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

Hien D. Nguyen¹,²

¹DECRA Research Fellow, Australian Research Council. ²Lecturer, 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
Framework

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

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

$$ \theta_0 \equiv \arg\max_{\theta \in \Theta} Q(\theta) $$

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

- We call

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

the **extremum estimator** of $\theta_0$. 
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:
  - 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 \textit{a posteriori} estimation.
- Minimum distance estimation (e.g. least-squares, least-absolute deviation, etc).
Statistical inference

- Since $\theta_0$ is defined as the maximum of $Q$, it must contain some information regarding the DGP of $\{Z_i\}$.

1. We hope that given $Q_n$, $\hat{\theta}_n$ will provide us with the same information regarding $Q$, provided that $n$ is large enough.

2. We also hope that $\hat{\theta}_n$ also has some DGP that is dependent on $\theta_0$, which allows us to assess a priori hypotheses regarding $\theta$.
Suppose that we observe independent and identically distributed (IID) data pairs $Z_i = (X_i, Y_i)$, where

$$Y_i = X_i^\top \theta^* + E_i,$$

where $\mathbb{E}(E_i) = 0$, and that the DGP of $Z_i$ is in some sense, well-behaved.

- $\theta^* \in \Theta \subset \mathbb{R}^p$ and $X_i \in \mathbb{X} \subset \mathbb{R}^p$, $p \in \mathbb{N}$, and $\{E_i\}$ is independent of $\{X_i\}$.

Define the (negative) sum-of-squares as

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

The least-squares estimator is defined as

$$\hat{\theta}_n \equiv \arg \max_{\theta \in \Theta} - \frac{1}{2} \sum_{i=1}^n \left( Y_i - X_i^\top \theta \right)^2.$$
Ordinary least squares (1B)

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

\[
\nabla Q_n = \sum_{i=1}^{n} X_i \left( Y_i - X_i^\top \theta \right) = 0
\]

\[
\implies \sum_{i=1}^{n} X_i X_i^\top \theta = \sum_{i=1}^{n} X_i Y_i
\]

\[
\implies \hat{\theta}_n = \left( \sum_{i=1}^{n} X_i X_i^\top \right)^{-1} \sum_{i=1}^{n} X_i Y_i.
\]

- More familiarly, if we put $X_i^\top$ into the $i$th row of $X_n \in \mathbb{R}^{n \times p}$ and put $Y_i$ into the $i$th position of $y_n \in \mathbb{R}^{n}$, then we can write

\[
\hat{\theta}_n = \left( X_n^\top X_n \right)^{-1} X_n^\top y_n.
\]
Ordinary least squares (1C)

- Since $\hat{\theta}_n$ is an estimate of $\theta_0$, we must determine if there is a sensible relationship between $Q_n$ and $\theta_0$.
- The following is a heuristic argument. Note that $\xrightarrow{p}$ denotes convergence in probability.

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

$$g(Z_i) = -\frac{1}{2} (Y_i - X_i^\top \theta)^2.$$

2. Since $Z_i$ is well-behaved, then a weak law of large numbers implies that

$$n^{-1}Q_n \xrightarrow{p} \mathbb{E}[g(Z_i)] = -\frac{1}{2} \mathbb{E} \left[ (Y_i - X_i^\top \theta)^2 \right] = Q$$
Ordinary least squares (1D)

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

\[
\nabla Q = \mathbb{E} \left[ X_i \left( Y_i - X_i^\top \theta \right) \right]
\]

\[
= \mathbb{E} \left[ X_i \left( X_i^\top \theta + E_i - X_i^\top \theta \right) \right]
\]

\[
= \mathbb{E} \left( X_i X_i^\top \right) \theta^* + \mathbb{E} (X_i E_i) - \mathbb{E} \left( X_i X_i^\top \right) \theta
\]

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

\[
0 = \mathbb{E} \left( X_i X_i^\top \right) \theta^* - \mathbb{E} \left( X_i X_i^\top \right) \theta
\]

\[
\implies \theta_0 = \arg \max_{\theta \in \Theta} Q = \theta^*
\]

Thus, in this case, we have found that \( \theta_0 \) is the generative parameter \( \theta^* \)!
Consistency

- We must now make precise the notion regarding how $\hat{\theta}_n$ and $\theta_0$ are related.

