Theory of statistical inference: a lazy approach to obtaining asymptotic results in parametric models

Hien D. Nguyen[1,2]

[1]DECRA Research Fellow, Australian Research Council. [2]Lecturer, Department of Mathematics and Statistics, La Trobe University, Melbourne Australia.
(Contact–Email: h.nguyen5@latrobe.edu.au, Twitter: @tresbienhien, Website: hiendn.github.io)

S4D, Caen, 2018 June 21

LA TROBE UNIVERSITY

Framework

- Suppose that we observe $\{\boldsymbol{Z}_i\}$ from some data generating process (DGP).
  - $i \in \{1, \ldots, n\}$.
- Define a function $Q_n(\boldsymbol{\theta})$ that depends on $\{\boldsymbol{Z}_i\}$.
  - $\boldsymbol{\theta} \in \Theta$, where Θ is a subset of a Euclidean space.
  - We call Q_n the **objective function** and $\boldsymbol{\theta}$ the **parameter vector**.
  - We say that Θ is the **parameter space**.

Extremum estimation

- Following the nomenclature of Amemiya (1985), we say that the vector

$$\boldsymbol{\theta}_0 \equiv \arg\max_{\boldsymbol{\theta}\in\Theta} \; Q(\boldsymbol{\theta})$$

  is the **extremum parameter** of Q, where $n^{-1}Q_n \to Q$ in some sense (to be defined).

- We call

$$\hat{\boldsymbol{\theta}}_n \equiv \arg\max_{\boldsymbol{\theta}\in\Theta} \; Q_n(\boldsymbol{\theta})$$

  the **extremum estimator** of $\boldsymbol{\theta}_0$.

A rose by any other name...

- We call the process of obtaining the extremum estimator: **extremum estimation**.
- Extremum estimation has appeared in the literature under numerous names:
    - Empirical risk minimization (Vapnik, 1998, 2000).
    - M-estimation (Huber, 1964; Serfling, 1980).
    - Minimum contrast estimation (Pfanzagl, 1969; Bickel and Docksum, 2000).

Some specific cases

- Important cases include:
  - Generalized method of moments.
  - Loss function minimization (e.g. fitting support vector machines, neural networks, etc.).
  - Maximum likelihood estimation (including empirical-, partial-, penalized-, pseudo-, quasi-, restricted-, etc).
  - Maximum *a posteriori* estimation.
  - Minimum distance estimation (e.g. least-squares, least-absolute deviation, etc).

Statistical inference

- Since $\boldsymbol{\theta}_0$ is defined as the maximum of Q, it must contain some information regarding the DGP of $\{\boldsymbol{Z}_i\}$.

1. We hope that given Q_n, $\hat{\boldsymbol{\theta}}_n$ will provide us with the same information regarding Q, provided that n is large enough.
2. We also hope that $\hat{\boldsymbol{\theta}}_n$ also has some DGP that is dependent on $\boldsymbol{\theta}_0$, which allows us to assess *a priori* hypotheses regarding $\boldsymbol{\theta}$.

Ordinary least squares (1A)

- Suppose that we observe independent and identically distributed (IID) data pairs $\boldsymbol{Z}_i = (\boldsymbol{X}_i, Y_i)$, where

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\theta}_* + E_i,$$

where $\mathbb{E}(E_i) = 0$, and that the DGP of $\boldsymbol{Z}_i$ is in some sense, well-behaved.
  - $\boldsymbol{\theta}_* \in \Theta \subset \mathbb{R}^p$ and $\boldsymbol{X}_i \in \mathbb{X} \subset \mathbb{R}^p$, $p \in \mathbb{N}$, and $\{E_i\}$ is independent of $\{\boldsymbol{X}_i\}$.
- Define the (negative) sum-of-squares as

$$Q_n(\boldsymbol{\theta}) = -\frac{1}{2}\sum_{i=1}^{n}\left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\theta}\right)^2.$$

- The least-squares estimator is defined as

$$\hat{\boldsymbol{\theta}}_n \equiv \arg\max_{\boldsymbol{\theta}\in\Theta} \; -\frac{1}{2}\sum_{i=1}^{n}\left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\theta}\right)^2.$$

Ordinary least squares (1B)

- We can obtain $\hat{\boldsymbol{\theta}}_n$ by solving the first-order condition (FOC)

$$\nabla Q_n = \sum_{i=1}^{n} \boldsymbol{X}_i \left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\theta} \right) = \mathbf{0}$$

$$\implies \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^\top \boldsymbol{\theta} = \sum_{i=1}^{n} \boldsymbol{X}_i Y_i$$

$$\implies \hat{\boldsymbol{\theta}}_n = \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^\top \right)^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i Y_i .$$

- More familiarly, if we put $\boldsymbol{X}_i^\top$ into the ith row of $\mathbf{X}_n \in \mathbb{R}^{n \times p}$ and put Y_i into the ith position of $\mathbf{y}_n \in \mathbb{R}^n$, then we can write

$$\hat{\boldsymbol{\theta}}_n = \left(\mathbf{X}_n^\top \mathbf{X}_n \right)^{-1} \mathbf{X}_n^\top \mathbf{y}_n .$$

Ordinary least squares (1C)

- Since $\hat{\boldsymbol{\theta}}_n$ is an estimate of $\boldsymbol{\theta}_0$, we must determine if there is a sensible relationship between Q_n and $\boldsymbol{\theta}_0$.
- The following is a **heuristic** argument. Note that $\xrightarrow{\text{p}}$ denotes **convergence in probability**.

1. Notice that $n^{-1}Q_n = n^{-1}\sum_{i=1}^{n} g(\boldsymbol{Z}_i)$, for some

$$g(\boldsymbol{Z}_i) = -\frac{1}{2}\left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\theta}\right)^2.$$

2. Since $\boldsymbol{Z}_i$ is well-behaved, then a weak law of large numbers implies that

$$\begin{aligned} n^{-1}Q_n \xrightarrow{\text{p}} \mathbb{E}[g(\boldsymbol{Z}_i)] &= -\frac{1}{2}\mathbb{E}\left[\left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\theta}\right)^2\right] \\ &\equiv Q \end{aligned}$$

Ordinary least squares (1D)

3. Suppose that we can exchange integration and differentiation, then the FOC implies that

$$\begin{aligned} \nabla Q &= \mathbb{E}\left[\boldsymbol{X}_i\left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\theta}\right)\right] \\ &= \mathbb{E}\left[\boldsymbol{X}_i\left(\boldsymbol{X}_i^\top \boldsymbol{\theta}_* + E_i - \boldsymbol{X}_i^\top \boldsymbol{\theta}\right)\right] \\ &= \mathbb{E}\left(\boldsymbol{X}_i\boldsymbol{X}_i^\top\right)\boldsymbol{\theta}_* + \mathbb{E}(\boldsymbol{X}_i E_i) - \mathbb{E}\left(\boldsymbol{X}_i\boldsymbol{X}_i^\top\right)\boldsymbol{\theta} \end{aligned}$$

4. Under the assumption that $\mathbb{E}(\boldsymbol{X}_i E_i) = \mathbf{0}$ (e.g. independence between $\{\boldsymbol{X}_i\}$ and $\{E_i\}$), we have

$$\begin{aligned} \mathbf{0} &= \mathbb{E}\left(\boldsymbol{X}_i\boldsymbol{X}_i^\top\right)\boldsymbol{\theta}_* - \mathbb{E}\left(\boldsymbol{X}_i\boldsymbol{X}_i^\top\right)\boldsymbol{\theta} \\ \Longrightarrow \boldsymbol{\theta}_0 &= \arg\max_{\boldsymbol{\theta}\in\Theta} Q = \boldsymbol{\theta}_* \end{aligned}$$

- Thus, in this case, we have found that $\boldsymbol{\theta}_0$ is the generative parameter $\boldsymbol{\theta}_*$!

