointwise convergence suffices for the convergence of the infima, while . While we do not provide a proof here, let us explicitly state this result as it will be useful later:

Theorem 3. *Suppose that:*

(a') Θ *is closed and convex.*

(b') c_n *is continuous and strictly convex with respect to* θ *almost surely for each sample realization.*

(c') *There exists a function* $c : \Theta \to \mathbb{R}$*, continuous and strictly convex, such that* $c_n \xrightarrow{p} c$*.*

(d') $u_n \xrightarrow{p} 0$*.*

(e') $\theta_0 \in \arg\min_{\theta\in\Theta} c(\theta)$*.*

Then, given (a')-(c') and (d'), we have:

$$\inf_{\theta\in\Theta} c_n(\theta) + u_n \xrightarrow{p} \inf_{\theta\in\Theta} c(\theta).$$

Moreover, if (e') holds, then $\theta_n \xrightarrow{p} \theta_0$*.*

We observe that in (e'), it is not explicitly necessary to assume that θ_0 uniquely minimizes the limiting function. Conditions (c') and (e') together ensure identification (why?). The continuity assumptions in the previous theorems can be further generalized; however, we will not elaborate on this.

Even more general results are possible in the context of continuous convergence in probability. For example, if Θ is not compact, yet the remaining conditions of Theorem 2 hold, while convexity fails, then weak

consistency would be the case, as long as it is further ensured that (a.) the sequence (θ_n) is tight, and (b.) that "well distinguishability" does not fail for θ_0; for any $\delta > 0$, $c(\theta_0) < \inf_{\theta \in B_\delta^c(\theta_0)} c(\theta)$.

The above are sufficient for examining the issue of consistency in our examples. Let us now analyze them:

Example 7. [LLS] We know that when $\Theta = \mathbb{R}^p$, then

$$\theta_n := (\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{X}_n'Y_n.$$

By substituting $Y_n = \mathbf{X}_n\theta_0 + \varepsilon_n$ and multiplying by $\frac{n}{n}$, we obtain the equivalent expression:[26]

$$\theta_n = \theta_0 + \left(\frac{1}{n}\mathbf{X}_n'\mathbf{X}_n\right)^{-1}\frac{1}{n}\mathbf{X}_n'\varepsilon_n.$$

This expression can be used to study consistency without resorting to the previous theorems. Thus, guided by the Continuous Mapping Theorem, we consider the following High-Order Conditions (HOCs):[27]

1. $\frac{1}{n}\mathbf{X}_n'\varepsilon_n \xrightarrow{\text{p}} \mathbf{0}_p$,
2. There exists a non-stochastic $p \times p$ positive definite matrix $M_{X'X}$ such that $\frac{1}{n}\mathbf{X}_n'\mathbf{X}_n \xrightarrow{\text{p}} M_{X'X}$,
3. $\text{rank}(M_{X'X}) = p$.

Since $\frac{1}{n}\mathbf{X}_n'\varepsilon_n$ is a vector of sample averages of the form $\frac{1}{n}\sum_{i=1}^n \mathbf{X}_{i,j}\varepsilon_i$, $j = 1, \ldots, p$, where $\mathbf{X}_{i,j}$ is the i,j-th entry of $\mathbf{X}_n$ and ε_i is the i-th entry of ε_n, Condition (1)-and given the assumption about the conditional mean of ε_n in the model-can be facilitated by the validity of a Law of Large Numbers, e.g., in iid settings or under stationarity and ergodicity, provided $\mathbb{E}(|\mathbf{X}_{0,j}\varepsilon_0|) < +\infty$, $\forall j = 1, \ldots, p$.

[26] This expression is not useful for estimation as it depends on the unknown θ_0 and the latent ε_n. However, it is useful for studying the properties of the estimator.

[27] These conditions do not specify the probabilistic properties of the involved random elements to ensure their validity.

Similarly, Condition (2) would hold under the aforementioned frameworks if $\mathbb{E}(\mathbf{X}_{0,j}^2) < +\infty$, $\forall j = 1, \ldots, p$. The matrix $M_{X'X}$ in such cases is the covariance matrix of the random vector formed by the first row of the matrix of independent variables. Due to its construction as a Gram matrix (see, e.g., [4]), Condition (3) will hold in the above frameworks if and only if the random variables forming this row are linearly independent.

Conditions (2), (3), and the Continuous Mapping Theorem[28] imply that $\left(\frac{1}{n}\mathbf{X}_n'\mathbf{X}_n\right)^{-1} \xrightarrow{\text{p}} M_{X'X}^{-1}$. Hence, by the Continuous Mapping Theorem and Condition (1), $\left(\frac{1}{n}\mathbf{X}_n'\mathbf{X}_n\right)^{-1}\frac{1}{n}\mathbf{X}_n'\varepsilon_n \xrightarrow{\text{p}} M_{X'X}^{-1}\mathbf{0}_p = \mathbf{0}_p$. Therefore, by the Continuous Mapping Theorem, Conditions (1)–(3) ensure the weak consistency of the OLSE when $\Theta = \mathbb{R}^p$.

The question that arises is whether this holds in other cases for this parameter space. Recall that, in general, the estimator equivalently satisfies

$$\theta_n \in \arg\min_{\theta\in\Theta} \left\|(\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{X}_n'Y_n - \theta\right\|_{\frac{\mathbf{X}_n'\mathbf{X}_n}{n}}.$$

We can attempt to apply the previous theorems in this general case. Note first that the previous result, combined with the Continuous Mapping Theorem,[29] implies that for any $\theta \in \Theta$,

$$\|(\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{X}_n'Y_n - \theta\|_{\frac{\mathbf{X}_n'\mathbf{X}_n}{n}} \xrightarrow{\text{p}} \|\theta_0 - \theta\|_{M_{X'X}}.$$

Furthermore, due to the triangle inequality-in its dual form,[30] for any $\theta, \theta_\star \in \Theta$,

$$\left|\|(\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{X}_n'Y_n - \theta\|_{\frac{\mathbf{X}_n'\mathbf{X}_n}{n}} - \|(\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{X}_n'Y_n - \theta_\star\|_{\frac{\mathbf{X}_n'\mathbf{X}_n}{n}}\right| \leq \|\theta_\star - \theta\|_{\frac{\mathbf{X}_n'\mathbf{X}_n}{n}}.$$

The above, combined with the submultiplicative property of the Frobe-

[28] Condition (3) implies that inversion of matrices near $M_{X'X}$ is a continuous operator-why?

[29] The root of the associated quadratic form is continuous both with respect to its argument and the matrix in the center.

[30] $|\|a\| - \|b\|| \leq \|a - b\|$.

nius norm and Condition (2), implies that Lemma 1 applies with $k_n = \|\frac{\mathbf{X}_n'\mathbf{X}_n}{n}\|$, and thus,

$$\|(\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{X}_n'Y_n - \theta\|_{\frac{\mathbf{X}_n'\mathbf{X}_n}{n}} \xrightarrow{\text{cp}} \|\theta_0 - \theta\|_{M_{X'X}}.$$

Additionally, since by Condition (3), $\theta_0 \in \Theta$, we have $\theta_0 = \arg\min_{\theta\in\Theta} \|\theta_0 - \theta\|_{M_{X'X}}$. Therefore, when Θ is compact, Theorem 2 tells us that Conditions (1)-(3) are also sufficient for weak consistency.

In the case where Θ is closed and convex, we also note that

$$\theta_n = \arg\min_{\theta\in\Theta} \left[(\theta - \theta_0)'\frac{\mathbf{X}_n'\mathbf{X}_n}{n}(\theta - \theta_0) - 2(\theta - \theta_0)\frac{\mathbf{X}_n'\varepsilon_n}{n}\right]$$

(**Exercise**: Show this! You may use the fact that $\frac{\varepsilon_n'\varepsilon_n}{n}$ is independent of θ). The objective function is strictly convex because $\text{rank}(\frac{\mathbf{X}_n'\mathbf{X}_n}{n}) = p$, and due to (1)-(2), it converges pointwise in probability to $(\theta - \theta_0)'M_{X'X}(\theta - \theta_0)$, which, due to (3), is also strictly convex. Therefore, Theorem 3 informs us that even in this case-where the parameter space is closed and convex-(1)–(3) are also sufficient for weak consistency.

> *Exersice*: Show that when θ_0 is an interior point of a closed and convex subset of Θ, then (1)-(3) are sufficient for weak consistency.

> *Exersice*: Show that when θ_0 is an interior point of a compact subset of Θ, then (1) and (3) are sufficient for weak consistency.

> *Exersice*: Use $(\theta - \theta_0)'\frac{\mathbf{X}_n'\mathbf{X}_n}{n}(\theta - \theta_0) - 2(\theta - \theta_0)\frac{\mathbf{X}_n'\varepsilon_n}{n}$ and Theorem 2 to derive the sufficiency of (1)-(3) in the case of compactness. Use Lemma 1, including-among other things-the fact that due to compactness, Θ is necessarily bounded.

The handling of the IVE is analogous to that of the previous example. It is outlined in the following exercises:

Exersice: Within the framework of Example 7.5, find the asymptotic theory of the IVE when $\Theta = \mathbb{R}^p$, under the high-order conditions:

1. $\frac{1}{n}\mathbf{W}_n'\varepsilon_n \xrightarrow{\text{p}} \mathbf{0}_p$,
2. There exists a non-stochastic square $p \times p$ matrix $M_{X'W}$, such that $\frac{1}{n}\mathbf{X}_n'\mathbf{W}_n \xrightarrow{\text{p}} M_{X'W}$, and
3. $\text{rank}(M_{X'W}) = p$.