- Earlier, we defined $\xrightarrow{p}$ to denote convergence in probability. We say that a random variable $U_n$ converges in probability to another random variable $U$, if for every $\varepsilon > 0$, we have
  \[
  \lim_{n \to \infty} \mathbb{P}(\|U_n - U\| > \varepsilon) = 0,
  \]
  where $\|\cdot\|$ is some appropriate norm (usually Euclidean, in our case).

- We say that $\hat{\theta}_n$ is a **consistent** estimator of $\theta_0$, if $\hat{\theta}_n \xrightarrow{p} \theta_0$. 

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

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

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

(D) $Q(\theta)$ obtains a unique global maximum at $\theta_0$. 
Under Assumptions (A)–(D), then the EE, defined as

\[ \hat{\theta}_n \equiv \arg \max_{\theta \in \Theta} Q_n(\theta), \]

is consistent, in the sense that \( \hat{\theta}_n \xrightarrow{p} \theta_0 \).

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

\[
\lim_{n \to \infty} \mathbb{P} \left( \sup_{\theta \in \Theta} \left| n^{-1} Q_n(\theta) - Q(\theta) \right| > \varepsilon \right) = 0.
\]
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 \( Q_n(\theta) = \sum_{i=1}^{n} g(Z_i; \theta) \) be a measurable function of the IID sequence \( \{Z_i\} \), where \( Z_i \) is supported in a Euclidean space, for each \( \theta \in \Theta \), where \( \Theta \) is compact and Euclidean. If \( \mathbb{E}\left[g(Z_i; \theta)\right] \) exists, and \( \mathbb{E}\left[\sup_{\theta \in \Theta} g(Z_i; \theta)\right] < \infty \), then \( n^{-1} Q_n(\theta) \) converges in probability uniformly to \( Q(\theta) = \mathbb{E}[g(Z_i; \theta)] \).
Ordinary least squares (2A)

- Make the following assumptions:

  (a) \( \{Z_i\} \) is an IID sequence and that the DGP of \( Z_i = (X_i, Y_i) \) is such that \( \mathbb{E}(X_iX_i^\top) \) exists and is positive definite, \( \mathbb{E}(E_i) = 0 \), \( \mathbb{E}(E_i^2) = \sigma^2 < \infty \), and \( \mathbb{E}(X_iE_i) = 0 \), where

\[
Y_i = X_i^\top \theta_* + E_i.
\]

(b) The parameter space is \( \Theta = [-L, L]^p \), where \( L \) is sufficiently large.
Ordinary least squares (2B)

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

$$-2g = \left(Y_i - X_i^T \theta\right)^2 = Y_i^2 + Y_i X_i^T \theta - \theta^T X_i X_i^T \theta$$

and

$$E \left[ \left(Y_i - X_i^T \theta\right)^2 \right] = E(Y_i^2) - 2E(Y_i X_i^T) \theta + \theta^T E(X_i X_i^T) \theta.$$
Ordinary least squares (2C)

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

\[
\mathbb{E}\left[\left(Y_i - X_i^\top \theta\right)^2\right] = \theta_\ast^\top \mathbb{E}(X_i X_i^\top) \theta_\ast + 2 \mathbb{E}(E_i X_i^\top) \theta_\ast \\
+ \mathbb{E}(E_i^2) - 2 \theta^\top \mathbb{E}(X_i X_i^\top) \theta_\ast \\
- 2 \mathbb{E}(E_i X_i^\top) \theta + \theta^\top \mathbb{E}(X_i X_i^\top) \theta \\
= \theta_\ast^\top \mathbb{E}(X_i X_i^\top) \theta_\ast - 2 \theta^\top \mathbb{E}(X_i X_i^\top) \theta_\ast \\
+ \theta^\top \mathbb{E}(X_i X_i^\top) \theta + \sigma^2.
\]

