STAT5610: Advanced Inference

Charles Christopher Hyland

Semester 2 2020

Abstract

Thank you for stopping by to read this. These are notes collated from lectures and tutorials as I took this course.

Contents

STAT5610: Advanced Inference

1. Parametric Estimation

1.1 Parametric Estimation

1.1.1 Motivation for Semiparametric Estimation

A semiparametric statistical model is a family of distributions indexed by two parameters

$$\{\mathbb{P}_{\theta,f} : \underset{\sim}{\theta} \in \underset{\sim}{\Theta}; f \in \mathcal{F}\} \tag{1.1}$$

where $\underset{\sim}{\theta}$ is an Euclidean parameter, i.e. $\underset{\sim}{\theta} \in \underset{\sim}{\Theta} \in \mathbb{R}^d$, and the parameter f is a function in the function space $\mathcal{F}$, so it can be *infinite dimensinoal*.

Our canonical example is given by the **canonical location model**.

Definition 1: Canonical Location Model

Suppose that $X_1, ..., X_n$ are i.i.d with common density

$$p(x : \theta, f) = f(x - \theta)$$

whereby $\theta \in \mathbb{R}$ and f is a density centered at 0.

The Euclidean (finite) parameter is what we will be primarly interested in. The issue is how to estimate this finite-dimensional parameter in the presence of the unknown infinite-dimensional **nuisance parameter** (nuisance function).

The central idea in semiparametric methods is to identify a least favourable parametric submodel, which lives inside the full model, whereby estimating the parameters in the submodel also estimates the parameters in the full model. Therefore, we first review some aspects of parametric theory.

1.1.2 Regular Parametric Models

We are interested in sequences of parametric models for the data $\underset{\sim}{x}$. From this, we will then define the LAN property which captures a particular aspect of regularity in parametric models. We will now review results from parametric estimation of parameters.

Definition 1.1 *(Convergence in distribution). Let X_n be a sequence of random variables. We say that X_n converges in distribution to X if*

$$\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$$

We will be taking alot of Taylor series expansion in this course.

Definition 1.2 *(Taylor Series). The Taylor series representation of a function $f(x)$ around a point c is given by*

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)(c)}}{n!}(x-c)^n$$

where the remaining terms converges to zero.

Definition 2: 3-Term Taylor Series Expansion

A 3-term Taylor series expansion with mean-value remainder of a three-times differentiable function $f(x)$ about 0 takes the form

$$f(x) = f(0) + xf^{'}(0) + \frac{x^2}{2!}f^{''}(0) + \frac{x^3}{3!}g^{'''}(\alpha x) \tag{1.2}$$

for some intermediate value $\alpha = \alpha(x) \in [0, 1]$.

Suppose we have data $\underset{\sim}{x_n}$ modelled as the value taken by a $\mathcal{X}_n$-valued random vector $\underset{\sim}{X_n}$ whereby n indicates the sample size.

Suppose that for each n, we have a family of distributions

$$\{\mathbb{P}_{n,\underset{\sim}{\gamma}} : \underset{\sim}{\gamma} \in \underset{\sim}{\Gamma}\}$$

of probability distributions and each $\mathbb{P}_{n,\underset{\sim}{\gamma}}$ has a density $p_{n,\underset{\sim}{\gamma}}(.)$ with respect to some dominating measure $\nu_n(\cdot)$ on $\mathcal{X}_n$. Generally, the dominating measure is the Lebesgue measure. We can now define the log-likelihood ratio.

Definition 3: Log Likelihood Ratio

Suppose that for each $n = 1, 2, ...$ and each $\underset{\sim}{\gamma} \in \Gamma \subseteq \mathbb{R}^d$, we have a probability distribution $\mathbb{P}_{n\underset{\sim}{\gamma}}$ on a sample space $\mathcal{X}_n$ whose density function with respect to the measure $\nu_n(.)$ is $p_{n\underset{\sim}{\gamma_0}}(.)$. Fix an interior point $\underset{\sim}{\gamma_0} \in \Gamma$.
Define the sequence of supports $A_n = A_n(\underset{\sim}{\gamma_0}) = \{\underset{\sim}{x} : \mathcal{X}_n : p_{n\underset{\sim}{\gamma}}(\underset{\sim}{x}) > 0\}$. Then, for any other $\underset{\sim}{\gamma_1} \in \Gamma$, we define the **log likelihood ratio** as

$$L_n(\underset{\sim}{x}; \underset{\sim}{\gamma_1} | \underset{\sim}{\gamma_0}) = \begin{cases} \frac{log\ p_{n\underset{\sim}{\gamma_1}}(\underset{\sim}{x})}{log\ p_{n\underset{\sim}{\gamma_0}}(\underset{\sim}{x})} & \text{for } \underset{\sim}{x} \in A_n(\underset{\sim}{\gamma_0}) \\ \\ 0 \quad \text{for } \underset{\sim}{x} \in A_n^c(\underset{\sim}{\gamma_0}) \end{cases}$$

where we define L_n to be zero outside of the support A_n.

Remark 1.3 *We can give a different interpretation of the log likelihood ratio. Suppose that $\mathbb{P}_{n\underset{\sim}{\gamma_1}}$ is absolutely continuous with respect to $\mathbb{P}_{n\underset{\sim}{\gamma_0}}$. Then, the log likelihood ratio is the logarithm of the Radon-Nikodym derivative of these two probability measures.*

We now want to do a ”local Pitman analysis” of the log likelihood ratio L_n. We want to restrict our analysis to a local area where $\underset{\sim}{\gamma_n} = \underset{\sim}{\gamma_0} + n^{-\frac{1}{2}}\underset{\sim}{h}$. We can think of this as we are taking a step $\underset{\sim}{h}$ away from the original point $\underset{\sim}{\gamma_0}$. Then, we can define the LAN property as a local property around $\underset{\sim}{\gamma_0}$.

Definition 4: Local Asymptotic Normality

Let $L_n(\utilde{x};\utilde{\gamma}_1|\utilde{\gamma}_0)$ be the log likelihood of $\utilde{\gamma}_1$ against the true value $\utilde{\gamma}_0$. We say that the **local asymptotic normality** property holds at $\utilde{\gamma}_0$ if

1. There exists a symmetric positive definite matrix $\utilde{J} = \utilde{J}(\utilde{\gamma}_0)$
2. There exists a random vector $\utilde{S}_n = \utilde{S}_n(\utilde{X}_n;\utilde{\gamma}_0)$ such that
$$\utilde{S}_n(\utilde{X}_n;\utilde{\gamma}_0) \xrightarrow{d} \mathcal{N}(\utilde{0},\utilde{J}) \tag{1.3}$$
where $\utilde{0}$ is the zero mean vector and covariance matrix $\utilde{J}$
3. Finally, when $\utilde{X}_n \sim \mathbb{P}_{n\utilde{\gamma}_0}$, for any d-dimensional vector $\utilde{h} \in \mathbb{R}^d$, the log likelihood ratio satisfies
$$L_n(\utilde{x};\utilde{\gamma}_0 + n^{-\frac{1}{2}}\utilde{h}|\utilde{\gamma}_0) = \utilde{h}^T\utilde{S}_n - \frac{1}{2}\utilde{h}^T\utilde{J}\utilde{h} + R_n \tag{1.4}$$
where the remainder $R_n = R_n(\utilde{\gamma}_0;\utilde{h}) \xrightarrow{p} 0$.

Definition 1.4 *(Vector of scores and information matrix). The random d-dimensional vector $\utilde{S}_n$ is the **Scores vector** and $\utilde{J}$ is the information matrix associated with $\utilde{\gamma}_0$.*

Remark 1.5 *The information matrix $\utilde{J}$ is usually the covariance matrix of the first derivatives of the log likelihood. Therefore, the information matrix can be interpreted to be the asymptotic variance of the scores.*

Remark 1.6 *We will later see conditions which imply the LAN property but it is worth noting that if the conditions which imply LAN holds, then we can show that $\utilde{S}_n(\utilde{X}_n;\utilde{\gamma}_0) \xrightarrow{d} \mathcal{N}(\utilde{0},\utilde{J})$ using the central limit theorem.*

The score function converges to the multivariate normal distribution through the central limit theorem. This leads to the following result.

Proposition 1: Log likelihood Ratio is asymptotically normal under LAN

The log likelihood ratio is asymptotically normal at $\utilde{\gamma}_0$ if the LAN property holds at $\utilde{\gamma}_0$. That is, if $\utilde{X}_n \sim \mathbb{P}_{n\utilde{\gamma}_0}$, then
$$L_n \xrightarrow{d} \mathcal{N}\Big(-\frac{\utilde{h}^T\utilde{J}\utilde{h}}{2}, \utilde{h}^T\utilde{J}\utilde{h}\Big)$$

1.1.3 Consequences of LAN

We are now interested in describing what are the consequences if the LAN property holds.

We first describe Le Cam's third lemma, which tells us the joint distribution of a statistic and the log-likelihood ratio under nearby alternatives.

Theorem 1: Le Cam's Third Lemma

Suppose that the LAN property holds at $\underset{\sim}{\gamma_0}$. Furthermore, assume that Y_n is a statistic such that under the value at which LAN holds, $\mathbb{P}_{n\underset{\sim}{\gamma_0}}$, we have the following limiting behaviour

$$\begin{pmatrix} Y_n \\ L_n \end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(\begin{pmatrix} 0 \\ -\frac{\delta^2}{2} \end{pmatrix}, \begin{pmatrix} \sigma^2 & \underset{\sim}{\Sigma_{YL}} \\ \underset{\sim}{\Sigma_{LY}} & \delta^2 \end{pmatrix} \Big) \tag{1.5}$$

where Y_n has limiting mean 0 and variance σ^2, $\delta^2 = \underset{\sim}{h}^T \underset{\sim}{J}\underset{\sim}{h}$, and $\underset{\sim}{\Sigma_{YL}}$ is the covariance between Y_n and L_n. Then, under a nearby alternative sequence $\mathbb{P}_{n(\underset{\sim}{\gamma_0}+n^{-1/2}\underset{\sim}{h})}$

$$\begin{pmatrix} Y_n \\ L_n \end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(\begin{pmatrix} \underset{\sim}{\Sigma_{YL}} \\ +\frac{\delta^2}{2} \end{pmatrix}, \begin{pmatrix} \sigma^2 & \underset{\sim}{\Sigma_{YL}} \\ \underset{\sim}{\Sigma_{LY}} & \delta^2 \end{pmatrix} \Big) \tag{1.6}$$

Remark 1.7 *More succinctly, if a statistic Y_n is asymptotically jointly normal with the log likelihood ratio L_n, then for a nearby alternative, Y_n is still asymptotically jointly normal with the log likelihood ratio with the limiting mean of Y_n being the covariance under $\underset{\sim}{\gamma_0}$. The nearby limiting sequence only has an effect on altering the limiting means.*

However, we are actually interested in **Scores version of Le Cam's third lemma**, whereby it changes the result from the log-likelihood ratio to the scores vector.

Proposition 2: Scores Version of Le Cam's Third Lemma

Suppose that LAN holds at $\underset{\sim}{\gamma_0}$. This implies that we have a scores vector $\underset{\sim}{S_n}$ and information matrix $\underset{\sim}{J}$. Furthermore, suppose that there exists a statistic $\underset{\sim}{Y_n}$ that is AJN (asymptotically jointly normal) with the scores $\underset{\sim}{S_n}$ for some positive semi-definite matrix $\underset{\sim}{\Sigma_Y}$, such that for $\underset{\sim}{X_n} \sim \mathbb{P}_{n,\underset{\sim}{\gamma_0}}$

$$\begin{pmatrix} \underset{\sim}{S_n} \\ \underset{\sim}{Y_n} \end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(\underset{\sim}{0}, \begin{pmatrix} \underset{\sim}{J} & \underset{\sim}{\Sigma_{SY}} \\ \underset{\sim}{\Sigma_{YS}} & \underset{\sim}{\Sigma_Y} \end{pmatrix} \Big) \tag{*}$$

Then, for any $\underset{\sim}{h} \in \mathbb{R}^d$, under a nearby alternative $\underset{\sim}{\gamma_n} = \underset{\sim}{\gamma_0} + n^{1/2}\underset{\sim}{h}$, we have that for $\underset{\sim}{X_n} \sim \mathbb{P}_{n,\gamma_0+n^{-1/2}\underset{\sim}{}h}$

$$\begin{pmatrix} \underset{\sim}{S_n} \\ \underset{\sim}{Y_n} \end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(\begin{pmatrix} \underset{\sim}{J}\underset{\sim}{h} \\ \underset{\sim}{\Sigma_{YS}}\underset{\sim}{h} \end{pmatrix}, \begin{pmatrix} \underset{\sim}{J} & \underset{\sim}{\Sigma_{SY}} \\ \underset{\sim}{\Sigma_{YS}} & \underset{\sim}{\Sigma_Y} \end{pmatrix} \Big) \tag{1.7}$$

Remark 1.8 *Only knowing the joint limiting behaviour under $\underset{\sim}{\gamma_0}$ and that LAN holds, we can deduce the joint limiting behaviour of both the score vector $\underset{\sim}{S_n}$ and statistic $\underset{\sim}{Y_n}$ under a nearby alternative $\underset{\sim}{\gamma_n} = \underset{\sim}{\gamma_0} + n^{1/2}\underset{\sim}{h}$.*

We see that in Score's version of Le Cam's third lemma, the asymptotic mean of the statistic $\utilde{Y}_n$ is now a linear combination of $\utilde{h}$ and $\utilde{\Sigma}_{YS}\utilde{h}$.

The Score's version of Le Cam's third lemma is important to us as many times, we will be analysing the joint distribution of a statistic and scores vector rather than the log-likelihood and statistic, especially when we introduce estimators that are asymptotically jointly normal with the scores vector.

1.1.4 Regular Estimators

We now restrict our attention to certain classes of estimators. We can derive under certain conditions what are the best estimators. In particular, we will see a limiting local version of the Cramér-Rao lower bound.

Definition 5: Regular Estimator

An estimator $\utilde{\tilde{\gamma}}_n$ is regular at $\utilde{\gamma}_0$ if for any $\utilde{h} \in \mathbb{R}^d$, if $\utilde{X}_n \sim \mathbb{P}_{n,\utilde{\gamma}_n}$ where $\utilde{\gamma}_n(\utilde{h}) = \utilde{\gamma}_0 + n^{-1/2}\utilde{h}$, then the rescaled estimation error

$$\sqrt{n}(\utilde{\tilde{\gamma}}_n - \utilde{\gamma}_n) \xrightarrow{d} \mathcal{N}(\utilde{0}, \utilde{\Sigma}(\utilde{\gamma}_0)) \tag{1.8}$$

That is, the asymptotic distribution of the estimation error possibly depends on $\utilde{\gamma}_0$ but does not depend on the local deviation $\utilde{h}$.

Remark 1.9 *As we move the nearby alternative $\utilde{\gamma}_n$ around near the true parameter $\utilde{\gamma}_0$, the distribution of the estimator error moves with $\utilde{\gamma}_n$ but only in its location. The shape and spread does not change. This is akin to asymptotic unbiasedness.*

Proposition 3: RAJN Estimators are analogous to unbiased estimators

Suppose that an estimator $\utilde{\tilde{\gamma}}_n$ is such that under $\utilde{\gamma}_0$, the rescaled estimation error

$$\utilde{Y}_n = \sqrt{n}(\utilde{\tilde{\gamma}}_n - \utilde{\gamma}_0)$$

is AJN with the scores. Then, the estimator $\utilde{\tilde{\gamma}}_n$ is also regular at $\utilde{\gamma}_0$ if and only if under $\utilde{X}_n \sim \mathbb{P}_{n,\utilde{\gamma}_0}$,

$$\begin{pmatrix} \utilde{S}_n \\ \utilde{Y}_n \end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(\utilde{0}, \begin{pmatrix} \utilde{J} & \utilde{\Sigma}_{SY} \\ \utilde{\Sigma}_{YS} & \utilde{\Sigma}_{Y} \end{pmatrix}\Big)$$

the cross-covariance $\utilde{\Sigma}_{S,Y}$ is a $d \times d$ identity matrix $\mathbb{I}$.

Proof: Follows easily from Le Cam's third lemma. ■

Remark 1.10 *First, recall that in the CRLB, any unbiased estimator has a covariance of 1 with the scores. This proposition is analogous to this.*

If the cross-covariance matrix is **not** the identity matrix, we will not get the correct shift under the alternative hypothesis $\mathbb{P}_{n,\underset{\sim}{\gamma_0}}$ whereby we will now have

$$\underset{\sim}{Y_n} \xrightarrow{d} \mathcal{N}(\Sigma_{\underset{\sim}{Y,S}}\underset{\sim}{h}, \Sigma_{\underset{\sim}{Y}})$$

and therefore the rescaled estimation error now **depends** on $\underset{\sim}{h}$ and is no longer regular.

STAT5610: Advanced Inference

2. Optimality of Estimators

2.1.5 Optimality of Estimators

We can now define estimators that combine both the property of regularity and being AJN with the scores. We will now restrict our attention to RAJN estimators.

Definition 6: RAJN Estimator

If an estimator $\underset{\sim}{\tilde{\gamma}_n}$ is AJN (i.e. its rescaled estimation error is AJN) and regular at $\underset{\sim}{\gamma_0}$, we call it RAJN.

We now show why RAJN estimators are desirable.

Theorem 2: Desirable properties of RAJN estimators

A RAJN estimator $\underset{\sim}{\tilde{\gamma}_n}$ satisfies

$$\begin{pmatrix} \underset{\sim}{S_n} \\ \sqrt{n}(\underset{\sim}{\tilde{\gamma}_n} - \underset{\sim}{\gamma_0}) \end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(\underset{\sim}{0}, \begin{pmatrix} \underset{\sim}{J} & \underset{\sim}{I} \\ \underset{\sim}{I} & \underset{\sim}{\Sigma_Y} \end{pmatrix}\Big) \tag{2.9}$$

for some positive semi-definite matrix $\underset{\sim}{\Sigma_Y}$.

Recall an important inequality relating the covariance and variance.

Definition 2.11 *(Correlation Inequality). The correlation inequality between an unbiased estimator $\hat{\theta}$ and score function ℓ°_θ is given by*

$$Cov_\theta\Big[\hat{\theta}, \ell^\circ_\theta\Big]^2 \leq Var_\theta(\hat{\theta}(\underset{\sim}{x}))Var_\theta(\ell^\circ_\theta(\underset{\sim}{x}))$$

Going back and looking at the 1-dimensional case for RAJN estimator, this immediately gives a local asymptotic version of the Cramér-Rao lower bound

$$\begin{pmatrix} \underset{\sim}{S_n} \\ \sqrt{n}(\underset{\sim}{\tilde{\gamma}_n} - \underset{\sim}{\gamma_0}) \end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} J & 1 \\ 1 & \sigma_Y^2 \end{pmatrix}\Big)$$

From this, we now have the correlation inequality as

$$\sigma_Y^2 \geq \frac{1}{J}$$

We now look at it for a general dimension d case. Again, by the correlation inequality, we have that

$$Cov_\theta\Big[\hat{\theta}, \ell^\circ_\theta\Big]^2 = \underset{\sim}{I} \leq \underset{\sim}{\Sigma_Y}\underset{\sim}{J} = Var_\theta(\hat{\theta}(\underset{\sim}{x}))Var_\theta(\ell^\circ_\theta(\underset{\sim}{x})$$

Therefore, we have that

$$\underset{\sim}{\Sigma_Y} \geq \underset{\sim}{J}^{-1}$$

and hence we can conclude that

$$\underset{\sim}{\Sigma_Y} - \underset{\sim}{J}^{-1}$$

is positive semi-definite.

However, this can be hard to interpret and therefore, we can try a different way to express it. First, we define a remainder as a linear combination of the Scores as follows:

$$\underset{\sim}{R_n} = \sqrt{n}(\underset{\sim}{\tilde{\gamma}_n} - \underset{\sim}{\gamma_0}) - \underset{\sim}{J}^{-1}\underset{\sim}{S_n}$$

We can therefore rewrite the above and say that the scaled estimation error can be decomposed into two terms. From this, we can now derive a result as a multivariate generalisation of the correlation inequality for the rescaled estimation error $\sqrt{n}(\underset{\sim}{\tilde{\gamma}_n} - \underset{\sim}{\gamma_0})$.

Theorem 3: Orthogonal Decomposition Theorem

Let $\underset{\sim}{\tilde{\gamma}_n}$ be a RAJN estimator. Then, write the rescaled estimation error as

$$\sqrt{n}(\underset{\sim}{\tilde{\gamma}_n} - \underset{\sim}{\gamma_0}) = \underset{\sim}{J}^{-1}\underset{\sim}{S_n} + \underset{\sim}{R_n} \tag{2.10}$$

Then, these two terms are asymptotically independent whereby the covariance of the two terms are zero

$$\begin{pmatrix} \underset{\sim}{J}^{-1}\underset{\sim}{S_n} \\ \underset{\sim}{R_n} \end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(\underset{\sim}{0}, \begin{pmatrix} \underset{\sim}{J}^{-1} & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{\Sigma_Y} - \underset{\sim}{J}^{-1} \end{pmatrix}\Big) \tag{2.11}$$

Proof:(Sketch). Recall that by definition $\underset{\sim}{\Sigma_Y} = Var(\sqrt{n}(\underset{\sim}{\tilde{\gamma}_n} - \underset{\sim}{\gamma_0})$ and therefore

$$Var(\underset{\sim}{R_n}) = Var(\sqrt{n}(\underset{\sim}{\tilde{\gamma}_n} - \underset{\sim}{\gamma_0}) - \underset{\sim}{J}^{-1}\underset{\sim}{S_n}) = \underset{\sim}{\Sigma_Y} - \underset{\sim}{J}^{-1}$$

■

As $\underset{\sim}{\Sigma_Y} - \underset{\sim}{J}^{-1}$ has to be positive semi-definite, the smallest that this extra added error term $\underset{\sim}{R_n}$ can be is the zero vector. Hence, all RAJN estimators are of the form

$$\sqrt{n}(\underset{\sim}{\tilde{\gamma}_n} - \underset{\sim}{\gamma_0}) = \underset{\sim}{J}^{-1}\underset{\sim}{S_n} + \underset{\sim}{R_n}$$

and to find the one with the least variance, we need to find the one without the extra error term $\underset{\sim}{R_n}$.