Exersice: Using the above, show that high-order conditions (I)-(III) are sufficient for the weak consistency of the IVE when $\Theta = \mathbb{R}^p$.

Exersice: Using the representation

$$\theta_n \in \arg\min_{\theta\in\Theta} \left\|(\mathbf{X}_n'\mathbf{W}_n V \mathbf{W}_n'\mathbf{X}_n)^{-1}\mathbf{X}_n'\mathbf{W}_n V \mathbf{W}_n' Y_n - \theta\right\|_{\frac{\mathbf{X}_n'\mathbf{W}_n}{n} V \frac{\mathbf{W}_n'\mathbf{X}_n}{n}},$$

and Theorem 2, show that high-order conditions (I)-(III) are sufficient for the weak consistency of the IVE when Θ is compact.

Exersice: Using Theorem 3, show that high-order conditions (I)-(III) are sufficient for the weak consistency of the IVE when Θ is closed and convex.

Exersice: Adapt the above results for the IVE to the more general case where the matrix V depends on n and/or is stochastic.

Example 8 (GARCH(1,1))**.** It is proven that the description of the example so far includes all the sufficient conditions for the weak consistency of the QMLE. The corresponding derivations-see, for example, Chapter 5 of

[10]-include the use of results for SREs (see the relevant part of the notes about stochastic processes), to show that

$$c_n(\theta) := \frac{1}{n}\sum_{t=1}^{n}\left(\ln(h_t^*(\theta)) + \frac{y_t^2}{h_t^*(\theta)}\right),$$

from which the estimator is derived, is approximated, with probability converging to one, by the stationary and ergodic (h_t). It is further shown that $\sup_{\theta\in\Theta}\left(\ln(h_0(\theta)) + \frac{y_0^2}{h_0(\theta)}\right)$ is integrable, ensuring the applicability of a uniform version of Birkhoff's Law of Large Numbers to $\frac{1}{n}\sum_{t=1}^{n}\left(\ln(h_t(\theta)) + \frac{y_t^2}{h_t(\theta)}\right)$. Through this, it is finally shown that

$$\frac{1}{n}\sum_{t=1}^{n}\left(\ln(h_t^*(\theta)) + \frac{y_t^2}{h_t^*(\theta)}\right) \xrightarrow{\text{cp}} \mathbb{E}\left(\ln(h_0(\theta)) + \frac{\sigma_0^2}{h_0(\theta)}\right),$$

which makes Theorem 2 applicable.[31]

4.2 Rate of Convergence and Asymptotic Distribution

Given the consistency of the estimator, we are concerned with formulating sufficient conditions to determine the rate of convergence and the asymptotic distribution when these are well-defined.

The conditions that led us to consistency pertained to the global behavior of the criterion c_n; they imposed restrictions on the asymptotic behavior of the sequence of stochastic processes forming the objective functions in neighborhoods of every point in the parameter space. By definition, they led to a property-weak consistency-that is equivalent to the estimator being within an arbitrary neighborhood of θ_0 with probability converging to one. This implies that the analysis leading to the rate of convergence and the asymptotic distribution will be more local, as it essentially concerns

[31] In fact, it is shown that the QMLE converges almost surely to θ_0; it is, as commonly stated, strongly consistent.

the asymptotic behavior of the process relative to sequences converging at suitable rates to θ_0. The approach below partially follows the standard practice in the literature.

The latter relies on a local quadratic approximation of the criterion in a suitable neighborhood of θ_0, derived from the local second-order Taylor expansion of c_n with respect to θ around θ_0. Therefore, it requires that θ_0 is in the interior of Θ.[32] It also requires the criterion to be twice continuously differentiable in some neighborhood of θ_0, so that the expansion is valid. Consequently, consistency implies that, at least with probability converging to one, the estimator-as the minimizer of the objective function-will satisfy first-order conditions. These are used to obtain, under suitable assumptions, an asymptotic representation of the estimator that leads to the determination of the rate of convergence and the asymptotic distribution.

The approach followed here is somewhat more general, as it allows θ_0 to lie on the boundary of the parameter space. In this case, first-order conditions may not hold-even asymptotically-unless the optimization problem can, at least asymptotically, be characterized by first- and second-order conditions involving some Lagrangian function. Our analysis will be slightly more general and will not involve such characterizations. However, it will involve a notion of an asymptotic parameter space constructed as some kind of limit of sets, essentially forming the support of the asymptotic distribution. Under assumptions analogous to those mentioned earlier, when $\theta_0 \in \Theta^o$ (interior of Θ), the analysis will determine the rate and characterize the asymptotic distribution through the minimizer of a strictly convex quadratic form over a convex set.

Before proceeding with the relevant formulation, we need the following concept of the limit of a sequence of Euclidean sets:

[32] Consequently, Θ cannot have an empty interior; for instance, it cannot be discrete.

Let (H_n) be a sequence of non-empty subsets of $\mathbb{R}^p$. The Painlevé-Kuratowski limit of this sequence-if it exists-is defined as the non-empty subset $H \subset \mathbb{R}^p$, consisting of all $x \in \mathbb{R}^p$ such that $x = \lim x_n$ for some sequence $x_n \in H_n$, $\forall n \in \mathbb{N}$, and simultaneously, there does not exist any accumulation point of a sequence (x_n) with $x_n \in H_n$, $\forall n \in \mathbb{N}$, that is not in H. By construction, when it exists, the limit is a closed subset of $\mathbb{R}^p$. For example, for $p = 1$ and $A_n = (0, n)$, the limit is $H = [0, +\infty)$. As a counterexample, consider $H_n := \{-n, n\}$. Clearly, the limit does not exist in this case. It can be shown that if H_n is convex and the sequence is increasing,then the limit exists and is also convex.

In what follows, for a sequence $r_n \to +\infty$, which will be specified later, we define $H_n := r_n(\Theta - \theta_0) = \{r_n(\theta - \theta_0) : \theta \in \Theta\}$. The H_n, given the aforementioned sequence, will act as a modified parameter space. It is constructed by translating Θ by θ_0 and scaling the Euclidean norm of the elements in the translated set by r_n. Since $\theta_0 \in \Theta$, we have $H_n \neq \emptyset$, as it will contain at least $\mathbf{0}_p$. If r_n represents the rate of convergence of the estimator, then H_n will include all possible values of $r_n(\theta_n - \theta_0)$.

Since we are interested in the asymptotic behavior of the latter, an argument based on the aforementioned Skorokhod representations (see, e.g., the proof of Theorem 2) will indicate that, if it exists, H, the Painlevé limit of H_n, will contain the possible values that the random element $r_n(\theta_n - \theta_0)$-if it exists-can asymptotically take in distribution.

We begin our analysis with the following assumption, which specifies the local (around θ_0) asymptotic behavior of the criterion process c_n, the asymptotic behavior of the modified parameter space H_n, and the optimization error involved in the general definition of θ_n.

Assumption 1. *Let the following conditions hold:*

1. *For any sequence $(\theta_n^\star)$ with values in Θ such that $\theta_n^\star \to \theta_0$,*

$$c_n(\theta_n^\star) - c_n(\theta_0) = (\theta_n^\star - \theta_0)' q_n(\theta_0) + (\theta_n^\star - \theta_0)' g_n(\theta_n^{\star\star})(\theta_n^\star - \theta_0),$$

with probability converging to 1, where $\theta_n^{\star\star}$ lies on the line segment connecting $\theta_n^{\star}$ and θ_0 within $\mathbb{R}^p$. The matrix g_n is a stochastic $p \times p$ matrix that is well-defined with probability converging to 1 at every point on the aforementioned line. The vector q_n is a random $p \times 1$ vector.

2. *For some non-negative sequence $r_n \to +\infty$, $r_n q_n(\theta_0) \rightsquigarrow z_{\theta_0}$, where z_{θ_0} is a random vector with a well-defined distribution that may depend on θ_0. Moreover, $g_n(\theta_n^{\star\star}) \xrightarrow{p} \breve{\mathbf{J}}_{\theta_0}$, where $\breve{\mathbf{J}}_{\theta_0}$ is a non-stochastic, strictly positive definite matrix that may depend on θ_0.*

3. *The Painlevé-Kuratowski limit H exists and is convex.*

4. *For the optimization error, we have $u_n = o_p(r_n^{-2})$.*

The first condition of the above assumption is trivially satisfied when θ_0 is an interior point and the criterion is twice continuously differentiable in a neighborhood of θ_0 (with probability converging to 1). In this case, the validity of the assumption follows from the corresponding local Taylor expansion of the criterion with a remainder term given by the Mean Value Theorem. When θ_0 is a boundary point, a similar condition can hold if, for example, c_n can be suitably extended to an open neighborhood of θ_0 within $\mathbb{R}^p$. Even when this is not feasible, the validity of the assumption may be ensured by a similar Taylor expansion involving appropriately one-sided or oriented derivatives (see, e.g., [1]). In such cases, the vector q_n corresponds to the relevant first-order derivatives, and the matrix g_n to the associated Hessian matrix. Note that by construction, $\theta_n^{\star\star} \to \theta_0$ (why?).

The second condition specifies the asymptotic behavior of the "derivatives"; the first part can be ensured by some Central Limit Theorem or an analysis such as the current one involving the asymptotic behavior of some primary auxiliary estimator. In many cases, $r_n = \sqrt{n}$ and z_{θ_0} follows some Gaussian distribution on $\mathbb{R}^p$. However, asymptotic normality may exist without the rate necessarily being $\sqrt{n}$. The second part involves a local (only for sequences converging to θ_0) version of the aforementioned

continuous convergence in probability and may be facilitated by some uniform Law of Large Numbers. The non-singularity of the limiting matrix may follow from some kind of asymptotic linear algebraic independence of the columns of g_n.