Consistency

- We must now make precise the notion regarding how $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_0$ are related.
- Earlier, we defined $\xrightarrow{\text{p}}$ to denote convergence in probability. We say that a random variable $\boldsymbol{U}_n$ **converges in probability** to another random variable $\boldsymbol{U}$, if for every $\varepsilon > 0$, we have

$$\lim_{n\to\infty} \mathbb{P}(\|\boldsymbol{U}_n - \boldsymbol{U}\| > \varepsilon) = 0,$$

  where $\|\cdot\|$ is some appropriate norm (usually Euclidean, in our case).
- We say that $\hat{\boldsymbol{\theta}}_n$ is a **consistent** estimator of $\boldsymbol{\theta}_0$, if $\hat{\boldsymbol{\theta}}_n \xrightarrow{\text{p}} \boldsymbol{\theta}_0$.

Proving consistency (1)

- We present the consistency result of Amemiya (1985, Thm. 4.1.1). See also van der Vaart (1998, Thm. 5.7).

Make the following assumptions:

(A) The parameter space Θ is a compact subset of a Euclidean space $\mathbb{R}^p$ ($p \in \mathbb{N}$).

(B) $Q_n(\boldsymbol{\theta})$ is a continuous function in $\boldsymbol{\theta}$ for all $\{\boldsymbol{Z}_i\}$, and measurable in $\{\boldsymbol{Z}_i\}$ for all $\boldsymbol{\theta}$.

(C) $n^{-1}Q_n(\boldsymbol{\theta})$ converges to a non-stochastic function $Q(\boldsymbol{\theta})$ in probability uniformly in $\boldsymbol{\theta}$ over Θ.

(D) $Q(\boldsymbol{\theta})$ obtains a unique global maximum at $\boldsymbol{\theta}_0$.

Proving consistency (2)

Under Assumptions (A)–(D), then the EE, defined as

$$\hat{\boldsymbol{\theta}}_n \equiv \arg\max_{\boldsymbol{\theta}\in\Theta} \; Q_n(\boldsymbol{\theta}),$$

is consistent, in the sense that $\hat{\boldsymbol{\theta}}_n \xrightarrow{\text{p}} \boldsymbol{\theta}_0$.

- Here, we say that $n^{-1}Q_n(\boldsymbol{\theta})$ **converges in probability uniformly** to $Q(\boldsymbol{\theta})$, if for any $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{\boldsymbol{\theta}\in\Theta} \left|n^{-1}Q_n(\boldsymbol{\theta}) - Q(\boldsymbol{\theta})\right| > \varepsilon\right) = 0.$$

Uniform weak law of large numbers

- The most difficult part, in general, of applying Amemiya (1985, Thm. 4.1.1) is checking assumption (C).
- The main traditional tool that we will apply is the **weak uniform law of large numbers** of Jennrich (1969) (see also Amemiya, 1985, Thm. 4.2.1):

Let $Q_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} g(\boldsymbol{Z}_i; \boldsymbol{\theta})$ *be a measurable function of the IID sequence* $\{\boldsymbol{Z}_i\}$*, where* $\boldsymbol{Z}_i$ *is supported in a Euclidean space, for each* $\boldsymbol{\theta} \in \Theta$*, where* Θ *is compact and Euclidean. If* $\mathbb{E}[g(\boldsymbol{Z}_i; \boldsymbol{\theta})]$ *exists, and* $\mathbb{E}[\sup_{\boldsymbol{\theta} \in \Theta} g(\boldsymbol{Z}_i; \boldsymbol{\theta})] < \infty$*, then* $n^{-1} Q_n(\boldsymbol{\theta})$ *converges in probability uniformly to* $Q(\boldsymbol{\theta}) = \mathbb{E}[g(\boldsymbol{Z}_i; \boldsymbol{\theta})]$*.*

Ordinary least squares (2A)

- Make the following assumptions:

(a) $\{\boldsymbol{Z}_i\}$ is and IID sequence and that the DGP of $\boldsymbol{Z}_i = (\boldsymbol{X}_i, Y_i)$ is such that $\mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\top\right)$ exists and is positive definite, $\mathbb{E}(E_i) = 0$, $\mathbb{E}\left(E_i^2\right) = \sigma^2 < \infty$, and $\mathbb{E}(\boldsymbol{X}_i E_i) = \mathbf{0}$, where

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\theta}_* + E_i.$$

(b) The parameter space is $\Theta = [-L, L]^p$, where L is sufficiently large.

Ordinary least squares (2B)

- By (b), Θ is a compact Euclidean space, thus (A) is validated.
- We can write $Q_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} g(\boldsymbol{Z}_i; \boldsymbol{\theta})$, where

$$-2g = \left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\theta}\right)^2 = Y_i^2 + Y_i \boldsymbol{X}_i^\top \boldsymbol{\theta} - \boldsymbol{\theta}^\top \boldsymbol{X}_i \boldsymbol{X}_i^\top \boldsymbol{\theta}$$

and

$$\begin{aligned} \mathbb{E}\left[\left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\theta}\right)^2\right] &= \mathbb{E}\left(Y_i^2\right) - 2\mathbb{E}\left(Y_i \boldsymbol{X}_i^\top\right)\boldsymbol{\theta} \\ &\quad + \boldsymbol{\theta}^\top \mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\top\right)\boldsymbol{\theta}. \end{aligned}$$

Ordinary least squares (2C)

- Continuing from the previous slide, and applying (a), we have:

$$\begin{aligned}
\mathbb{E}\left[\left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\theta}\right)^2\right] &= \boldsymbol{\theta}_*^\top \mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\top\right) \boldsymbol{\theta}_* + 2\mathbb{E}\left(E_i \boldsymbol{X}_i^\top\right) \boldsymbol{\theta}_* \\
&\quad + \mathbb{E}\left(E_i^2\right) - 2\boldsymbol{\theta}^\top \mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\top\right) \boldsymbol{\theta}_* \\
&\quad - 2\mathbb{E}\left(E_i \boldsymbol{X}_i^\top\right) \boldsymbol{\theta} + \boldsymbol{\theta}^\top \mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\top\right) \boldsymbol{\theta} \\
&= \boldsymbol{\theta}_*^\top \mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\top\right) \boldsymbol{\theta}_* - 2\boldsymbol{\theta}^\top \mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\top\right) \boldsymbol{\theta}_* \\
&\quad + \boldsymbol{\theta}^\top \mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\top\right) \boldsymbol{\theta} + \sigma^2.
\end{aligned}$$

- Since $\mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\top\right)$ exists, Q_n is measurable, and g is quadratic in $\boldsymbol{\theta}$, thus it is continuous and we have the validation of (B).