Proposition 4: Sufficient Condition for RAJN Estimator

Suppose that the estimator $\underset{\sim}{\tilde{\theta}_n}$ can be written in the form

$$\sqrt{n}(\underset{\sim}{\tilde{\gamma}_n} - \underset{\sim}{\gamma_0}) = \underset{\sim}{J}^{-1}\underset{\sim}{S_n} + \underset{\sim}{R_n}$$

Then, the estimator $\underset{\sim}{\tilde{\theta}_n}$ is a RAJN estimator.

Proof: (Sketch). To show that the estimator $\underset{\sim}{\tilde{\theta}_n}$ is **AJN** with the scores, first, note that the rescaled estimation error is a linear combination of the Score vector $\underset{\sim}{S_n}$

$$\sqrt{n}(\underset{\sim}{\tilde{\gamma_n}} - \underset{\sim}{\gamma_0}) = \underset{\sim}{J}^{-1}\underset{\sim}{S_n}$$

therefore, this is clearly asymptotically jointly normal with the scores vector.

To show that the estimator $\underset{\sim}{\tilde{\theta}_n}$ is **regular**, we can apply the Scores version of Le Cam's third lemma whereby under the alternative $\underset{\sim}{\theta_0} + n^{-1/2}\underset{\sim}{h}$

$$\sqrt{n}(\underset{\sim}{\tilde{\gamma_n}} - \underset{\sim}{\gamma_0}) \xrightarrow{d} \mathcal{N}(\underset{\sim}{0}, \underset{\sim}{J}^{-1})$$

■

This is analogous to restricting to a class of unbiased estimators. Since $\underset{\sim}{\Sigma_Y} - \underset{\sim}{J}^{-1}$ must be positive semi-definite, the *smallest* this *extra added error term* $\underset{\sim}{R_n}$ can be is asymptotically a zero vector. Then, to find the estimator with the minimum variance, we need to find the one without the $\underset{\sim}{R_n}$ remainder.

Definition 7: Asymptotically Efficient Estimator

Let $\underset{\sim}{\hat{\theta}_n}$ be a RAJN estimator. Then $\underset{\sim}{\hat{\theta}_n}$ is asymptotically efficient at $\underset{\sim}{\theta_0}$ if we can write the rescaled estimation error as

$$\sqrt{n}(\underset{\sim}{\tilde{\gamma_n}} - \underset{\sim}{\gamma_0}) = \underset{\sim}{J}^{-1}\underset{\sim}{S_n} + o_p(1) \xrightarrow{d} \mathcal{N}(\underset{\sim}{0}, \underset{\sim}{J}^{-1})$$

Remark 2.12 *This definition follows from the orthogonal decomposition theorem whereby the remainder $\underset{\sim}{R_n}$ now gets put into the $o_p(1)$ term.*

Remark 2.13 *We can therefore say that RAJN estimators are equivalent to unbiased estimators and that if it is asymptotically efficient, then it is similar to being the minimum variance estimator.*

2.1.6 Estimating in the presence of Nuisance Parameters

In general, we may not be interested in estimating the whole parameter vector $\underset{\sim}{\gamma}$, rather only a smooth lower-dimensional function

$$\underset{\sim}{\theta} : \mathbb{R}^d \to \mathbb{R}^{d_\theta}$$

We can write it as a function of the parameter

$$\underset{\sim}{\theta} = \underset{\sim}{\theta}(\underset{\sim}{\gamma})$$

In that case, it turns out that the form of A.E estimators of $\underset{\sim}{\theta}$ follows from an application of the delta method in probability.

Suppose $\underset{\sim}{\hat{\gamma}_n}$ is asymptotically efficient for $\underset{\sim}{\gamma}$ at $\underset{\sim}{\gamma_0}$. Then, by definition of asymptotically efficient estimators, we have that

$$\sqrt{n}\left(\underset{\sim}{\hat{\gamma}_n} - \underset{\sim}{\gamma_0}\right) \xrightarrow{d} \mathcal{N}(\underset{\sim}{0}, \underset{\sim}{J}^{-1}(\underset{\sim}{\gamma_0}))$$

Then, the estimator $\underset{\sim}{\hat{\theta}_n} = \underset{\sim}{\theta}(\underset{\sim}{\hat{\gamma}_n})$ turns out to be A.E if $\underset{\sim}{\theta}(\cdot)$ has a smooth gradient/Jacobian matrix $\underset{\sim}{\dot{\theta}}(\cdot)$.

Proposition 5: Smooth Function of efficient estimators are efficient

Suppose that $\underset{\sim}{\theta}$ is an asymptotically efficient estimator. Then, by the delta method in probability, the new estimator defined by taking a smooth function $\underset{\sim}{\theta}$ of the estimator, denoted by $\underset{\sim}{\underset{\sim}{\hat{\gamma}_n}}$ is an asymptotically efficient estimator of $\underset{\sim}{\theta}(\underset{\sim}{\gamma})$.

Proof: We use the definition of asymptotically efficient estimator and apply the delta method in probability

$$\begin{aligned} \sqrt{n}[\underset{\sim}{\hat{\theta}_n} - \underset{\sim}{\theta}(\underset{\sim}{\gamma_0})] &= \sqrt{n}[\underset{\sim}{\theta}(\underset{\sim}{\hat{\gamma}_n}) - \underset{\sim}{\theta}(\underset{\sim}{\gamma_0})] \qquad (2.12)\\ &= \underset{\sim}{\dot{\theta}}(\underset{\sim}{\gamma_0})\left\{\sqrt{n}(\underset{\sim}{\hat{\gamma}_n}) - \underset{\sim}{\gamma_0}\right\} + o_p(1) \qquad (2.13)\\ &\xrightarrow{d} \mathcal{N}(\underset{\sim}{0}, \underset{\sim}{\dot{\theta}}(\underset{\sim}{\gamma_0})\underset{\sim}{J}^{-1}\underset{\sim}{\dot{\theta}}(\underset{\sim}{\gamma_0})^T) \qquad (2.14)\end{aligned}$$

■

Now, we have seen that from the previous proposition, we have that for a function of an asymptotically efficient estimator

$$\sqrt{n}[\underset{\sim}{\hat{\theta}_n} - \underset{\sim}{\theta}(\underset{\sim}{\gamma_0})] \xrightarrow{d} \mathcal{N}(\underset{\sim}{0}, \underset{\sim}{\dot{\theta}}(\underset{\sim}{\gamma_0})\underset{\sim}{J}^{-1}\underset{\sim}{\dot{\theta}}(\underset{\sim}{\gamma_0})^T)$$

Definition 8: Effective Information Matrix

The inverse of the limiting covariance matrix is called the **effective information**

$$J_{\theta}^{*} = \left[\underset{\sim}{\dot{\theta}}(\underset{\sim}{\gamma_0})\underset{\sim}{J}^{-1}\underset{\sim}{\dot{\theta}}(\underset{\sim}{\gamma_0})^T\right]^{-1} \qquad (2.15)$$

2.1.7 Special Case of Estimating in presence of nuisance parameters

The canonical special case is where $\underset{\sim}{\theta}(\cdot)$ returns a sub-vector of $\underset{\sim}{\gamma}$. This is a smooth function. In that case, suppose we may partition $\underset{\sim}{\gamma}$ according to

$$\underset{\sim}{\gamma} = \begin{pmatrix} \underset{\sim}{\theta} \\ \underset{\sim}{\eta} \end{pmatrix}$$

whereby $\underset{\sim}{\theta}$ is $d_\theta-$dimensional, $\underset{\sim}{\eta}$ is d_η dimensional and $d = d_\theta + d_\eta$.

If the LAN property holds at $\underset{\sim}{\gamma_0} = \begin{pmatrix} \underset{\sim}{\theta_0} \\ \underset{\sim}{\eta_0} \end{pmatrix}$, we may partition the score and information as

$$\underset{\sim}{S_n} = \begin{pmatrix} \underset{\sim}{S_\theta} \\ \underset{\sim}{S_\eta} \end{pmatrix}$$

and

$$\underset{\sim}{J} = \begin{bmatrix} \underset{\sim}{J_{\theta\theta}} & \underset{\sim}{J_{\theta\eta}} \\ \underset{\sim}{J_{\eta\theta}} & \underset{\sim}{J_{\eta\eta}} \end{bmatrix}$$

We can now state a very important result which we will use in semiparametric estimation.

Proposition 6: RAJN Estimators are independent of nuisance parameters

Assume that the LAN property holds at $\underset{\sim}{\gamma_0} = \begin{pmatrix} \underset{\sim}{\theta_0} \\ \underset{\sim}{\eta_0} \end{pmatrix}$. Now, suppose that we have an AJN estimator of $\underset{\sim}{\hat{\theta}_n}$ of $\underset{\sim}{\theta}$ at $\begin{pmatrix} \underset{\sim}{\theta_0} \\ \underset{\sim}{\eta_0} \end{pmatrix}$ so that under $\underset{\sim}{\gamma_0} = \begin{pmatrix} \underset{\sim}{\theta_0} \\ \underset{\sim}{\eta_0} \end{pmatrix}$ the scaled estimation error $\underset{\sim}{Y_n} = \sqrt{n}(\underset{\sim}{\tilde{\theta}_n} - \underset{\sim}{\theta_0})$ satisfies

$$\begin{bmatrix} \underset{\sim}{S_\theta} \\ \underset{\sim}{S_\eta} \\ \underset{\sim}{Y_n} \end{bmatrix} \xrightarrow{d} \mathcal{N}\Big(\underset{\sim}{0}, \begin{bmatrix} \underset{\sim}{J_{\theta,\theta}} & \underset{\sim}{J_{\theta,\eta}} & \Sigma_{Y,\theta} \\ \underset{\sim}{J_{\eta,\theta}} & \underset{\sim}{J_{\eta,\eta}} & \Sigma_{Y,\eta} \\ \Sigma_{Y,\theta} & \Sigma_{Y,\eta} & \Sigma_{Y,Y} \end{bmatrix}\Big)$$

for some positive semi-definite matrix $\Sigma_{Y,Y}$.
If in addition, $\underset{\sim}{\tilde{\theta}_n}$ is regular at $\underset{\sim}{\theta_0}$, this forces both

$$\underset{\sim}{\Sigma_{Y,\theta}} = \underset{\sim}{I}$$

and

$$\underset{\sim}{\Sigma_{Y,\eta}} = \underset{\sim}{0}$$

Such a $\underset{\sim}{Y_n}$ is called **asymptotically uncorrelated** with the nuisance scores. We also call it **orthogonal**.

Remark 2.14 *Intuitively, a small change in the nuisance parameter $\underset{\sim}{\eta}$ has no effect on the estimator of $\underset{\sim}{\theta}$ as $\underset{\sim}{\Sigma_{Y,\eta}} = \underset{\sim}{0}$.*

We are now interested in defining an asymptotically efficient estimator under this setting. First, we require a few definitions.

Definition 9: Effective Scores

The effective scores $\underset{\sim}{S_\theta}^*$ is given by by a linear combination of the scores of the parameters of interest $\underset{\sim}{S_\theta}$ and scores of the nuisance parameters $\underset{\sim}{S_\eta}$

$$\underset{\sim}{S_\theta}^* = \underset{\sim}{S_\theta} - \underset{\sim}{J_{\theta,\eta}} J_{\eta,\eta}^{-1} \underset{\sim}{S_\eta} \tag{2.16}$$

Remark 2.15 *We can interpret the effective scores as the residual from regression the scores of the parameters of interest on the nuisance parameters scores. This linear combination of $\underset{\sim}{S_\theta}$ and $\underset{\sim}{S_\eta}$ is asymptotically uncorrelated with the nuisance scores.*

Remark 2.16 *When the co-information $J_{\theta,\eta} = 0$, then the effective scores is identical to the regular scores. That is, the nuisance parameters η do not matter.*

Definition 10: Effective Information

The effective information is given by

$$\underset{\sim}{J_\theta}^* = \left(\underset{\sim}{J_{\theta,\theta}} - \underset{\sim}{J_{\theta,\eta}} \underset{\sim}{J_{\eta,\eta}}^{-1} \underset{\sim}{J_{\eta,\theta}} \right) \tag{2.17}$$

Remark 2.17 *The limiting covariance matrix of $\underset{\sim}{S_\theta}^*$ is the effective information.*

We can now define what an asymptotically efficient estimator under this setting is.

Definition 11: Asymptotically Efficient Estimator in presence of nuisance parameters

An A.E estimator $\underset{\sim}{\hat{\theta}_n}$ in the presence of nuisance parameters is of the form

$$\sqrt{n}(\underset{\sim}{\hat{\theta}_n} - \underset{\sim}{\theta_0}) = (J_\theta^*)^{-1} S_\theta^* + o_p(1) \tag{2.18}$$

Furthermore, the limiting distribution is given by

$$(J_\theta^*)^{-1} S_\theta^* + o_p(1) \xrightarrow{d} \mathcal{N}(\underset{\sim}{0}, (J_\theta^*)^{-1})$$

Recall the definition of the effective information

$$\underset{\sim}{J_\theta}^* = \left(\underset{\sim}{J_{\theta,\theta}} - \underset{\sim}{J_{\theta,\eta}} \underset{\sim}{J_{\eta,\eta}}^{-1} \underset{\sim}{J_{\eta,\theta}} \right)$$

The larger the information matrix (where we use the partial ordering that $\underset{\sim}{M} \geq \underset{\sim}{N}$ iff $\underset{\sim}{M} - \underset{\sim}{N}$ is positive semi-definite), the better the estimator as we get a smaller limiting covariance matrix. That is, if $\underset{\sim}{M} - \underset{\sim}{N}$ is large, then $(\underset{\sim}{M} - \underset{\sim}{N})^{-1}$ will be small.

It is useful to compare the 2 cases where the nuisance parameter is known and unknown.

1. If $\underset{\sim}{\eta_0}$ is known, the information for $\underset{\sim}{\theta}$ is $\underset{\sim}{J_{\theta,\theta}}$

2. If $\underset{\sim}{\eta_0}$ is unknown, the effective information for $\underset{\sim}{\theta}$ is $\underset{\sim}{J_\theta}^* \leq \underset{\sim}{J_{\theta,\theta}}$.

Therefore, $\underset{\sim}{J_\theta}^*$ will be bigger if $\underset{\sim}{\eta_0}$ is known. However, if $\underset{\sim}{\eta_0}$ is unknown, then $\underset{\sim}{\eta_0}$ will be smaller.

We can conclude that not knowing $\underset{\sim}{\eta_0}$ in general leads to a loss of information unless $\underset{\sim}{S_\theta}$ and $\underset{\sim}{S_\eta}$ are asymptotically uncorrelated.

2.1.8 i.i.d case of parametric estimation

We are now interested in the special case where the data $\underset{\sim}{X_n}$ consists of i.i.d random vectors

$$\underset{\sim}{X_n} = (\underset{\sim}{Y_1}, \cdots, \underset{\sim}{Y_n})^T \tag{2.19}$$

where $\underset{\sim}{Y_i}$ is a vector for an observation.

The **log-likelihood ratio** can now be written as a sum

$$\underset{\sim}{L_n}(\underset{\sim}{X_n}|\underset{\sim}{\gamma_1};\underset{\sim}{\gamma_0}) = \sum_{i=1}^{n} log \left\{ \frac{p(\underset{\sim}{Y_i};\underset{\sim}{\gamma_1})}{p(\underset{\sim}{Y_i};\underset{\sim}{\gamma_0})} \right\} \tag{2.20}$$

whereby $p(\cdot;\underset{\sim}{\gamma})$ is the common density for each observation $\underset{\sim}{Y_i}$.

Definition 12: Score Function

We define the **score function** to be the gradient of the log density

$$\underset{\sim}{\dot{\ell}}(\underset{\sim}{Y};\underset{\sim}{\gamma}) = \begin{pmatrix} \frac{\partial}{\partial\gamma_1} log\ p(\underset{\sim}{Y};\underset{\sim}{\gamma}) \\ \cdots \\ \frac{\partial}{\partial\gamma_d} log\ p(\underset{\sim}{Y};\underset{\sim}{\gamma}) \end{pmatrix}^T \tag{2.21}$$

The **score vector** can also be written as a sum of the **score function**

$$\underset{\sim}{S_n} = \underset{\sim}{S_n}(\underset{\sim}{\gamma_0}) = n^{-1/2} \sum_{i=1}^{n} \underset{\sim}{\ell}^{\circ}(\underset{\sim}{Y_i};\underset{\sim}{\gamma_0}) \tag{2.22}$$

The **information matrix** can also be expressed in terms of the score function

$$\underset{\sim}{J} = \int \underset{\sim}{\ell}^{\circ} \underset{\sim}{\ell}^{\circ\, T} d\mathbb{P}_{\underset{\sim}{\gamma_0}} = \mathbb{E}_{\underset{\sim}{\gamma_0}}(\underset{\sim}{\ell}^{\circ} \underset{\sim}{\ell}^{\circ\, T}) \tag{2.23}$$

where $\mathbb{P}_{\underset{\sim}{\gamma_0}}$ is the probability measure defined on the sample space $\mathcal{Y}_i$, defined for each $\underset{\sim}{Y_i}$.

An asymptotically efficient (AE) estimator $\underset{\sim}{\hat{\gamma_n}}$ now satisfies

$$\sqrt{n}\Big(\underset{\sim}{\hat{\gamma_n}} - \underset{\sim}{\gamma_0}\Big) = \underset{\sim}{J}^{-1}\underset{\sim}{S_n} + o_p(1) = n^{-1/2}\sum_{i=1}^{n} \underset{\sim}{J}^{-1}\,\underset{\sim}{\ell^\circ}\,\Big(\underset{\sim}{Y_i};\underset{\sim}{\gamma_0}\Big) + o_p(1) \tag{2.24}$$

$$= n^{-1/2}\sum_{i=1}^{n} \underset{\sim}{\tilde{\ell}}(\underset{\sim}{Y_i};\underset{\sim}{\gamma_0}) + o_p(1) \tag{2.25}$$

The expression

$$\underset{\sim}{\tilde{\ell}}(\cdot;\underset{\sim}{\gamma_0}) = \underset{\sim}{J}^{-1}\underset{\sim}{\tilde{\ell}}(\cdot;\underset{\sim}{\gamma_0}) \tag{2.26}$$

is known as the **efficient influence function**.

Any estimator $\underset{\sim}{\tilde{\gamma_n}}$ which can be written in the form

$$\sqrt{n}\Big[\underset{\sim}{\tilde{\gamma_n}} - \underset{\sim}{\gamma_0}\Big] = n^{-1/2}\sum_{i=1}^{n} g(\underset{\sim}{Y_i};\underset{\sim}{\gamma_0}) + o_p(1) \tag{2.27}$$

is said to be **asymptotically linear**.

Proposition 2.18 *Any asymptotically linear estimator is an AJN estimator.*

The function $g(\cdot;\underset{\sim}{\gamma_0})$ is called the **influence function** of the estimator $\underset{\sim}{\hat{\gamma_n}}$.

Estimators which are smooth functions of sample moments or quantiles are asymptotically linear.

If we have a nuisance parameter $\underset{\sim}{\gamma} = \begin{bmatrix}\underset{\sim}{\theta}\\ \underset{\sim}{\gamma}\end{bmatrix}$, we can write the score function

$$\underset{\sim}{\ell^\circ} = \begin{pmatrix}\underset{\sim}{\ell^\circ_\theta}\\ \underset{\sim}{\ell^\circ_\eta}\end{pmatrix} \tag{2.28}$$

The **efficient score function** for $\underset{\sim}{\theta}$ is

$$\underset{\sim}{\ell_\theta}^{\circ\,*} = \underset{\sim}{\ell_\theta}^{\circ} - \underset{\sim\theta\eta}{J}\,\underset{\sim\eta\eta}{J}^{-1}\underset{\sim\eta\theta}{J}\,\underset{\sim}{\ell_\eta}^{\circ} \tag{2.29}$$

and the **efficient influence function** for estimating $\underset{\sim}{\theta}$ is

$$\underset{\sim}{\ell_\theta}^{\circ\,*} = \Big(\underset{\sim\theta}{J^*}\Big)^{-1}\underset{\sim}{\ell_\theta}^{\circ\,*} \tag{2.30}$$

STAT5610: Advanced Inference

3. Construction of A.E estimators

3.1.9 Construction of A.E estimators

We have seen a characterisation of optimal estimators. We are now interested in being able to construct optimal estimates. First, recall that an A.E estimator $\underset{\sim}{\hat{\theta_n}}$ of $\underset{\sim}{\theta}$ at $\underset{\sim}{\gamma_0} = \begin{pmatrix} \underset{\sim}{\theta_0} \\ \underset{\sim}{\eta_0} \end{pmatrix}$ if it satisfies

$$\sqrt{n}(\underset{\sim}{\hat{\theta_n}} - \underset{\sim}{\theta_0}) = (\underset{\sim}{J_\theta}^*)^{-1}\underset{\sim}{S_\theta}^* + o_p(1) \tag{3.31}$$

or alternatively

$$\begin{aligned} \underset{\sim}{\hat{\theta_n}} &= \underset{\sim}{\theta_0} + n^{-1/2}(\underset{\sim}{J_\theta}^*)^{-1}\underset{\sim}{S_\theta}^* + o_p(n^{-1/2}) && (3.32) \\ &= \underset{\sim}{\theta_0} + n^{-1/2}(\underset{\sim}{J_\theta}^*(\underset{\sim}{\theta_0}, \underset{\sim}{\eta_0}))^{-1}\underset{\sim}{S_\theta}^*(\underset{\sim}{\theta_0}, \underset{\sim}{\eta_0}) + o_p(n^{-1/2}) && (3.33) \end{aligned}$$

remembering that the effective information and score vector are functions of both the data and the parameters.

It turns out that under extra regularity conditions, that is, the LAN condition now holds in a neighbourhood of the true estimate, alongside other conditions, we can plug in a reasonable $\sqrt{n}-$consistent initial guess estimator into 3.33 and obtain an A.E estimator.