The third condition states that H_n must have a convex limit. It is trivially satisfied when θ_0 is an interior point, in which case $H = \mathbb{R}^p$. As mentioned earlier, it will hold when there is convexity and monotonicity. In any case, it presupposes that Θ has a non-empty interior. An example where the limit exists but is not convex is as follows: let Θ be a countable union of line segments, each centered at θ_0 with the same finite radius. Clearly, the parameter set is not convex unless these are collinear, in which case there is only one segment. The limit exists and is a countable union of one-dimensional lines. It will not be convex unless all segments are collinear.

The last condition, given the second, places restrictions on how slowly the optimization error can converge to zero in probability.

Given the above, we obtain the following fundamental result:

Theorem 4. *Let the estimator be weakly consistent and suppose that Assumptions 1.1, 1.2, and 1.4 hold. Then,*

$$r_n\left(\theta_n - \theta_0\right) = O_p\left(1\right). \tag{4}$$

If, in addition, Assumption 1.3 holds, then

$$r_n\left(\theta_n - \theta_0\right) \rightsquigarrow \tilde{h}_{\theta_0}, \tag{5}$$

where

$$\tilde{h}_{\theta_0} := \arg\min_{h \in H} \left(h + \frac{1}{2}\breve{\mathbf{J}}_{\theta_0}^{-1} z_{\theta_0}\right)' \breve{\mathbf{J}}_{\theta_0} \left(h + \frac{1}{2}\breve{\mathbf{J}}_{\theta_0}^{-1} z_{\theta_0}\right).$$

Proof. From the definition of θ_n and Assumption 1.4, we have:

$$c_n\left(\theta_n\right) - c_n\left(\theta_0\right) \leq o_p\left(r_n^{-2}\right).$$

Due to consistency and Assumptions 1.1 and 1.2, we have:

$$h_n' r_n q_n(\theta_0) + h_n' g_n(\theta_n^{\star\star}) h_n \leq o_p(1),$$

where $h_n := r_n(\theta_n - \theta_0)$ and $\theta_n^{\star\star}$ is stochastic and arises as in Assumption 1.1. By consistency, we have:

$$h_n' r_n q_n(\theta_0) + h_n' \left(\breve{\mathbf{J}}_{\theta_0} + o_p(1)\right) h_n \leq o_p(1).$$

Assumption 1.2 then implies that there exists $c > 0$ such that:

$$O_p(\|h_n\|) - c\|h_n\|^2 + \|h_n\|^2 o_p(1) \geq o_p(1),$$

which implies:

$$O_p(1) \geq \|h_n\|^2 (1 + o_p(1)) - 2\|h_n\| (1 + o_p(1)) O_p(1) + O_p(1).$$

Thus:

$$\|h_n\| (1 + o_p(1)) \leq O_p(1),$$

which proves (4).

Using Assumption 1, we have, with probability tending to 1,[33]

$$\varpi_n(h) := r_n^2 \left(c_n\left(\theta_0 + \frac{h}{r_n}\right) - c_n(\theta_0)\right) = h' r_n q_n(\theta_0) + h' g_n(\theta_n^{\star\star}) h.$$

From the first part of this proof and Assumption 1.2, for any arbitrary compact non-empty subset A of $\mathbb{R}^p$ and for every $h \in A$, $A \ni h_n \to h$, we have:

$$\varpi_n(h) \rightsquigarrow h' z_{\theta_0} + h' \breve{\mathbf{J}}_{\theta_0} h.$$

Thus, by the Continuous Mapping Theorem (see also the proof of Theorem

[33] Clearly, for any $h \in \mathbb{R}^p$, $\theta_0 + \frac{h}{r_n} \to \theta_0$, so Assumption 1 is applicable.

2):

$$\inf_{h\in A} \varpi_n(h) \rightsquigarrow \inf_{h\in A} \left(h' z_{\theta_0} + h' \breve{\mathbf{J}}_{\theta_0} h \right). \tag{6}$$

If F is a closed, non-empty subset of $\mathbb{R}^p$ and $h_n \in F$, then for sufficiently large n, $H_n \cap F \neq \varnothing$. In any case, due to the definition of θ_n, ϖ_n, and the fact that by assumption $u_n = o_p(r_n^{-2})$, we have:

$$\inf_{h\in H_n\cap F} \varpi_n(h) \leq \inf_{h\in H_n} \varpi_n(h) + o_p(1),$$

and therefore, due to the Continuous Mapping Theorem,

$$\mathbb{P}\left(h_n \in F\right) \leq \mathbb{P}\left(\inf_{h\in H_n\cap F} \varpi_n(h) \leq \inf_{h\in H_n} \varpi_n(h) + o_p(1) \right).$$

$$\leq \mathbb{P}\left(\inf_{h\in H_n\cap F} \varpi_n(h) \leq \inf_{h\in H_n} \varpi_n(h) \right) + o(1).$$

Equation 6 and the Continuous Mapping Theorem imply that Lemma 7.13.2-3 in [12] is applicable, and therefore, the preceding probability is less than or equal to:

$$\mathbb{P}\left(\inf_{h\in H\cap F} \varpi_n(h) \leq \inf_{h\in H} \varpi_n(h) + o_p(1) \right).$$

$$\leq \mathbb{P}\left(\inf_{h\in H\cap F} \varpi_n(h) \leq \inf_{h\in H} \varpi_n(h) \right) + o(1),$$

where the last inequality arises from the Continuous Mapping Theorem. From 6, the Continuous Mapping Theorem, and the Portmanteau Theorem (see the part of the notes about convergence in distribution), the $\limsup$ of the probability on the right-hand side of the previous expression is bounded above by:

$$\mathbb{P}\left(\inf_{h\in H\cap F} h' z_{\theta_0} + h' \breve{\mathbf{J}}_{\theta_0} h \leq \inf_{h\in H} h' z_{\theta_0} + h' \breve{\mathbf{J}}_{\theta_0} h \right),$$

which is equal to:

$$\mathbb{P}\left(\inf_{h\in H\cap F} 2h'\breve{\mathbf{J}}_{\theta_0}\breve{\mathbf{J}}_{\theta_0}^{-1}\frac{1}{2}z_{\theta_0}+h'\breve{\mathbf{J}}_{\theta_0}h\pm\frac{1}{4}z'_{\theta_0}\breve{\mathbf{J}}_{\theta_0}^{-1}z_{\theta_0}\leq\inf_{h\in H}2h'\breve{\mathbf{J}}_{\theta_0}\breve{\mathbf{J}}_{\theta_0}^{-1}\frac{1}{2}z_{\theta_0}+\frac{1}{2}h'\breve{\mathbf{J}}_{\theta_0}h\pm\frac{1}{4}z'_{\theta_0}\breve{\mathbf{J}}_{\theta_0}^{-1}z_{\theta_0}\right),$$

which is in turn equal to:

$$\mathbb{P}\left(\inf_{h\in H\cap F}\left(h+\frac{1}{2}\breve{\mathbf{J}}_{\theta_0}^{-1}z_{\theta_0}\right)'\breve{\mathbf{J}}_{\theta_0}\left(h+\frac{1}{2}\breve{\mathbf{J}}_{\theta_0}^{-1}z_{\theta_0}\right)\leq\inf_{h\in H}\left(h+\frac{1}{2}\breve{\mathbf{J}}_{\theta_0}^{-1}z_{\theta_0}\right)'\breve{\mathbf{J}}_{\theta_0}\left(h+\frac{1}{2}\breve{\mathbf{J}}_{\theta_0}^{-1}z_{\theta_0}\right)\right).$$

Since H is closed and convex, and $\breve{\mathbf{J}}_{\theta_0}$ is positive definite, $\tilde{h}_{\theta_0}$ is unique. Hence, when:

$$\inf_{h\in H\cap F}\left(h+\frac{1}{2}\breve{\mathbf{J}}_{\theta_0}^{-1}z_{\theta_0}\right)'\breve{\mathbf{J}}_{\theta_0}\left(h+\frac{1}{2}\breve{\mathbf{J}}_{\theta_0}^{-1}z_{\theta_0}\right)\leq\inf_{h\in H}\left(h+\frac{1}{2}\breve{\mathbf{J}}_{\theta_0}^{-1}z_{\theta_0}\right)'\breve{\mathbf{J}}_{\theta_0}\left(h+\frac{1}{2}\breve{\mathbf{J}}_{\theta_0}^{-1}z_{\theta_0}\right),$$

we have:

$$\tilde{h}_{\theta_0}\in H\cap F.$$

Thus, the probability is at most:

$$\mathbb{P}\left(\tilde{h}_{\theta_0}\in H\cap F\right)\leq\mathbb{P}\left(\tilde{h}_{\theta_0}\in F\right).$$

Therefore, we have shown that:

$$\limsup_{n\to\infty}\mathbb{P}\left(h_n\in F\right)\leq\mathbb{P}\left(\tilde{h}_{\theta_0}\in F\right),$$

and hence 5 follows from the Portmanteau Theorem and the fact that F was chosen arbitrarily. □

Given the quadratic expansion in Assumption 1.1, the rate in 1.2, and the non-singularity of the limiting matrix in the same assumption, the rate of the estimator is determined as r_n. In many cases-but not always-this rate is the classical $\sqrt{n}$. The asymptotic distribution is well-defined because the random element $\tilde{h}_{\theta_0}$ is uniquely determined due to the strict convexity of

the involved quadratic form and the convexity of the asymptotic parameter space H.