- Since \( \mathbb{E}(X_i X_i^\top) \) exists, \( Q_n \) is measurable, and \( g \) is quadratic in \( \theta \), thus it is continuous and we have the validation of (B).
Ordinary least squares (2D)

- Write $Q_n = \sum_{i=1}^{n} g(Z_i; \theta)$, where
  
  \[
g_i(Z_i; \theta) = -\frac{1}{2} \left( Y_i - X_i^\top \theta \right)^2.
  \]

- From the previous slide, we have the fact that
  
  \[
  \mathbb{E}[g(Z_i; \theta)] = -\frac{1}{2} \theta^\top \mathbb{E} \left( X_i X_i^\top \right) \theta + \theta^\top \mathbb{E} \left( X_i X_i^\top \right) \theta
  - \frac{1}{2} \mathbb{E} \left( X_i X_i^\top \right) \theta - \sigma^2.
  \]

- By (b), $\Theta$ is compact, and we have established that $g$ is continuous. Thus, via the Weierstrass extreme value theorem,
  
  \[
  \mathbb{E} \left[ \sup_{\theta \in \Theta} g(Z_i; \theta) \right] \leq M < \infty.
  \]
Ordinary least squares (2E)

- Via the theorem of Jennrich (1969), we have the conclusion that $n^{-1}Q_n$ converges in probability uniformly to $\mathbb{E}[g(Z_i; \theta)]$.

- Finally, we observe that $\mathbb{E}[g(Z_i; \theta)]$ is a concave quadratic in $\theta$ since $\mathbb{E}(X_iX_i^\top)$ is positive definite (it may be linear otherwise), so $\mathbb{E}[g(Z_i; \theta)]$ has a unique global maximum and thus (D) is validated.

  - The global maximum is $\theta_0 = \theta^*$.

- We have validated (A)–(D), and thus can conclude that $\hat{\theta}_n$ is a consistent estimator for $\theta_0$. 
Asymptotic normality

- We would now like to establish, in a more precise manner, how $\hat{\theta}_n$ fluctuates around $\theta_0$ as it converges.
- In most cases, $n^{1/2} \left( \hat{\theta}_n - \theta_0 \right) \xrightarrow{d} N(0, \Sigma)$.

  - We write $\xrightarrow{d}$ to denote convergence in distribution.
  - We write $N(\mu, \Sigma)$ to denote the multivariate normal distribution with mean vector $\mu$ and covariance matrix $\Sigma$.

- 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 $U_n$ converges to the distribution of $U$ if and only if the characteristic function of $U_n$ converges point-wise to that of $U$ (cf. van der Vaart, 1998, Thm. 2.13).
We now present the asymptotic normality result of Amemiya (1985, Thm. 4.1.6).

Make the following assumptions:

(A1) The parameter $\theta_0$ is in the interior (an open subset) of the Euclidean parameter space $\Theta$.

(B1) The objective $Q_n(\theta)$ is continuous and measurable with respect to $\{Z_i\}$, for all $\theta \in \Theta$, and the partial derivative $(\nabla Q_n)(\theta)$ exists and is continuous in an open neighborhood $N_1$ of $\theta_0$.

(C1) There exists an open neighborhood $N_2$ of $\theta_0$, where $n^{-1}Q_n(\theta)$ converges in probability uniformly to a non-stochastic function $Q(\theta)$ in $N_2$, and $Q(\theta)$ attains a strict local maximum at $\theta_0$. 

Proving asymptotic normality (1)
Proving asymptotic normality (2)

- Make the further assumptions:

(A2) The Hessian matrix \((HQ_n)(\theta) \equiv \partial^2 Q_n/\partial\theta\partial\theta^\top\) exists and is continuous in an open and convex neighborhood of \(\theta_0\).