We now state a condition needed whereby we assume the behaviour of the scores vector in a nearby neighbourhood around the point of interest.

Proposition 7: Regular Scores Condition

The full score vector

$$\underset{\sim}{S_n}(\underset{\sim}{\gamma} = \begin{pmatrix} S_\theta(\underset{\sim}{\theta}, \underset{\sim}{\eta}) \\ S_\eta(\underset{\sim}{\theta}, \underset{\sim}{\eta}) \end{pmatrix}$$

satisfies the following condition that for any $0 < M < \infty$

$$\sup_{|h| \leq M} |\underset{\sim}{S_n}(\underset{\sim}{\gamma_0} + n^{-1/2}\underset{\sim}{h}) - \underset{\sim}{S_n}(\underset{\sim}{\gamma_0}) + \underset{\sim}{J}(\underset{\sim}{\gamma_0})\underset{\sim}{h}) \xrightarrow{P} |0 \tag{3.34}$$

Remark 3.19 *We can interpret this as taking the biggest remainder such that h is bounded and seeing it tends to 0.*

Proposition 8: Continuous Information

The information matrix

$$\underset{\sim}{J}(\underset{\sim}{\gamma})$$

is continuous at $\underset{\sim}{\gamma_0}$.

We can now define the reasonable estimator we shall plug in.

Definition 13: $\sqrt{n}-$consistent estimator

An estimator $\underset{\sim}{\tilde{\gamma_n}} = \begin{pmatrix} \underset{\sim}{\tilde{\theta_0}} \\ \underset{\sim}{\tilde{\eta_0}} \end{pmatrix}$ is said to be $\sqrt{n}-$consistent if

$$\sqrt{n}(\underset{\sim}{\tilde{\gamma_n}} - \underset{\sim}{\gamma_0}) = O_p(1) \tag{3.35}$$

whereby $O_p(1)$ is bounded in probability such that for all $\epsilon > 0$, there exists $0 < M_\epsilon < \infty$ such that $lim\ \sup \mathbb{P}(|X_n|> M) < \epsilon$.

Remark 3.20 *Finding a $\sqrt{n}-$consistent estimator is a mild condition. They are easy to find and compute.*

Theorem 4: A.E. Estimator Construction Theorem

Under the regular scores condition and continuous information, an asymptotically efficient estimator $\underset{\sim}{\hat{\gamma_n}}$ can be constructed from a $\sqrt{n}-$consistent estimator by the formula

$$\underset{\sim}{\hat{\gamma_n}} = \underset{\sim}{\tilde{\gamma_n}} + n^{-1/2}\underset{\sim}{J}(\underset{\sim}{\tilde{\gamma_n}})^{-1}\underset{\sim}{S_n}(\underset{\sim}{\tilde{\gamma_n}}) \tag{3.36}$$

3.1.10 Hypothesis Testing

A parallel theory of hypothesis testing exists. We have the notion of **regular tests** whereby the limiting distribution of test statistics under local alternatives do not depend on a change in the nuisance parameters, only in changes of the parameters of interest. We also have quadratic form test statistics which are essentially quadratic form of AJN random variables.

Theorem 5: Optimal Test Statistic

Optimal tests have test statistics of the form

$$T_n = (\underset{\sim}{S_\theta}^*)^T(\underset{\sim}{J_\theta}^*)^{-1}(\underset{\sim}{S_\theta}^*) + o_p(1) \tag{3.37}$$

Under appropriate regularity conditions, we can construct optimal test statistics by plugging in $\sqrt{n}-$consistent estimates of nuisance parameters into equation 3.37.

Many common test statistics such as the log-likelihood ratio, Rao-Score test, and Wald test can be shown to satisfy equation 3.37.

STAT5610: Advanced Inference

4. Introduction to Semiparametric Models

4.2 Semiparametric Estimation

4.2.1 Introduction to Semiparametric Models

The canonical example of a semiparametric model is the i.i.d location model.

Definition 14: i.i.d Location Model

Let $\underset{\sim}{X}_n = (X_1, \cdots, X_n)^T$ consists of i.i.d random variables whose common density is given by

$$p(x; \theta, f) = f(x - \theta) \tag{4.38}$$

whereby $\theta \in \mathbb{R}$ and $f(\cdot)$ is a probability density on $\mathbb{R}$ with respect to the Lebesgue measure that is *centered*.

Here, we now have **two** parameters θ and $f(\cdot)$ which to estimate.

Different interpretations of *centered* lead to different models.

4.2.2 Constraint-Defined Location Model

The first interpretation of centered will involve the parameter $f(\cdot)$ satisfying an integral constraint.

Definition 15: Constraint Function

Suppose that there exists a function $w(\cdot)$ such that

1. $$\int_{-\infty}^{\infty} w(x) f(x) dx = 0$$

2. $$\int_{-\infty}^{\infty} w^2(x) f(x) dx < \infty$$

This is known as the constraint function.

We now give some examples of such constraint functions $w(\cdot)$.

Example 4.21 *We can define the function $w(\cdot)$ to be*

$$w(x) = x \tag{4.39}$$

This implies that the function parameter $f(\cdot)$ has mean 0 and the θ parameter is the mean.

Example 4.22 *We can define the function $w(\cdot)$ to be*

$$w(x) = sign(x) = 1\{x > 0\} - 1\{x < 0\} \tag{4.40}$$

This implies that the function parameter $f(\cdot)$ has median 0.

Example 4.23 *We can define the function $w(\cdot)$ to be*

$$w(x) = \begin{cases} c & x > c \\ x & |x| \leq c \\ -c & x < -c \end{cases} \tag{4.41}$$

for some constant $0 < c < \infty$. This implies that the function parameter $f(\cdot)$ has a Winsorised mean at $c = 0$.

Remark 4.24 *This function $w(\cdot)$ is used in robust statistics whereby we set a threshold c and pull outliers in the data to the threshold c.*

4.2.3 Nuisance Scores

We wish to determine how well we can estimate the location parameter θ in the presence of the unknown density $f(\cdot)$, where the true value $f(\cdot)$ is $f_0(\cdot)$.

As foreshadowed, we shall consider *regular parametric submodels*. What we do here is to now take the arbitrary nuisance function $f \in \mathcal{F}$ and index it by an Euclidean parameter $\underset{\sim}{\eta} \in \mathbb{R}^{d_{\underset{\sim}{\eta}}}$.

Definition 16: Regular parametric submodels

A regular parametric submodel is given by

$$p(x;\theta,\underset{\sim}{\eta}) = f_{\underset{\sim}{\eta}}(x-\theta) \tag{4.42}$$

for $\theta \in \mathbb{R}$ and a nuisance parameter $\underset{\sim}{\eta} \in \mathcal{H} \subseteq \mathbb{R}^{d_{\underset{\sim}{\eta}}}$ whereby $\{f_{\underset{\sim}{\eta}} : \underset{\sim}{\eta} \in \mathcal{H}\}$. Furthermore, this parametric family of densities satisfies 3 conditions.
First, the $\{f_{\underset{\sim}{\eta}} : \underset{\sim}{\eta} \in \mathcal{H}\}$ is a parametric family of densities such that

1. $$\int_{-\infty}^{\infty} w(x) f_{\underset{\sim}{\eta}}(x)dx = 0$$

2. $$\int_{-\infty}^{\infty} w^2(x) f_{\underset{\sim}{\eta}}(x)dx = \sigma_w^2(\underset{\sim}{\eta}) < \infty$$

   for all values of $\underset{\sim}{\eta} \in \mathcal{H}$.

Secondly, for some parameter value that corresponds to the true density $\underset{\sim}{\eta_0} \in \mathcal{H}$, the true density is $f_0(\cdot) = f_{\underset{\sim}{\eta_0}}(\cdot)$.
Finally, the LAN condition holds at $\underset{\sim}{\eta_0}$ for the family

$$\{p_n^H(\underset{\sim}{\eta}) : \underset{\sim}{\eta} \in \mathcal{H}\}$$

whereby $p_n^H(\underset{\sim}{\eta})$ is the joint distribution of $\underset{\sim}{Y_n} = (Y_1, \cdots, Y_n)^T$ is i.i.d $f_{\underset{\sim}{\eta}}(\cdot)$.

Remark 4.25 *If we ignore the location θ, we have regular parametric density whereby LAN holds at $\underset{\sim}{\eta_0}$.*

Due to the LAN condition holding in point 3, there exists a vector of **nuisance score functions**

$$\ell^\circ_{\underset{\sim}{\eta}}(\cdot;\underset{\sim}{\eta_0}) = (\ell^\circ_1(\cdot;\underset{\sim}{\eta_0}), \cdots, \ell^\circ_{d_\eta}(\cdot;\underset{\sim}{\eta_0}))^T \tag{4.43}$$

and a corresponding **nuisance score vector**

$$\underset{\sim}{S_{\underset{\sim}{\eta}}} = n^{-1/2}\sum_{i=1}^{n} \ell_{\underset{\sim}{\eta}}{}^\circ(X_i;\underset{\sim}{\eta_0}) = n^{-1/2}\sum_{i=1}^{n}\begin{pmatrix} \ell^\circ_1(X_i;\underset{\sim}{\eta_0}) \\ \cdots \\ \ell^\circ_{d_\eta}(X_i;\underset{\sim}{\eta_0}) \end{pmatrix} \tag{4.44}$$

Here, it is important to note that this is a parametric submodel as $d_{\underset{\sim}{\eta}} < \infty$!

4.2.4 Properties of Nuisance Scores

We will now focus on the properties of the nuisance scores. We will see that all nuisances scores will be orthogonal to the constraint function $w(\cdot)$. From this, this tells us how to estimate θ since in the parametric model, the effective scores is orthogonal to the nuisance scores. This will show us that the optimal score function is a multiple of the constraint function $w(\cdot)$ as $w(\cdot)$ is the only function orthogonal to all possible nuisance scores. Therefore, we wish to show, under mild conditions, the nuisance scores are orthogonal to $w(\cdot)$.

First, we require a theorem that gives us necessary and sufficient conditions for the sum of independent random variables and in a triangular array to be asymptotically normal.

Theorem 4.26 *Suppose that $X_{1,n}, ..., X_{n,n}$ are i.i.d with distribution F_n, whereby the distribution depends on the sample size n. Then, we have that*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_{i,n} \xrightarrow{d} \mathcal{N}(\mu, \sigma^2) \tag{4.45}$$

if and only if for each $\epsilon > 0$

1. $n\mathbb{P}(|X_{1,n} \geq \epsilon\sqrt{n}) \to 0$
2. $Var[X_{1,n}1\{|X_{1,n}|< \epsilon\sqrt{n}\}] \to 0$
3. $\sqrt{n}\mathbb{E}[X_{1,n}1\{|X_{1,n}|< \epsilon\sqrt{n}\}] \to \mu$

Corollary 4.27 *Suppose that $X_{1,n}, ..., X_{n,n}$ are i.i.d F_n and it is known that*

1. $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \xrightarrow{d} \mathcal{N}(\mu, \sigma^2)$ *for some* $\mu \in \mathbb{R}$
2. *The* $\mathbb{E}[X_{i,n}] = 0$
3. *The limiting variance* $\lim_{n\to\infty} Var(X_{1,n}) = \lim_{n\to\infty} (\mathbb{E}[X_{1,n}^2]) = \sigma^2$

This forces the mean $\mu = 0$.

We now come to the most important result which we shall use through the course. That is, the covariance of the nuisance scores and the constraint function $w(\cdot)$ are uncorrelated.

Theorem 6: Covariance of nuisance scores and constaint function is 0

In the constraint-defined location model, for any regular parametric submodel $\{f_{\underset{\sim}{\eta}} : \underset{\sim}{\eta} \in \mathcal{H}\}$ of centered densities for shape, if $\sigma_w^2(\underset{\sim}{\eta}) = \int_{-\infty}^{\infty} w^2(x) f_{\underset{\sim}{\eta}}(x)dx$ is continuous in $\underset{\sim}{\eta}$ then

$$\mathbb{E}_{f_{\underset{\sim}{\eta_0}}}[w(X)\ell_n^\circ[X]] = \int_{-\infty}^{\infty} w(x)\, \underset{\sim}{\ell_n^\circ}\, (x; \underset{\sim}{\eta_0}) f_{\underset{\sim}{\eta_0}}(x)dx = \underset{\sim}{0} \tag{4.46}$$

Proof:(Sketch). Apply the previous corollary as due to the ordinary CLT

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} w(X_i) \xrightarrow{d} \mathcal{N}(0, \sigma_w^2(\underset{\sim}{\eta_0})) \tag{4.47}$$

■

STAT5610: Advanced Inference

5. Semiparametric Information Bound

5.2.5 Constrained Location Model: Semiparametric Information Bound

Recall that the constrained location model is a location family that is centered by an integral constrained. We are now interested in investigating the information bound for such a model.

First, we receap the idea of information in estimation when no nuisance parameters are present. When estimating a parameter vector $\underset{\sim}{\theta}$ in a regular parametric model (i.e LAN holds)

$$\mathcal{P}_n = \{\mathbb{P}_{n,\underset{\sim}{\theta}}\}$$

whereby the distributions are indexed by the parameter $\underset{\sim}{\theta}$ and depends on the sample size n, if there are no nuisance parameters, the limiting variance of an asymptotically normal *regular* estimator is greater than or equal to

$$(n\underset{\sim}{J_{\theta\theta}})^{-1}$$

where $\underset{\sim}{J_{\theta\theta}}$ is the limiting covariance matrix of the Scores.

Now, if nuisance parameters $\underset{\sim}{\eta}$ are present, then in general, there is a loss of information whereby $\underset{\sim}{J_{\theta\theta}}$ is replaced by the effective information

$$\underset{\sim}{J_\theta}^* = \underset{\sim}{J_{\theta\theta}} - \underset{\sim}{J_{\theta\eta}}\underset{\sim}{J_{\eta\eta}}^{-1}\underset{\sim}{J_{\eta\theta}} \leq \underset{\sim}{J_{\theta\theta}} \tag{5.48}$$

Here, the matrix

$$\begin{pmatrix} \underset{\sim}{J_{\theta\theta}} & \underset{\sim}{J_{\theta\eta}} \\ \underset{\sim}{J_{\eta\theta}} & \underset{\sim}{J_{\eta\eta}} \end{pmatrix}$$

is the limiting covariance matrix of the Scores vector

$$\begin{pmatrix} \underset{\sim}{S_\theta} \\ \underset{\sim}{S_\eta} \end{pmatrix}$$

The reverse is also true: if some or all nuisance parameters are held fixed or regarded as known, the information in general increases. That is, our model has nuisance parameters and the more of the nuisance parameters that we know or fix, the higher the effective information. Equivalently, this is akin to that as we fix more parameters in our model, the smaller the model will be and hence the higher the information.

From this, we introduce the notation for the effective information for $\underset{\sim}{\theta}$ in the presence of nuisance parameters as

$$\underset{\sim}{J_\theta}^*(\mathcal{P}_n|\mathbb{P}_{n\underset{\sim}{\theta_0},\underset{\sim}{\eta_0}}) \tag{5.49}$$

the effective information for $\underset{\sim}{\theta}$ at $\mathbb{P}_{n\underset{\sim}{\theta_0}\underset{\sim}{\eta_0}}$ within $\mathcal{P}_n$. This matches the intuition that the effective information matrix depends on 2 things

1. The actual model we are playing with $\mathcal{P}_n$

2. The specific value of the model we are looking at $\mathbb{P}_{n\underset{\sim}{\theta_0}\underset{\sim}{\eta_0}}$

We now take what we discussed thus far and look at the **semiparametric case**, whereby the nuisance parameter η is now the density function $f(\cdot)$ that satisfies the integral constraint with $w(\cdot)$.

First, we write $\mathbb{P}_{n\theta f}$ for the joint distribution of $X_1, ..., X_n$ which are i.i.d with the density given by

$$p(x;\theta,f) = f(x-\theta) \tag{5.50}$$

On top of this, we now denote

$$\mathcal{P}_n = \{\mathbb{P}_{n\theta f}\} \tag{5.51}$$

to be the semiparametric model which we are interested in exploring.

Definition 17: Lower Information Bound for semiparametric model

We define the effective information for estimating θ within the semiparametric model $\mathcal{P}_n = \{\mathbb{P}_{n\theta f}\}$ by

$$J^*_\theta(\mathcal{P}_n|\mathbb{P}_{n\theta_0 f_0}) = \inf\{J^*_\theta(\mathcal{Q}_n|Q_{n\theta_0\underset{\sim}{\eta_0}}) : \mathcal{Q}_n \subseteq \mathcal{P}_n\} \tag{5.52}$$

whereby $\mathcal{Q}_n$ is a regular parametric submodel of the semiparametric model $\mathcal{P}_n$, and $Q_{n\theta_0\underset{\sim}{\eta_0}} = \mathbb{P}_{n\theta_0 f_0}$.

Remark 5.28 *Here, the effective information for estimating θ in the semiparametric model $\mathcal{P}_n$ is the greatest lower bound of the effective information for all parametric submodels $\mathcal{Q}_n$ that are subsets of the semiparametric model $\mathcal{P}_n$. This is due to the fact that as the semiparametric model always have more unknown parameters compared to the parametric submodels, it will always have a lower effective information compared to the parametric submodels.*

From this definition, we can see that any regular parametric submodel can be embedded into semiparametric models such that we can identify a density function as a member of both. Furthermore, each parametric submodel has an information for estimating θ at that point. The information for the semiparametric model can be no more than the information for any of the parametric submodels. This follows the intuition that larger models with more parameters will have less information. That is, the semiparametric model $\mathcal{P}_n$ is much larger than all the parametric submodels $\mathcal{Q}_n$ and therefore the semiparametric model will have a smaller effective information compared to all the parametric submodels $\mathcal{Q}_n$. From this, we can conclude that the effective information for the semiparametric model is the infimum of the effective information of all the parametric submodels $Q_{n\theta_0\underset{\sim}{\eta_0}}$ which coincide with the semiparametric model $\mathbb{P}_{n\theta_0 f_0}$

Our strategy for determining the effective information of the semiparametric model is as follows.

Proposition 9: Strategy to determine effective information for semiparametric model $\mathcal{P}_n$

There is a two step process to determining the effective information for the semiparametric model $\mathcal{P}_n$.

1. Obtain an upper bound to the infimum of the effective information of parametric submodels $J^*_\theta(\mathcal{Q}_n|\underset{\sim}{Q_{n\theta_0\eta_0}})$ by constructing a suitable sequence of regular parametric submodels $\mathcal{Q}_n$ which is equivalent to finding a lower bound to the limiting variance of asymptotically normal regular estimators.
2. Identify a suitable estimator and/or influence function which obtains the bound.

Remark 5.29 *First, recall that when we construct a sequence of regular parametric submodels, computing the effective information in the submodel is equivalent to finding the lower bound to the limiting variance of an asymptotically normal regular estimator.*

In the second point, a suitable estimator here is one that is RAJN estimator whereby its covariance with the Scores of the parameter of interest is 1 and orthogonal to all nuisance scores. From this, we want to find a lower bound to the limiting variance of all such regular estimators. Therefore, if we can find a regular estimator that obtains this lower bound, we have found an optimal estimator.

5.2.6 The structure of the space of Score functions

We now look at some results from functional analysis as we wish to be able to describe the function space that the Score functions live in.

Regular parametric i.i.d (sub)-models may be identified with their Score functions $\underset{\sim}{\ell}^\circ(\cdot)$. This is because due to the i.i.d condition, we are able to write their Score vectors as

$$S_\theta = n^{-1/2}\sum_{i=1}^{n}\underset{\sim}{\ell}^\circ(X_i) \tag{5.53}$$

Proposition 10: Properties of Score Functions

For i.i.d models under the LAN property, the Score functions satisfy two properties

1. The Score function has mean zero
$$\mathbb{E}_0\left[\underset{\sim}{\ell}^\circ(X)\right] = \int_{-\infty}^{\infty}\underset{\sim}{\ell}^\circ(x)f_0(x)dx = 0 \tag{5.54}$$
2. The Score function has finite variance due to having a finite second moment
$$\mathbb{E}_0\left[\underset{\sim}{\ell}^\circ(X)^2\right] = \int_{-\infty}^{\infty}\ell^\circ(x)^2 f_0(x-\theta_0)dx < \infty. \tag{5.55}$$

Proof: Recall that under LAN, the Score vector has property

$$\underset{\sim}{S_n} \xrightarrow{d} \mathcal{N}(0, \underset{\sim}{J})$$

whereby the covariance matrix $\underset{\sim}{J}$ is finite. ■

For convenience, we shall denote

$$p_0(x) = f_0(x - \theta_0).$$

We can now define an important class of functions.

Definition 18: Space of square-integrable functions

We define the space of square-integrable function under $p_0(\cdot)$ as

$$L_2 = L_2(p_0) = \{a(\cdot) : \int a^2(x)p_0(x)dx < \infty\} \tag{5.56}$$

Remark 5.30 *This is in fact the space of equivalence classes of functions that are square-integrable with respect to the density $p_0(\cdot)$. That is, all the function $a(\cdot)$ that has finite second moment with respect to the density $p_0(\cdot)$.*

The space L_2 has certain structures we can exploit.

Proposition 11: Norm and inner product in L_2-space of functions

Let $L_2 = L_2(p_0)$ be the space of square-integrable functions with respect to p_0. We define the L_2 inner product to be

$$\langle a, b\rangle_0 = \int_{-\infty}^{\infty} a(x)b(x)p_0(x)dx \tag{5.57}$$

We then define the L_2-norm to be

$$||a||_0 = \sqrt{\langle a, a\rangle_0} \tag{5.58}$$

$$= \sqrt{\int_{-\infty}^{\infty} a^2(x)p_0(x)dx} \tag{5.59}$$

The norm $||a||_0$ behaves much like the length of a vector whereby the inner $\langle a, b\rangle_0$ behaves like a dot product.

