Asymptotic Distribution
of Variable Importance
Wenying Deng

Lemma 1. Functional Delta Method (univariate). *Suppose $\mathcal{P}_n$ is the empirical distribution of a random sample $X_1, \dots, X_n$ from a distribution P, and ϕ is a function that maps the distribution of interest into some space. Define the Gateaus derivative*

$$\phi'_P(\delta_x - P) = \frac{d}{dt}\,|_{t=0}\ \phi((1-t)P + t\delta_x) = IF_{\phi,P}(x),$$

which is also the Influence Function, and $\gamma^2 = \int IF_{\phi,P}(x)^2 dP$. If integration and differentiation can be exchanged, then

$$\int \phi'_P(\delta_x - P)dP = 0.$$

Further, if $\sqrt{n}R_n \xrightarrow{P} 0$, where

$$R_n = \phi(\mathcal{P}_n) - \phi(P) - \frac{1}{n}\sum_i \phi'_P(\delta_{x_i} - P),$$

then from the Central Limit Theory that

$$\sqrt{n}(\phi(\mathcal{P}_n) - \phi(P)) \xrightarrow{d} N(0, \gamma^2).$$

Lemma 2. Functional Delta Method (multivariate). *Suppose $\mathcal{P}_n$ is the empirical distribution of a random sample $X_1, \dots, X_n$ from a distribution P, and $\boldsymbol{\phi} : \mathcal{R}^d \to \mathcal{R}^k$. Define the Gateaus derivative*

$$\boldsymbol{\phi}'_P(\delta_x - P) = \frac{d}{dt}\,|_{t=0}\ \boldsymbol{\phi}((1-t)P + t\delta_x) = IF_{\boldsymbol{\phi},P}(x),$$

which is also the Influence Function, and $[\mathbf{V}_0]_{i,j} = \int \langle [IF_{\boldsymbol{\phi},P}(x)]_i, [IF_{\boldsymbol{\phi},P}(x)]_j \rangle dP$. If integration and differentiation can be exchanged, then

$$\int \boldsymbol{\phi}'_P(\delta_x - P)dP = 0.$$

Further, if $\sqrt{n}\mathbf{R}_n \xrightarrow{P} 0$, where

$$\mathbf{R}_n = \boldsymbol{\phi}(\mathcal{P}_n) - \boldsymbol{\phi}(P) - \frac{1}{n}\sum_i \boldsymbol{\phi}'_P(\delta_{x_i} - P),$$

then from the Central Limit Theory that

$$\sqrt{n}(\boldsymbol{\phi}(\mathcal{P}_n) - \boldsymbol{\phi}(P)) \xrightarrow{d} MVN(0, \mathbf{V}_0).$$

Theorem 1. Asymptotic Distribution of Variable Importance (univariate). *Suppose* $y_i = f_0(x_i) + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2),\ i = 1, \dots, n$. *Denote* $D_p : f \to \frac{\partial}{\partial x_p} f$ *the differentiation operator,* $H_p = D_p^\top D_p$ *the inner product of* D_p*, define*

$$\psi_p(f) = \|D_p(f)\|_n^2 = \frac{1}{n}\langle D_p f, D_p f\rangle = \frac{1}{n} f^\top H_p f, \tag{1}$$

if $H_p = o_p(\sqrt{n})$*, then*

$$\sqrt{n}(\psi_p(\hat{f}_n) - \psi_p(f_0)) \xrightarrow{d} N(0, 4\sigma^2 \|H_p f_0\|_n^2).$$

Proof. From the definition in (1), we have

$$\psi_p'(f) = \frac{\partial}{\partial f}\psi_p(f) = \frac{2}{n} H_p f.$$

Define a mean function $m : F \to E(F)$, where F is the distribution. Then in our case, $f_0 = E(F) = m(F)$. According to **Lemma 1**, we have