(B2) For any sequence \(\theta_n\), such that \(\theta_n \xrightarrow{p} \theta_0\), \(n^{-1}(HQ_n)(\theta_n)\) converges in probability to

\[
A(\theta_0) \equiv \lim_{n \to \infty} \mathbb{E} \left[ n^{-1}(HQ_n)(\theta_0) \right].
\]

(C2) \(n^{-1/2} (\nabla Q_n)(\theta_0) \xrightarrow{d} N(0, B(\theta_0))\), where

\[
B(\theta_0) \equiv \lim_{n \to \infty} \mathbb{E} \left[ n^{-1} (\nabla Q_n)(\theta_0)(\nabla Q_n)^\top(\theta_0) \right].
\]
Define $\tilde{\Theta}_n$ to be the set

$$\tilde{\Theta}_n = \{ \theta_n : (\nabla Q_n)(\theta_n) = 0 \}.$$ 

Under Assumptions (A1)–(C1) and (A2)–(C2), if $\hat{\theta}_n$ is a sequence of local maximizers taking values in $\Theta_n$, such that $\hat{\theta}_n \xrightarrow{p} \theta_0$, then

$$n^{1/2} \left( \hat{\theta}_n - \theta_0 \right) \xrightarrow{d} N(0, A^{-1}(\theta_0)B(\theta_0)A^{-1}(\theta_0)).$$
Ordinary least squares (3A)

- Make the following assumptions.

(a) \( \{Z_i\} \) is an IID sequence and that the DGP of \( Z_i = (X_i, Y_i) \) is such that \( \mathbb{E}(X_i X_i^\top) \) exists and is positive definite, \( \mathbb{E}(E_i) = 0 \), \( \mathbb{E}(E_i^2) = \sigma^2 < \infty \), and \( \mathbb{E}(X_i E_i) = 0 \), where

\[
Y_i = X_i^\top \theta^* + E_i.
\]

(b*) The parameter space is \( \Theta = [-L, L]^p \), where \( L \) is sufficiently large, and \( \theta_0 \) is in the interior of \( \Theta \).

Under (a) and (b*), we have the fulfillment of Assumptions (A1)–(C1).
Ordinary least squares (3B)

- Recall that

\[
\nabla Q_n = \sum_{i=1}^{n} x_i \left( y_i - x_i^\top \theta \right)
\]

\[
= \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i x_i^\top \theta
\]

\[
\implies (HQ_n)(\theta) = - \sum_{i=1}^{n} x_i x_i^\top.
\]

- Thus, we observe that \((HQ_n)(\theta)\) is constant for any \(\theta\) and is thus continuous, which fulfills (A2).
Ordinary least squares (3C)

- At $\theta_0$, we have
  $$(\nabla g)(\nabla g)^\top = X_i \left( Y_i - X_i^\top \theta_0 \right) \left( Y_i - X_i^\top \theta_0 \right)^\top X_i^\top$$

- Recalling that $\theta_0 = \theta_*$, the parentheses equate to
  $$Y_i - X_i^\top \theta_0 = X_i^\top \theta_* - X_i^\top \theta_0 + E_i$$
  $$= X_i^\top \theta_0 - X_i^\top \theta_0 + E_i$$
  $$= E_i.$$ 

- Therefore, we have $(\nabla g)(\nabla g)^\top = E_i^2 X_i X_i^\top$ and therefore, the expectation is
  $$\mathbb{E} \left[ (\nabla g)(\nabla g)^\top \right] = \mathbb{E} \left( E_i^2 X_i X_i^\top \right)$$
  $$= \mathbb{E} \left( E_i^2 \right) \mathbb{E} \left( X_i X_i^\top \right) = \sigma^2 \mathbb{E} \left( X_i X_i^\top \right).$$
By Assumption (a), \( \{Z_i\} \) is IID, and by definition of \( \theta_0 \), we have

\[
\mathbb{E} \left( \frac{1}{n} \nabla Q_n \right) = \mathbb{E} [\nabla g (Z; \theta_0)] = 0.
\]

Again, since \( \{Z_i\} \) is IID, we have

\[
\text{cov} \left( n^{-1} \nabla Q_n \right) = \mathbb{E} \left[ (n^{-1} \nabla Q_n)(n^{-1} \nabla Q_n)^\top \right] = \mathbb{E} \left[ \left( \sum_{i=1}^{n} \nabla g \right) \left( \sum_{i=1}^{n} \nabla g \right)^\top \right] = \mathbb{E} \left[ (\nabla g)(\nabla g)^\top \right],
\]

which exists!
Ordinary least squares (3E)

- We now need to establish the fact that

\[ n^{-1/2} \nabla Q_n = n^{-1/2} \sum_{i=1}^{n} g(\mathbf{Z}_i; \mathbf{\theta}_0) \]

converges in distribution to \( \mathcal{N}\left(0, \sigma^2 \mathbb{E}\left[\mathbf{X}_i\mathbf{X}_i^\top\right]\right) \).