Definition 19: Orthogonal Functions

Let L_2 be the space of square integrable functions. Let $a(\cdot), b(\cdot) \in L_2$. Then, we say that the functions a and b are **orthogonal** if

$$\langle a, b\rangle_0 = 0. \tag{5.60}$$

We have seen earlier that the expectation of the Score function is 0

$$\mathbb{E}_0\left[\underset{\sim}{\ell}^\circ(X)\right] = \int_{-\infty}^{\infty} \ell^\circ(x)p_0(x)dx = 0$$

In fact, this is equivalent to the Score function being orthogonal to the constant function 1

$$\mathbb{E}_0\left[\underset{\sim}{\ell}^\circ(X)\right] = \int_{-\infty}^{\infty} \ell^\circ(x)p_0(x)dx = \int_{-\infty}^{\infty} \ell^\circ(x)(1)p_0(x)dx = \langle \underset{\sim}{\ell}^\circ, 1\rangle_0 = 0$$

We can therefore define a subspace of the space of square-integrable functions L_2 that satisfies such a property.

Definition 20: Space of square-integrable functions orthogonal to constants

We define the space

$$L_2^0 = \{a \in L_2 : \int a(x)p_0(x)dx = \langle 1, a\rangle_0 = 0\} \tag{5.61}$$

to be space of square integrable functions $a(\cdot)$ such that they are orthogonal to constant functions 1. Alternatively, we can interpret this as the space of all square integrable functions $a(\cdot)$ such that $a(\cdot)$ has mean zero

$$\mathbb{E}_0[a(X)] = 0 \tag{5.62}$$

if $X \sim p_0$.

Therefore, we can interpret that a square integrable function having mean zero is equivalent to being orthogonal to the constant function.

Proposition 12: Score function is orthogonal to constants

Let L_2^0 be the space of square integrable functions orthogonal to constants. Then, the Score function

$$\underset{\sim}{\ell}^\circ \in L_2^0 \tag{5.63}$$

If we now restrict our attention to the space L_2^0, the norm and inner product has statistical interpretation.

Proposition 13: Statisitcal interpretation of norm and inner product in L_2^0

Define the function L_2^0 to be the space of square integrable functions that are orthogonal to constants. Let $a(\cdot), b(\cdot) \in L_2^0$. The inner product has the interpretation of being the covariance

$$\langle a, b\rangle_0 = Cov[a(X), b(X)] \tag{5.64}$$

if $X \sim p_0$. Furthermore, the norm of a function has the interpretation of being the standard deviation

$$||a||_0 = SD[a(X)] \tag{5.65}$$

Therefore, we can interpret functions $a(X), b(X)$ being orthogonal in L_2^0 as the functions $a(X), b(X)$ being **uncorrelated**.

We now show why this interpretation is important to us.

Corollary 5.31 *(Covariance and variance of Score functions) Consider the i.i.d 2-parameter model $\{\mathbb{P}_{n\theta\eta}\}$ and let ℓ_θ° and ℓ_η° be the Score function for the parameter of interest and nuisance parameter respectively.*

The covariance between these two Score functions can be expressed as the inner product

$$J_{\theta\eta} = \mathbb{E}[\ell_\theta^\circ \ell_\eta^\circ] = \langle \ell_\theta^\circ, \ell_\eta^\circ \rangle \tag{5.66}$$

The variance of a Score function be expressed as the squared norm

$$J_{\theta\theta} = ||\ell_\theta^\circ||_0^2; \qquad J_{\eta\eta} = ||\ell_\eta^\circ||_0^2 \tag{5.67}$$

We now want to construct a special type of basis for the function space L_2.

Definition 21: Complete Orthonormal Basis

Define the function space L_2. A complete orthonormal basis to L_2 is a countable collection of functions $\{b_j(\cdot)\}$ such that

1. It is **orthonormal**
$$\langle b_j, b_k \rangle_0 = \begin{cases} 0 & j \neq k \\ 1 & j = k \end{cases}$$
2. It is **complete** whereby the only function $a(\cdot)$ such that $\langle a, b_j \rangle = 0$ for all j is the zero almost everywhere function.

Using this basis, we are able to represent every function in L_2 as a linear combination of this basis.

Proposition 14: Basis representation of functions

Let L_2 be the function space of square integrable functions. Then, any function $a(\cdot) \in L_2$ may be represented as a linear combination of the basis functions

$$a(x) = \sum_j a_j b_j(x) \tag{5.68}$$

whereby $a_j = \langle a, b_j \rangle$.

This basis representation of a function induces a co-ordinate system whereby **we can now represent each function by a sequence of coefficients**. That is, each function $a(\cdot) \in L_2$ may be identified with its (square-summable) sequence of coefficients $a_1, a_2, ...$ whereby $\sum_k a_k < \infty$.

We can also represent the norm of a function by its coefficients.

Proposition 15: Coefficient representation of the norm of a function

A function $a \in L_2$ has a norm $||a||_0$ such that

$$||a||_0^2 = \sum_j a_j^2 \tag{5.69}$$

whereby $a_j = \langle a, b_j \rangle_0$ with $\{b_j\}$ being a complete orthonormal basis function of L_2.

Proof:(Sketch). First recall that

$$||a||_0^2 = \int a^2(x) p_0(x) dx = \int (\sum_j \langle \sum_j a_j b_j(x))^2 p_0(x) dx$$

where $a_j = \langle a, b_j \rangle$ and then use the fact that $b_j b_k = b_j^2 = 1$ by being orthonormal. ■

STAT5610: Advanced Inference

6. Properties of Score Functions

We will now discuss projecting the Score functions in the $L_2(\cdot)$ space and from this, a geometric interpretation of the effective Scores and effective information.

6.2.7 Projection of Score Functions

Consider an i.i.d 2-parameter model $\{\mathbb{P}_{n\theta\eta}\}$ with Score functions $\ell^\circ_\theta(\cdot)$ and $\ell^\circ_\eta(\cdot)$ at the true density $p(\cdot;\theta_0,\eta_0)$. Furthermore, recall that the effective score function was defined to be

$$\ell^{\circ *}_\theta = \ell^\circ_\theta - J_{\theta\eta}J^{-1}_{\eta\eta}\ell^\circ_\eta \tag{6.70}$$

Futhermore, recall that from the previous lecture, the covariance between the Score functions $J_{\theta\eta}$ can be written as the inner product of the respective Score functions $\langle \ell^\circ_\theta, \ell^\circ_\eta \rangle_0$. Finally, the variance of the Score function can be written as the squared norm $||\ell^\circ_\eta||^2_0$. Putting all this together, we now have a new interpretation of equation 6.70 in the L_2 space.

Proposition 16: Effective Score Function in L_2

The effective Score function $\ell^{\circ *}_\theta$ in L_2 is given by

$$\ell^{\circ *}_\theta = \ell^\circ_\theta - \langle \ell^\circ_\theta, \ell^\circ_\eta \rangle_0 \frac{1}{||\ell^\circ_\eta||^2_0}\ell^\circ_\eta \tag{6.71}$$

in the i.i.d 2-parameter model $\{\mathbb{P}_{n\theta\eta}\}$.

Remark 6.32 *Written this way, we can interpret the term*

$$\frac{\langle \ell^\circ_\theta, \ell^\circ_\eta \rangle_0}{||\ell^\circ_\eta||^2_0}\ell^\circ_\eta$$

to be the projection of the Score function of interest ℓ°_θ onto the 1-dimensional linear subspace spanned by ℓ°_η.

This means that we can interpret the effective score function $\ell^{\circ *}_\theta$ to be the **residual after the projection** of the Score function of the parameter of interest ℓ°_θ onto ℓ°_η.

Geometrically, we can think about the Score function for θ, ℓ°_θ being decomposed into 2 orthogonal components

1. The projection onto the nuisance Score, which is parallel to the nuisance Score
2. The other component orthogonal to the nuisance Score.

Therefore, we can interpret the effective score to be the orthogonal component of ℓ°_θ to ℓ°_η.

In the general case of a *vector-valued* nuisance parameter $\underset{\sim}{\eta}$, this interpretation carries through unchanged.

Corollary 6.33 *(Effective Score in the presence of vector-valued nuisance parameter). The effective score is the component of ℓ°_θ which is orthogonal to the nuisance Scores space. that is, the linear span of the nuisance Score functions*

$$\underset{\sim}{\ell_\eta}^\circ = (\ell^\circ_{\eta_1} \cdots \ell^\circ_{\eta_{d_\eta}})^T$$

We shall soon see, that when we have regular estimators, their influence function, which is a multiple of the Score function, will have to be orthogonal to all nuisance scores.

We are more interested in the **effective information**. First, recall the definition of the effective information

$$J^*_\theta = J_{\theta\theta} - J_{\theta\underset{\sim}{\eta}} J^{-1}_{\underset{\sim}{\eta}\underset{\sim}{\eta}} J_{\underset{\sim}{\eta}\theta}$$

We can now define the effective information in terms of inner products and norms.

Proposition 17: Effective Information in $L_2(\cdot)$ space

The effective information in the 1-dimensional θ and $d_\eta-$dimensional nuisance parameter $\underset{\sim}{\eta}$ is given by

$$J^*_\theta = ||\ell^\circ_\theta||_0^2 - \langle \ell^\circ_\theta, \underset{\sim}{\ell_\eta}^\circ \rangle_0^T \langle \underset{\sim}{\ell_\eta}^\circ, (\underset{\sim}{\ell_\eta}^\circ)^T \rangle_0^{-1} \langle \ell^\circ_\theta, \underset{\sim}{\ell_\eta}^\circ \rangle_0 \tag{6.72}$$

The effective information quantity J^*_θ is what we are trying to compute. The idea for the future is that we will construct a family of parametric submodels and find a upper bound to the infimum of the effective information of the parametric submodel.

6.2.8 Construction of regular parametric submodels in the constrained location model

We now have the necessary components to describe our first step to finding the information bound for our semiparametric model. That is, recall that our first step is to be able to construct a sequence of **regular parametric submodels** $\mathcal{Q}_n$ of the semiparametric model $\mathcal{P}_n$. We will construct such a sequence of regular parametric submodel and in turn, the effective information for each submodel will be an upper bound to the effective information for the semiparametric model.

We now describe the basis functions we will use to construct parametric submodels.

First, recall that in the constrained location model, we have that $\mathbb{P}_{n\theta f}$ is the joint distribution which are i.i.d with the density given by

$$p(x, \theta, f) = f(x - \theta)$$

whereby the first moment of the constraint function $w(\cdot)$ with respect to $p(x, \theta, f)$ is zero and the second moment is finite.

Suppose we can find a complete orthonormal basis for $L_2(p_0)$ of the form

$$\{1, \frac{w(\cdot)}{||w||_0}, b_1(\cdot), b_2(\cdot), \cdots\} \tag{6.73}$$

whereby 1 is the constant function and $\frac{w(\cdot)}{||w||_0}$ is the normalization of the constraint function $w(\cdot)$ of our location model. Finally, $b_j(\cdot)$ are other basis functions that are bounded.

For each fixed integer $k \geq 1$, consider the parametric submodel

$$\mathcal{Q}_{nk} = \{Q^{(k)}_{n\theta\underset{\sim}{\eta}}\} \tag{6.74}$$

whereby $Q^{(k)}_{n\theta\underset{\sim}{\eta}}$ is the joint distribution of n i.i.d random variables $X_1, ..., X_n$ with the joint density

$$q(x; t, \underset{\sim}{\eta}) = p_0(x-t)[1 + \sum_{j=1}^{k} \eta_j b_j(x-t)]$$

whereby $\underset{\sim}{\eta} = (\eta_1, \cdot\cdot\cdot, \eta_k)^T$ are the k nuisance parameters we fixed.

For all t and bounded nuisance parameters $|\eta_j| \leq \frac{1}{k}\sup_y|b_j(y)|$, then the density function is always non-negative and since

$$\mathbb{E}_0[b_j(X)] = \int p_0(x) b_j(x) dx = 0$$

for all j, this defines a proper probability density function as

$$\int q(x; t, \underset{\sim}{\eta}) dx = 1$$

for all t and sufficiently small $\underset{\sim}{\eta}$.

Under minimal extra conditions, the LAN condition holds at $t = 0, \underset{\sim}{\eta} = \underset{\sim}{0}$ with the Score functions

$$\begin{pmatrix} S_\theta \\ S_{\underset{\sim}{\eta}} \end{pmatrix} = n^{-1/2} \sum_{i=1}^{n} \begin{pmatrix} \psi(X_i) \\ \underset{\sim}{b}(X_i) \end{pmatrix} \tag{6.75}$$

whereby $\psi(\cdot)$ is the Score function for the location model and $\underset{\sim}{b}(\cdot)$ is the vector of basis functions

$$\underset{\sim}{b}(\cdot) = \begin{pmatrix} b_1(\cdot) \\ \cdot \\ \cdot \\ \cdot \\ b_k(\cdot) \end{pmatrix}$$

and the information matrix is given by

$$\begin{bmatrix} J_{\theta\theta} & J_{\underset{\sim}{\theta\eta}} \\ J_{\underset{\sim}{\eta\theta}} & \underset{\sim}{I_k} \end{bmatrix}$$

whereby

$$J_{\theta\theta} = \int \psi^2(x) p_0(x) dx = ||\psi||_0^2$$

is the squared length of ψ and

$$J_{\underset{\sim}{\eta\theta}}{}^T = J_{\underset{\sim}{\theta\eta}} = \int \psi(x) \underset{\sim}{b}(x) p_0(x) dx = \langle \psi, \underset{\sim}{b} \rangle_0$$

has the covariance of the Score functions being the inner product and

$$\underset{\sim}{I_k} = \langle \underset{\sim}{b}, \underset{\sim}{b}^T \rangle_0$$

is the k by k identity matrix, which holds due to the basis functions being orthonormal.

Proposition 18: Effective Score Function in semiparametric submodel

The effective Score function in the semiparametric submodel is

$$\underset{\sim}{\ell_\theta}^{\circ *} = \psi - \langle \psi, \underset{\sim}{b} \rangle_0^T \underset{\sim}{b}$$

Proof: Recall that the effective score is given by

$$\underset{\sim}{\ell_\theta}^{\circ *} = \ell_\theta^\circ - J_{\theta\eta} J_{\eta\eta}^{-1} \ell_\eta^\circ$$

Therefore, we have that

$$\underset{\sim}{\ell_\theta}^{\circ *} = \psi - \langle \psi, \underset{\sim}{b} \rangle_0^T \langle \underset{\sim}{b}, \underset{\sim}{b}^T \rangle_0^{-1} \underset{\sim}{b}$$

However, as $\underset{\sim}{b}$ is a orthonormal basis, we have that $\langle \underset{\sim}{b}, \underset{\sim}{b}^T \rangle_0^{-1} = \underset{\sim}{I_k}$. Therefore, we have that

$$\psi - \langle \psi, \underset{\sim}{b} \rangle_0^T \underset{\sim}{b}$$

■

Proposition 19: Effective Information in parametric submodel

The effective information matrix in the parametric submodel

$$J_\theta^* = ||\psi||_0^2 - \langle \psi, \underset{\sim}{b}_0 \rangle_0^T \langle \psi, \underset{\sim}{b} \rangle_0$$

Proof: Recall that the effective information matrix is given by

$$J_\theta^* = J_{\theta\theta} - J_{\theta\underset{\sim}{\eta}} J_{\underset{\sim}{\eta}\underset{\sim}{\eta}}^{-1} J_{\underset{\sim}{\eta}\theta}$$

Therefore, we have that

$$J_\theta^* = ||\psi||_0^2 - \langle \psi, \underset{\sim}{b} \rangle_0^T \langle \underset{\sim}{b}, \underset{\sim}{b}^T \rangle_0^{-1} \langle \psi, \underset{\sim}{b} \rangle_0$$

whereby $\underset{\sim}{b}$ is a orthonormal basis and therefore $\langle \underset{\sim}{b}, \underset{\sim}{b}^T \rangle_0^{-1} = \underset{\sim}{I_k}$. Hence, we have

$$= ||\psi||_0^2 - \langle \psi, \underset{\sim}{b}_0 \rangle_0^T \langle \psi, \underset{\sim}{b} \rangle_0$$

■

Now, it is important to note that the inner product of the Score function ψ with the complete orthonormal basis $\underset{\sim}{b}$ is

$$\langle \psi, \underset{\sim}{b} \rangle_0^T \langle \psi, \underset{\sim}{b} \rangle_0 = \langle \psi, b_1 \rangle_0^2 + \langle \psi, b_2 \rangle_0^2 + \cdots + \langle \psi, b_k \rangle_0^2 \tag{6.76}$$

Now, recall that we can write the norm squared of any function $\psi \in L_2$ space as the sum of its squared coefficients in the basis expansion

$$||\psi||_0^2 = \sum_j \psi_j^2 \tag{6.77}$$

$$= \langle \psi, 1 \rangle_0^2 + \frac{\langle \psi, w \rangle_0^2}{||w||_0^2} + \langle \psi, b_1 \rangle_0^2 + \langle \psi, b_2 \rangle_0^2 + \cdots \tag{6.78}$$

Now. recall that all score functions have mean zero, which is equivalent to being orthogonal to constants

$$\langle \psi, 1 \rangle_0^2 = 0$$

Furthermore, $\langle \psi, \underset{\sim}{b} \rangle_0^T$ is k terms of the inner product $\langle \psi, b_j \rangle_0^T$, whereby we subtract these k terms of $||\psi||_0^2$.

Proposition 20: Effective Information in parametric submodel

For a fixed integer $k \geq 1$, the effective information in the parametric submodel of the integral-constrained location model is

$$J_\theta^* = \frac{\langle \psi, w \rangle_0^2}{||w||_0^2} + \langle \psi, b_{k+1} \rangle_0^2 + \langle \psi, b_{k+2} \rangle_0^2 + \cdots \tag{6.79}$$

Remark 6.34 *It is important to note that the bigger we set the integer $k \geq 1$, the smaller the remaining terms*

$$\langle \psi, b_{k+1} \rangle_0^2 + \langle \psi, b_{k+2} \rangle_0^2 + \cdots$$

becomes as the coefficients are square-summable, that is, as $k \to \infty$,

$$\langle \psi, b_{k+1} \rangle_0^2 \to 0.$$

Using notation from previous lectures, we can write the **effective information of the parametric submodel** as

$$J_\theta^*(\mathcal{Q}_{nk}|\{\underset{\sim}{Q_{n00}^{(k)}}\}) = J_\theta^*$$

Proposition 21: Upper bound to effective information of semiparametric full model

The upper bound to the effective information of the full semiparametric model is given by

$$inf\left\{ J_\theta^*(\mathcal{Q}_{nk}|\{\underset{\sim}{Q_{n00}^{(k)}}\}); k = 1, 2, ...\right\} = \frac{\langle \psi, w \rangle_0^2}{||w||_0^2} \tag{6.80}$$

Remark 6.35 *By choosing k big enough and consequently including more nuisance parameters, we make all the remaining terms $\langle \psi, b_{k+1} \rangle_0^2$ small.*

As we increase k, the number of nuisance parameters in our parametric submodel increases, and as a result, the effective information J_θ^* gets smaller. We eventually reach the lower bound as we let $k \to \infty$. Therefore our model gets worse as we increase k and eventaully, the parametric submodel will have an effective information that is the same as the effective information of the full semiparametric model. This is because the effective information of the full semiparametric model will be smaller (which is worse) than the effective information for each parametric submodel.

The upper bound to the effective information in the full semi-parametric submodel is given by

$$J_\theta^*(\mathcal{P}_n|\mathbb{P}_{n,\theta_0 p_0}) \leq \frac{\langle \psi, w \rangle_0^2}{||w||_0^2} \tag{6.81}$$

We have satisfied the first step of the strategy whereby we obtain an upper bound to the infimum of the constructed sequence of regular parametric submodels. With this, we now have a bound on the limiting variance of a regular estimator by taking the inverse of our effective information.

Proposition 22: Limiting variance of regular estimators in full semiparametric model

Suppose that $\tilde{\theta}_n$ is a regular estimator, that is

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

Then, the lower bound on the limiting variance of the rescaled estimation error of $\tilde{\theta}_n$ is given by

$$\sigma^2 \geq J_\theta^*(\mathcal{P}_n|\mathbb{P}_{n,\theta_0 p_0})^{-1} \geq \frac{||w||_0^2}{\langle \psi, w\rangle_0^2}$$

Proof: We have that the upper bound to the effective information of the full semiparametric model is the smallest effective information from all our parametric submodels

$$J_\theta^*(\mathcal{P}_n|\mathbb{P}_{n,\theta_0 p_0}) \leq \frac{\langle \psi, w\rangle_0^2}{||w||_0^2}$$

Now, recall that the limiting variance of a regular estimator is $(J_\theta^*)^{-1}$. Therefore, taking inverses reverse the inequality above. ■

Remark 6.36 *Recall that the limiting variance of the regular estimator is the inverse of the effective information. As the effective information for the full semiparametric model will be small (compared to the effective information in the regular parametric submodels), the limiting variance of the regular estimator will be high.*

6.2.9 Regular Estimators in the parametric case

First, we recall the definition and properties of regular estimators in the parametric setting.

A **regular** estimator $\tilde{\theta}_n$ in a parametric model is one where under a nearby sequence

$$\underset{\sim}{\gamma_n} = \begin{pmatrix} \underset{\sim}{\theta_n} \\ \underset{\sim}{\eta_n} \end{pmatrix} = \begin{pmatrix} \underset{\sim}{\theta_0} \\ \underset{\sim}{\eta_0} \end{pmatrix} + n^{-1/2} \begin{pmatrix} \underset{\sim}{h_\theta} \\ \underset{\sim}{h_\eta} \end{pmatrix}$$

the limiting distribution of the rescaled estimation error

$$\underset{\sim}{\gamma_n} = \sqrt{n}(\underset{\sim}{\tilde{\theta}_n} - \underset{\sim}{\theta_n}) \tag{6.82}$$

$$= \sqrt{n}(\underset{\sim}{\tilde{\theta}_n} - \underset{\sim}{\theta_0} - n^{-1/2}\underset{\sim}{h_\theta}) \tag{6.83}$$

does not depend on the nuisance local deviation $\underset{\sim}{h_\eta}$.