Ordinary least squares (2D)

- Write $Q_n = \sum_{i=1}^{n} g\left(\boldsymbol{Z}_i; \boldsymbol{\theta}\right)$, where

$$g_i\left(\boldsymbol{Z}_i; \boldsymbol{\theta}\right) = -\frac{1}{2}\left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\theta}\right)^2.$$

- From the previous slide, we have the fact that

$$\begin{aligned}\mathbb{E}\left[g\left(\boldsymbol{Z}_i; \boldsymbol{\theta}\right)\right] &= -\frac{1}{2}\boldsymbol{\theta}_*^\top \mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\top\right)\boldsymbol{\theta}_* + \boldsymbol{\theta}^\top \mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\top\right)\boldsymbol{\theta}_* \\ &\quad -\frac{1}{2}\boldsymbol{\theta}^\top \mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\top\right)\boldsymbol{\theta} - \sigma^2.\end{aligned}$$

- By (b), Θ is compact, and we have established that g is continuous. Thus, via the Weierstrass extreme value theorem,

$$\mathbb{E}\left[\sup_{\boldsymbol{\theta} \in \Theta} g\left(\boldsymbol{Z}_i; \boldsymbol{\theta}\right)\right] \leq M < \infty.$$

Ordinary least squares (2E)

- Via the theorem of Jennrich (1969), we have the conclusion that $n^{-1}Q_n$ converges in probability uniformly to $\mathbb{E}[g(\boldsymbol{Z}_i;\boldsymbol{\theta})]$.
- Finally, we observe that $\mathbb{E}[g(\boldsymbol{Z}_i;\boldsymbol{\theta})]$ is a concave quadratic in $\boldsymbol{\theta}$ since $\mathbb{E}\left(\boldsymbol{X}_i\boldsymbol{X}_i^\top\right)$ is positive definite (it may be linear otherwise), so $\mathbb{E}[g(\boldsymbol{Z}_i;\boldsymbol{\theta})]$ has a unique global maximum and thus (D) is validated.
    - The global maximum is $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_*$.
- We have validated (A)–(D), and thus can conclude that $\hat{\boldsymbol{\theta}}_n$ is a consistent estimator for $\boldsymbol{\theta}_0$.

Asymptotic normality

- We would now like to establish, in a more precise manner, how $\hat{\boldsymbol{\theta}}_n$ fluctuates around $\boldsymbol{\theta}_0$ as it converges.
- In most cases, $n^{1/2}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \xrightarrow{\mathrm{d}} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$.
  - We write $\xrightarrow{\mathrm{d}}$ to denote **convergence in distribution**.
  - We write $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
- Convergence in distribution can be characterized in numerous ways (cf. the famous **Portmanteau Lemma**; see, e.g. van der Vaart, 1998, Lem. 2.2).
- By the **Levy continuity Theorem** states that $\boldsymbol{U}_n$ converges to the distribution of $\boldsymbol{U}$ if and only if the characteristic function of $\boldsymbol{U}_n$ converges point-wise to that of $\boldsymbol{U}$ (cf. van der Vaart, 1998, Thm. 2.13).

Proving asymptotic normality (1)

- We now present the asymptotic normality result of Amemiya (1985, Thm. 4.1.6).
- Make the following assumptions:

(A1) The parameter $\boldsymbol{\theta}_0$ is in the interior (an open subset) of the Euclidean parameter space Θ.

(B1) The objective $Q_n(\boldsymbol{\theta})$ is continuous and measurable with respect to $\{\boldsymbol{Z}_i\}$, for all $\boldsymbol{\theta} \in \Theta$, and the partial derivative $(\nabla Q_n)(\boldsymbol{\theta})$ exists and is continuous in an open neighborhood N_1 of $\boldsymbol{\theta}_0$.

(C1) There exists an open neighborhood N_2 of $\boldsymbol{\theta}_0$, where $n^{-1}Q_n(\boldsymbol{\theta})$ converges in probability uniformly to a non-stochastic function $Q(\boldsymbol{\theta})$ in N_2, and $Q(\boldsymbol{\theta})$ attains a strict local maximum at $\boldsymbol{\theta}_0$.

Proving asymptotic normality (2)

- Make the further assumptions:

(A2) The Hessian matrix $(\mathbf{H}Q_n)(\boldsymbol{\theta}) \equiv \partial^2 Q_n / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top$ exists and is continuous in an open and convex neighborhood of $\boldsymbol{\theta}_0$.

(B2) For any sequence $\boldsymbol{\theta}_n$, such that $\boldsymbol{\theta}_n \xrightarrow{\text{p}} \boldsymbol{\theta}_0$, $n^{-1}(\mathbf{H}Q_n)(\boldsymbol{\theta}_n)$ converges in probability to

$$\mathbf{A}(\boldsymbol{\theta}_0) \equiv \lim_{n \to \infty} \mathbb{E}\left[n^{-1}(\mathbf{H}Q_n)(\boldsymbol{\theta}_0)\right].$$

(C2) $n^{-1/2}(\nabla Q_n)(\boldsymbol{\theta}_0) \xrightarrow{\text{d}} \mathrm{N}(\mathbf{0}, \mathbf{B}(\boldsymbol{\theta}_0))$, where

$$\mathbf{B}(\boldsymbol{\theta}_0) \equiv \lim_{n \to \infty} \mathbb{E}\left[n^{-1}(\nabla Q_n)(\boldsymbol{\theta}_0)(\nabla Q_n)^\top(\boldsymbol{\theta}_0)\right].$$

Proving asymptotic normality (3)

- Define $\bar{\Theta}_n$ to be the set

$$\bar{\Theta}_n = \{\boldsymbol{\theta}_n : (\nabla Q_n)(\boldsymbol{\theta}_n) = \mathbf{0}\}.$$

Under Assumptions (A1)–(C1) and (A2)–(C2), if $\hat{\boldsymbol{\theta}}_n$ is a sequence of local maximizers taking values in Θ_n, such that $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$, then

$$n^{1/2}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\big(\mathbf{0}, \mathbf{A}^{-1}(\boldsymbol{\theta}_0)\,\mathbf{B}(\boldsymbol{\theta}_0)\,\mathbf{A}^{-1}(\boldsymbol{\theta}_0)\big).$$

Ordinary least squares (3A)

- Make the following assumptions.

(a) $\{\boldsymbol{Z}_i\}$ is and IID sequence and that the DGP of $\boldsymbol{Z}_i = (\boldsymbol{X}_i, Y_i)$ is such that $\mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\top\right)$ exists and is positive definite, $\mathbb{E}(E_i) = 0$, $\mathbb{E}\left(E_i^2\right) = \sigma^2 < \infty$, and $\mathbb{E}(\boldsymbol{X}_i E_i) = \mathbf{0}$, where

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\theta}_* + E_i.$$

(b*) The parameter space is $\Theta = [-L, L]^p$, where L is sufficiently large, and $\boldsymbol{\theta}_0$ is in the interior of Θ.

Under (a) and (b*), we have the fulfillment of Assumptions (A1)–(C1).

Ordinary least squares (3B)

- Recall that

$$\begin{aligned} \nabla Q_n &= \sum_{i=1}^{n} \boldsymbol{X}_i \left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\theta} \right) \\ &= \sum_{i=1}^{n} \boldsymbol{X}_i Y_i - \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^\top \boldsymbol{\theta} \\ \implies (\mathbf{H} Q_n)(\boldsymbol{\theta}) &= - \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^\top . \end{aligned}$$

- Thus, we observe that $(\mathbf{H} Q_n)(\boldsymbol{\theta})$ is constant for any $\boldsymbol{\theta}$ and is thus continuous, which fulfills (A2).