It can be shown that if Assumption 1.2 did not hold with the positive definiteness of the asymptotic matrix, then extending the analysis to a higher-order polynomial-if feasible-would imply that the convergence rate would not equal r_n but would depend on r_n and the degree of this polynomial (see, for instance, [5]). When θ_0 is an interior point, then due to the aforementioned observation, we have $\tilde{h}_{\theta_0} = -\frac{1}{2}\breve{\mathbf{J}}_{\theta_0}^{-1} z_{\theta_0}$, giving us an asymptotic expression for the estimator.[34]

In this case, if $r_n = \sqrt{n}$ and $z_{\theta_0} \sim N(\mathbf{0}_p, V_{\theta_0})$ for some positive definite V_{θ_0}, then the standard asymptotic theory is recovered:

$$\sqrt{n}(\theta_n - \theta_0) \rightsquigarrow N\left(\mathbf{0}_p, \frac{1}{4}\breve{\mathbf{J}}_{\theta_0}^{-1} V_{\theta_0} \breve{\mathbf{J}}_{\theta_0}^{-1}\right),$$

indicating that the estimator is asymptotically normal with asymptotic variance $\frac{1}{4}\breve{\mathbf{J}}_{\theta_0}^{-1} V_{\theta_0} \breve{\mathbf{J}}_{\theta_0}^{-1}$.

However, when $r_n = \sqrt{n}$ and $z_{\theta_0} \sim N(\mathbf{0}_p, V_{\theta_0})$, but θ_0 is a boundary point, the asymptotic distribution is not normal but some form of a projection of $N\left(\mathbf{0}_p, \frac{1}{4}\breve{\mathbf{J}}_{\theta_0}^{-1} V_{\theta_0} \breve{\mathbf{J}}_{\theta_0}^{-1}\right)$ onto H. The estimator has a smaller asymptotic variance[35] than $\frac{1}{4}\breve{\mathbf{J}}_{\theta_0}^{-1} V_{\theta_0} \breve{\mathbf{J}}_{\theta_0}^{-1}$, in the sense that the difference between the latter and the former is necessarily a positive definite matrix.

This indicates that when there is external information about θ_0, it may

[34] Recall that we typically do not have the analytical form of the estimator as a function of the sample. This analysis provides a relative expression based on the local characteristics of the criterion.

[35] The projection is determined by the optimization problem described in the theorem. When $-\frac{1}{2}\breve{\mathbf{J}}_{\theta_0}^{-1} z_{\theta_0} \in H$, we have $\tilde{h}_{\theta_0} = -\frac{1}{2}\breve{\mathbf{J}}_{\theta_0}^{-1} z_{\theta_0}$. Otherwise, $\tilde{h}_{\theta_0}$ equals the unique-due to the related convexities-element of H with the smallest possible distance, in terms of the quadratic form with respect to the matrix $\breve{\mathbf{J}}_{\theta_0}$, from $-\frac{1}{2}\breve{\mathbf{J}}_{\theta_0}^{-1} z_{\theta_0}$. Thus, the resulting distribution assigns the probability given by $N\left(\mathbf{0}_p, \frac{1}{4}\breve{\mathbf{J}}_{\theta_0}^{-1} V_{\theta_0} \breve{\mathbf{J}}_{\theta_0}^{-1}\right)$ to any measurable subset of H entirely within its interior, zero probability to anything disjoint from H, and probabilities for measurable portions of the boundary derived from the corresponding probabilities assigned by $N\left(\mathbf{0}_p, \frac{1}{4}\breve{\mathbf{J}}_{\theta_0}^{-1} V_{\theta_0} \breve{\mathbf{J}}_{\theta_0}^{-1}\right)$ to what projects there via the aforementioned process.

be worthwhile-concerning the criterion of asymptotic variance and if the above holds-to incorporate it into the choice of Θ when this implies that θ_0 will be a boundary point. It can be shown that the boundaries of subsets of $\mathbb{R}$ have zero Lebesgue measure. Therefore, in the absence of additional information, such scenarios are not expected to occur "frequently." However, when a related theory suggests that the target parameter value is likely to be on the boundary, it may be advantageous to use this information in terms of asymptotic variance.

4.2.1 Examples and Exercises

Let us now examine how the above concepts apply to our examples:

Example 9 (Linear Model)**.** We first consider the case where $\Theta = \mathbb{R}^p$. To the assumptions (1)-(3) developed in Example 7.7, we add the following high-order assumption:

4. Assume there exists $z \sim N(\mathbf{0}_p, M_{X'X})$ such that $\frac{1}{\sqrt{n}}\mathbf{X}'_n\varepsilon_n \rightsquigarrow z$.

Given the specification assumption $\text{Var}(\varepsilon_n) = \mathbf{I}_n$ and Assumption (2), (4) can be derived in an iid setting or more generally within the framework of our established Central Limit Theorem if $\mathbb{E}(|\mathbf{X}_{0,j}\varepsilon_0|^k) < +\infty$, $\forall j = 1, \ldots, p$, for some $k > 2$, due to the uniform integrability property.

Given (2)-(4) and the expression $\theta_n = \theta_0 + (\frac{1}{n}\mathbf{X}'_n\mathbf{X}_n)^{-1}\frac{1}{n}\mathbf{X}'_n\varepsilon_n$, which equivalently and under (4) can be restated as

$$\sqrt{n}(\theta_n - \theta_0) = (\frac{1}{n}\mathbf{X}'_n\mathbf{X}_n)^{-1}\frac{1}{\sqrt{n}}\mathbf{X}'_n\varepsilon_n.$$

Slutsky's Lemma, the Continuous Mapping Theorem, and assumptions (2)-(4) imply that the right-hand side of the above expression converges in distribution to $M_{X'X}^{-1}z$. Hence, (2)-(4) imply that[36]

$$\sqrt{n}(\theta_n - \theta_0) \rightsquigarrow M_{X'X}^{-1}z \sim N(\mathbf{0}_p, M_{X'X}^{-1}M_{X'X}M_{X'X}^{-1}) = N(\mathbf{0}_p, M_{X'X}^{-1}).$$

[36]Recall that $M_{X'X}$ and its inverse are symmetric matrices.

How does the above change when the parameter space does not coincide with $\mathbb{R}^p$? What do (2)-(4) imply about the asymptotic theory of the OLSE? Recall that in this case, the estimator is equivalently the minimizer of $c_n(\theta) = (\theta - \theta_0)' \frac{\mathbf{X}_n' \mathbf{X}_n}{n} (\theta - \theta_0) - 2(\theta - \theta_0) \frac{\mathbf{X}_n' \varepsilon_n}{n}$, so Assumption 1.1 is satisfied with $q_n(\theta_0) = -2\frac{\mathbf{X}_n' \varepsilon_n}{n}$-independent of θ_0-and $g_n(\theta^{\star\star}) = \frac{\mathbf{X}_n' \mathbf{X}_n}{n}$-also independent of the parameter in this case. Clearly, (2)-(4) imply Assumption 1.2.

If Assumption 1.3 is also satisfied and the estimator is weakly consistent, then we have

$$\sqrt{n}(\theta_n - \theta_0) \rightsquigarrow \arg\min_{h \in H} \left(h - M_{X'X}^{-1} z\right)' M_{X'X} \left(h - M_{X'X}^{-1} z\right),$$

where H is the convex limit of the corresponding $\sqrt{n}(\Theta - \theta_0)$ if it exists. When θ_0 is an interior point, we recover the standard asymptotic theory of the unrestricted estimator, as in this case $H = \mathbb{R}^p$. When θ_0 is on the boundary of the parameter space, we obtain a projection of the above distribution with "reduced variance."

> *Exersice*: Show that within the framework of the previous example, (4) implies (1).

> *Exersice*: Let $p = 1$, the matrix $\mathbf{X}_n$ consists entirely of ones, the elements of ε_n are iid with mean 0 and unit variance, $\Theta = [0, +\infty)$, and $\theta_0 = 0$. Find the rate of convergence and the asymptotic distribution of the OLSE.

The handling of the IVE is analogous to that of the previous example. As in the previous section, it is explored through the following exercises:

> *Exersice*: Within the framework of Example 7.5, derive the asymptotic theory of the IVE under the assumptions:
>
> 2. There exists a non-stochastic square $p \times p$ matrix $M_{X'W}$ such that

$\frac{1}{n}\mathbf{X}'_n\mathbf{W}_n \xrightarrow{\text{p}} M_{X'W}$,

3. $\text{rank}(M_{X'W}) = p$, and,

4. $\frac{1}{\sqrt{n}}\mathbf{W}'_n\varepsilon_n \rightsquigarrow z \sim N(\mathbf{0}_p, V_{W'\varepsilon})$, where $V_{W'\varepsilon}$ is a strictly positive-definite matrix.

Exersice: Under the above assumptions, derive the corresponding asymptotic theory when the IVE is consistent and Θ does not necessarily coincide with $\mathbb{R}^p$. When is the asymptotic theory independent of V?

Exersice: Adapt the above results concerning the IVE to the more general case where the matrix V depends on n and/or is stochastic.

Example 10 (GARCH(1,1))**.** It can be shown-see, for example, Chapter 5 of [10]-using among others our general CLT and Theorem 4, the theory of SREs for the existence, uniqueness, and approximability of solutions to recursive relations derived from first- and second-order differentiations of the recursion defining $(h_t(\theta))$, and under the conditions $\mathbb{E}(z_0^4) < +\infty$, the algebraic linear independence of the random variables forming the random vector $(1, y_0, \sigma_0^2)$, and that the optimization error satisfies the asymptotic behavior prescribed in Assumption 1.4, that for the QMLE:

$$\sqrt{n}(\theta_n - \theta_0) \rightsquigarrow \arg\min_{h\in H}\left(h - \breve{\mathbf{J}}_{\theta_0}^{-1}z\right)'\breve{\mathbf{J}}_{\theta_0}\left(h - \breve{\mathbf{J}}_{\theta_0}^{-1}z\right),$$

where $\breve{\mathbf{J}}_{\theta_0} := \mathbb{E}\left[\frac{h'_0(\theta_0)h'_0(\theta_0)^T}{\sigma_0^4}\right]$, $z \sim N\left(\mathbf{0}_3, (\mathbb{E}(z_0^4) - 1)\,\breve{\mathbf{J}}_{\theta_0}\right)$, and H is the convex limit of the corresponding $\sqrt{n}(\Theta - \theta_0)$ if it exists. Here, $(h'_t(\theta))$ represents the stationary and ergodic solution to the recursive relation derived by differentiating with respect to the parameter in the recursion defining $h_t(\theta)$.