By Le Cam's third lemma, this means that under the true values

$$\begin{pmatrix} \theta_0 \\ \underset{\sim}{\eta_0} \end{pmatrix}$$

, the limiting distribution is

$$\begin{pmatrix} \underset{\sim}{Y_n} \\ \underset{\sim}{S_\theta} \\ \underset{\sim}{S_\eta} \end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(\begin{pmatrix} \underset{\sim}{0} \\ \underset{\sim}{0} \\ \underset{\sim}{0} \end{pmatrix}, \begin{pmatrix} \Sigma_Y & I & \underset{\sim}{0} \\ I & \underset{\sim}{J_{\theta\theta}} & \underset{\sim}{J_{\theta\eta}} \\ \underset{\sim}{0} & \underset{\sim}{J_{\eta\theta}} & \underset{\sim}{J_{\eta\eta}} \end{pmatrix} \Big)$$

First, recall that the covariance of $\underset{\sim}{Y_n}$ with the scores of the parameter of interest $\underset{\sim}{S_\theta}$ is the identity matrix I and the covariance with the nuisance scores is $\underset{\sim}{0}$ due to $\underset{\sim}{\tilde{\theta}_n}$ being a regular estimator.

6.2.10 Regular Estimators in the semiparametric case

In the semi-parametric setting, we now simply extend the conditions described in the parametric setting to any nearby sequence within any *regular parametric submodel.*

Proposition 23: Limiting distribution of estimation error under true values

Let $\tilde{\theta}_n$ be a **regular** estimator . In the constrained location model, given any complete orthonormal basis $\{1, \frac{w(\cdot)}{||w||_0}, b_1(\cdot), b_2(\cdot), \cdots\}$, we must have for any $j \geq 1$, under the true density $p_0(\cdot)$

$$\begin{pmatrix} Y_n \\ \underset{\sim}{S_\theta} \\ \underset{\sim}{S_\eta} \end{pmatrix} = \begin{pmatrix} Y_n \\ n^{-1/2}\sum_{i=1}^n \psi(X_i) \\ n^{-1/2}\sum_{i=1}^n b_j(X_i) \end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(\begin{pmatrix} \underset{\sim}{0} \\ \underset{\sim}{0} \\ \underset{\sim}{0} \end{pmatrix}, \begin{pmatrix} \Sigma_Y^2 & 1 & 0 \\ 1 & ||\psi||_0^2 & \langle \psi, b_j \rangle_0 \\ 0 & \langle \psi, b_j \rangle_0 & 1 \end{pmatrix} \Big) \tag{6.84}$$

Proof: First, recall that any nuisance score can be expressed as any linear combination of the complete orthonormal basis functions $\{b_j(\cdot)\}_{j\geq 1}$, which have mean zero and is orthogonal to the constraint function $w(\cdot)$. Furthermore, as established before, the covariance of the estimation error of a regular estimator with the scores of interest will be 1 and the covariance with the nuisance scores will be 0. ■

Now, we recall the definition of an asymptotically linear estimator.

Definition 22: Asymptotically linear Estimator

An estimator $\underset{\sim}{\hat{\theta}_n}$ is asymptotically linear if it can be written in the form

$$\sqrt{n}(\underset{\sim}{\hat{\theta}_n} - \theta_0) = n^{-1/2}\sum_{i=1}^n \tilde{\ell}(X_i) + o_p(1) \tag{6.85}$$

where $\tilde{\ell}(\cdot)$ is the **influence function**.

We can see that if a complete orthonormal basis exists, we can find an asymptotically linear estimator in our parametric submodel.

Proposition 24: Influence function of parametric submodel

Suppose that a complete orthonormal basis $\{1, \frac{w(\cdot)}{||w||_0}, b_1(\cdot), b_2(\cdot), \cdots\}$ exists for bounded $\{b_j(\cdot)\}_{j\geq 1}$. Then, the *optimal* influence function of an asymptotically linear estimator $\underset{\sim}{\hat{\theta}_n}$ is

$$\tilde{\ell}(X) = \frac{w(X)}{\langle w, \psi \rangle_0} \tag{6.86}$$

Remark 6.37 *Recall that if an estimator has an optimal influence function, the estimator is optimal.*

If an estimator has the optimal influence function

$$\tilde{\ell}(X) = \frac{w(X)}{\langle w, \psi \rangle_0}$$

then, it is a **regular** estimator since

1. $$\langle \tilde{\ell}, \psi \rangle_0 = \langle \frac{w}{\langle w, \psi \rangle_0}, \psi \rangle_0 = 1$$

2. $$\langle \tilde{\ell}, b_j \rangle_0 = \frac{\langle w, b_j \rangle_0}{\langle w, \psi \rangle_0} = 0$$

   for j = 1, 2, ...

Furthermore, the estimator attains the lower bound to the variance of a regular estimator

$$\sqrt{n}(\hat{\theta_n} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \frac{||w||^2}{\langle w, \psi \rangle_0^2}) \tag{6.87}$$

whereby $\langle w, \psi \rangle_0^2$ is the lower bound from the constructed sequence of parametric submodels.

We are now interested in figuring out whether **can** we actually find an estimator with the influence function

$$\hat{\psi}(X) = \frac{w(X)}{\langle w, \psi \rangle_0}$$

Proposition 25: Finding an estimator with optimal influence function

If the constraint function $w(\cdot)$ is suitable regular, then the solution to the estimating equation

$$\sum_{i=1}^{n} w(X_i - \theta) = 0$$

provides an estimator with the optimal influence function

$$\tilde{\ell}(x) = \frac{w(x)}{\langle w, \psi \rangle_0}$$

STAT5610: Advanced Inference

7. Basis expansion

Previously, we talked about constructing parametric submodels with orthonormal bases, and from that, we obtain the information bound. The goal of today is to describe how can we find such an orthonormal basis.

7.2.11 The Haar Wavelet Basis

A convenient choice of a complete orthonormal basis is provided by the Haar Wavelet basis with respect to the density $p_0(\cdot)$.

Definition 23: Haar Wavelet Basis

First, define the family of functions

$$u_{k,j}(x) = 2^{k/2}\Big[1\{j2^{-(k+1)} \leq x < (j+1)2^{-(k+1)}\} - 1\{(j+1) < 2^{-(k+1)} \leq x < (j+2)2^{-(k+1)}\}\Big] \tag{7.88}$$

for k = 0,1,2,... and $0 \leq j < 2^k$. Here, j is the shift parameter and k determines the scale. That is, for each fixed value of k, there is a range for the integer j.
The Haar wavelet basis with respect to $p_0(\cdot)$ is given by the collection

$$\{1\} \cup \{u_{k,j} : k = 0, 1, 2, \cdots ; 0 \leq j < 2^k\}$$

It is important to note that the *standard* Haar wavelet function

$$u_{0,0}(x) = 1\{0 \leq x < \frac{1}{2}\} - 1\{\frac{1}{2} \leq x < 1\} = -sign(x)$$

Theorem 7.38 *(Complete Orthonormal Basis for $L_2(U(0,1))$). The Haar wavelet basis*

$$\{1\} \cup \{u_{k,j} : k = 0, 1, 2, \cdots ; 0 \leq j < 2^k - 1\}$$

form a complete orthonormal basis of $L_2(U[0,1])$ where $U[0,1]$ is the uniform distribution.

Theorem 7: Complete Orthonormal Basis for $L_2(F)$.

For any strictly positive density $f(\cdot)$, with a continuous, strictly increasing CDF $F(\cdot)$, if we define

$$b_{k,j}(x) = u_{k,j}(F(x)) \tag{7.89}$$

the collection

$$\{1\} \cup \{b_{k,j}\}$$

forms a complete orthonormal basis for $L_2(f)$.

Proof: Recall that the definition for a complete ON basis in $L_2(U(0,1))$ is that we require that the only function $g(\cdot)$ such that it is orthogonal to all the Haar wavelets and the constant function

$$\int_0^1 g(y)u_{k,j}(y)dy = 0$$

and

$$\int_0^1 g(y)dy = 0$$

is the zero function $g(\cdot) = 0$ almost everywhere.

We want to show that the only function $h(\cdot)$ such that

$$\int_{-\infty}^{\infty} h(y)b_{k,j}(y)f(y)dy = 0$$

$$\int_{-\infty}^{\infty} h(y)dy = 0$$

is the zero function $h(\cdot) = 0$ almost everywhere in order to show that

$$\{1\} \cup \{b_{k,j} : k = 0, 1, 2, ...; 0 \leq j < 2^k - 1\}$$

is a complete orthonormal basis for $L_2(F)$.

This can be done by assuming such a function exists and then applying the change of variables

$$y = F(x) \Rightarrow dy = f(x)dx$$

onto the assumption that there exists a function $h(\cdot)$

$$\int_{-\infty}^{\infty} h(x)b_{k,j}(x)f(x)dx = 0$$

whereby we can appeal to the fact that $\{u_{k,j}\}$ form a complete ON basis

$$\int_0^1 h(F^{-1}(y))u_{k,j}(y)dy = 0$$

to conclude that $h(\cdot) = 0$ almost everywhere. ■

Remark 7.39 *Here, the function $u_{k,j}(F(x))$ maps from the unit interval to $\mathbb{R}$.*

As indicated before, we have that if our constraint function is

$$w(x) = sign(x) = -b_{0,0}(x)$$

whereby the constraint condition is $\int w(x)f(x)dx = 0$, then this implies that the constraint function $w(x)$ is orthogonal to all of $\{b_{k,j}(\cdot) : k \geq \mathbf{1}, 0 \leq j < 2^k\}$.

Proposition 26: Complete ON basis includes the constaint function

Let $w(x) = sign(x)$ be the constraint function. Then, the set

$$\{1\} \cup \{w(\cdot)\} \cup \{b_{k,j}\} \tag{7.90}$$

forms a complete orthonormal basis for $L_2(F)$.

Lemma 7.40 *Each basis function $b_{k,j}(\cdot)$ is a bounded function.*

Then, the conditions of the construction of lecture 5 are satisfied for this given constraint function $w(\cdot)$. That is, for any integral constrained location model with constraint function $w(x) = sign(x)$, if the true density has a strictly increasing CDF over its domain, we can conclude that

1. The effective information is $4f_0(\cdot)^2$

2. Any regular estimator (in the semiparametric sense) $\tilde{\theta_n}$ is such that the limiting variance of the rescaled estimation error
$$\sqrt{n}(\tilde{\theta_n} - \theta_0) \geq \frac{1}{4f_0(0)^2}$$

3. The sample median is asymptotically efficient.

7.2.12 Orthogonal Polynomials

We now wish to look at cases where the constraint function is **not** $w(x) = x$. In order to do so, we need to define a different set of bases.

Consider the monomials $\{1, x, x^2, \cdot\cdot\cdot, x^k\}$ for some positive integer k. Then, if all of these are linearly independent with respect to a density $f(\cdot)$, then we can construct a set of orthonormal functions with respect to $f(\cdot)$ using the Gram-Schmidt orthogonalisation procedure.

We shall denote the following

$$\langle g, h \rangle = \int g(x)h(x)f(x)dx \qquad ||g||= \langle g, g \rangle^{1/2}$$

Furthermore, we shall denote each monomial as

$$p_i(x) = x^i$$

whereby $p_0(x) = 1$.

We now describe the procedure of constructing a set of orthonormal functions.

Proposition 27: Construction of orthonormal functions

Let $\{1, x, x^2, \cdot\cdot\cdot, x^k\}$ be a set of monomials for some positive integer k. Furthermore, define the density $f(\cdot)$. Let $p_i(x) = x^i$. Then, we can construct a sequence of orthonormal functions whereby we first orthogonalise the function

$$\tilde{p}_k = p_k - \sum_{i=0}^{k-1} \langle p_k, p_i^* \rangle p_i^* \tag{7.91}$$

and then we normalise

$$p_k^* = \frac{\tilde{p}_k}{||\tilde{p}_k||} \tag{7.92}$$

Then, by construction, $\{p_0^*, \cdot\cdot\cdot, p_k^*\}$ are orthonormal with respect to $f(\cdot)$.

We will now look at a special case in order to construct a complete orthonormal basis for L_2.

Definition 24: Hermite Polynomial

Define the set of monomials $\{1, x, x^2, \cdots, x^k\}$ for some positive integer k. Define the density

$$f(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

to be the $\mathcal{N}(0,1)$ density. Then, following the construction of orthonormal functions using $f(\cdot)$ gives us the **Hermite Polynomials**.

The un-normalised Hermite polynomials $\{He_j(\cdot)\}$ satisfy the relationship

$$\frac{d^j}{dx^j}(e^{-\frac{x^2}{2}}) = (-1)^j He_j(x) e^{-\frac{x^2}{2}}$$

and they are orthogonal with respect to the normal density

$$\int He_j(x) He_k(x) \phi(x) dx = \begin{cases} j! & j = k \\ 0 & j \neq k \end{cases}$$

Finally, we can construct the normalised Hermite polynomials $\{p_j(\cdot)\}$ by

$$p_j(x) = \frac{He_j(x)}{\sqrt{j!}}$$

Proposition 28: Hermite Polynomials in $L_2(f)$

The normalised Hermite polynomials $\{p_j(x)\}$ where

$$p_j(x) = \frac{He_j(x)}{\sqrt{j!}}$$

forms a complete orthonormal basis of $L_2(\phi)$.

With this construction of orthonormal polynomials as a complete orthonormal basis, we now come to an important theorem on the integral-constrained location model.

Proposition 29: Estimating the Mean in integral-constrained location model

Suppose that the constraint function is $w(x) = x$. Furthermore, assume that the density $f(\cdot)$ is such that the 1-parameter location model with common density $f(x - \theta)$ satisfies the LAN property at $\theta = 0$ with Score function $\psi(\cdot)$ and information J. Then, there exists orthogonal polynomials $\{p_j^*(\cdot)\}$ forming a complete orthonormal basis for $L_2(F)$ and

$$p_0^*(x) = 1 \quad p_1^*(x) = \frac{w(x)}{||w||_f}$$

Then, the effective information for estimating θ in the corresponding integral constrained location model is $\frac{1}{\sigma^2}$ whereby

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_{-\infty}^{\infty} w(x)^2 f(x)dx \tag{7.93}$$

Furthermore, the sample mean is an asymptotically efficient estimator.

Proof:(Sketch). Use the construction from question 1 tutorial 3 for the parametric submodel of zero-mean densities for the shape parameter of the form

$$q(x; \underset{\sim}{\eta})^{(k)} = \frac{f(x)L\{\sum_{j=2}^{k} \eta_j p_j^*(x)\}}{\int f(x)L\{\sum_{j=2}^{k} \eta_j p_j^*(x)dx}$$

We have that the vector of Score functions is given by

$$\underset{\sim}{p^*} = \begin{pmatrix} p_2^* \\ \cdots \\ p_k^* \end{pmatrix}$$

Now, looking at the combined model by including the location parameter, we then have the parametric submodel is given by

$$q(x; \theta, \underset{\sim}{\eta})^{(k)} = \frac{f(x-\theta)L\{\sum_{j=2}^{k} \eta_j p_j^*(x-\theta)\}}{\int f(x)L\{\sum_{j=2}^{k} \eta_j p_j^*(x-\theta)dx} \tag{7.94}$$

This can be shown to satisfy the LAN conditions at $\theta = \theta_0$ and $\underset{\sim}{\eta} = \underset{\sim}{0}$ with the vector of Score functions

$$\begin{pmatrix} \psi \\ \underset{\sim}{p^*} \end{pmatrix} = \begin{pmatrix} \psi \\ p_2^* \\ \cdots \\ p_k^* \end{pmatrix}$$

and information matrix

$$\begin{bmatrix} J & \underset{\sim}{J_{\theta\eta}} \\ \underset{\sim}{J_{\eta\theta}} & \underset{\sim}{I} \end{bmatrix}$$

The effective Score function is given by

$$\ell_\theta^* = \ell_\theta^\circ - \underset{\sim}{J_{\theta\eta}}^T \underset{\sim}{J_{\eta\eta}}^{-1} \underset{\sim}{\ell_\eta}^\circ = \psi - \langle \psi, \underset{\sim}{p^*} \rangle^T \underset{\sim}{p^*}$$

and the effective information is given by

$$J^* = J_{\theta\theta} - \underset{\sim}{J_{\theta\eta}}^T \underset{\sim}{J_{\eta\eta}}^{-1} \underset{\sim}{J_{\eta\theta}} = J - \langle \psi, \underset{\sim}{p}^* \rangle^T \langle \psi, \underset{\sim}{p}^* \rangle \tag{7.95}$$

$$= \sum_j \langle \psi, p_j^* \rangle^2 - \langle \psi, p_1^* \rangle^2 - \cdots - \langle \psi, p_k^* \rangle^2 \tag{7.96}$$

$$= \langle \psi, \frac{w}{||w||_f} \rangle^2 + \langle \psi, p_{k+1}^* \rangle^2 + \langle \psi, p_{k+2}^* \rangle^2 + ... \tag{7.97}$$

and hence we can see that as $k \to \infty$, we get the infimum of the effective information for our regular parametric submodels or the effective information of the semiparametric full model

$$\langle \psi, \frac{w}{||w||_f} \rangle^2 = \frac{\langle \psi, w \rangle^2}{||w||_f^2}$$

but recall that

$$||w||_f^2 = \int w^2(x) f(x) dx = \int x^2 f(x) dx = \sigma^2$$

which is the variance. Furthermore, we can show that $\langle \psi, w \rangle^2 = 1$. From this, we can see that the sample mean is clearly a regular estimator that obtains such a bound. ■

STAT5610: Advanced Inference

8. Symmetric Location Model

8.2.13 Symmetric Location Model

We now move onto a new model setup.

Consider the semiparametric model where $X_1, \cdots, X_n$ are i.i.d with common density

$$p(n; \theta, f) = f(x - \theta)$$

whereby

1. $\theta \in \mathbb{R}$
2. f is symmetric about 0

We are interested in estimating θ in the presence of the unknown nuisance parameter function $f(\cdot)$.

As usual, we consier parametric submodels whereby the common density is of the form

$$q(x; \theta, \underset{\sim}{\eta}) = f_{\underset{\sim}{\eta}}(x - \theta)$$

whereby the base density without the location parameter θ, $\{f_{\underset{\sim}{\eta}}(\cdot) : \underset{\sim}{\eta} \in \mathcal{H} \subseteq \mathbb{R}^d\}$ is a **regular** parametric family of densities where

1. All density elements $f_{\underset{\sim}{\eta}}(\cdot)$ is symmetric about 0
2. When $\underset{\sim}{\eta} = \underset{\sim}{0}$, $f_{\underset{\sim}{0}}(\cdot)$ is the true density $f_0(\cdot)$.

By a regular parametric family, we mean that the LAN holds at the true value $\underset{\sim}{\eta} = \underset{\sim}{0}$ for the sub-family of models where $Y_1, \cdots, Y_n$ is i.i.d $f_{\underset{\sim}{\eta}}(\cdot)$ for some vector of Score functions

$$\underset{\sim}{\ell_{\eta}}^{\circ} = \begin{pmatrix} \ell^{\circ}_{\eta_1}(\cdot) \\ \cdot \\ \cdot \\ \cdot \\ \ell^{\circ}_{\eta_{d_\eta}}(\cdot) \end{pmatrix}$$

and information matrix $\underset{\sim}{J_{\eta\eta}}$.

Remark 8.41 *Recall from tutorial that we are able to construct any parametric submodel using $\underset{\sim}{\eta}$ to get any arbitrary Score function we want.*

We now come to two very imporant fundamental facts for our symmetric parametric submodel

$$q(x;\theta,\underset{\sim}{\eta}) = f_{\underset{\sim}{\eta}}(x-\theta)$$

Proposition 30: Symmetry of nuisance Score functions

For any regular parametric submodel, each nuisance Score function

$$\ell^{\circ}_{\eta_j} = \ell^{\circ}_{\eta_j}(\cdot)$$

is **symmetric** about zero for $1 \leq \eta_j \leq d_\eta$.

Proposition 31: Antisymmetry of Score function of parameter of interest

For any regular parametric submodel, the Score function

$$\ell^{\circ}_{\theta} = \psi(\cdot)$$

for the parameter of interest θ is antisymmetric.

Proof:(Sketch). First, recall that the Score function for a symmetric differentiable $f_0(\cdot)$ is given by

$$\psi(x) = -\frac{f^{'}(x)}{f(x)}$$

Then, the slope of a unimodal symmetric density is positive on the left of the mode and negative on the right of the mode. Hence, the derivative is antisymmetric. ∎

Remark 8.42 *For the Laplace density, the Score function is the sign function. In the normal density, the Score function is x.*

Hence, we have that for the regular parametric submodel of the semiparametric symmetric location model,

1. $\ell_{\underset{\sim}{\eta}}{}^{\circ}$ is symmetric;
2. ℓ°_{θ} is antisymmetric.

Now, recall that any asymptotically linear estimator $\hat{\theta_n}$ can be written as

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \tilde{\ell} = J^{-1}_{\theta\theta} S_\theta.$$

That means, for any estimator $\hat{\theta_n}$ satisfying

$$\sqrt{n}(\hat{\theta_n} - \theta_0) = n^{-1/2} J^{-1}_{\theta\theta} \sum_{i=1}^{n} \psi(x_i - \theta) + o_p(1)$$

that is, the estimator $\hat{\theta_n}$ is asymptotically linear with influence function

$$\tilde{\ell}(x) = J^{-1}_{\theta\theta}\psi(x - \theta_0)$$

which is the optimal influence function when f is known, is also a regular estimator in the semiparametric sense.

Remark 8.43 *Note that as $\psi(\cdot)$ is antisymmetric, the influence function $\tilde{\ell}$ is also antisymmetric.*

Now, we note two things regarding this influence function $\tilde{\ell}(x) = J_{\theta\theta}^{-1}\psi(x - \theta_0)$.

Theorem 8.44 *(Estimator with influence function is asymptotically efficient).*

The influence function $\tilde{\ell}(\cdot) = J_{\theta\theta}^{-1}\psi(x - \theta_0)$ is orthogonal to any possible symmetric nuisance Score function $\underset{\sim}{\ell_\eta}^\circ$ since the product

$$\tilde{\ell}(x)\ell_{\eta_j}^\circ(x)$$

is also antisymmetric and integrates to zero.