- The multivariate Lindeberg-Lévy **central limit theorem** (CLT; van der Vaart, 1998, Thm. 2.18) states that if \( \{\mathbf{U}_i\} \) is an IID sequence that has finite mean vector \( \mathbf{\mu} \) and covariance matrix \( \mathbf{\Sigma} \), then

\[ n^{1/2} \left( n^{-1} \sum_{i=1}^{n} \mathbf{U}_i - \mathbf{\mu} \right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}). \]

- Since \( n^{-1/2} \sum_{i=1}^{n} g(\mathbf{Z}_i; \mathbf{\theta}_0) = n^{1/2} \left( n^{-1} \sum_{i=1}^{n} g(\mathbf{Z}_i; \mathbf{\theta}_0) - \mathbf{0} \right) \), we have the desired result, and (C2) is validated with \( \mathbf{B}(\mathbf{\theta}_0) = \sigma^2 \mathbb{E}\left[\mathbf{X}_i\mathbf{X}_i^\top\right] \).
Ordinary least squares (3F)

- Lastly,

\[ n^{-1} (H Q_n)(\theta_n) = n^{-1} \left( - \sum_{i=1}^{n} X_i X_i^\top \right). \]

- By independence, we have \[ \mathbb{E} \left[ n^{-1} (H Q_n)(\theta_0) \right] = \mathbb{E} (X_i X_i^\top), \]
  and via the weak law of large numbers, we have

\[ n^{-1} (H Q_n)(\theta_n) \xrightarrow{p} A(\theta_0), \]

where

\[ A(\theta_0) = -\mathbb{E} (X_i X_i^\top). \]

- Thus, (B2) is validated.
Finally, compute the matrix:

\[
A^{-1}BA^{-1} = \left[ \mathbb{E} \left( X_i X_i^\top \right) \right]^{-1} \left[ \sigma^2 \mathbb{E} \left( X_i X_i^\top \right) \right] \left[ \mathbb{E} \left( X_i X_i^\top \right) \right]^{-1} = \sigma^2 \left[ \mathbb{E} \left( X_i X_i^\top \right) \right]^{-1}.
\]

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

\[
n^{1/2} \left( \hat{\theta}_n - \theta_0 \right) \xrightarrow{d} \mathcal{N} \left( 0, \sigma^2 \left[ \mathbb{E} \left( X_i X_i^\top \right) \right]^{-1} \right).
\]
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 \( \{\hat{\theta}_n\} \) is a sequence of local maximizers that take values in \( \bar{\Theta}_n = \{\theta_n : (\nabla Q_n)(\theta_n) = 0\} \), then for any \( \varepsilon > 0 \)

\[
\lim_{n \to \infty} \mathbb{P}\left( \inf_{\theta_n \in \bar{\Theta}_n} \|\theta_n - \theta_0\| > \varepsilon \right) = 0.
\]

We read this as “there exists a consistent sequence of locally maximal roots \( \hat{\theta}_n \), 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

\[
f (z_i; \mu, \pi, \sigma) = \sum_{j=1}^{m} \pi_j \phi (z_i; \mu_j, \sigma_j^2),
\]

where \( \mu \in [-L, L]^m \), \( \sigma \in [S^{-1}, S]^m \), and

\[
\pi \in S_{m-1} = \left\{ (\pi_1, \ldots, \pi_m) : \pi_j \geq 0, \sum_{j=1}^{m} \pi_j = 1 \right\},
\]

for large \( L \) and \( S > 1 \).