Note that in this example, unlike the previous ones, the involved z (via the variance of its distribution) and $\breve{\mathbf{J}}_{\theta_0}$ are related to θ_0. When $\mathbb{E}(z_0^4) = +\infty$,

and if, as $M \to +\infty$, the truncated moment $\mathbb{E}(z_0^4 \mathbf{1}(|z_0| \leq M))$ is asymptotically proportional to $C \ln(M)$ for some positive constant C, it is shown that there exists a positive constant $C^\star$ such that

$$\sqrt{\frac{n}{\ln(n)}}(\theta_n - \theta_0) \rightsquigarrow \arg\min_{h \in H} \left(h - \breve{\mathbf{J}}_{\theta_0}^{-1} z\right)' \breve{\mathbf{J}}_{\theta_0} \left(h - \breve{\mathbf{J}}_{\theta_0}^{-1} z\right),$$

where $\breve{\mathbf{J}}_{\theta_0}$ is as before, and $z \sim N\left(\mathbf{0}_3, C^\star \breve{\mathbf{J}}_{\theta_0}\right)$-see, for instance, [2] for a similar result in a different constrained heteroskedasticity model. Proving such a result cannot rely on a result like our general CLT and requires a partial generalization of it (to be considered along with tools like the Wold device). This result illustrates the possibility of asymptotic normality with a convergence rate different from the standard $\sqrt{n}$, even when the criterion has a non-degenerate local quadratic expansion.

5 Asymptotic Hypothesis Tests

Consider, in relation to the framework we have developed for hypothesis tests in the previous notes that for some $\theta^\star \in \Theta$, we are interested in testing the hypothesis structure:

$$\begin{aligned} H_0 &: \theta_0 = \theta^\star, \\ H_1 &: \theta_0 \in \Theta - \{\theta^\star\}, \end{aligned} \tag{7}$$

which contains a simple null and a composite alternative hypothesis.

The structure of the preceding sections allows us to construct asymptotic hypothesis tests using test statistics derived from the criterion c_n and the estimator, with rejection regions based on the aforementioned asymptotic theory.

One such approach is described below.[37] Under the framework described

[37] This is not an exhaustive approach.

by Theorems 2, 3, and 4, and using an argument that shows the asymptotic independence of the estimator (which, as consistent, converges to something non-stochastic) and $r_n^2(c_n(\theta) - c_n(\theta_0))$, it follows-tracing the proof of Theorem 4-that:

$$r_n^2(c_n(\theta_0) - c_n(\theta_n)) \rightsquigarrow \mathbf{Z}(\theta_0) := z'_{\theta_0} \frac{1}{4} \mathbf{J}_{\theta_0}^{-1} z_{\theta_0} - \min_{h \in H} \left(h + \frac{1}{2} \breve{\mathbf{J}}_{\theta_0}^{-1} z_{\theta_0} \right)' \breve{\mathbf{J}}_{\theta_0} \left(h + \frac{1}{2} \mathbf{J}_{\theta_0}^{-1} z_{\theta_0} \right).$$

Moreover, when θ_0 is an interior point, the second term in the above limit vanishes, since $H = \mathbb{R}^p$ and thus h is free to admit every value $-\frac{1}{2}\mathbf{J}_{\theta_0}^{-1} z_{\theta_0}$ admits, and thus we have:

$$r_n^2(c_n(\theta_0) - c_n(\theta_n)) \rightsquigarrow z'_{\theta_0} \frac{1}{4} \mathbf{J}_{\theta_0}^{-1} z_{\theta_0}.$$

In the case where $z_{\theta_0} \sim N(\mathbf{0}_p, 4\breve{\mathbf{J}}_{\theta_0})$, the asymptotic distribution of $r_n^2(c_n(\theta_0) - c_n(\theta_n))$ has a special relationship with the standard normal distributions:

> If k is a strictly positive integer, the chi-squared distribution with k degrees of freedom-χ_k^2-is defined as the distribution on $\mathbb{R}$ with support $[0, +\infty)$ and density function:
>
> $$f(x; k) = \begin{cases} 0, & x < 0, \\ \frac{x^{\frac{k}{2}-1}}{2^{k/2}\Gamma(\frac{k}{2})} \exp\left(-\frac{x}{2}\right), & x \geq 0, \end{cases}$$
>
> where $\Gamma(x) := \int_0^{+\infty} t^{x-1} \exp(-t) dt$ is the gamma function. **Exercise:** Show that this is a well-defined density function. It can be shown that if $\mathbf{x} \sim N(\mathbf{0}_k, V)$, with V non-singular, then the quadratic form $\mathbf{x}' V^{-1} \mathbf{x} \sim \chi_k^2$. **Exercise:** Prove this!

Thus, in this case, by the Continuous Mapping Theorem, $r_n^2(c_n(\theta_0) - c_n(\theta_n)) \rightsquigarrow \chi_p^2$.

Exersice: Prove this!

When θ_0 is a boundary point, and $z_{\theta_0} \sim N(\mathbf{0}_p, 4\breve{\mathbf{J}}_{\theta_0})$, then-and due to that $\min_{h\in H}\left(h + \frac{1}{2}\breve{\mathbf{J}}_{\theta_0}^{-1} z_{\theta_0}\right)' \breve{\mathbf{J}}_{\theta_0}\left(h + \frac{1}{2}\mathbf{J}_{\theta_0}^{-1} z_{\theta_0}\right)$ is non-negative (why?)-it can be shown that the distribution of the limit is stochastically dominated in the first-order sense by χ^2_p; the cumulative distribution function of the latter is pointwise less than or equal to that of the former.

5.1 The Testing Procedure

Consider the previously mentioned hypothesis structure:

Algorithm 2 A Testing Procedure Based on the Criterion

1. Use as test statistic $\mathcal{L}_n := r_n^2(c_n(\theta^\star) - c_n(\theta_n))$.
2. Choose a significance level $\alpha \in (0, 1)$.
3. Based on α, determine:

$$q_\alpha := \inf\left\{x \in (0, +\infty) : \int_0^x \frac{z^{\frac{k}{2}-1}}{2^{k/2}\Gamma(\frac{k}{2})} \exp\left(-\frac{z}{2}\right)dz \geq 1 - \alpha\right\}.$$[38]

4. Define the rejection region for H_0 as the interval $(q_{1-\alpha}, +\infty)$.
5. Evaluate the test statistic on the sample, and reject H_0 if and only if $\mathcal{L}_n \in (q_{1-\alpha}, +\infty)$.

Note that the above procedure relies on the knowledge of the convergence rate r_n. In many econometric settings, this rate is known and equals $\sqrt{n}$, so the test statistic simplifies to $\mathcal{L}_n := n(c_n(\theta^\star) - c_n(\theta_n))$. When the rate is unknown, it may sometimes be approximated through statistical inference, or alternatively, the test statistic may be modified using a known sample-based function that asymptotically mimics the rate.[39] Addition-

[39] In such cases, it is referred to as self-normalized.

ally, computing the statistic requires the extraction of the estimator. The rejection region is based on χ^2_p, which-under the stated framework-is the asymptotic distribution of the test statistic under the null hypothesis when θ_0 is an interior point; *this exemplifies the characterization of the procedure as asymptotic: the decision is based on the limiting properties of the test statistic*.

If θ_0 is not an interior point, then due to the previously mentioned dominance relationship between the two distributions, using this rejection region results in an asymptotic probability of rejecting the null hypothesis that is smaller than the nominal level α.[40] The distribution of the test statistic under the null hypothesis belongs to a broader family and is related to the chi-squared distribution. Direct use of this distribution may be challenging-for instance, it may depend on the potentially latent H. Nevertheless, it is possible to construct tests with stochastic rejection regions that asymptotically approximate those derived using this distribution by employing resampling techniques-see, for example, [8].

Based on the above discussion and given that under the null hypothesis $\theta^\star = \theta_0$, i.e. when it is true, $\mathcal{L}_n \rightsquigarrow \mathbf{Z}(\theta_0)$. Furthermore, When θ_0 is an interior point, the second term in $\mathbf{Z}(\theta_0)$ vanishes. Due to the construction of the rejection region and Portmanteau Theorem, we have:

$$\lim_{n\to\infty} \mathbb{P}(\text{Reject } H_0 \mid H_0 \text{ true}) = \alpha.$$

When θ_0 is not an interior point, it can be shown (**Exercise**: Prove this using Footnote 3.) that the above limit is strictly less than α.

If the null hypothesis is false, i.e., $\theta^\star \neq \theta_0$, we have:

$$r_n^2(c_n(\theta^\star) - c_n(\theta_n)) = r_n^2(c_n(\theta_0) - c_n(\theta_n)) + r_n^2(c_n(\theta^\star) - c_n(\theta_0)).$$

[40] It is relatively straightforward to show that if F, G satisfy the first-order stochastic dominance relationship $F(x) \leq G(x)$, $\forall x \in \mathbb{R}$, then for any $\alpha \in (0,1)$, $\inf\{x \in \mathbb{R} : F(x) \geq 1-\alpha\} \leq \inf\{x \in \mathbb{R} : G(x) \geq 1-\alpha\}$.