Ordinary least squares (3C)

- At $\boldsymbol{\theta}_0$, we have

$$(\nabla g)(\nabla g)^\top = \boldsymbol{X}_i \left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\theta}_0 \right) \left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\theta}_0 \right)^\top \boldsymbol{X}_i^\top$$

- Recalling that $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_*$, the parentheses equate to

$$\begin{aligned} Y_i - \boldsymbol{X}_i^\top \boldsymbol{\theta}_0 &= \boldsymbol{X}_i^\top \boldsymbol{\theta}_* - \boldsymbol{X}_i^\top \boldsymbol{\theta}_0 + E_i \\ &= \boldsymbol{X}_i^\top \boldsymbol{\theta}_0 - \boldsymbol{X}_i^\top \boldsymbol{\theta}_0 + E_i \\ &= E_i. \end{aligned}$$

- Therefore, we have $(\nabla g)(\nabla g)^\top = E_i^2 \boldsymbol{X}_i \boldsymbol{X}_i^\top$ and therefore, the expectation is

$$\begin{aligned} \mathbb{E}\left[(\nabla g)(\nabla g)^\top \right] &= \mathbb{E}\left(E_i^2 \boldsymbol{X}_i \boldsymbol{X}_i^\top \right) \\ &= \mathbb{E}\left(E_i^2\right) \mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\top \right) = \sigma^2 \mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\top \right). \end{aligned}$$

Ordinary least squares (3D)

- By Assumption (a), $\{\boldsymbol{Z}_i\}$ is IID, and by definition of $\boldsymbol{\theta}_0$, we have
$$\mathbb{E}\left(\frac{1}{n}\nabla Q_n\right) = \mathbb{E}\left[\nabla g\left(\boldsymbol{Z};\boldsymbol{\theta}_0\right)\right] = \mathbf{0}.$$
- Again, since $\{\boldsymbol{Z}_i\}$ is IID, we have
$$\begin{aligned}
\operatorname{cov}\left(n^{-1}\nabla Q_n\right) &= \mathbb{E}\left[\left(n^{-1}\nabla Q_n\right)\left(n^{-1}\nabla Q_n\right)^\top\right] \\
&= \mathbb{E}\left[\left(n^{-1}\sum_{i=1}^{n}\nabla g\right)\left(n^{-1}\sum_{i=1}^{n}\nabla g\right)^\top\right] \\
&= \mathbb{E}\left[\left(\nabla g\right)\left(\nabla g\right)^\top\right],
\end{aligned}$$

which exists!

Ordinary least squares (3E)

- We now need to establish the fact that
$$n^{-1/2}\nabla Q_n = n^{-1/2}\sum_{i=1}^{n} g\left(\boldsymbol{Z}_i;\boldsymbol{\theta}_0\right)$$
converges in distribution to $\mathrm{N}\left(\mathbf{0},\sigma^2\mathbb{E}\left[\boldsymbol{X}_i\boldsymbol{X}_i^\top\right]\right)$.
- The multivariate Lindeberg-Lévy **central limit theorem** (CLT; van der Vaart, 1998, Thm. 2.18) states that if $\{\boldsymbol{U}_i\}$ is an IID sequence that has finite mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, then
$$n^{1/2}\left(n^{-1}\sum_{i=1}^{n}\boldsymbol{U}_i-\boldsymbol{\mu}\right)\xrightarrow{\mathrm{d}}\mathrm{N}\left(\mathbf{0},\boldsymbol{\Sigma}\right).$$
- Since $n^{-1/2}\sum_{i=1}^{n} g\left(\boldsymbol{Z}_i;\boldsymbol{\theta}_0\right) = n^{1/2}\left(n^{-1}\sum_{i=1}^{n} g\left(\boldsymbol{Z}_i;\boldsymbol{\theta}_0\right)-\mathbf{0}\right)$, we have the desired result, and (C2) is validated with $\mathbf{B}\left(\boldsymbol{\theta}_0\right)=\sigma^2\mathbb{E}\left(\boldsymbol{X}_i\boldsymbol{X}_i^\top\right)$.

Ordinary least squares (3F)

- Lastly,

$$n^{-1}\left(\mathbf{H}Q_n\right)\left(\boldsymbol{\theta}_n\right)=n^{-1}\left(-\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^\top\right).$$

- By independence, we have $\mathbb{E}\left[n^{-1}\left(\mathbf{H}Q_n\right)\left(\boldsymbol{\theta}_0\right)\right]=\mathbb{E}\left(\boldsymbol{X}_i\boldsymbol{X}_i^\top\right)$, and via the weak law of large numbers, we have

$$n^{-1}\left(\mathbf{H}Q_n\right)\left(\boldsymbol{\theta}_n\right)\xrightarrow{\mathrm{p}}\mathbf{A}\left(\boldsymbol{\theta}_0\right),$$

  where

$$\mathbf{A}\left(\boldsymbol{\theta}_0\right)=-\mathbb{E}\left(\boldsymbol{X}_i\boldsymbol{X}_i^\top\right).$$

- Thus, (B2) is validated.

Ordinary least squares (3G)

- Finally, compute the matrix:

$$\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1} = \left[\mathbb{E}\left(\boldsymbol{X}_i\boldsymbol{X}_i^\top\right)\right]^{-1}\left[\sigma^2\mathbb{E}\left(\boldsymbol{X}_i\boldsymbol{X}_i^\top\right)\right]\left[\mathbb{E}\left(\boldsymbol{X}_i\boldsymbol{X}_i^\top\right)\right]^{-1}$$
$$= \sigma^2\left[\mathbb{E}\left(\boldsymbol{X}_i\boldsymbol{X}_i^\top\right)\right]^{-1}.$$

Under Assumptions (a) and (b), the ordinary least squares estimator is asymptotically normal, in the sense that*

$$n^{1/2}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(\mathbf{0}, \sigma^2\left[\mathbb{E}\left(\boldsymbol{X}_i\boldsymbol{X}_i^\top\right)\right]^{-1}\right).$$

A bonus prize

- Under Assumptions (A1)–(C1) Amemiya (1985, Thm. 4.1.2) states the **Wald-consistency** result (cf. Wald, 1949). See also van der Vaart (1998, Thm. 5.14).

If (A1)–(C1) hold, and $\left\{\hat{\boldsymbol{\theta}}_n\right\}$ *is a sequence of local maximizers that take values in* $\bar{\Theta}_n = \{\boldsymbol{\theta}_n : (\nabla Q_n)(\boldsymbol{\theta}_n) = \mathbf{0}\}$*, then for any* $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\left(\inf_{\boldsymbol{\theta}_n \in \bar{\Theta}_n} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_0\| > \varepsilon\right) = 0.$$

- We read this as "there exists a consistent sequence of locally maximal roots $\hat{\boldsymbol{\theta}}_n$, taking values in $\bar{\Theta}_n$".