We write \( \theta \in \Theta \) as the concatenation of \( \mu, \pi, \) and \( \sigma \).
Mixture of normal distributions (2)

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

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

- 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 \( \theta_0 \) is in the interior (an open subset) of the Euclidean parameter space \( \Theta \).

(B1) The objective \( Q_n(\theta) \) is continuous and measurable with respect to \( \{Z_i\} \), for all \( \theta \in \Theta \), and the partial derivative \( (\nabla Q_n)(\theta) \) exists and is continuous in an open neighborhood \( N_1 \) of \( \theta_0 \).

(C1) There exists an open neighborhood \( N_2 \) of \( \theta_0 \), where \( n^{-1}Q_n(\theta) \) converges in probability uniformly to a non-stochastic function \( Q(\theta) \) in \( N_2 \), and \( Q(\theta) \) attains a strict local maximum at \( \theta_0 \).
Mixture of normal distributions (4)

- Clearly, $\Theta = [-L, L]^m \times [S^{-1}, S]^m \times \mathbb{S}_{m-1}$ is Euclidean. We thus must simply make the assumption that (a1) $\theta_0$ is in the interior of $\Theta$. 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

$$\mathbb{E} \left[ \log \sum_{j=1}^{m} \pi_j \phi (Z_i; \mu_j, \sigma_j^2) \right] < \infty.$$
Mixture of normal distributions (5)

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

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

- We can write

\[
\log \phi \left( z_i; \mu_i, \sigma_i^2 \right) = -\frac{1}{2} \log (2\pi) - \frac{1}{2} \log \sigma_i^2 - \frac{1}{2\sigma_i^2} (z_i - \mu_i)^2
\]

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 $\theta$, we also have the existence of a continuous $\nabla Q_n$, 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 $\Theta$ 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 $\hat{\theta}_n$ is a sequence from $\bar{\Theta}_n$. 
Mixture of normal distributions (7)

Assume that $\theta_0$ 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 $\hat{\theta}_n$ is a sequence of locally maximal roots from the set

$$
\tilde{\Theta}_n = \left\{ \theta_n : (\nabla Q_n)(\theta_n) = 0 \right\}.
$$

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

$$
\lim_{n \to \infty} \mathbb{P} \left( \inf_{\theta_n \in \tilde{\Theta}_n} \| \theta_n - \theta_0 \| > \varepsilon \right) = 0.
$$

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 $\theta_0$, 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:

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

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

Here

$$Y_i = X_i^\top \theta_S + E_i,$$

where $\mathbb{E}(E_i) = 0$, $\mathbb{E}(E_i^2) = \sigma^2 < \infty$, and $\mathbb{E}(X_iX_i^\top)$ exists and is positive definite.

We say that $\theta$ is $q$-sparse ($q \in \mathbb{N}$, $q < p$) in the sense that

$$\theta_S = (\theta_1, \theta_2, \ldots, \theta_q, 0, \ldots, 0).$$
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 $\Theta$ is a compact subset of a Euclidean space $\mathbb{R}^p$ ($p \in \mathbb{N}$).

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

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

(D) $Q(\theta)$ obtains a unique global maximum at $\theta_0$. 
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
  \[
g(Z_i; \theta) = -\frac{1}{2} \left( Y_i - X_i^T \theta \right)^2 - \lambda \sum_{j=1}^{p} |\theta_j|.
  \]
- 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.
Again, we know that $\mathbb{E} \left[ (Y_i - X_i^\top \theta)^2 \right]$ 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

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

Finally, by note that the square and absolute value functions are both strictly convex (under the positive definiteness of $\mathbb{E} \left[ X_i X_i^\top \right]$), and thus $Q$ has a strict global maximum $\theta_0 \in \Theta$. 
A consistency result? (4)

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

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

converge in probability to some $\theta_0 \in \Theta$ that globally maximizes $Q$.