The first term on the right-hand side of the equality converges in distribution (why?) to $\mathbf{Z}(\theta_0)$. The second term, due to the assumption of continuous convergence in probability of the criterion and the uniqueness of θ_0 as the minimizing point of the asymptotic criterion, diverges to $+\infty$.

Therefore, it can be proven that under H_1, not only the test statistic is not tight, but diverges to infinity with probability approaching one. Consequently:

$$\lim_{n\to\infty} \mathbb{P}(\text{Accept } H_0 \mid H_0 \text{ false}) = 0.$$

The above outline serves as the proof of the following theorem:

Theorem 5. *Under the assumptions of Theorem 4 and the described hypothesis testing procedure, and if* $z_{\theta_0} \sim N(\mathbf{0}_p, 4\breve{\mathbf{J}}_{\theta_0})$*:*

1. *If* θ_0 *is an interior point of* Θ*, then the asymptotic size of the test equals the nominal significance level* α*, i.e.,*

$$\lim_{n\to\infty} \mathbb{P}(\textit{Reject } H_0 \mid H_0 \textit{ true}) = \alpha.$$

2. *If* θ_0 *is not an interior point of* Θ*, the asymptotic size of the test is strictly less than* α*, i.e.,*

$$\lim_{n\to\infty} \mathbb{P}(\textit{Reject } H_0 \mid H_0 \textit{ true}) < \alpha.$$

3. *If the null hypothesis is false, the test is consistent, i.e.,*

$$\lim_{n\to\infty} \mathbb{P}(\textit{Accept } H_0 \mid H_0 \textit{ false}) = 0.$$

Remark 1. If $\mathrm{Var}(z_{\theta_0})$ is not necessarily $4\breve{\mathbf{J}}_{\theta_0}$, then it can be proven that the distribution of the quadratic form $z'_{\theta_0}\frac{1}{4}\mathbf{J}_{\theta_0}^{-1}z_{\theta_0}$ is that of the random variable $\sum_{i=1}^{p}\lambda_i W_i$, where the $W_i \sim \chi^2_1$ and are independent across $i = 1, \ldots, p$, and λ_i is the i^{th} eigenvalue of $\frac{1}{4}\mathbf{J}_{\theta_0}^{-1}\mathrm{Var}(z_{\theta_0})$. If the testing procedure is performed as designed above, i.e. using the χ^2_p quantile, and $\max_i \lambda_i \leq 1$, then it

can be easily proven that the procedure is conservative, even in the case where θ_0 is an interior point. In any case the procedure is consistent (why?). Notice that the design of a modified procedure that approximates the quantile of the distribution of $\sum_{i=1}^{p} \lambda_i W_i$, via the consistent estimation of the eigenvalues, and/or some resampling procedure is feasible; it would result in an asymptotically exact test in the interior case.

Example 11. In the context of LS Example, and for a hypothesis structure like the one described above, the test statistic is given by:

$$\mathcal{L}_n = (Y_n - \mathbf{X}_n\theta^\star)'(Y_n - \mathbf{X}_n\theta^\star) - e_n'e_n,$$

where $e_n := Y_n - \mathbf{X}_n\theta_n$. However, if $\mathrm{Var}(\varepsilon_n) \neq \mathbf{I}_n$, then the null limiting distribution of the test statistic is not necessarily the χ^2_p, since generally $\mathrm{Var}(z_{\theta_0}) \neq 4\breve{\mathbf{J}}_{\theta_0}$. In the more general case of homoskedasticity and zero correlations case, where $\mathrm{Var}(\varepsilon_n) = \sigma^2\mathbf{I}_n$ for some latent $\sigma^2 > 0$, a simple modification of $\mathcal{L}_n$ works; under the high-order assumptions for the LS model (Assumption (4) would be appropriately modified for the presence of σ^2 in the variance of z), $\frac{e_n'e_n}{n} \xrightarrow{p} \sigma^2$, and hence under the null $\mathcal{L}_n^\star := \frac{n}{e_n'e_n}\mathcal{L}_n \rightsquigarrow \chi^2_p$ due to Slutsky's Lemma-notice that $\mathcal{L}_n^\star = n(\frac{Y_n - \mathbf{X}_n\theta^\star)'(Y_n - \mathbf{X}_n\theta^\star)}{e_n'e_n} - 1)$.

Example 12. The following is the Python/Matlab code for the implementation of such a test in a simple version of the exponential/linear NLLS example (notice that under the high level assumptions adopted for this model the χ^2_p null limiting distribution for the statistic is also the case here (why?) since $\mathrm{Var}(\varepsilon_n) \neq \mathbf{I}_n$. The $\mathrm{Var}(\varepsilon_n) = \sigma^2\mathbf{I}_n$ could be handled analogously to the previous example):

```
Python Code:

import numpy as np
from scipy.optimize import minimize
from scipy.stats import chi2
import matplotlib.pyplot as plt
```

```
7
8  # Generate synthetic data for NLLS
9  def generate_nlls_data(n, beta_true, sigma, seed=42):
10     np.random.seed(seed)
11     x = np.linspace(0, 10, n)
12     y =  np.exp(beta_true[0] +beta_true[1] * x) + sigma * np.random.
       randn(n)
13     return x, y
14
15 # Define the NLLS model
16 def model(x, beta):
17     return np.exp(beta[0] +beta[1] * x)
18
19 # Sum of squared residuals (SSR) function
20 def ssr(beta, x, y):
21     return np.sum((y - model(x, beta))**2)
22
23 # Chi-squared test using SSR difference
24 def chi_squared_test_ssr(x, y, beta_null, alpha=0.05):
25     q = len(beta_null)  # Number of restrictions
26
27     # SSR under the null hypothesis
28     ssr_null = ssr(beta_null, x, y)
29
30     # SSR under the alternative hypothesis
31     res = minimize(ssr, beta_null, args=(x, y), method='BFGS')
32     ssr_alt = res.fun
33
34     # Chi-squared test statistic
35     chi_stat = ssr_null - ssr_alt
36
37     # Critical value and p-value
38     critical_value = chi2.ppf(1 - alpha, df=q)
39     p_value = 1 - chi2.cdf(chi_stat, df=q)
40
41     return chi_stat, critical_value, p_value, res.x
42
```

```
# Main script
n = 100  # Sample size
beta_true = [2.0, 0.3]  # True parameters
sigma = 0.5  # Noise standard deviation
alpha = 0.05  # Significance level

# Generate synthetic data
x, y = generate_nlls_data(n, beta_true, sigma)

# Hypothesis: Test beta_null = [1.5, 0.2]
beta_null = [1.5, 0.2]

# Perform Chi-squared test
chi_stat, critical_value, p_value, beta_est = chi_squared_test_ssr(x
    , y, beta_null, alpha)

# Output results
print(f"Chi-squared Statistic: {chi_stat:.4f}")
print(f"Critical Value (chi-squared): {critical_value:.4f}")
print(f"p-value: {p_value:.4f}")
print(f"Estimated Parameters under Alternative: {beta_est}")

# Visualization of fit
plt.figure(figsize=(8, 6))
plt.scatter(x, y, label='Data', color='blue', alpha=0.7)
plt.plot(x, model(x, beta_null), label=f'Null Hypothesis Model: beta
    ={beta_null}', color='red', linestyle='--')
plt.plot(x, model(x, beta_est), label=f'Estimated Model: beta={
    beta_est}', color='green')
plt.xlabel('x')
plt.ylabel('y')
plt.title('NLLS Model Fit and Chi-Squared Test')
plt.legend()
plt.show()
```

```
Matlab Code:

% Generate synthetic data for NLLS
```

```
function [x, y] = generate_nlls_data(n, beta_true, sigma, seed)
    rng(seed);
    x = linspace(0, 10, n)';
    y =  exp(beta_true(1)+beta_true(2) * x) + sigma * randn(n, 1);
end

% Define the NLLS model
function y_model = model(x, beta)
    y_model =  exp(beta(1)+beta(2) * x);
end

% Sum of squared residuals (SSR) function
function ssr_val = ssr(beta, x, y)
    residuals = y - model(x, beta);
    ssr_val = sum(residuals .^ 2);
end

% Chi-squared test using SSR difference
function [chi_stat, critical_value, p_value, beta_est] =
   chi_squared_test_ssr(x, y, beta_null, alpha)
    q = length(beta_null); % Number of restrictions

    % SSR under the null hypothesis
    ssr_null = ssr(beta_null, x, y);

    % SSR under the alternative hypothesis (unrestricted)
    options = optimset('Display', 'off'); % Suppress output
    beta_init = beta_null; % Use beta_null as the starting point
    [beta_est, ssr_alt] = fminsearch(@(b) ssr(b, x, y), beta_init,
   options);

    % Chi-squared statistic
    chi_stat = ssr_null - ssr_alt;

    % Critical value and p-value
    critical_value = chi2inv(1 - alpha, q);
    p_value = 1 - chi2cdf(chi_stat, q);
```

```
end

% Main Script
n = 100; % Sample size
beta_true = [2.0, 0.3]; % True parameters
sigma = 0.5; % Noise standard deviation
alpha = 0.05; % Significance level

% Generate synthetic data
[x, y] = generate_nlls_data(n, beta_true, sigma, 42);

% Hypothesis: Test beta_null = [1.5, 0.2]
beta_null = [1.5, 0.2];

% Perform Chi-squared test
[chi_stat, critical_value, p_value, beta_est] = chi_squared_test_ssr
    (x, y, beta_null, alpha);

% Output results
fprintf('Chi-squared Statistic: %.4f\n', chi_stat);
fprintf('Critical Value (chi-squared): %.4f\n', critical_value);
fprintf('p-value: %.4f\n', p_value);
fprintf('Estimated Parameters under Alternative: [%.4f, %.4f]\n',
    beta_est);

% Visualization of fit
figure;
scatter(x, y, 'b', 'DisplayName', 'Data');
hold on;
plot(x, model(x, beta_null), 'r--', 'LineWidth', 1.5, 'DisplayName',
     sprintf('Null Hypothesis Model: beta = [%0.2f, %0.2f]',
    beta_null));
plot(x, model(x, beta_est), 'g', 'LineWidth', 1.5, 'DisplayName',
    sprintf('Estimated Model: beta = [%0.2f, %0.2f]', beta_est));
xlabel('x');
ylabel('y');
title('NLLS Model Fit and Chi-Squared Test');
```

```
legend('Location', 'Best');
grid on;
hold off;
```