Mixture of normal distributions (1)

- We say that the IID random sequence $\{Z_i\}$ arises from an m-component mixture of normal distributions, if it has a DGP characterized by the PDF

$$f\left(z_i; \boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{\sigma}\right) = \sum_{j=1}^{m} \pi_j \phi\left(z_i; \mu_i, \sigma_i^2\right),$$

where $\boldsymbol{\mu} \in [-L, L]^m$, $\boldsymbol{\sigma} \in \left[S^{-1}, S\right]^m$, and

$$\boldsymbol{\pi} \in \mathbb{S}_{m-1} = \left\{(\pi_1, \dots, \pi_m) : \pi_j \geq 0, \sum_{j=1}^{m} \pi_j = 1\right\},$$

for large L and $S > 1$.

- We write $\boldsymbol{\theta} \in \Theta$ as the concatenation of $\boldsymbol{\mu}$, $\boldsymbol{\pi}$, and $\boldsymbol{\sigma}$.

Mixture of normal distributions (2)

- Upon observing $\{Z_i\}$, we would wish to estimate the parameter vector $\boldsymbol{\theta}$ via maximization of the log-likelihood function

$$Q_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left[\sum_{j=1}^{m} \pi_j \phi\left(z_i; \mu_i, \sigma_i^2\right) \right].$$

- Unfortunately, it is well-known that Q_n has multiple global maxima, due to lack of identifiability (cf. Titterington et al., 1985, Sec. 3.1)!
- For example, consider that

$$\pi_1 \phi\left(z_i; \mu_1, \sigma_1^2\right) + \pi_2 \phi\left(z_i; \mu_2, \sigma_2^2\right)$$

is the same as

$$\pi_2 \phi\left(z_i; \mu_2, \sigma_2^2\right) + \pi_1 \phi\left(z_i; \mu_1, \sigma_1^2\right).$$

Mixture of normal distributions (3)

- Since Q_n does not have a unique global maximum, we can't apply Amemiya (1985, Thm. 4.1.1).
- We can use the Wald consistency theorem by checking:

(A1) The parameter $\boldsymbol{\theta}_0$ is in the interior (an open subset) of the Euclidean parameter space Θ.

(B1) The objective $Q_n(\boldsymbol{\theta})$ is continuous and measurable with respect to $\{\boldsymbol{Z}_i\}$, for all $\boldsymbol{\theta} \in \Theta$, and the partial derivative $(\nabla Q_n)(\boldsymbol{\theta})$ exists and is continuous in an open neighborhood N_1 of $\boldsymbol{\theta}_0$.

(C1) There exists an open neighborhood N_2 of $\boldsymbol{\theta}_0$, where $n^{-1}Q_n(\boldsymbol{\theta})$ converges in probability uniformly to a non-stochastic function $Q(\boldsymbol{\theta})$ in N_2, and $Q(\boldsymbol{\theta})$ attains a strict local maximum at $\boldsymbol{\theta}_0$.

Mixture of normal distributions (4)

- Clearly, $\Theta = [-L, L]^m \times \left[S^{-1}, S\right]^m \times \mathbb{S}_{m-1}$ is Euclidean. We thus must simply make the assumption that (a1) $\boldsymbol{\theta}_0$ is in the interior of Θ. This validates (A1).
- Since the normal PDF is continuous, Q_n is continuous (since it is a convex combination of normal PDFs).
- We now need to validate the measurability of Q_n by showing that
$$\mathbb{E}\left[\log \sum_{j=1}^{m} \pi_j \phi\left(Z_i; \mu_j, \sigma_j^2\right)\right] < \infty.$$

Mixture of normal distributions (5)

- Luckily, by Atienza et al. (2007), we have

$$\left|\log\sum_{j=1}^{m}\pi_j\phi\left(z_i;\mu_j,\sigma_j^2\right)\right| \le \sum_{j=1}^{m}\left|\log\phi\left(z_i;\mu_j,\sigma_j^2\right)\right|.$$

- We can write

$$\begin{aligned}\log\phi\left(z_i;\mu_i,\sigma_i^2\right) &= -\frac{1}{2}\log(2\pi)-\frac{1}{2}\log\sigma_i^2\\ &\quad -\frac{1}{2\sigma_i^2}\left(z_i-\mu_i\right)^2\end{aligned}$$

which is quadratic in z_i!

- So $\mathbb{E}\log\phi\left(z_i;\mu_i,\sigma_i^2\right)$ exists, since normal random variables have second moments. Thus, we have the measurability of Q_n.

Mixture of normal distributions (6)

- Since the PDF f is smooth in all parameter components $\boldsymbol{\theta}$, we also have the existence of a continuous ∇Q_n, and thus (B1).
- Now recall that we have already proved that

$$\mathbb{E}\left[\log \sum_{j=1}^{m} \pi_j \phi\left(Z_i; \mu_j, \sigma_j^2\right)\right] < \infty.$$

- Since $\{Z_i\}$ is IID and Θ is compact, we can directly apply the weak uniform law of large numbers to obtain the convergence of $n^{-1} Q_n$ to $\mathbb{E}\left[\log \sum_{j=1}^{m} \pi_j \phi\left(Z_i; \mu_j, \sigma_j^2\right)\right]$, uniformly in probability. We therefore have (C1) if we also assume that $\hat{\boldsymbol{\theta}}_n$ is a sequence from $\bar{\Theta}_n$.

Mixture of normal distributions (7)

Assume that $\boldsymbol{\theta}_0$ is a locally maximal root of $\mathbb{E}\left[\log \sum_{j=1}^{m} \pi_j \phi\left(Z_i; \mu_j, \sigma_j^2\right)\right]$, and that $\hat{\boldsymbol{\theta}}_n$ is a sequence of locally maximal roots from the set

$$\bar{\Theta}_n = \left\{\boldsymbol{\theta}_n : (\nabla Q_n)(\boldsymbol{\theta}_n) = \mathbf{0}\right\}.$$

If $\{Z_i\}$ is an IID sequence from a model with density $f\left(z_i; \boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{\sigma}\right)$, then for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\inf_{\boldsymbol{\theta}_n \in \bar{\Theta}_n} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_0\| > \varepsilon\right) = 0.$$

- An interpretation of the result is that: if you enumerated all of the local maxima of Q_n at each n, then one of the sequences of local maxima will converge to the parameter vector $\boldsymbol{\theta}_0$, in probability.

A modern problem

- Consider the LASSO problem of Tibshirani (1996) (see also Hastie et al., 2015), where we maximize the negative regularized sum-of-squares:

$$Q_n(\boldsymbol{\theta}) = -\frac{1}{2}\sum_{i=1}^{n}\left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\theta}\right)^2 - n\lambda \sum_{j=1}^{p} |\theta_j|,$$

  where $\boldsymbol{\theta} \in \Theta = [-L, L]^p$ for large L, $\lambda > 0$, and $\{\boldsymbol{Z}_i\}$ is an IID sequence with $\boldsymbol{Z}_i = (\boldsymbol{X}_i, Y_i)$.
- Here

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\theta}_\mathsf{S} + E_i,$$

  where $\mathbb{E}(E_i) = 0$, $\mathbb{E}\left(E_i^2\right) = \sigma^2 < \infty$, and $\mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\top\right)$ exists and is positive definite.
- We say that $\boldsymbol{\theta}$ is q-sparse ($q \in \mathbb{N}$, $q < p$) in the sense that

$$\boldsymbol{\theta}_\mathsf{S} = (\theta_1, \theta_2, \dots, \theta_q, 0, \dots, 0).$$

A consistency result? (1)

- We can check the following assumptions to prove consistency via the result of Amemiya (1985, Thm. 4.1.1):

(A) The parameter space Θ is a compact subset of a Euclidean space $\mathbb{R}^p$ ($p \in \mathbb{N}$).