- But does $\theta_0 = \theta_s$?
  - Unless $\lambda$ is sufficiently small, the answer is no, since the regularization $\lambda$ 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

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

and $\theta_s \in \Theta \setminus \Theta_\kappa$, then $\theta_0 \neq \theta_s$. 
A consistency result? (5)

Figure: Schematic of the parameter spaces $\Theta_k$ and $\Theta$. 
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 $\theta_0 \in \Theta$ be the parameter of interest, and let $\Theta$ be a compact Euclidean space.
  - At each $n \in \mathbb{N}$, define the compact set $\Theta_n$ as the **sieve space**, where
    \[ \Theta_n \subset \Theta_{n+1} \subset \cdots \subset \Theta. \]
  - Define the **sieve estimator**, at $n$, as
    \[ \hat{\theta}_n \equiv \arg \max_{\theta \in \Theta_n} Q_n(\theta), \]
    where $Q_n$ is constructed from the data $\{Z_i\}$. 
Consistency of the sieve estimator (1)

Let \( \Pi_n \) be a (loosely defined) projection operator into the set \( \Theta_n \) and make the following assumptions:

(A3) The parameter space \( \Theta \) is compact and \( Q_n(\theta) \) is continuous with respect to \( \theta \in \Theta \). There exists a \( Q \), such that \( \theta_0 \) is the unique global maximizer of \( Q \), and \( Q(\theta_0) > -\infty \).

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

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

(D3) For each \( k \geq 1 \), \( Q_n \) converges in probability uniformly to \( Q \), in the sieve space \( \Theta_k \).
Consistency of the sieve estimator (2)

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

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

$$\tilde{\theta}_n \overset{p}{\to} \theta_0.$$

- As a note, (A3)–(D3) are one set of many possible sets of assumptions that result in the same theorem.
A simple oracle (1)

Make the following assumptions:

(a*) \( \{Z_i\} \) is an IID sequence and that the DGP of \( Z_i = (X_i, Y_i) \) is such that \( \mathbb{E}(X_i X_i^\top) \) exists and is positive definite, \( \mathbb{E}(E_i) = 0 \), \( \mathbb{E}(E_i^2) = \sigma^2 < \infty \), and \( \mathbb{E}(X_i E_i) = 0 \), where

\[ Y_i = X_i^\top \theta_S + E_i. \]

(b**) The parameter space is \( \Theta = [-L, L]^p \), where \( L \) is sufficiently large, and \( \theta_S \) is in \( \Theta \).
A simple oracle (2)

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

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

- Clearly, $\Theta_n \subset \Theta_{n+1} \subset \Theta$, for each $n$, and $\Theta_n$ is compact.

- Define $\Pi_n \theta = \arg\min_{\theta_n \in \Theta_n} \| \theta_n - \theta \|.$

- For sufficiently large $N$, $\Theta_N = \Theta$, and thus $\Pi_N \theta = \theta$, and thus $\Pi_n \theta \to \theta$, for all $\theta \in \Theta$.

- We have therefore fulfilled Assumption (B3).

- We also note that $\theta_0 = \theta_S$, due to Assumption (B3).
A simple oracle (3)

Define, $\lambda (\kappa (n))$ fulfill the relationship

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

such that the problem

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

is equivalent to the problem

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

Under the assumptions on the model, The first problem is strictly concave and thus has a unique global maximizer $\hat{\theta}_n$, 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 $k$,

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

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

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

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

then the sieve estimator $\tilde{\theta}_n = \arg \max_{\theta \in \Theta_n} -\frac{1}{2} \sum_{i=1}^{n} \left( Y_i - X_i^\top \theta \right)^2$ is a consistent estimator of $\theta_0 = \theta_S$. 
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*) \{Z_i\} is an IID sequence and that the DGP of \(Z_i = (X_i, Y_i)\) is such that \(\mathbb{E}(X_iX_i^\top)\) exists and is positive definite, \(\mathbb{E}(E_i) = 0\), \(\mathbb{E}(E_i^2) = \sigma^2 < \infty\), and \(\mathbb{E}(X_iE_i) = 0\), where

  \[Y_i = X_i^\top \theta_S + E_i.\]

  (b**) The parameter space is \(\Theta = [-L, L]^p\), where \(L\) is sufficiently large, and \(\theta_S\) is in \(\Theta\).
A different kind of oracle (1B)

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

$$\hat{\Theta}_k = \left\{ \hat{\theta} : \hat{\theta} = \arg \max_{\theta \in \Theta_k} \mathbb{E}[g(Z_i; \theta)] \right\},$$

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

and $k \in \{1, \ldots, q, \ldots K \}$.