Furthermore, the influence function $\tilde{\ell}(\cdot)$ has a covariance of 1 with the Score function for the parameter of interest

$$\langle\tilde{\ell}, \psi\rangle_f = \langle J_{\theta\theta}^{-1}\psi, \psi\rangle = J_{\theta\theta}^{-1}||\psi||_f^2 = J_{\theta\theta}^{-1}J_{\theta\theta} = 1$$

8.2.14 Construction of estimator

We are now interested in constructing an estimator.

First, we suppose that we have a complete orthonormal basis with the following structure

$$\{1\} \cup \{s_j(\cdot)\} \cup \{a_j(\cdot)\}$$

where each $s_j(\cdot)$ is symmetric and $a_j(\cdot)$ is antisymmetric.

Such a basis can be constructed using the Haar wavelet basis since any basis function $u_{k,j}(\cdot)$ for $k \geq 1$ can form a matching pair with another function on the same resolution level (i.e the same k).

Proposition 8.45 *Let $\{u_{k,j}\}$ for $k \geq 1$ be functions of the Haar wavelet basis. Then, the function*

$$\frac{u_{k,j} - u_{k,2^k-1-j}}{\sqrt{2}} \tag{8.98}$$

is symmetric and

$$\frac{u_{k,j} + u_{k,2^k-1-j}}{\sqrt{2}} \tag{8.99}$$

is antisymmetric.

Proposition 32: Basis expansion of symmetric and antisymmetric functions

Any symmetric function in $L_2(f_0)$ has a basis expansion only involving symmetric functions $s_j(\cdot)'s$ whilst any antisymmetric function has a basis expansion involving the $a_j(\cdot)'s$.

We can now put all our results together for figuring out the effective score and effective information for the symmetric location model with the basis $\{1\} \cup \{s_j(\cdot)\} \cup \{a_j(\cdot)\}$.

Theorem 8: Estimating parameter in symmetric location model

Consider the semiparametric model $X_1, ..., X_n$ that are i.i.d with common density

$$p(n; \theta, f) = f(x - \theta)$$

where $\theta \in \mathbb{R}$ and $f(\cdot)$ is unknown and symmetric about 0.
Suppose we can construct a parametric submodel with common density

$$q(x; \theta, \underset{\sim}{\eta}) = f_{\underset{\sim}{\eta}}(x - \theta)$$

where $f_{\underset{\sim}{\eta}}(\cdot)$ is symmetric for all $\underset{\sim}{\eta}$ and is a regular parametric family. Also, let us use the complete orthonormal basis for $L_2(f_0)$ as

$$\{1\} \cup \{s_j(\cdot)\} \cup \{a_j(\cdot)\}$$

Then, LAN holds at true values $\theta = \theta_0, \underset{\sim}{\eta} = \underset{\sim}{0}$ with vector of Score functions

$$\begin{pmatrix} \psi \\ s_1(\cdot) \\ \cdots \\ s_k(\cdot) \end{pmatrix}$$

where $\{s_1, \cdot \cdot \cdot, s_k\}$ are the first k of symmetric basis functions of the nuisance score $S_{\underset{\sim}{\eta}}$.
The effective Score is the parameter score function

$$\ell^*_\theta = \psi = \ell^\circ_\theta$$

and the corresponding effective information is the ordinary information

$$J^* = J_{\theta\theta} = \langle \psi, \psi \rangle_0^2$$

Proof:(Sketch). The effective Score is given by

$$\begin{aligned} \ell^*_\theta &= \ell^\circ_\theta - \underset{\sim}{J_{\theta\eta}}^T \underset{\sim}{J_{\eta\eta}}^{-1} \underset{\sim}{\ell_\eta}^\circ && (8.100) \\ &= \psi - \langle \psi, \underset{\sim}{s} \rangle^T \underset{\sim}{I}\,\underset{\sim}{s} && (8.101) \\ &= \psi && (8.102) \end{aligned}$$

as ψ is antisymmetric and $\underset{\sim}{s}$ is symmetric.

The effective information is given by

$$\begin{aligned} J^* &= J_{\theta\theta} - \underset{\sim}{J_{\theta\eta}}^T \underset{\sim}{J_{\eta\eta}}^{-1} \underset{\sim}{J_{\eta\theta}} && (8.103) \\ &= J_{\theta\theta} - \langle \psi, \underset{\sim}{s} \rangle^T \underset{\sim}{I} \langle \psi, \underset{\sim}{s} \rangle && (8.104) \\ &= J_{\theta\theta} = \sum_j \langle \psi, a_j \rangle a_j && (8.105) \end{aligned}$$

■

This is an extremely important results as under the assumption of symmetric for the density function, when

f_0 is unknown, we can estimate θ just as well as if $f(\cdot)$ was known. The term **adaptive** is used to describe such cases.

8.2.15 Practical Implementation

We now describe the process of constructing optimal adaptive estimator.

When $f_0(\cdot)$ is known, in general, we can try to solve the Score equation for θ

$$\sum_{i=1}^{n} \psi(x_i - \theta) = 0 \tag{8.106}$$

Under a regular Scores assumption, if the root $\hat{\theta_n}$ is $\sqrt{n}-$consistent, then it is also an asymptotically efficient estimator.

However, as we do not know $f_0(\cdot)$, we can construct an estimate $\hat{\psi_n}(\cdot)$ of the Score function and try to solve the Score equation

$$\sum_{i=1}^{n} \hat{\psi}(x_i - \theta) = 0 \tag{8.107}$$

STAT5610: Advanced Inference

9. Integral Constrained Semiparametric Location-Scale Model

9.2.16 Integral Constrained Semiparametric Location-Scale Model

We are now interested in our final model for semiparametric estimation. We now consider the density characterised by both a scale and location parameter.

Definition 25: Location-Scale Model

Suppose $X_1, ..., X_n$ are i.i.d with common density

$$p(x; \underset{\sim}{\theta}, f) = p(x; (\mu, \sigma)^T, f) = \frac{1}{\sigma} f(\frac{x-\mu}{\sigma})$$

where

1. $\underset{\sim}{\theta} = (\mu, \sigma)^T$ where $\mu \in \mathbb{R}$ and $\sigma > 0$
2. f is a probability density function on $\mathbb{R}$ which is centered and of unit scale in that it satisfies two constraints
$$\int_{-\infty}^{\infty} u(x)f(x)dx = \int_{-\infty}^{\infty} v(x)f(x)dx = 0$$
for suitable constraint functions $u(\cdot)$ and $v(\cdot)$.

Here, the density function $f(\cdot)$ is orthogonal to the constraint functions $u(\cdot)$ and $v(\cdot)$.

We give some examples of such constraint functions.

Example 9.46 *Let $u(x) = x$ and $v(x) = x^2 - 1$. Then the density $f(\cdot)$ has mean zero and variance 1 as*

$$\int v(x)f(x) = \int (x^2 - 1)f(x) = 0 \Rightarrow \mathbb{E}[X^2] = 1.$$

Therefore, the parameters μ, σ are respectively the mean and standard deviation of X_1.

Example 9.47 *We define the constraint functions*

$$u(x) = sign(x)$$
$$v(x) = sign(|x|-1)$$

Then the density $f(\cdot)$ has median zero and median absolute value 1.

Example 9.48 *We have that $f(\cdot)$ has -1 as the lower quartile and 1 as the upper quartile*

$$u(x) = 1\{x \leq -1\} - \frac{1}{4}$$
$$v(x) = 1\{x \geq 1\} - \frac{1}{4}$$

The parameters μ, σ become respectively the average of the quartiles and half the IQR of X_1.

Now, we shall impose **two assumptions** on the integral-constrained semiparametric location-scale model.

1) We assume that the 2-parameter location-scale model where $f(\cdot)$ is known based on $Y_1, ..., Y_n$ is i.i.d with common density $\frac{1}{\sigma}f(\frac{y-\mu}{\sigma})$ satisfies the LAN property at the true values μ_0, σ_0 with Score functions given by

$$\underset{\sim}{\ell_\theta}{}^\circ(y) = \begin{pmatrix} \frac{1}{\sigma_0}\psi(\frac{y-\mu_0}{\sigma_0}) \\ \\ \frac{1}{\sigma_0}\chi(\frac{y-\mu_0}{\sigma_0}) \end{pmatrix} \tag{9.108}$$

where under differentiability conditions, the Score functions are

$$\begin{cases} \psi(x) = \frac{-f^{'}(x)}{f(x)} \\ \\ \chi(x) = x\psi(x) - 1 \end{cases}$$

and the information matrix is

$$J_{\underset{\sim}{\theta},\underset{\sim}{\theta}} = \mathbb{E}[\underset{\sim}{\ell_\theta}{}^\circ(X_1)\underset{\sim}{\ell_\theta}{}^\circ(X_1)^T] \tag{9.109}$$

2) We assume that there exists a complete orthonormal basis for $L_2(f)$ of the form

$$\{1, u^*, v^*\} \cup \{b_j\}$$

where we have the orthonormalised constraint functions as part of the basis

$$u^* = \frac{u}{||u||_f} \qquad v^* = \frac{v - \frac{\langle v,u\rangle_f u}{||u||_f^2}}{\sqrt{||v||_f^2 - \frac{\langle v,u\rangle_f^2}{||u||_f^2}}} \tag{9.110}$$

We now shall only focus on the nuisance parameter $f(\cdot)$. We shall consider the parametric submodels for the shape (no location or scale parameter)

$$\{f_{\underset{\sim}{\eta}}(\cdot) : \underset{\sim}{\eta} \in \mathcal{H} \subseteq \mathbb{R}^{d_{\underset{\sim}{\eta}}}\}$$

for the shape f, which includes the true density f as $f_{\underset{\sim}{0}}$ where $\underset{\sim}{\eta}$ are the Nuisance parameters.

We **assume** that the LAN condition holds at $\underset{\sim}{\eta} = \underset{\sim}{0}$ with the (nuisance) Score Functions

$$\underset{\sim}{\ell_\eta^\circ}$$

and information

$$J_{\underset{\sim}{\eta\eta}}$$

As in the integral-constrained location-scale model, we must have that the nuisance Scores for this shape model is orthogonal to the constraint functions

$$\langle \underset{\sim}{\ell_\eta^\circ}, u\rangle \int \underset{\sim}{\ell_\eta^\circ}(x)u(x)f(x)dx = 0$$

$$\langle \underset{\sim}{\ell_\eta^\circ}, v\rangle \int \underset{\sim}{\ell_\eta^\circ}(x)v(x)f(x)dx = 0$$

This is the conditions needed for the nuisance shape model.

Definition 26: Parametric submodel for the integral-constrained location-scale semiparametric model

or each $k \geq 1$, for the integral-constrained location-scale semiparametric model, we can construct a parametric submodel with common density of the form

$$q(x; \mu, \sigma, \underset{\sim}{\eta}) = f_{\underset{\sim}{\eta}}(\frac{x-\mu}{\sigma}) \tag{9.111}$$

whereby $f_{\underset{\sim}{\eta}}$ is the parametric model for the nuisance shape model.

Theorem 9: LAN condition for location-scale parametric submodel

The LAN conditions hold for the location-scale parametric submodel with Score functions

$$\begin{pmatrix} \underset{\sim}{\ell_\theta}^{\circ}(x) \\ \frac{1}{\sigma} b_1(\frac{x-\mu}{\sigma}) \\ \cdots \\ \frac{1}{\sigma} b_k(\frac{x-\mu}{\sigma}) \end{pmatrix} \tag{9.112}$$

whereby $\underset{\sim}{\ell_\theta}^{\circ}(\cdot)$ is a bivariate vector containing the Score functions ψ and χ for μ and σ respectively. Furthermore, $b_j(\cdot)$ are the basis functions being the Nuisance scores.
The information matrix is given by

$$\begin{bmatrix} \underset{\sim}{J_{\theta\theta}} & \underset{\sim}{J_{\theta\eta}} \\ \underset{\sim}{J_{\eta\theta}} & \sigma^{-2}\underset{\sim}{I} \end{bmatrix} \tag{9.113}$$

as the basis functions $b_j(\cdot)$ are orthonormal with scaling parameter $1/\sigma$ and $\underset{\sim}{J_{\eta\theta}}$ is the $2 \times k$-matrix

$$\mathbb{E}[\underset{\sim}{\ell_\theta}^{\circ}(X_1)\frac{1}{\sigma}\underset{\sim}{b}(\frac{X_1-\mu}{\sigma})]$$

for $\underset{\sim}{b} = (b_1, \cdot\cdot\cdot, b_k)^T$.

Let $\mu_0 = 1$ and $\sigma_0 = 1$ to simplify notation. Then, the **effective score** is given by

$$\underset{\sim}{\ell_\theta}^{*} = \underset{\sim}{\ell_\theta}^{\circ} - \underset{\sim}{J_{\theta\eta}}\underset{\sim}{J_{\eta\eta}}^{-1}\underset{\sim}{\ell_\eta}^{\circ} \tag{9.114}$$

$$= \underset{\sim}{\ell_\theta}^{\circ} - \langle \underset{\sim}{\ell_\theta}^{\circ}, \underset{\sim}{b}^T \rangle \underset{\sim}{b} \tag{9.115}$$

and the **effective information** is

$$\underset{\sim}{J_\theta}^{*} = \underset{\sim}{J_{\theta\theta}} - \langle \underset{\sim}{\ell_\theta}^{\circ}, \underset{\sim}{b}^T \rangle^T \langle \underset{\sim}{\ell_\theta}^{\circ}, \underset{\sim}{b}^T \rangle$$

as $\underset{\sim}{J_{\eta\eta}} = \underset{\sim}{I}$.

Now, recall that the Score function $\underset{\sim}{\ell_\theta}^{\circ}$ can be written in terms of the basis $\{1, u^*, v^*\} \cup \{b_j\}$, which is

$$\underset{\sim}{\ell_\theta}^{\circ} = \langle \underset{\sim}{\ell_\theta}^{\circ}, u^* \rangle u^* + \langle \underset{\sim}{\ell_\theta}^{\circ}, v^* \rangle v^* + \sum_{j=1}^{\infty} \langle \underset{\sim}{\ell_\theta}^{\circ}, b_j \rangle b_j$$

We can therefore deduce the effective information.

Theorem 10: Effective information for semiparametric integral-constrained location-scale model

The effective information for the semiparametric integral-constrained location-scale model is

$$\underset{\sim}{J_\theta}^{*} = \langle u^*, \underset{\sim}{\ell_\theta}^{\circ} \rangle \langle \underset{\sim}{\ell_\theta}^{\circ T}, u^* \rangle + \langle v^*, \underset{\sim}{\ell_\theta}^{\circ} \rangle \langle \underset{\sim}{\ell_\theta}^{\circ T}, v^* \rangle = \underset{\sim}{M} \tag{9.116}$$

Proof: First, recall that the effective information is

$$\underset{\sim}{J_\theta}^{*} = \underset{\sim}{J_{\theta\theta}} - \langle \underset{\sim}{\ell_\theta}^{\circ}, \underset{\sim}{b}^T \rangle^T \langle \underset{\sim}{\ell_\theta}^{\circ}, \underset{\sim}{b}^T \rangle$$

Then, we have that the ordinary information is

$$\underset{\sim}{J_{\theta\theta}} = \langle \underset{\sim}{\ell_\theta}^{\circ}, \underset{\sim}{\ell_\theta}^{\circ T} \rangle = \langle u^*, \underset{\sim}{\ell_\theta}^{\circ} \rangle \langle \underset{\sim}{\ell_\theta}^{\circ T}, u^* \rangle + \langle v^*, \underset{\sim}{\ell_\theta}^{\circ} \rangle \langle \underset{\sim}{\ell_\theta}^{\circ T}, v^* \rangle + \sum_{j=1}^{\infty} \langle b_j, \underset{\sim}{\ell_\theta}^{\circ} \rangle \langle \underset{\sim}{\ell_\theta}^{\circ T}, b_j \rangle$$

Additionally, the coinformation is given by

$$\langle \underset{\sim}{\ell_\theta}^{\circ}, \underset{\sim}{b}^T \rangle = \sum_{j=1}^{k} \langle b_j, \underset{\sim}{\ell_\theta}^{\circ T} \rangle \langle \underset{\sim}{\ell_\theta}^{\circ}, b_j \rangle$$

Therefore, as we let $k \to \infty$, we have that the terms from the coinformation will cancel out with terms from the ordinary information and hence we have that

$$\underset{\sim}{J_\theta}^{*} = \langle u^*, \underset{\sim}{\ell_\theta}^{\circ} \rangle \langle \underset{\sim}{\ell_\theta}^{\circ T}, u^* \rangle + \langle v^*, \underset{\sim}{\ell_\theta}^{\circ} \rangle \langle \underset{\sim}{\ell_\theta}^{\circ T}, v^* \rangle + \sum_{j=k+1}^{\infty} \langle b_j, \underset{\sim}{\ell_\theta}^{\circ} \rangle \langle \underset{\sim}{\ell_\theta}^{\circ T}, b_j \rangle$$

which is bounded from below by the first two terms

$$\geq \langle u^*, \underset{\sim}{\ell_\theta}^{\circ} \rangle \langle \underset{\sim}{\ell_\theta}^{\circ T}, u^* \rangle + \langle v^*, \underset{\sim}{\ell_\theta}^{\circ} \rangle \langle \underset{\sim}{\ell_\theta}^{\circ T}, v^* \rangle = \underset{\sim}{M}.$$

■

Therefore, as we have seen, the effective information for the semiparametric location-scale model is

$$\langle u^*, \underset{\sim}{\ell_\theta}^{\circ} \rangle \langle \underset{\sim}{\ell_\theta}^{\circ T}, u^* \rangle + \langle v^*, \underset{\sim}{\ell_\theta}^{\circ} \rangle \langle \underset{\sim}{\ell_\theta}^{\circ T}, v^* \rangle$$

which is the parameter Score function $\underset{\sim}{\ell_\theta}^{\circ}$ projected onto the constraint functions u^*, v^*.

The lower bound to the limiting covariance matrix of the scaled estimation error of regular estimators is the inverse of $\underset{\sim}{M}$

We note that $\underset{\sim}{M}$ is the covariance matrix of the linear combination of the orthonormalised constraint functions u^*, v^*

$$\langle \underset{\sim}{\ell_\theta}^\circ, u^* \rangle u^* + \langle \underset{\sim}{\ell_\theta}^\circ, v^* \rangle v^*$$

whereby the effective Score function was

$$\underset{\sim}{\ell_\theta}^\circ = \langle \underset{\sim}{\ell_\theta}^\circ, u^* \rangle u^* + \langle \underset{\sim}{\ell_\theta}^\circ, v^* \rangle v^* + \sum_{j=1}^{\infty} \langle \underset{\sim}{\ell_\theta}^\circ, b_j \rangle b_j$$

where the effective information is the first two terms of the effective score function.

We can write

$$\langle \underset{\sim}{\ell_\theta}^\circ, u^* \rangle u^* + \langle \underset{\sim}{\ell_\theta}^\circ, v^* \rangle v^* = \begin{pmatrix} \langle \psi, u^* \rangle & \langle \chi, u^* \rangle \\ \langle \chi, u^* \rangle & \langle \psi, u^* \rangle \end{pmatrix} \begin{pmatrix} u^* \\ v^* \end{pmatrix} = \underset{\sim}{C}^T \begin{pmatrix} u^* \\ v^* \end{pmatrix}$$

as $\langle \underset{\sim}{\ell_\theta}^\circ, u^* \rangle$ is

$$\begin{pmatrix} \langle \psi, u^* \rangle \\ \langle \chi, u^* \rangle \end{pmatrix}$$

where u^*, v^* has the identity matrix as their covariance matrix.

Therefore, we have that

$$\underset{\sim}{M} = \underset{\sim}{C}^T \underset{\sim}{C}$$

We can then describe the influence function needed for an asymptotically efficient estimator.

Proposition 33: Efficient Influence Function for location-scale model

The efficient influence function for the location-scale semiparametric model is a multiple of the constraint functions

$$\underset{\sim}{\tilde{\ell}} = (\underset{\sim}{C}^T \underset{\sim}{C})^{-1} \underset{\sim}{C}^T \begin{pmatrix} u^* \\ v^* \end{pmatrix}$$

Proof: We see that any estimator that has $\underset{\sim}{\tilde{\ell}}$ as its influence function will obtain the asymptotic lower bound as

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \underset{\sim}{\tilde{\ell}}(X_i) \xrightarrow{d} \mathcal{N}(0, \underset{\sim}{M}^{-1})$$

■

We now want to know how to find estimators with such influence functions.

Proposition 34: Finding optimal estimators

Solving the estimating equation

$$\sum_{i=1}^{n} \begin{pmatrix} u(\frac{x_i - \mu}{\sigma}) \\ v(\frac{x_i - \mu}{\sigma}) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

will yield estimators $\tilde{\mu}, \tilde{\sigma}$ which will have

$$\underset{\sim}{\tilde{\ell}} = (\underset{\sim}{C}^T \underset{\sim}{C})^{-1} \underset{\sim}{C}^T \begin{pmatrix} u^* \\ v^* \end{pmatrix}$$

as their influence function.

We can now generalise what we have described so far where now we have the true values μ_0 and σ_0.

Proposition 35: Complete Orthonormal basis for general location-scale model

The complete orthonormal basis for the general location-scale model is given by

$$\left\{1, \frac{1}{\sigma_0}u^*(\frac{x-\mu_0}{\sigma_0}), \frac{1}{\sigma_0}v^*(\frac{x-\mu_0}{\sigma_0})\right\} \cup \left\{\frac{1}{\sigma_0}b_j(\frac{x-\mu_0}{\sigma_0})\right\} \tag{9.117}$$

Proposition 36: Effective information for general location-scale model

The effective information for the general location-scale model is given by

$$\frac{1}{\sigma^2}\underset{\sim}{M} \tag{9.118}$$

Corollary 9.49 *Optimal estimators that obtains the limiting variance $\sigma^2\underset{\sim}{M}^{-1}$ can be obtained by solving the estimating equations*

$$\sum_{i=1}^{n}\begin{pmatrix}u(\frac{x_i-\mu}{\sigma})\\ v(\frac{x_i-\mu}{\sigma})\end{pmatrix} = \begin{pmatrix}0\\0\end{pmatrix} \tag{9.119}$$

STAT5610: Advanced Inference

10. Optimal Semiparametric Testing

10.3 Semiparametric Testing

Suppose we have a semiparametric model $\{\mathbb{P}_{n\underset{\sim}{\theta} f} : \underset{\sim}{\theta} \in \Theta; f \in \mathcal{F}\}$ where $\mathbb{P}_{n\underset{\sim}{\theta} f}$ is the joint distribution of $X_1, ..., X_n$ that is i.i.d with common density $p(x; \underset{\sim}{\theta}, f)$ and we wish to test $\underset{\sim}{\theta} = \underset{\sim}{\theta_0}$.