5.1.1 Hypothesis Testing with a bit more complicated Null

Exercise: Suppose $\Theta^\star$ is more generally a non-empty closed subset of Θ, and we aim to test the following more general hypothesis structure:

$$H_0 : \theta_0 \in \Theta^\star, \tag{8}$$

$$H_1 : \theta_0 \in \Theta - \Theta^\star. \tag{9}$$

Using arguments similar to those in the current section, show that the following testing procedure described by:

Algorithm 3 The Testing Procedure for LS and a non simple Null Hypothesis

1. Given that the convergence rate r_n is known, the test statistic is $\mathcal{L}_n := r_n^2 \min_{\theta\in\Theta^\star}(c_n(\theta) - c_n(\theta_n)) = r_n^2(\min_{\theta\in\Theta^\star} c_n(\theta) - \min_{\theta\in\Theta} c_n(\theta))$.
2. Choose a significance level $\alpha \in (0, 1)$.
3. For the given α, find:

$$q_\alpha := \inf\left\{x \in (0, +\infty) : \int_0^x \frac{z^{\frac{k}{2}-1}}{2^{k/2}\Gamma(\frac{k}{2})} \exp(-\frac{z}{2})dz \geq 1 - \alpha\right\}.$$

4. Define the rejection region for H_0 as the interval $(q_{1-\alpha}, +\infty)$.
5. Compute the test statistic from the sample and reject H_0 if and only if $\mathcal{L}_n \in (q_{1-\alpha}, +\infty)$.

is, when the conditions of Theorem 2 or 3, and Assumption 1 hold, with $z_{\theta_0} \sim N(\mathbf{0}_p, 4\breve{\mathbf{J}}_{\theta_0})$:

(a) *Asymptotically exact*, when θ_0 is an interior point of Θ.

(b) *Asymptotically conservative*, when θ_0 is a boundary point of Θ.

(c) *Consistent*.

Hint: Using the previous asymptotic framework, show that:

$$r_n^2(c_n(\theta) - \min_{\theta\in\Theta} c_n(\theta)) \rightsquigarrow \{^{\mathbf{Z}(\theta_0),\, \theta=\theta_0}_{+\infty,\, \theta\neq\theta_0}.$$

Then, conclude that under the null hypothesis:

$$r_n^2(\min_{\theta\in\Theta^\star} c_n(\theta) - \min_{\theta\in\Theta} c_n(\theta)) \rightsquigarrow \mathbf{Z}(\theta_0),$$

and similarly, under the alternative hypothesis, that:

$$r_n^2(\min_{\theta\in\Theta^\star} c_n(\theta) - \min_{\theta\in\Theta} c_n(\theta)) \rightsquigarrow +\infty.$$

5.2 Wald-type Tests with Unknown Asymptotic Variance

Consider the hypothesis structure in 7 and the assumptions described in Theorem 5. Another hypothesis testing procedure can be constructed using the limit theory of the estimator. Suppose $\theta^\star$ is an interior point. Under the null hypothesis, the quadratic form

$$r_n^2(\theta_n - \theta^\star)' \left(\frac{1}{4}\breve{\mathbf{J}}_{\theta_0}^{-1} V_{\theta_0} \breve{\mathbf{J}}_{\theta_0}^{-1}\right)^{-1} (\theta_n - \theta^\star)$$

converges in distribution to χ_p^2, and can therefore be used as a test statistic in a similar manner.

The challenge with directly using this quadratic form is the relatively common scenario where the asymptotic variance is unknown, and only its functional form is known. In such cases, having a weakly consistent estimator of the asymptotic variance would be helpful. If there exists a stochastic matrix V_n that converges in probability to the asymptotic variance $\frac{1}{4}\breve{\mathbf{J}}_{\theta_0}^{-1} V_{\theta_0} \breve{\mathbf{J}}_{\theta_0}^{-1}$, then, by Slutsky's Lemma and the Continuous Mapping

Theorem (explain!), the modified quadratic form

$$r_n^2(\theta_n - \theta^\star)' V_n^{-1}(\theta_n - \theta^\star)$$

converges in distribution under the null hypothesis to χ_p^2.

Remark 2. Notice that the assumption $z_{\theta_0} \sim N(\mathbf{0}_p, 4\breve{\mathbf{J}}_{\theta_0})$ is not needed here since the quadratic form is *constructed* so as to be asymptotically χ_p^2. It however requires the consistent estimattion of the asymptotic variance of the estimator, something that the previous procedure avoided.

[41]

> *Exersice*: Construct an asymptotic hypothesis testing procedure based on the modified quadratic form, and show that it satisfies the properties of asymptotic exactness and consistency, given that $\theta^\star$ is an interior point.

> *Exersice*: What happens to these properties if $\theta^\star$ lies on the boundary of Θ?

[41] For instance, in the context of the linear model developed so far, and given Assumptions (2)-(3) and the Continuous Mapping Theorem, the matrix $(\frac{\mathbf{X}_n'\mathbf{X}_n}{n})^{-1}$ is by construction a consistent estimator of the asymptotic variance $M_{X'X}$. Thus, the modified test statistic in this case could take the form $(\theta_n - \theta^\star)'\mathbf{X}_n'\mathbf{X}_n(\theta_n - \theta^\star)$. Generally, note that the asymptotic variance might be unknown either as a function of θ or because this function is evaluated at the unknown θ_0. Therefore, estimating the asymptotic variance can often be decomposed into: (a) the existence of an estimator for the function itself-this estimator should converge in probability to the function evaluated at θ_0 when applied to any sequence converging to θ_0, representing a version of the aforementioned continuous convergence in probability at θ_0; and (b) the existence of a consistent estimator for θ_0-such an estimator already exists in θ_n. Given (a) and (b), the estimator of the function evaluated at θ_0 would provide a consistent estimator of the asymptotic variance. In the specific case of the linear model, (b) is unnecessary, and the continuous convergence trivially follows from the assumptions since the asymptotic variance is a fixed function of θ_0.

Exersice: Building on the previous exercise but considering the hypothesis structure in 8, construct an asymptotic hypothesis test that uses as the test statistic

$$r_n^2 \min_{\theta \in \Theta^\star} (\theta_n - \theta)' V_n^{-1} (\theta_n - \theta),$$

and prove that the test satisfies the properties of asymptotic exactness and consistency, given that θ_0 is an interior point.

5.2.1 Using the Delta Method

Suppose that θ_0 is an interior point, and for a function $f : \mathbb{R}^p \to \mathbb{R}^q$, f is continuously differentiable in a neighborhood of θ_0, with $\frac{\partial f(\theta_0)}{\partial \theta'}$ being the Jacobian matrix evaluated at θ_0.[42] By the mean value theorem, and since the estimator is consistent under our assumptions, it holds with probability tending to 1 that

$$f(\theta_n) - f(\theta_0) = \frac{\partial f(\theta_n^\star)}{\partial \theta'} (\theta_n - \theta_0),$$

where $\theta_n^\star$ lies on the line segment in $\mathbb{R}^p$ joining θ_n and θ_0. Consequently, $\theta_n^\star \xrightarrow{p} \theta_0$, and thus, by the Continuous Mapping Theorem, $\frac{\partial f(\theta_n^\star)}{\partial \theta'} \xrightarrow{p} \frac{\partial f(\theta_0)}{\partial \theta'}$. Therefore, under the framework of the previous exercise, and by Slutsky's Lemma and the Continuous Mapping Theorem (explain!), the asymptotic distribution for the transformed estimator $f(\theta_n)$ is given by:

$$r_n \left(f(\theta_n) - f(\theta_0) \right) \rightsquigarrow N \left(\mathbf{0}_q, \frac{1}{4} \frac{\partial f(\theta_0)}{\partial \theta'} \breve{\mathbf{J}}_{\theta_0}^{-1} V_{\theta_0} \breve{\mathbf{J}}_{\theta_0}^{-1} \frac{\partial f'(\theta_0)}{\partial \theta} \right).$$

This result is a special case of the Delta Method, which, among other things, is useful for deriving the asymptotic theory of sufficiently smooth transformations of estimators when the transformed parameter of interest

[42] Recall that this is the $q \times p$ matrix containing all the partial derivatives of f at θ_0.

is a function of θ_0. [43] For instance, in the framework of the linear model, let $q = 1$ and f is the function that select the first component of the associated vector.[44] Then, under our assumptions, it follows that $\sqrt{n}(\theta_{n,1} - \theta_{0,1}) \rightsquigarrow N(0, M_{X'X,(1,1)})$ (why?).