(B) $Q_n(\boldsymbol{\theta})$ is a continuous function in $\boldsymbol{\theta}$ for all $\{\boldsymbol{Z}_i\}$, and measurable in $\{\boldsymbol{Z}_i\}$ for all $\boldsymbol{\theta}$.

(C) $n^{-1}Q_n(\boldsymbol{\theta})$ converges to a non-stochastic function $Q(\boldsymbol{\theta})$ in probability uniformly in $\boldsymbol{\theta}$ over Θ.

(D) $Q(\boldsymbol{\theta})$ obtains a unique global maximum at $\boldsymbol{\theta}_0$.

A consistency result? (2)

- Clearly, (A) is validated since $\Theta = [-L, L]^p$.
- Both the quadratic and absolute value functions are continuous and thus Q_n is continuous.
- Write
$$g(\boldsymbol{Z}_i; \boldsymbol{\theta}) = -\frac{1}{2}\left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\theta}\right)^2 - \lambda \sum_{j=1}^{p} |\theta_j| .$$
- By the same argument as for the ordinary least squares, the first part is measurable. The second part is a constant, and is therefore also measurable. (B) is therefore validated.

A consistency result? (3)

- Again, we know that $\mathbb{E}\left[\left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\theta}\right)^2\right]$ exists, and since $\lambda \sum_{j=1}^p |\theta_j|$ is constant for each n, the expectation also exists. We can apply the weak uniform law of large numbers to prove (C): that Q_n converges uniformly in probability to

$$Q = \mathbb{E}\left[g\left(\boldsymbol{Z}_i; \boldsymbol{\theta}\right)\right] = -\frac{1}{2}\mathbb{E}\left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\theta}\right)^2 - \lambda \sum_{j=1}^p |\theta_j|.$$

- Finally, by note that the square and absolute value functions are both strictly convex (under the positive definiteness of $\mathbb{E}\left[\boldsymbol{X}_i \boldsymbol{X}_i^\top\right]$), and thus Q has a strict global maximum $\boldsymbol{\theta}_0 \in \Theta$.

A consistency result? (4)

We have therefore proved that under the assumptions of the model, the sequence of global maximal values $\hat{\boldsymbol{\theta}}_n$ of

$$Q_n = -\frac{1}{2}\sum_{i=1}^{n}\left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\theta}\right)^2 - n\lambda \sum_{j=1}^{p} |\theta_j|,$$

converge in probability to some $\boldsymbol{\theta}_0 \in \Theta$ that globally maximizes Q.

- But does $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_S$?
  - Unless λ is sufficiently small, the answer is no, since the regularization λ enforces an l_1 ball constraint.

A consistency result? (5)

- Consider the l_1 ball, for $\kappa > 0$,

$$\sum_{j=1}^{p} |\theta_i| \leq \kappa.$$

- From Osborne et al. (2000), we have the result that

$$\lambda(\kappa) \equiv \lambda = C_1 - C_2\kappa,$$

  for real constant C_1 and positive constant C_2.

- So if $\lambda(\kappa)$ is such that

$$\Theta_{\lambda(\kappa)} \equiv \left\{ \boldsymbol{\theta} : \sum_{j=1}^{p} |\theta_i| \leq \kappa \right\} \subsetneq \Theta,$$

  and $\boldsymbol{\theta}_S \in \Theta \backslash \Theta_\kappa$, then $\boldsymbol{\theta}_0 \neq \boldsymbol{\theta}_S$.

A consistency result? (5)

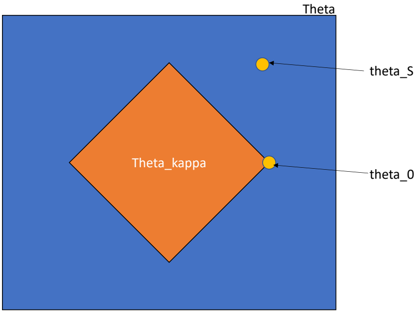

Figure: Schematic of the parameter spaces Θ_κ and Θ.

The method of sieves

- The **method of sieves** is a general estimation philosophy that was first introduced in Grenander (1981, Ch. 8).
- The modern interpretation of the method of sieves is as follows (cf. Chen, 2007):
  - Let $\boldsymbol{\theta}_0 \in \Theta$ be the parameter of interest, and let Θ be a compact Euclidean space.
  - At each $n \in \mathbb{N}$, define the compact set Θ_n as the **sieve space**, where

$$\Theta_n \subset \Theta_{n+1} \subset \cdots \subset \Theta.$$

  - Define the **sieve estimator**, at n, as

$$\tilde{\boldsymbol{\theta}}_n \equiv \arg\max_{\boldsymbol{\theta} \in \Theta_n} Q_n(\boldsymbol{\theta}),$$

    where Q_n is constructed from the data $\{\boldsymbol{Z}_i\}$.

Consistency of the sieve estimator (1)

- Let Π_n be a (loosely defined) projection operator into the set Θ_n and make the following assumptions:

(A3) The parameter space Θ is compact and $Q_n(\boldsymbol{\theta})$ is continuous with respect to $\boldsymbol{\theta} \in \Theta$. There exists a Q, such that $\boldsymbol{\theta}_0$ is the unique global maximizer of Q, and $Q(\boldsymbol{\theta}_0) > -\infty$.

(B3) For all $k \geq 1$, $\Theta_k \subset \Theta_{k+1} \subset \Theta$ is compact, and for any $\boldsymbol{\theta} \in \Theta$, there exists a $\Pi_k \boldsymbol{\theta} \in \Theta_k$, such that $\lim_{k \to \infty} \|\boldsymbol{\theta} - \Pi_k \boldsymbol{\theta}\| = 0$.

(C3) Q_n is measurable with respect to $\{\boldsymbol{Z}_i\}$ for all $\boldsymbol{\theta} \in \Theta_k$, and Q_n is continuous for every $\{\boldsymbol{Z}_i\}$.

(D3) For each $k \geq 1$, Q_n converges in probability uniformly to Q, in the sieve space Θ_k.

Consistency of the sieve estimator (2)

- Theorem 3.1 of Chen (2007) states the provides the following result.

Under Assumptions (A3)–(D3), the sieve estimator is consistent in the sense that

$$\tilde{\boldsymbol{\theta}}_n \xrightarrow{\text{p}} \boldsymbol{\theta}_0.$$

- As a note, (A3)–(D3) are one set of many possible set of assumptions that results in the same theorem.

A simple oracle (1)

Make the following assumptions:

(a*) $\{\boldsymbol{Z}_i\}$ is and IID sequence and that the DGP of $\boldsymbol{Z}_i = (\boldsymbol{X}_i, Y_i)$ is such that $\mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\top\right)$ exists and is positive definite, $\mathbb{E}(E_i) = 0$, $\mathbb{E}\left(E_i^2\right) = \sigma^2 < \infty$, and $\mathbb{E}(\boldsymbol{X}_i E_i) = \mathbf{0}$, where

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\theta}_\mathsf{S} + E_i.$$

(b**) The parameter space is $\Theta = [-L, L]^p$, where L is sufficiently large, and $\boldsymbol{\theta}_\mathsf{S}$ is in Θ.