As before, we can construct a sequence of regular parametric submodels and run tests on each of them.

Proposition 37: Power of test for semiparametric model

The power of a test for the semiparametric model is bounded above by the highest power for any parametric submodel, which in this case is $\underset{\sim}{h_\theta}^T \underset{\sim}{J_\theta}^* \underset{\sim}{h_\theta}$.

Remark 10.50 *Hence, in regular parametric models, the best power of the test is bounded by the effective information matrix.*

Generally, we tend to use an asymptotically efficient estimator.

Proposition 38: Optimal Test Statistic

The regular quadratic test statistic with the highest power in the regular parametric submodel is given by

$$T_n = (\underset{\sim}{S_\theta}^*)^T (\underset{\sim}{J_\theta}^*)^{-1} \underset{\sim}{S_\theta}^* + o_p(1) \tag{10.120}$$

From this, we can define a quadratic form test statistic for each regular parametric submodel of the semiparametric model and the limit of such a sequence will gives us information about the semiparametric test.

Proposition 39: Optimal Test Statistic in Integral Constrained Location Model

An optimal test statistic of $H : \theta = \theta_0$ is of the form

$$T_n = \frac{\left(n^{-1/2} \sum_{i=1}^n w(X_i - \theta_0)\right)^2}{Var_{\theta_0, f_0}[w(X_1, \theta_0)]} + o_p(1) \tag{10.121}$$

Proposition 40: Optimal Test Statistic in Integral Constrained Symmetric Model

An optimal test statistic of $H : \theta = \theta_0$ is of the form

$$T_n = \frac{\left(n^{-1/2} \sum_{i=1}^n \psi_0(X_i - \theta_0)\right)^2}{\int \psi_0^2(x) f(x) dx]} + o_p(1) \tag{10.122}$$

STAT5610: Advanced Inference

11. Probability Theory

11.4 Probability Theory

11.4.1 Modes of convergence

Definition 11.51 *(Convergence in distribution). $X_n \xrightarrow{D} X$ if for each continuous point x of $F(x)$,*

$$F_n(x) \to F(x)$$

where $F_x(x) = P(X_n \leq x)$ and $F(x) = P(X \leq x)$.

Definition 11.52 *(Convergence almost surely). $X_n \xrightarrow{a.s.} X$ if*

$$P(\omega : \lim_{n\to\infty} X_n = X) = 1$$

Definition 11.53 *(Convergence in probability). $X_n \xrightarrow{p} X$ if for every $\epsilon > 0$,*

$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$$

Definition 11.54 *(Convergence in pth moment). For a real number $p > 0$ and $X_n \in \mathcal{L}^p$. Then, $X_n \xrightarrow{L_p} X$ if*

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0$$

The relationships between them are as follows.

Proposition 11.55 *(Relationships between modes of convergence).*

1. *$X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{p} X$*
2. *If for some $p > 0$, $X_n \xrightarrow{L_p} X \Rightarrow X_n \xrightarrow{p} X$*
3. *$X_n \xrightarrow{p} X \Rightarrow X_n \xrightarrow{D} X$*

Theorem 11.56 *Suppose that $X_n \xrightarrow{p} X$ and $|X_n| \leq Y$ where $Y \in \mathcal{L}^p$ for $p > 0$. Then*

$$X_n \xrightarrow{L_p} X.$$

Theorem 11.57 *(Borel-Cantelli Lemma). If for every $\epsilon > 0$,*

$$\sum_{n=1}^{\infty} P(|X_n - X| \geq \epsilon) < \infty$$

then

$$X_n \xrightarrow{a.s.} X.$$

Proposition 11.58 *Assume $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$. Then*

$$X_n + Y_n \xrightarrow{p} X + Y$$

and

$$X_n Y_n \xrightarrow{p} XY$$

Theorem 11.59 *(Chebychev Inequality). Suppose $\psi(x)$ is a function such that:*

1. $\psi(x) > 0$
2. $\psi(x) \uparrow \infty$
3. $\mathbb{E}[\psi(|X|)] < \infty$

Then, for any $x > 0$

$$P(|X| \geq x) \leq \frac{1}{\psi(x)} \mathbb{E}[\psi(x)]$$

Proposition 11.60 *(Helly-Bray Theorem). The sequence $X_n \xrightarrow{D} X$ if and only if $\mathbb{E}[g(X_n)] \to \mathbb{E}[g(X)]$ for all continuous bounded functions g.*

Definition 11.61 *(Characteristic Function). The characteristic function of random variable X is defined by*

$$\psi_X(t) = \mathbb{E}[exp(itX)]$$

for all $t \in \mathbb{R}$ and $i = \sqrt{-1}$.

Theorem 11.62 *(Continuity Theorem). By the Helly-Bray theorem, we have that*

$$X_n \xrightarrow{D} X \leftrightarrow \psi_{X_n}(t) \to \psi_{X(t)}$$

Lemma 11.63 *For a constant C, $X_n \xrightarrow{D} C$ if and only if $X_n \xrightarrow{p} C$.*

Proposition 11.64 *(Slutsky's Theorem). Suppose that $X_n \xrightarrow{D} X, Y_n \xrightarrow{p} Y, Z_n \xrightarrow{p} b$ where a and b are constants. Then*

$$(X_n + Y_n)Z_n \xrightarrow{D} (X_n + a)b$$

Remark 11.65 *This does* **not** *hold that $X_n + Y_n \xrightarrow{D} X + Y$ if $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{D} Y$.*

Corollary 11.66 *Suppose that $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{Y}$. Furthermore, suppose that X_n and Y_n are* ***independent sequences****. Then*

$$(X_n, Y_n) \xrightarrow{D} (X, Y).$$

Proposition 11.67 *(Continuous Mapping Theorem). Let $f(x)$ be a continuous function. Suppose that $X_n \xrightarrow{D} X$, then*

$$f(X_n) \xrightarrow{D} f(X).$$

Proposition 11.68 *(Delta Method). Suppose there exists a sequence of random variables X_n such that*

$$d_n(X_n - \theta) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$

where $0 < d_n \to \infty$ and θ, σ^2 are constants. Then, for any function $g(.)$ such that $g^{'}(\theta)$ exists and $g^{'}(\theta) \neq 0$ then

$$d_n(g(X_n) - g(\theta)) \xrightarrow{D} \mathcal{N}(0, \sigma^2[g^{'}(\theta)]^2)$$

11.4.2 Sequences of independent random variables

Theorem 11.69 *(Weak law of large numbers). Let $(X_n : n \geq 1)$ be i.i.d random variables. Let $S_n = \sum_{j=1}^n X_j$. Then, if $\mathbb{E}[|X|] < \infty$, then*

$$S_n \xrightarrow{p} \mathbb{E}[X_j].$$

Theorem 11.70 *(Strong law of large numbers). Let $(X_n : n \geq 1)$ be i.i.d random variables. Let $S_n = \sum_{j=1}^n X_j$. Then, if $\mathbb{E}[|X|] < \infty$, then*

$$S_n \xrightarrow{a.s.} \mathbb{E}[X_j].$$

This is a useful inequality.

Theorem 11.71 *(Kolmogorov Inequality). Let $(X_n : n \geq 1)$ be independent with $\mathbb{E}[X_j] = 0$ and $\sigma_j^2 = \mathbb{E}[X_j^2] < \infty$. Let $S_n = \sum_{j=1}^n X_j$. For all $A > 0$ and $N \geq 1$*

$$P[\max_{1 \leq j \leq N} |S_j| \geq A] \leq \frac{1}{A^2} \sum_{j=1}^N \sigma_j^2.$$

Theorem 11.72 *(Central Limit Theorem). Let $(X_j : j \geq 1)$ be i.i.d random variables with mean μ and finite variance σ^2. Then*

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{D} \mathcal{N}(0, \sigma^2).$$

We can now relax the identically distributed assumption for the central limit theorem. That is, the only condition we need is independent random variables.

Definition 11.73 *(Triangular array). A triangular array (X_{nj}) is when (X_{nj}) is independent for a fixed n.*

We now introduce an important condition.

Definition 27: Lindeberg Condition

For each n = 1,2,..., let X_{nj} for each $1 \leq j \leq n$ be independent random variables with $\mathbb{E}[X_{nj}] = 0$ and $Var[X_{nj}] = \sigma_{nj}^2$. Denote $B_n^2 = Var[\sum_{j=1}^n X_{nj}] = \sum_{j=1}^n \sigma_{nj}^2$. The Lindeberg condition states that for every $\epsilon > 0$

$$\lim_{n \to \infty} \frac{1}{B_n^2} \sum_{j=1}^n \mathbb{E}[X_{nj}^2 I\{|X_{nj}| \geq \epsilon B_n\}] = 0$$

We now state two important results from Lindeberg condition.

Claim 11.74 *Suppose that Lindeberg condition holds. Then*

$$\frac{1}{B_n^2}\max_{1\leq j\leq n}\sigma_{nj}^2 \to 0.$$

Claim 11.75 *Suppose that Lindeberg condition holds. Then*

$$\sum_{j=1}^{n} P[|X_{nj}|>\epsilon] \to 0$$

for all $\epsilon > 0$.

However, this is hard to check for. We can instead check Lyapunov theorem.

Definition 28: Lyapunov Theorem

For each n = 1,2,..., let X_{nj} for each $1 \leq j \leq n$ be independent random variables with $\mathbb{E}[X_{nj}] = 0$ and $Var[X_{nj}] = \sigma_{nj}^2$. Denote $B_n^2 = Var[\sum_{j=1}^n X_{nj}] = \sum_{j=1}^n \sigma_{nj}^2$. Suppose there exists a $\delta > 0$ such that

$$\lim_{n\to\infty}\frac{1}{B_n^{2+\delta}}\sum_{j=1}^{n}\mathbb{E}[|X_{nj}|^2] = 0$$

Theorem 11: Sufficient Conditions for Lindeberg Condition

If Lyapunov's theorem holds, this implies that Lindeberg's conditions hold.

Proof: Fix $\epsilon > 0$ and we want to show Lindeberg condition holds

$$\frac{1}{B_n^2}\sum_{j=1}^{n}\mathbb{E}[X_{nj}^2 I\{|X_{nj}|\geq \epsilon B_n\}] = \frac{1}{B_n^2}\sum_{j=1}^{n}\int_{|X_{nj}|\geq\epsilon B_n} X_{nj}^2 d\mu_j$$

Clearly, $1 \leq (\frac{|X_{nj}|}{\epsilon B_n})^\delta$ and therefore

$$\leq \frac{1}{\epsilon^\delta}\frac{1}{B_n^{2+\delta}}\sum_{j=1}^{n}\int_{|X_{nj}|\geq\epsilon B_n}|X_{nj}|^{2+\delta}d\mu_j \to 0$$

by Lyapunov theorem. ■

Theorem 12: Lindeberg-Feller Theorem

For each n = 1,2,..., let X_{nj} for each $1 \leq j \leq n$ be independent random variables with $\mathbb{E}[X_{nj}] = 0$ and $Var[X_{nj}] = \sigma_{nj}^2$. Denote $B_n^2 = Var[\sum_{j=1}^n X_{nj}] = \sum_{j=1}^n \sigma_{nj}^2$. Suppose that Lindeberg condition holds, then

$$\frac{\sum_{j=1}^n X_{nj}}{B_n} \xrightarrow{D} \mathcal{N}(0,1).$$

Definition 11.76 *(Uniform asymptotic negligibility). The u.a.n condition is*

$$\lim_{n\to\infty}\max_{1\leq j\leq n} P(|X_{nj}|\geq\epsilon) = 0$$

for all $\epsilon > 0$.

This leads us for another way to check if Lindeberg condition holds.

Proposition 11.77 *(Lindeberg condition is necessary and sufficient). Suppose that the u.a.n condition holds and* $\frac{\sum_{j=1}^{n} X_{nj}}{B_n} \xrightarrow{D} \mathcal{N}(0,1)$. *Then, the Lindeberg condition holds.*

11.4.3 Big-oh notation

Definition 29: Big-oh

For $n \geq 0$, we say that $a_n = \mathcal{O}(1)$ if

$$|a_n| \leq C$$

for some constant C. We say that $a_n = \mathcal{O}(b_n)$ if

$$\frac{a_n}{b_n} \to 0$$

as $n \to \infty$.

Definition 30: Bounded in probability

A sequence of random variables $\{X_k\}_{k\geq 1}$ is said to be bounded in probability if for every $\epsilon > 0$, there exists a constant $M = M(\epsilon)$ and rank $N = N(\epsilon)$ such that

$$\mathbb{P}(|X_k| \geq M) < \epsilon \tag{11.123}$$

for all $k \geq N$.

Definition 11.78 *(Big-oh in probability). The sequence of random variables* $\{X_n\}$ *is* $\mathcal{O}_p(1)$ *if it is bounded in probability. Additionally, if* $\{X_n\}$ *is* $\mathcal{O}_p(Y_n)$ *if* $\frac{X_n}{Y_n}$ *is* $\mathcal{O}_p(1)$.

Definition 11.79 *(Small-oh in probability). The sequence of random variables* $\{X_n\}$ *is* $o_p(1)$ *if* $X_n \xrightarrow{P} 0$. *Additionally,* $\{X_n\}$ *is* $o_p(Y_n)$ *if* $\frac{X_n}{Y_n} \xrightarrow{P} 0$.

STAT5610: Advanced Inference

12. Kernel Density Estimation

12.5 Kernel Density Estimation

12.5.1 Introduction to Kernel Density Estimation

The goal is to estimate the density function $f(x)$ of a random variable X. We do not want to specify a parameteric distribution as if it is mispecified, then this will lead to many issues.

Definition 31: Kernel Density Estimator

The general Kernel estimator has the form

$$\hat{f}(x) = \frac{1}{nh}\sum_{k=1}^{n} K(\frac{X_k - x}{h})$$

whereby $0 < h = h_n \to 0$ is called the bandwidth and $K(u)$ is a Kernel function satisfying

1. $K(u) \geq 0$
2. $\int K(u)du = 1$

We impose some assumptions that we shall assume throughout the lecture. First, assume that $X_1, ..., X_n$ are iid with density function $f(x)$. Second, we assume that $f^{''}(x)$ is continuous and bounded in the neighbourhood of x. Third, we assume that the Kernel function $K(x)$ satisfies that

1. $K(u) \geq 0$
2. $\int K(u)du = 1$
3. $K(u)$ is a symmetric function whereby $\int uK(u)du = 0 \Rightarrow K(-u) = K(u)$
4. $K(x) = 0$ for $|x| \geq A$ where A is a compact support
5. $h \to 0$ and $nh \to \infty$

proposition˙exam

12.5.2 Properties of Kernel Density Estimator

We now describe some useful properties of the kernel density estimator (KDE).

Proposition 12.80 *The Kernel density estimator is a valid probability density function.*

Proposition 41: Bias of KDE

The bias of the KDE is given by

$$Bias(\hat{f}(x)) = \mathbb{E}[\hat{f}(x)] - f(x) = \frac{1}{2}h^2 f^{''}(x) \int u^2 K(u)du + o(h^2) \tag{12.124}$$

Proof: (Sketch). By the i.i.d assumption, we can arrive at

$$\mathbb{E}[\hat{f}(x)] - f(x) = \frac{1}{h}\int K(\frac{x-x_0}{h})f(x)dx - f(x_0)$$

and then use a change of variables $y = (x - x_0)h^{-1}$ and $hdy = dx$ to get

$$= \int K(y)f(x_0 + hy)dy - f(x_0)$$

We can then apply a second order Taylor expansion to $f(x_0 + hy)$ to get

$$= \int K(y)[f(x_0) - hyf^{'}(x_0) + \frac{1}{2}h^2y^2f^{''}(x_0) + o(h^2)]dy - f(x_0)$$

and using the fact that $\int K(y)dy = 1$ and $\int K(y)ydy = 0$, we arrive at the desired result. ■

Remark 12.81 *This has the interesting interpretation that as $h \to 0$, the bias of the KDE shrinks at rate $O(h^2)$. Therefore, the bias is affected by the curvature of the density function $f^{''}$.*

Proposition 42: Variance of KDE

The variance of the KDE is given by

$$Var[\hat{f}(x)] = \frac{1}{nh}f(x)\int K^2(u)du + o(1/nh) \tag{12.125}$$

Proof:(Sketch). We bound the variance by the second moment

$$Var[\hat{f}(x)] = Var[\frac{1}{nh}\sum_{i=1}^{n} K(\frac{X_i - x_0}{h})] \leq \frac{1}{nh}\int K^2(\frac{x-x_0}{h})f(x)dx$$

Let $y = (x - x_0)h^{-1}$ and $hdy = dx$ to arrive at

$$= \frac{1}{nh}\int K^2(y)f(x_0 + hy)dy$$

and apply a **first order** Taylor expansion

$$= \frac{1}{nh}\int K^2(y)[f(x_0) + hyf^{'}(x_0) + o(h)]dy$$

and expand out to arrive at the desired result. ■

Remark 12.82 *The variance shrinks at rate $O((nh)^{-1})$ when $n \to \infty$ and $h \to 0$. Additionally, the variance is large when the density value is large.*

We now describe a pointwise measure known as the mean squared error (MSE) of the estimator.

Proposition 43: MSE of KDE

The MSE of the KDE is given by

$$MSE(\hat{f}(x)) = \frac{1}{nh}f(x)\int K^2(u)du + \frac{1}{4}h^4[f^{''}(x)\tau_2]^2 + o(\frac{1}{nh} + h^4) \tag{12.126}$$

whereby $\tau_j = \int u^j K(u)du$.

Proof:(Sketch). We use the fact that the MSE can be decomposed as

$$MSE(\hat{f}(x)) = bias(\hat{f}(x))^2 + Var[\hat{f}(x)]$$

and use the previous results we derived. ■

However, such a measure is only a pointwise measure. We can define the mean integrated squared error (MISE) to get an idea of the global accuracy of the KDE $\hat{f}(x)$.

Proposition 44: MISE of KDE

The MISE of the KDE is given by

$$MISE(\hat{f}(x)) = \frac{1}{nh}\int K^2(u)du + \frac{1}{4}h^4\tau_2^2\int [f^{''}(x)]^2dx + o(\frac{1}{nh} + h^4) \tag{12.127}$$

whereby $\tau_j = \int u^j K(u)du$.

Proof:(Sketch). Use the fact that MISE is

$$MISE(\hat{f}(x)) = \int MSE(\hat{f}(x))dx = \int [\frac{1}{nh}f(x)\int K^2(u)du + \frac{1}{4}h^4[f^{''}(x)\tau_2]^2 + o(\frac{1}{nh} + h^4)]dx$$

and that $\int f(x)dx = 1$ to get the desired result. ■

We can see that the $MISE(\hat{f}(x)) = \mathcal{O}(h^4) + \mathcal{O}(\frac{1}{nh})$.

The MISE can be made small if we balance out between the bias and variance term. From MISE of the KDE, these are balanced out if $(nh)^{-1} \sim h^4$ which implies that the optimal bandwidth $h \sim n^{-1/5}$. If this is the case, then the MISE will be $n^{-4/5}$.

Theorem 12.83 *(Optimal bandwidth). Minimising the MISE, the optimal bandwidth h_{opt} is given by*

$$h_{opt} = \Big(\frac{1}{n}\frac{\int K^2(u)dy}{\tau_2^2 \int [f''(x)]^2 dx}\Big)^{1/5} = Cn^{-1/5}$$

where C is a constant.

A big issue is that we do not know the true density $f(x)$ and therefore, it is impossible for us to know what $f^{''}(x)$ is. If we did, we wouldn't need to use nonparametric methods!

A **plug-in** method involves first setting an initial value of h to estimate $\int [f^{''}(x)]^2dx$ and then use this estimate to estimate h_{opt}. For example, if we assume $f(x)$ belongs to a Gaussian distribution with variance

σ^2, then it can be shown that our initial estimate will be

$$h_{initial} \approx 1.06\sigma n^{-1/5}$$

which we then plug into $\int [f^{''}(x)]^2 dx$ and estimate h_{opt}. However, sometimes the initial estimate is used as the bandwidth and is known as the **rule of thumb bandwidth**.

12.5.3 Asymptotic Results

We now give the first result describing the limiting distribution of the KDE.

Proposition 12.84 *(Asymptotic normality). Suppose that $nh \to \infty$ and $nh^5 \to 0$, then*

$$\sqrt{nh}[\hat{f}(x) - f(x)] \xrightarrow{D} \mathcal{N}(0, f(x)\int K^2(u)dy)$$

Remark 12.85 *An issue is that the optimal bandwidth $h = cn^{-1/5}$ which means we are unable to use our optimal estimate for this result.*

We can relax the bandwidth condition.

Proposition 45: Relaxed asymptotic normality

Suppose that $nh \to \infty$ and $nh^5 = O(1)$. Then

$$\sqrt{nh}[\hat{f}(x) - f(x) - \frac{1}{2}h^2 f^{''}(x)\int u^2 K(u)dy] \xrightarrow{D} \mathcal{N}(0, f(x)\int K^2(u)dy)$$

In fact, we can show that the KDE at different points are asymptotically independent.

That is, for $x_1 \neq x_2$, if we let $Y_{n1} = \sqrt{nh}(\hat{f}(x_1) - f(x_1))$ and $Y_{n2} = \sqrt{nh}(\hat{f}(x_2) - f(x_2))$, then

$$(Y_{n1}, Y_{n2}) \xrightarrow{D} (Y_1, Y_2)$$

whereby $(Y_1, Y_2) \sim \mathcal{N}(0, \Sigma)$ whereby

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

.