- Recall that $g(Z_i; \theta) = - (Y_i - X_i^T \theta)^2 / 2$.

- Is there an estimation method for using the sequence $\hat{\theta}_k$ (or the estimate sequence $\hat{\theta}_{k,n}$) 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 \( \{ \Theta^M_k \} \) to be a collection of models \( \Theta^M_k \subset \mathbb{R}^{d_k} \), where \( k = \{1, 2, \ldots, K\} \), and \( d_1 \leq d_2 \leq \cdots \leq d_K \in \mathbb{N} \).

- Let \( Q_n(\theta) = \sum_{i=1}^n g(Z_i; \theta) \) for the sequence of data \( \{Z_i\} \) be such that \( \theta \in \bigcup \Theta^M_k \).

- Define \( \hat{\Theta}_k \subset \hat{\Theta}^M_k \), with

\[
\hat{\Theta}_k = \left\{ \hat{\theta}_k : \hat{\theta}_k = \arg \max_{\theta \in \Theta^M_k} \mathbb{E}[g(Z_i; \theta)] \right\}.
\]

- 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

\[ k_0 = \min \left\{ \arg \max_{k \in \{1, \ldots, K\}} \mathbb{E} \left[ g \left( Z_i; \hat{\theta}_k \right) \right] \right\}. \]

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

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

and

\[ n^{-1} Q_n \left( \hat{\theta}_{k,n} \right) \xrightarrow{p} \mathbb{E} \left[ g \left( Z_i; \hat{\theta}_k \right) \right]. \]
A model selection result (3)

(C4) We can define a penalty function \( \text{pen}(k, n) \), such that
\[
\text{pen}(k, n) > 0, \quad \lim_{n \to \infty} \text{pen}(k, n) = \infty,
\]
and \( n \left[ \text{pen}(k_2, n) - \text{pen}(k_1, n) \right] \xrightarrow{p} \infty \), when \( k_2 > k_1 \).

(D4) For any \( \hat{k} \in \arg\max_{k \in \{1, \ldots, K\}} \mathbb{E} \left[ g \left( \mathbf{Z}_i; \hat{\theta}_k \right) \right] \),
\[
Q_n \left( \hat{\theta}_{k_0, n} \right) - Q_n \left( \hat{\theta}_{\hat{k}, n} \right) = O_p(1).
\]

Under (A4)–(D4), \( \lim_{n \to \infty} \mathbb{P} \left( \hat{k}_n \neq k_0 \right) = 0 \), where
\[
\hat{k}_n = \min \left\{ \arg\min_{k \in \{1, \ldots, K\}} - n^{-1} Q_n \left( \hat{\theta}_k \right) + \text{pen}(k, n) \right\}.
\]
A model selection result (4)

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

(c) Some conditions that suffice are:

- $g$ is twice continuously differentiable.
- $\Theta^M_k$ is compact for each $k$.
- $\{Z_i\}$ is a sequence of bounded random variables.
- The Hessian $(H \mathbb{E}g)(\hat{\theta}_{k0})$ is nonsingular.
A different kind of oracle (2A)

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

\[ k_0 = \min \left\{ \arg \max_{k \in \{1, \ldots, K\}} \mathbb{E} \left[ - \left( Y_i - X_i^\top \hat{\theta}_k \right)^2 / 2 \right] \right\}. \]

- 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

\[ \text{pen}(n, k) = k \log n / n \]

satisfies the criteria of (C4).

- Clearly, \( k \geq 1 \) and \( n \geq 1 \), so \( \text{pen}(n, k) \geq 0 \).
- \( k_2 \log n - k_1 \log n = (k_2 - k_1) \log n \to \infty \), 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 $\|X_i\| \leq C_2$, 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

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

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

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

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

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


References III


References IV


Thank you for your attention!

Email: h.nguyen5@latrobe.edu.au

Twitter: @tresbienhien

Website: https://hiendn.github.io