A simple oracle (2)

- Let $\kappa(n) \equiv \kappa$, be a non-zero and strictly increasing function of n, and define the set

$$\Theta_n = \left\{ \boldsymbol{\theta} : \sum_{j=1}^{p} |\theta_i| \leq \kappa(n) \right\} \cap \Theta.$$

- Clearly, $\Theta_n \subset \Theta_{n+1} \subset \Theta$, for each n, and Θ_n is compact.
- Define $\Pi_n \boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}_n \in \Theta_n} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}\|$.
- For sufficiently large N, $\Theta_N = \Theta$, and thus $\Pi_N \boldsymbol{\theta} = \boldsymbol{\theta}$, and thus $\Pi_n \boldsymbol{\theta} \to \boldsymbol{\theta}$, for all $\boldsymbol{\theta} \in \Theta$.
- We have therefore fulfilled Assumption (B3).
- We also note that $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_S$, due to Assumption (B3).

A simple oracle (3)

- Define, $\lambda\left(\kappa(n)\right)$ fulfill the relationship $\lambda\left(\kappa(n)\right) = C_1 - C_2\kappa(n)$, such the problem

$$\max_{\boldsymbol{\theta}\in\Theta} Q_n = -\frac{1}{2}\sum_{i=1}^{n}\left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\theta}\right)^2 - n\lambda\left(\kappa(n)\right)\sum_{j=1}^{p}|\theta_j|$$

is equivalent to the problem

$$\max_{\boldsymbol{\theta}\in\Theta_n} -\frac{1}{2}\sum_{i=1}^{n}\left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\theta}\right)^2.$$

- Under the assumptions on the model, The first problem is strictly concave and thus has a unique global maximizer $\hat{\boldsymbol{\theta}}_n$, which implies the satisfaction of Assumption (A3).

A simple oracle (4)

- We have already proved that Q_n is measurable and continuous, previously, and thus (C3) is fulfilled.
- For each constant k,

$$\mathbb{E}\left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\theta}\right)^2$$

  is finite, since Θ_k is compact, and since $\mathbb{E}\left(E_i^2\right) < \infty$ and $\mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\top\right)$ exists. Thus (D3) is fulfilled.

Under (a) and (b**), if $\kappa(n)$ is a non-zero and strictly increasing function of n, and*

$$\Theta_n = \left\{\boldsymbol{\theta} : \sum_{j=1}^{p} |\theta_i| \le \kappa(n)\right\} \cap \Theta,$$

then the sieve estimator $\tilde{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\Theta_n} -\frac{1}{2}\sum_{i=1}^n \left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\theta}\right)^2$ is a consistent estimator of $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_S$.

A simple oracle (5)

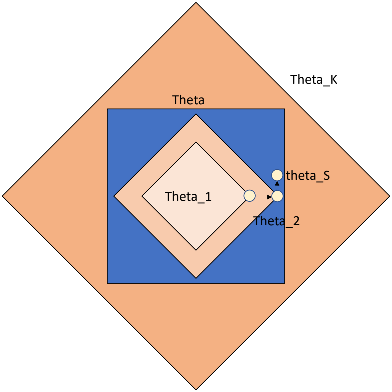

Figure: Schematic of the behaviour of the sieve estimator.

A different kind of oracle (1A)

- Make the same assumptions as the previous example:

(a*) $\{\boldsymbol{Z}_i\}$ is and IID sequence and that the DGP of $\boldsymbol{Z}_i = (\boldsymbol{X}_i, Y_i)$ is such that $\mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\top\right)$ exists and is positive definite, $\mathbb{E}(E_i) = 0$, $\mathbb{E}\left(E_i^2\right) = \sigma^2 < \infty$, and $\mathbb{E}(\boldsymbol{X}_i E_i) = \boldsymbol{0}$, where

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\theta}_\mathsf{S} + E_i.$$

(b**) The parameter space is $\Theta = [-L, L]^p$, where L is sufficiently large, and $\boldsymbol{\theta}_\mathsf{S}$ is in Θ.

A different kind of oracle (1B)

- Suppose now that we want to estimate the q-sparse parameter $\boldsymbol{\theta}_{\mathsf{S}}$ again, but by estimating a sequence of estimators $\hat{\boldsymbol{\theta}}_k \in \hat{\Theta}_k^{\mathsf{S}}$, where

$$\hat{\Theta}_k^{\mathsf{S}} = \left\{ \hat{\boldsymbol{\theta}} : \hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta_k^{\mathsf{S}}} \mathbb{E}\left[g\left(\boldsymbol{Z}_i; \boldsymbol{\theta}\right)\right] \right\},$$

$$\Theta_k^{\mathsf{S}} = \left\{ \boldsymbol{\theta} \in \Theta : \boldsymbol{\theta} \text{ is } k\text{-sparse (has } k \text{ non-zero elements)} \right\},$$

  and $k \in \{1, \dots, q, \dots K\}$.
  - Recall that $g\left(\boldsymbol{Z}_i; \boldsymbol{\theta}\right) = -\left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\theta}\right)^2 / 2$.
- Is there an estimation method for using the sequence $\hat{\boldsymbol{\theta}}_k$ (or the estimate sequence $\hat{\boldsymbol{\theta}}_{k,n}$) in order to selection the correct k, say $\hat{k}_n$, where $\hat{k}_n$ goes to q in n, in some sense?

A model selection result (1)

- Define $\left\{\Theta_k^{\mathrm{M}}\right\}$ to be a collection of models $\Theta_k^{\mathrm{M}} \subset \mathbb{R}^{d_k}$, where $k = \{1, 2, \ldots, K\}$, and $d_1 \leq d_2 \leq \cdots \leq d_K \in \mathbb{N}$.
- Let $Q_n(\boldsymbol{\theta}) = \sum_{i=1}^n g(\boldsymbol{Z}_i; \boldsymbol{\theta})$ for the sequence of data $\{\boldsymbol{Z}_i\}$ be such that $\boldsymbol{\theta} \in \cup \Theta_k^{\mathrm{M}}$.
- Define $\hat{\boldsymbol{\theta}}_k \in \hat{\Theta}_k^{\mathrm{M}}$, with

$$\hat{\Theta}_k = \left\{ \hat{\boldsymbol{\theta}}_k : \hat{\boldsymbol{\theta}}_k = \arg\max_{\boldsymbol{\theta} \in \Theta_k^{\mathrm{M}}} \mathbb{E}\left[g(\boldsymbol{Z}_i; \boldsymbol{\theta})\right] \right\}.$$

- The following results arises from Theorem 8.1 of Baudry (2015).