We can also show that the KDE is consistent.

Proposition 46: Consistency of KDE

For a fixed x, it can be shown that

$$\hat{f}(x) \xrightarrow{p} f(x)$$

or equivalently

$$\hat{f}(x) - f(x) = \mathcal{O}_p(h^2 + (nh)^{-1/2}).$$

We can in fact show an even stronger form of convergence under additional assumptions.

Theorem 12.86 *Define the Fourier transform of the KDE as*

$$\hat{K}(x) = \int e^{itx} K(t) dt$$

and assume that $\hat{K}(x)$ is uniformly integrable. Furthermore, assume that $f(x)$ is uniformly continuous on $\mathbb{R}$ and $nh^2 \to \infty$. Then

$$\sup_x |\hat{f}(x) - f(x)| \xrightarrow{p} 0$$

We can also estimate the cumulative distribution function $F(x)$ from the KDE.

Definition 32: CDF Estimation

Let $\hat{f}(x)$ be the KDE. Then, assuming the Kernel is symmetric $(K(-u) = K(u))$, the CDF estimate is

$$\hat{F}(x) = \int_{-\infty}^{x} \hat{f}(t) dt = \frac{1}{n} \sum_{k=1}^{n} G(\frac{x - X_k}{h})$$

whereby $G(x) = \int_{-\infty}^{x} K(u) du$.

STAT5610: Advanced Inference

13. Nonparametric Regression

13.6 Nonparametric Regression

13.6.1 Nadaraya-Watson Estimator

Let (X,Y) be a random vector. Our goal is to estimate the regression function $m(x) = \mathbb{E}[Y|X = x]$ from a random sample $(x_1, y_1), ...(x_n, y_n)$. The reason we want to estimate the conditional expectation since $m(x) = \mathbb{E}[Y|X]$ minmises the MSE $\mathbb{E}[(Y - m(X))^2]$.

The regression model is defined by

$$y_i = m(x_i) + \epsilon_i \quad \text{for } i = 1, 2, ..., n. \tag{13.128}$$

Definition 33: Nadaraya-Watson Estimator

The Nadaraya-Watson estimator is defined to be an estimate of the conditional expectation

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K(\frac{x_i - x}{h}) y_i}{\sum_{i=1}^{n} K(\frac{x_i - x}{h})} \tag{13.129}$$

whereby $h = h(n)$ is the bandwidth and $K(\cdot)$ is a kernel function satisfying $K(u) \geq 0$ and $\int K(u)du = 1$.

We shall assume that $x_1, ..., x_n$ are iid with density f(x) and $\epsilon_1, ..., \epsilon_n$ are iid with $\mathbb{E}[\epsilon_1] = 0$ and $\sigma^2 = \mathbb{E}[\epsilon_i^2] < \infty$. Furthermore, x_i and ϵ_i are independent. We also assume that $m^{''}$ and $f^{''}(x)$ are continuous and bounded in the neighbourhood of x. Finally, we assume that $K(-u) = K(u)$ and $K(x) = 0$ for $|x| \geq A$ where A is a compact set.

Proposition 47: Asymptotic Normality

Suppose that $nh \to \infty$ and $nh^3 \to 0$. Assume that $\mathbb{E}[|\epsilon_1|^{2+\delta}] < \infty$ for some $\delta > 0$. Then

$$\sqrt{nh}(\hat{m}(x) - m(x)) \xrightarrow{D} \mathcal{N}(0, f^{-1}(x)\sigma^2 \int K^2(u)du) \tag{13.130}$$

We can impose further restrictions to remove the bias.

Proposition 48: Asymptotic Normality without bias

Suppose that $nh \to \infty$ and $nh^5 = o(1)$ and $\mathbb{E}[|\epsilon_1|^{2+\delta}] < \infty$ for some $\delta > 0$. Then

$$\sqrt{nh}(\hat{m}(x) - m(x) - h^2 B_x \int u^2 K(u) du) \xrightarrow{D} \mathcal{N}(0, f^{-1}(x)\sigma^2 \int K^2(u) du) \tag{13.131}$$

whereby $B_x = \frac{1}{2} m^{''}(x) + \frac{f^{'}(x) m^{'}(x)}{f(x)}$.

Proposition 13.87 *(MSE). The MSE of the NW-estimator is given by*

$$MSE(\hat{m}(x)) \sim \frac{1}{nh} \frac{\sigma^2 \int K^2 du}{f(x)} + h^4 B_x^2 (\int u^2 K(u) du)^2$$

STAT5610: Advanced Inference

14. Martingale Central Limit Theorems

14.7 Martingale Central Limit Theorems

14.7.1 Tightness and Weak Convergence

Definition 14.88 *(Relatively Compact). A family of distribution functions $\{F_n : n \geq 1\}$ is relatively compact if for every sequence of distribution functions $\{F_n : n \geq 1\}m$ there exists a subsequence $(M_k)_{k\geq 1}$ that weakly converges to a generalised distribution function G*

$$F_{m_k} \xrightarrow{d} G.$$

Definition 34: Tightness

A family of measures $\{\mu_\alpha : \alpha \in I\}$ is tight if for every $\epsilon > 0$, there exists a compact set $K \subseteq \mathbb{R}$ such that

$$\mu_\alpha(K) \geq 1 - \epsilon$$

for all $\alpha \in I$.

Remark 14.89 *Alternatively, we can think of tightness as that the measure is uniformly small outside the compact set*

$$\mu_\alpha(\mathbb{R} \setminus K) \leq \epsilon$$

Theorem 14.90 *(Prokhorov's Theorem). A family of probability measures $\{\mu_\alpha : \alpha \in I\}$ is relatively compact if and only if it is tight.*

Lemma 14.91 *Let $\{\mu_\alpha : \alpha \in I\}$ be a tight sequence. If every weakly convergent subsequence has the same limit μ, then*

$$\mu_n \xrightarrow{w} \mu.$$

Lemma 14.92 *Let $\{\mu_\alpha : \alpha \in I\}$ be a tight sequence. Define the characteristic function associated to each μ_n*

$$\psi_n(t) = \int_{-\infty}^{\infty} e^{itx} d\mu_n(x).$$

Then, $\mu_n \xrightarrow{\mu}$ if and only if for all t, $\psi_n(t)$ has the same limit as $n \to \infty$.

14.7.2 Uniform Integrability

We first give a motivation for uniform integrability. Assume that $X \in L^p$. For a fixed $C \in \mathbb{R}$, define the event

$$\{|X|\geq C\}$$

It is clear that this is a decreasing sequence of events, that is, for $C_1 < C_2$

$$\{|X| \geq C_2\} \subseteq \{|X| \geq C_1\}$$

Furthermore, if we define the random variable

$$1\{|X| \geq C\}$$

this is an **nonincreasing** random variable in C.

Lemma 14.93 *Let $X \in L^p$. The random variable*

$$1\{|X| \geq C\} \xrightarrow{a.s.} 0$$

as $C \to \infty$.

Proof: As $X \in L^p$, we have that $\{|X| \geq C\} \downarrow \emptyset$. Therefore, as $C \to \infty$, the random variable X cannot be larger than C. ■

From this, we can also define the random variable

$$|X|1\{|X| \geq C\}$$

whereby it is clear that

$$|X|1\{|X| \geq C\} \leq |X|$$

with $\mathbb{E}[|X|^p] < \infty$. Using a similar line of reasoning, we have that

$$|X|1\{|X| \geq C\} \xrightarrow{a.s.} 0$$

as $C \to 0$. This leads us to an important result.

Lemma 14.94 *Let $X \in L^p$. Then,*

$$\lim_{C \to \infty} \mathbb{E}[|X|1\{|X| \geq C\}] = 0$$

Proof: Since $|X|1\{|X| \geq C\} \xrightarrow{a.s.} 0$, we can apply the dominating convergence theorem as $X \in L^p$

$$\lim_{C \to \infty} \mathbb{E}[|X|1\{|X| \geq C\}] = \mathbb{E}[\lim_{C \to \infty} |X|1\{|X| \geq C\}] = 0$$

■

This leads us to an important definition.

Definition 35: Uniformly integrable

Define the family of random variables $\{X_t : t \in I\}$ such that $X_t \in L^p$ for $t \in I$. Then, the family of random variables $\{X_t : t \in I\}$ is **uniformly integrable** if

$$\lim_{C \to \infty} \sup_{t \in I} \int_{|X_t| \geq C} |X_t| dP = 0$$

or equivalently

$$\lim_{C \to \infty} \sup_{t \in I} \mathbb{E}[|X_t| I\{|X_t| \geq C\}] = 0$$

Lemma 14.95 *If $\{X_t : t \in I\}$ is uniformly integrable, then*

$$\sup_{t\in I} \mathbb{E}[|X_t|] < \infty$$

Proposition 14.96 *Suppose that $X_n \xrightarrow{D} X$ and $\{|X_n|: n \geq 1\}$ is uniformly integrable. Then*

$$\mathbb{E}[X_n] \to \mathbb{E}[X].$$

We now state an important result showing how to arrive at convergence in mean from convergence in probability.

Proposition 49: Modes of convergence

Suppose that $X_n \in L^p$ and $X_n \xrightarrow{p} X$. Then, the following are equivalent:

1. $X_n \xrightarrow{L^p} X$
2. X_n is uniformly integrable
3. $||X||_p \to ||X||_p$

Remark 14.97 *This is an useful result because if we have convergence in probability and we have shown that the sequence $(X_n)_{n\geq 1}$ is uniformly integrable, then we are guaranteed convergence in p-th mean.*

14.7.3 Martingale Review

Definition 14.98 *(Filtration). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A filtration $\{\mathcal{F}_n\}_{n\geq 1}$ is a sequence of $\sigma-$fields such that*

1. *$\mathcal{F}_n$ is a $\sigma-$field for all n*
2. *$\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$ for all n.*

Definition 36: Martingale

A discrete-time martingale is a stochastic process $(X_n)_{n\geq 1}$ such that

1. X_n is adapted to $\mathcal{F}_n$
2. $\mathbb{E}[|X_n|] < \infty$ for all n
3. $\mathbb{E}[X_n|\mathcal{F}_{n-1}] = X_{n-1}$.

We can extend this idea further in the following definition.

Definition 37: Martingale Difference Sequence

An adapted sequence $\{X_t, \mathcal{F}_t\}$ is a Martingale Difference Sequence if

$$\mathbb{E}[|X_n|] < \infty$$

and

$$\mathbb{E}[X_{n+1}|\mathcal{F}_{t-1}] = 0 \quad a.s.$$

for all t.

Therefore, if Y_t is a Martingale, then $X_t = Y_t - Y_{t-1}$ is a Martingale Difference Sequence. This is useful because it imposes much milder conditions on the memory of sequence rather than independence.

14.8 Martingale Central Limit Theorem

Let $\{X_{ni}\}_{n\geq 1, i\geq 1}$ be an array of random variables. For each $n \geq 1$, let $\{\mathcal{F}_{ni}\}_{i\geq 1}$ be a sequence of $\sigma-$fields such that

$$\mathcal{F}_{ni} \subseteq \mathcal{F}_{n,i+1} \quad i \geq 1$$

and $\{X_{ni}\}_{i\geq 1}$ is adapted to $\{\mathcal{F}_{ni}\}_{i\geq 1}$. We are interested in the convergence of

$$S_n = \sum_{i=1}^{n} X_{ni}.$$

Theorem 14.99 *(Original martingale central limit theorem). Suppose that*

1.

$$\max_{1\leq i\leq n} |X_{ni}| \xrightarrow{p} 0 \tag{14.132}$$

2.

$$\sum_{i=1}^{n} \left| \mathbb{E}[X_{ni} 1\{|X_{ni}| \leq 1\}]|\mathcal{F}_{n,i-1} \right| \xrightarrow{p} 0 \tag{14.133}$$

3.

$$\sum_{i=1}^{n} X_{ni}^2 \xrightarrow{p} 1 \tag{14.134}$$

Then, we conclude that

$$S_n \xrightarrow{D} \mathcal{N}(0,1)$$

Define

$$T_n(t) = \prod_{k=1}^{n} (1 + itX_{n,k})$$

where $i^2 = 1$.

Theorem 14.100 *(Modified martingale central limit theorem). Suppose that*

1.

$$\max_{1 \leq i \leq n} |X_{ni}| \xrightarrow{p} 0 \tag{14.135}$$

2.

$$\mathbb{E}[T_n(t)] \to 1 \tag{14.136}$$

3. $T_n(t)$ *is uniformly integrable for any* t

4.

$$\sum_{i=1}^{n} X_{ni}^2 \xrightarrow{p} 1 \tag{14.137}$$

Lemma 14.101 *If* $X_n \xrightarrow{p} X$ *and* $|X_n|$ *is uniformly integrable, then*

$$\mathbb{E}[X_n] \to \mathbb{E}[X]$$

Assume that $\{X_{n,i}, \mathcal{F}_{ni}\}_{i \geq 1}$ forms a martingale difference and

$$S_n = \sum_{i=1}^{n} X_{ni}$$

Theorem 13: Sufficient conditions for Martingale Limit Theorem

Suppose one of the following conditions set holds. Then $S_n \xrightarrow{D} \mathcal{N}(0,1)$.

1. $\mathbb{E}[max_{1 \leq i \leq n} |X_{ni}|] \to 0$ and $\sum_{i=1}^{n} X_{ni}^2 \to 1$
2. The **conditional Lindeberg condition holds** $\sum_{i=1}^{n} \mathbb{E}[|X_{ni}|^2 1\{|X_{ni}| \geq \epsilon\} | \mathcal{F}_{n,i-1}] \xrightarrow{p} 0$ for all $\epsilon > 0$ amd
$$\sum_{i=1}^{n} X_{ni}^2 \xrightarrow{p} 1$$
or
$$\sum_{i=1}^{n} \mathbb{E}[X_{ni}^2 | \mathcal{F}_{n,i-1}] \xrightarrow{p} 1$$
3. $\max_{1 \leq i \leq n} |X_{ni}| \xrightarrow{p} 0, \sum_{i=1}^{n} X_{ni}^2 \xrightarrow{p} 1$ and
$$\sum_{i=1}^{n} \mathbb{E}[|X_{ni}|^2 1\{|X_{ni}| \geq 1\} | \mathcal{F}_{n,i-1}] \xrightarrow{p} 0$$
4. $\max_{1 \leq i \leq n} |X_{ni}| \xrightarrow{p} 0, \sum_{i=1}^{n} X_{ni}^2 \xrightarrow{p} 1$ and $\sum_{i=1}^{n} \mathbb{E}[X_{ni}^2] = 1$

Proposition 50: Additional conditions for Martingale Limit Theorems

1. $\max_{1\leq i\leq n}|X_{ni}|\xrightarrow{p} 0$ if and only if

$$\sum_{i=1}^{n}|X_{ni}|^m 1\{|X_{ni}|\geq \epsilon\} \xrightarrow{p} 0$$

for all $\epsilon > 0$ where $m \geq 1$ is an integer, and if and only if

$$\sum_{i=1}^{n} P(|X_{ni}|\geq \epsilon|\mathcal{F}_{n,i-1}) \xrightarrow{p} 0$$

for all $\epsilon > 0$

2. If $\mathbb{E}[\max_{1\leq i\leq n}|X_{ni}|^m|\rightarrow 0$ where $m \geq 1$ is an integer, then

$$\sum_{i=1}^{n}\mathbb{E}[|X_{ni}|^m 1\{|X_{ni}|\geq \epsilon|\mathcal{F}_{n,i-1}\}] \xrightarrow{p} 0 \tag{14.138}$$

for all $\epsilon > 0$.

3. If 14.138 holds with $m = 2$, then $\max_{1\leq i\leq n}|X_{ni}|\xrightarrow{p} 0$ and

$$\sum_{i=1}^{n}\mathbb{E}[X_{ni}^2|\mathcal{F}_{n,i-1}] \xrightarrow{p} 1$$

is equivalent to

$$\sum_{i=1}^{n} X_{ni}^2 \xrightarrow{p} 1$$

STAT5610: Advanced Inference

15. Martingale Central Limit Theorems: Applications

15.8.1 Applications of MLT

We show how the martingale central limit theorem can be applied to many examples in time series. The trick is to take a time series, represent it in terms of a martingale difference sequence, and then apply the martingale central limit theorem to it.

Generally, a sequence of stationary variables $(w_t)_{t\geq 1}$ may not be a martingale difference. However, under certain conditions, we can write it as

$$w_t = v_t + z_{t-1} - z_t \quad t = 1, 2, ...$$

where v_t is a martingale difference and z_t is another sequence of stationary variables. Then, we can see that

$$\sum_{k=1}^{n} w_k = \sum_{k=1}^{n} v_k + (z_0 - z_n)$$

and hence, we see that the martingale $\sum_{k=1}^{n} v_k$ provides an approximation to $\sum_{k=1}^{n} w_k$ with their limiting behaviour coinciding.

Definition 15.102 *(Strictly Stationary). A process $\{X_t : t \in T\}$ is called **strictly stationary** if the joint distribution is invariant under shifts of t*

$$(X_{t_1}, ..., X_{t_n}) \overset{d}{=} (X_{t_1+h}, ..., X_{t_n+h})$$

for all possible choices of times $t_1, ..., t_n \in T$ for $n \geq 1$ and h such that $t_1 + h, ..., t_n + h \in T$.

Definition 15.103 *(Weakly Stationary). A process $\{X_t : t \in T\}$ is called **weakly stationary** if its mean and covariance functions are invariant under shifts of index t*

$$\mathbb{E}[X_{t+h}] = \mathbb{E}[X_t]$$

and

$$Cov(X_{s+h}, X_{t+h}) = Cov(X_s, X_t)$$

for all possible choices of times $s, t \in T$ and h such that $s + h, t + h \in T$.

Definition 38: α-mixing process

A process $\{X_t : t \geq 1\}$ is called $\alpha-$mixing if

$$\alpha_k = \sup_{A \in \mathcal{F}_{-\infty}^{t}, B \in \mathcal{F}_{t+k}^{\infty}} |P(AB) - P(A)P(B)| \to 0$$

as $k \to \infty$ where $\mathcal{F}_{-\infty}^{t} = \sigma(X_t, X_{t-1}, ...)$ and $\mathcal{F}_{t+k}^{\infty} = \sigma(X_{t+k}, X_{t+k+1}, ...)$.

Definition 39: ϕ-mixing process

A process $\{X_t : t \geq 1\}$ is called $\phi-$mixing if

$$\alpha_k = \sup_{A \in \mathcal{F}^t_{-\infty}, B \in \mathcal{F}^\infty_{t+k}} |P(A|B) - P(A)| \to 0$$

as $k \to \infty$ where $\mathcal{F}^t_{-\infty} = \sigma(X_t, X_{t-1}, ...)$ and $\mathcal{F}^\infty_{t+k} = \sigma(X_{t+k}, X_{t+k+1}, ...)$.

We now give a powerful result whereby if a time series is strictly stationary and is $\alpha-$mixing, then we may approximate it by a martingale difference sequence. From this, we can then apply the martingale central limit theorem to conclude that the sum of the time series converges in distribution to the normal distribution.

Theorem 14: MLT for $\alpha-$mixing processes

Suppose that $\{X_t : t \geq 1\}$ is

1. Strictly stationary
2. $\alpha-$mixing
3. Mean zero $\mathbb{E}[X_1] = 0$
4. $\mathbb{E}[|X_1|^{2+\delta}] < \infty$ for some $\delta > 0$
5. $\sum_{k=1}^\infty \alpha_k^{\delta/(2+\delta)} < \infty$

Then, we may write that

$$X_t = v_t + z_{t-1} - z_t$$

for $t = 1, 2, ...$ where v_t is a stationary martingale difference with $\mathbb{E}[|v_t|^{2+\delta}] < \infty$, z_t is a stationary sequence with $\mathbb{E}[z_t^2] < \infty$ and

$$\frac{1}{n}\sum_{k=1}^n \mathbb{E}[v_t^2|\mathcal{F}_t] \xrightarrow{p} \sigma^2 := Var[X_1] + 2\sum_{k=1}^\infty Cov(X_1, X_{1+k})$$

where $\mathcal{F}_t = \sigma(x_t, x_{t-1}, ...)$ for $t \geq 2$ and $\mathcal{F}_t = \sigma(\phi, \Omega)$ for $t \leq 0$.
As a consequence, we have that

$$\frac{1}{\sqrt{n}}\sum X_t \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$

Lemma 15.104 *(Maximum inequality).*

$$\max_{1 \leq i \leq n} |X_{ni}| \leq (\sum |X_{ni}|^2)^{1/2}$$

15.8.2 Casual Processes

We now show that casual processes can also be approximated by a martingale difference sequence.

Definition 40: Casual Process

Suppose that $\{\eta_i\}_{i\in\mathbb{Z}}$ are i.i.d random variable with $\mathbb{E}[\eta_i] = 0$ and $\mathbb{E}[\eta_i^2] = 1$. Then, let F be a measurable function and define the random variable

$$w_k = F(..., \eta_{k-1}, \eta_k), \quad k \in \mathbb{Z}$$

such that w_k are well-defined stationary random variables with $\mathbb{E}[w_k] = 0$ and $\mathbb{E}[w_k^2] < \infty$. Then, $\{w_k\}_{k\in\mathbb{Z}}$ is known as a **stationary casual process**.

We can approximate stationary casual processes by martingale difference sequences.

Theorem 15.105 *Define $\mathcal{P}_k Z = \mathbb{E}[Z|\mathcal{F}_k] - \mathbb{E}[Z|\mathcal{F}_{k-1}]$. Suppose that $\sum_{i=1}^{\infty} i\mathbb{E}[|\mathcal{P}_k w_{i+k}|^2] < \infty$. Then, we can decompose the stationary casual process by*

$$w_k = v_k + z_{k-1} + z_k$$

whereby z_k, v_k are stationary processes such that $\mathbb{E}[|v_0|^2 + |z_0|^2] < \infty$. From this,

$$\frac{1}{\sqrt{n\sigma}} \sum_{k=1}^{n} w_k \xrightarrow{D} \mathcal{N}(0,1)$$

where $\sigma^2 = \mathbb{E}[v_0^2]$.