Now let $\phi^\star \in \mathbb{R}^q$, and consider the hypothesis structure:[45]

$$\begin{aligned} H_0 &: f(\theta_0) = \phi^\star, \\ H_1 &: f(\theta_0) \neq \phi^\star. \end{aligned} \tag{10}$$

Assume further that f is continuously differentiable everywhere, and that $\frac{\partial f(\theta)}{\partial \theta'}$ is of rank $\min\{p, q\}$ for every $\theta \in \Theta$. Within the framework of the previous exercise, construct an asymptotic hypothesis test using as the test statistic

$$r_n^2 \left(f(\theta_n) - \phi^\star\right)' \left(\frac{\partial f(\theta_n)}{\partial \theta'} V_n \frac{\partial f'(\theta_n)}{\partial \theta}\right)^{-1} \left(f(\theta_n) - \phi^\star\right),$$

and prove that the test satisfies the properties of asymptotic exactness and consistency, given that θ_0 is an interior point.[46] How would the results change if θ_0 were on the boundary of the parameter space?

Example 13. The following is the Python/Matlab code for the implementation of the Wald test in the framework of Example 11:

```
Python Code:

import numpy as np
from scipy.optimize import minimize
from scipy.stats import chi2
```

[43] Such estimators may arise from reparameterizations of the statistical model according to f, or from covariance properties inherent to θ_n due to its definition, e.g., when f is bijective, appropriately monotonic, etc.

[44] In this case, the Jacobian is constant and equal to $(1, 0, \ldots, 0)$—why?

[45] Does this fall under the general hypothesis structure discussed in the previous chapter?

[46] For the linear model mentioned earlier, the test statistic becomes $(\theta_{n,1} - \phi^\star)^2 \sum_{i=1}^n \mathbf{X}_{n,1}^2$, which corresponds to the square of the usual t-statistic appropriately adapted to this model.

```
import matplotlib.pyplot as plt

# Generate synthetic data for NLLS
def generate_nlls_data(n, beta_true, sigma, seed=42):
    np.random.seed(seed)
    x = np.linspace(0, 10, n)
    y = np.exp(beta_true[0]+beta_true[1] * x) + sigma * np.random.randn(n)
    return x, y

# Define the NLLS model
def model(x, beta):
    return np.exp(beta[0]+beta[1] * x)

# Sum of squared residuals (SSR) function
def ssr(beta, x, y):
    return np.sum((y - model(x, beta))**2)

# Wald test function
def wald_test(x, y, beta_null, alpha=0.05):
    n = len(y)
    k = len(beta_null)  # Number of parameters
    q = len(beta_null)  # Number of restrictions

    # Estimate the parameters under the alternative hypothesis
    res = minimize(ssr, beta_null, args=(x, y), method='BFGS')
    beta_hat = res.x
    ssr_alt = res.fun

    # Estimate the covariance matrix of beta_hat (inverse Hessian approximation)
    hessian_inv = res.hess_inv
    cov_beta_hat = np.diag(hessian_inv)  # Diagonal covariance matrix approximation

    # Compute the Wald statistic
    diff = beta_hat - beta_null
```

```
    wald_stat = diff.T @ np.linalg.inv(np.diag(cov_beta_hat)) @ diff

    # Critical value and p-value
    critical_value = chi2.ppf(1 - alpha, df=q)
    p_value = 1 - chi2.cdf(wald_stat, df=q)

    return wald_stat, critical_value, p_value, beta_hat

# Main script
n = 100  # Sample size
beta_true = [2.0, 0.3]  # True parameters
sigma = 0.5  # Noise standard deviation
alpha = 0.05  # Significance level

# Generate synthetic data
x, y = generate_nlls_data(n, beta_true, sigma)

# Hypothesis: Test beta_null = [1.5, 0.2]
beta_null = [1.5, 0.2]

# Perform Wald test
wald_stat, critical_value, p_value, beta_est = wald_test(x, y,
    beta_null, alpha)

# Output results
print(f"Wald Statistic: {wald_stat:.4f}")
print(f"Critical Value (chi-squared): {critical_value:.4f}")
print(f"p-value: {p_value:.4f}")
print(f"Estimated Parameters under Alternative: {beta_est}")

# Visualization of fit
plt.figure(figsize=(8, 6))
plt.scatter(x, y, label='Data', color='blue', alpha=0.7)
plt.plot(x, model(x, beta_null), label=f'Null Hypothesis Model: beta
    = {beta_null}', color='red', linestyle='--')
plt.plot(x, model(x, beta_est), label=f'Estimated Model: beta = {
    beta_est}', color='green')
```

```
plt.xlabel('x')
plt.ylabel('y')
plt.title('NLLS Model Fit and Wald Test')
plt.legend()
plt.show()
```

```
Matlab Code:

% Generate synthetic data for NLLS
function [x, y] = generate_nlls_data(n, beta_true, sigma, seed)
    rng(seed);
    x = linspace(0, 10, n)';
    y =  exp(beta_true(1)+beta_true(2) * x) + sigma * randn(n, 1);
end

% Define the NLLS model
function y_model = model(x, beta)
    y_model = exp(beta(1)+beta(2) * x);
end

% Sum of squared residuals (SSR) function
function ssr_val = ssr(beta, x, y)
    residuals = y - model(x, beta);
    ssr_val = sum(residuals .^ 2);
end

% Wald test function
function [wald_stat, critical_value, p_value, beta_est] = wald_test(
   x, y, beta_null, alpha)
    q = length(beta_null); % Number of restrictions

    % Estimate the parameters under the alternative hypothesis
    options = optimset('Display', 'off', 'HessUpdate', 'bfgs'); %
   BFGS approximation
    beta_init = beta_null; % Use beta_null as the starting point
    [beta_est, ssr_alt, exitflag, output, hessian_inv] = fminunc(@(b
   ) ssr(b, x, y), beta_init, options);

```

```
    % Covariance matrix approximation
    cov_beta_hat = diag(hessian_inv);

    % Compute the Wald statistic
    diff = beta_est - beta_null';
    wald_stat = diff' * inv(diag(cov_beta_hat)) * diff;

    % Critical value and p-value
    critical_value = chi2inv(1 - alpha, q);
    p_value = 1 - chi2cdf(wald_stat, q);
end

% Main script
n = 100; % Sample size
beta_true = [2.0, 0.3]; % True parameters
sigma = 0.5; % Noise standard deviation
alpha = 0.05; % Significance level

% Generate synthetic data
[x, y] = generate_nlls_data(n, beta_true, sigma, 42);

% Hypothesis: Test beta_null = [1.5, 0.2]
beta_null = [1.5, 0.2];

% Perform Wald test
[wald_stat, critical_value, p_value, beta_est] = wald_test(x, y,
    beta_null, alpha);

% Output results
fprintf('Wald Statistic: %.4f\n', wald_stat);
fprintf('Critical Value (chi-squared): %.4f\n', critical_value);
fprintf('p-value: %.4f\n', p_value);
fprintf('Estimated Parameters under Alternative: [%.4f, %.4f]\n',
    beta_est);

% Visualization of fit
figure;
```

```
scatter(x, y, 'b', 'DisplayName', 'Data');
hold on;
plot(x, model(x, beta_null), 'r--', 'LineWidth', 1.5, 'DisplayName',
    sprintf('Null Hypothesis Model: beta = [%0.2f, %0.2f]',
   beta_null));
plot(x, model(x, beta_est), 'g', 'LineWidth', 1.5, 'DisplayName',
   sprintf('Estimated Model: beta = [%0.2f, %0.2f]', beta_est));
xlabel('x');
ylabel('y');
title('NLLS Model Fit and Wald Test');
legend('Location', 'Best');
grid on;
hold off;
```

6 Epilogue

The structure of statistical models can imply variational (i.e. optimization based) properties for the approximation of the unknown parameter values. These properties may be exploited for the design of statistical inference procedures related to mathematical optimization. Objective functions may incorporate "geometric" properties of the distributions that are members of the statistical model or their empirical (i.e. sample-based) approximations. More generally, it is possible to construct statistical inference procedures using functions that represent notions of distance between probability distributions. The finer properties of these procedures will generally depend on—potentially local—properties of these functions. Where these finer properties are optimized may depend on the "geometric" characteristics of the underlying probability distributions. The computational implementation of these statistical inference procedures may be non-trivial and could involve advanced operations research methods.

References

[1] Donald WK Andrews. "Estimation when a parameter is on a boundary". In: *Econometrica* 67.6 (1999), pp. 1341–1383.

[2] Stelios Arvanitis and Alexandros Louka. "A CLT for martingale transforms with infinite variance". In: *Statistics & Probability Letters* 119 (2016), pp. 116–123.

[3] Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press, 2021.

[4] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[5] Jeankyung Kim and David Pollard. "Cube root asymptotics". In: *The Annals of Statistics* (1990), pp. 191–219.

[6] Erwin Klein and Anthony C Thompson. *Theory of correspondences: including applications to mathematical economics*. Vol. 2. Wiley-Interscience, 1984.

[7] Ilya S Molchanov and Ilya S Molchanov. *Theory of random sets*. Vol. 19. 2. Springer, 2005.

[8] Dimitris N Politis, Joseph P Romano, and Michael Wolf. *Subsampling*. Springer Science & Business Media, 1999.

[9] Gabriella Salinetti and Roger JB Wets. "On the relations between two types of convergence for convex functions". In: *Journal of Mathematical Analysis and applications* 60.1 (1977), pp. 211–226.

[10] Daniel Straumann. "Maximum Likelihood Estimation in Conditionally Heteroscedastic Time Series Models". In: *Estimation in Conditionally Heteroscedastic Time Series Models* (2005), pp. 141–168.

[11] AW van der Vaart and Jon A Wellner. *Weak convergence and empirical processes with applications to statistics*. Springer Series in Statistics, 1997.

[12] Aad W Van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge university press, 2000.