A model selection result (2)

- Make the assumptions:

(A4) Suppose that there exists some

$$k_0 = \min \left\{ \underset{k \in \{1,\dots,K\}}{\arg\max}\, \mathbb{E}\left[g\left(\boldsymbol{Z}_i ; \hat{\boldsymbol{\theta}}_k \right) \right] \right\}.$$

(B4) For all k, $\hat{\boldsymbol{\theta}}_{k,n} \in \Theta_k^{\mathrm{M}}$ is such that

$$Q_n\left(\hat{\boldsymbol{\theta}}_{k,n} \right) \geq Q_n\left(\hat{\boldsymbol{\theta}}_k \right)$$

and

$$n^{-1} Q_n\left(\hat{\boldsymbol{\theta}}_{k,n} \right) \xrightarrow{\mathrm{p}} \mathbb{E}\left[g\left(\boldsymbol{Z}_i ; \hat{\boldsymbol{\theta}}_k \right) \right].$$

A model selection result (3)

(C4) We can define a penalty function $\text{pen}(k,n)$, such that

$\text{pen}(k,n) > 0$,

$$\lim_{n\to\infty} \text{pen}(k,n) = \infty,$$

and $n[\text{pen}(k_2,n) - \text{pen}(k_1,n)] \xrightarrow{\text{p}} \infty$, when $k_2 > k_1$.

(D4) For any $\hat{k} \in \arg\max_{k\in\{1,\dots,K\}} \mathbb{E}\left[g\left(\boldsymbol{Z}_i; \hat{\boldsymbol{\theta}}_k\right)\right]$,

$$Q_n\left(\hat{\boldsymbol{\theta}}_{k_0,n}\right) - Q_n\left(\hat{\boldsymbol{\theta}}_{\hat{k},n}\right) = \text{O}_{\text{p}}(1).$$

Under (A4)–(D4), $\lim_{n\to\infty} \mathbb{P}\left(\hat{k}_n \neq k_0\right) = 0$, *where*

$$\hat{k}_n = \min\left\{\arg\min_{k\in\{1,\dots,K\}} -n^{-1} Q_n\left(\hat{\boldsymbol{\theta}}_k\right) + pen(k,n)\right\}.$$

A model selection result (4)

- The most difficult assumption to prove in general is (D4).
- A set of conditions for for guaranteeing (D4) is provided in Corollary 8.2 of Baudry (2015).

(c) Some conditions that suffice are:

- g is twice continuously differentiable.
- Θ_k^{M} is compact for each k.
- $\{\boldsymbol{Z}_i\}$ is a sequence of bounded random variables.
- The Hessian $(\mathbf{H}\ \mathbb{E}g)\left(\hat{\boldsymbol{\theta}}_{k_0}\right)$ is nonsingular.

A different kind of oracle (2A)

- (A4) must be assumed, and we will restate it as the existence of

$$k_0 = \min\left\{\underset{k\in\{1,\dots,K\}}{\arg\max}\ \mathbb{E}\left[-\left(Y_i - \boldsymbol{X}_i^\top \hat{\boldsymbol{\theta}}_k\right)^2/2\right]\right\}.$$

- We have proved (B4) in all of the previous examples (since Q_n is still concave, and the law of large numbers still applies).
- We must propose a penalty that has the properties that we desire. We can check that the penalty

$$\text{pen}\left(n,k\right) = k\frac{\log n}{n}$$

satisfies the criteria of (C4).

  - Clearly, $k \geq 1$ and $n \geq 1$, so $\text{pen}\left(n,k\right) \geq 0$.
  - $k_2 \log n - k_1 \log n = \left(k_2 - k_1\right)\log n \to \infty$, since $k_2 > k_1$.

A different kind of oracle (2B)

- Assumption (c) only requires us to assume that each $|Y_i| \leq C_1$ and $\|\boldsymbol{X}_i\| \leq C_2$, for some C_1 and C_2, and so we make these extra assumptions and validate (D4).
- We therefore have the following result:

For each k, define the k-sparse parameter space to be

$$\Theta_k^S = \{\boldsymbol{\theta} \in \Theta : \boldsymbol{\theta} \text{ is } k\text{-sparse (has } k \text{ non-zero elements)}\}.$$

Assume that (a), (b**), and (c) hold. If*

$$\hat{\boldsymbol{\theta}}_{k,n} = \arg\max_{\boldsymbol{\theta} \in \Theta_k^S} -\frac{1}{2}\sum_{i=1}^{n}\left(Y_i - \boldsymbol{X}_i^\top \hat{\boldsymbol{\theta}}_{k,n}\right)^2,$$

then $\lim_{n\to\infty} \mathbb{P}\left(\hat{k}_n \neq k_0\right) = 0$, *where*

$$\hat{k}_n = \min\left\{\arg\min_{k\in\{1,\dots,K\}} \left[\frac{1}{2n}\sum_{i=1}^{n}\left(Y_i - \boldsymbol{X}_i^\top \hat{\boldsymbol{\theta}}_{k,n}\right)^2 + k\frac{\log n}{n}\right]\right\}.$$

Some final notes

- Note that there is a distinct lack of independence assumptions in the main theorems: Amemiya (1985, Thms. 4.1.1, 4.12, 4.1.6), Chen (2007, Thm. 3.1), and Baudry (2015, Thm. 8.1).
- Each of the theorems rely on the use of some law of large numbers, uniform law of large numbers, or central limit theorems.
- Generic law of large numbers for non-IID data can be found in Davidson (1994), Potscher and Prucha (1997), and White (2001).
- Generic uniform laws can be found in Andrews (1992), Potscher and Prucha (1997), and Jenish and Prucha (2009).

References I

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press, Cambridge.

Andrews, D. W. K. (1992). Generic uniform convergence. *Econometric Theory*, 8:241–257.

Atienza, N., Garcia-Heras, J., Munoz-Pichardo, J. M., and Villa, R. (2007). On the consistency of MLE in finite mixture models of exponential families. *Journal of Statistical Planning and Inference*, 137:496–505.

Baudry, J.-P. (2015). Estimation and model selection for model-based clustering with the conditional classification likelihood. *Electronic Journal of Statistics*, 9:1041–1077.

References II

Bickel, P. J. and Docksum, K. A. (2000). *Mathematical Statistics: Basic Ideas and Selected Topics*, volume 1. Prentice Hall, Upper Saddle River.

Chen, X. (2007). *Handbook of Econometrics*, volume 6B, chapter Large sample sieve estimation of semi-nonparametric models, pages 5549–5632. Elsevier.

Davidson, J. (1994). *Stochastic Limit Theory*. Oxford University Press, Oxford.

Grenander, U. (1981). *Abstract Inference*. Wiley, New York.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, Boca Raton.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:73–101.

References III

Jenish, N. and Prucha, I. R. (2009). Central limit theorems and uniform laws of large nnumber for arrays of random fields. *Journal of Econometrics*.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Annals of Mathematical Statistics*, 40:633–643.

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20:389–404.

Pfanzagl, J. (1969). On the measurability and consistency of minimum contast estimates. *Metrika*, 14:249–272.

Potscher, B. M. and Prucha, I. R. (1997). *Dynamic Nonlinear Econometric Models: Asymptotic Theory*. Springer, Berlin.

Serfling, R. J. (1980). *Approximation Theorems Of Mathematical Statistics*. Wiley, New York.

References IV

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B*, 58:267–288.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis Of Finite Mixture Distributions*. Wiley, New York.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Vapnik, V. (2000). *The Nature of Statistical Learning Theory*. Springer, New York.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley, New York.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, 20:595–601.

White, H. (2001). *Asymptotic Theory For Econometricians*. Academic Press, San Diego.

Thank you for your attention!

Email: **h.nguyen5@latrobe.edu.au**

Twitter: **@tresbienhien**

Website: **https://hiendn.github.io**