$$\psi_p(f_0) = \psi_p(E(F)) = \psi_p(m(F)) = \phi(F),$$

i.e., $\phi(\cdot) = \psi_p(m(\cdot))$. Therefore,

$$\begin{aligned}
\phi_F'(\delta_y - F) &= \psi_p'(m(\delta_y - F)) \\
&= \frac{d}{dt}\ |_{t=0}\ \psi_p(m((1-t)F + t\delta_y)) \\
&= \frac{d}{dt}\ |_{t=0}\ \psi_p((1-t)f_0 + ty) \\
&= \frac{d}{dt}\ |_{t=0}\ \frac{1}{n}[(1-t)f_0 + ty]^\top H_p [(1-t)f_0 + ty] \\
&= \frac{2}{n}(y - f_0)^\top H_p f_0 \\
&= IF_{\phi,F}(y).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\gamma^2 &= \int IF_{\phi,F}(y)^2 dF \\
&= 4\int \frac{1}{n} \cdot f_0^\top H_p (y - f_0)(y - f_0)^\top H_p f_0 \cdot \frac{1}{n} dF \\
&= 4\sigma^2 \|H_p f_0\|_n^2.
\end{aligned}$$

Moreover, we have

$$\int \phi_F'(\delta_y - F) dF = \frac{2}{n}\int (y - f_0)^\top H_p f_0 dF = 0,$$

and

$$
\begin{aligned}
\sqrt{n}R_n &= \sqrt{n}[\phi(\mathcal{F}_n) - \phi(F) - \frac{1}{n}\sum_i \phi'_F(\delta_{y_i} - F)] \\
&= \sqrt{n}[\psi_p(\hat{f}_n) - \psi_p(f_0) - \frac{1}{n}\cdot\frac{2}{n}\sum_i (y_i - f_{0,i})^\top [H_p f_0]_i] \\
&= \sqrt{n}[\frac{1}{n}\cdot(\hat{f}_n^\top H_p \hat{f}_n - f_0^\top H_p f_0) - \frac{1}{n}\cdot\frac{2}{n}(y - f_0)^\top H_p f_0] \\
&= \frac{1}{\sqrt{n}}[\hat{f}_n^\top H_p \hat{f}_n - \hat{f}_n^\top H_p f_0 + \hat{f}_n^\top H_p f_0 - f_0^\top H_p f_0 - \frac{2}{n}(y - f_0)^\top H_p f_0] \\
&= \frac{1}{\sqrt{n}}[(\hat{f}_n - f_0)^\top H_p(\hat{f}_n + f_0) - \frac{2}{n}(y - f_0)^\top H_p f_0] \\
&= \frac{1}{\sqrt{n}}[(\hat{f}_n - f_0)^\top H_p(\hat{f}_n - f_0) + 2(\hat{f}_n - f_0)^\top H_p f_0 - \frac{2}{n}(y - f_0)^\top H_p f_0] \\
&= \frac{1}{\sqrt{n}}[\hat{\epsilon}_n^\top H_p \hat{\epsilon}_n + 2(\hat{f}_n - f_0)^\top H_p f_0 - \frac{2}{n}(y - f_0)^\top H_p f_0] \qquad (2) \\
&= \frac{1}{\sqrt{n}} o_p(\sqrt{n}) \qquad (3) \\
&= o_p(1) \xrightarrow{P} 0,
\end{aligned}
$$

where $\hat{\epsilon}_n = \hat{f}_n - f_0$. The derivation from (2) to (3) is because we can perform eigendecomposition of H_p, where $H_p = U_p \Lambda U_p^\top$, with $U_p = [\mathbf{u}_1, \dots, \mathbf{u}_n]$ a $n \times n$ orthonormal matrix and Λ a diagonal matrix with elements $\{\lambda_i\}_{i=1}^n$ being the eigenvalues of H_p, then define

$$
\mathbf{v} = U_p^\top \hat{\epsilon}_n = \begin{bmatrix} \mathbf{u}_1^\top \hat{\epsilon}_n \\ \vdots \\ \mathbf{u}_n^\top \hat{\epsilon}_n \end{bmatrix}.
$$

Therefore,

$$
\hat{\epsilon}_n^\top H_p \hat{\epsilon}_n = \mathbf{v}^\top \Lambda \mathbf{v} = \sum_{i=1}^n \lambda_i \mathbf{v}_i^\top \mathbf{v}_i.
$$

Further, we have

$$
E(\mathbf{u}_i^\top \hat{\boldsymbol{\epsilon}}) = 0, \quad Var(\mathbf{u}_i^\top \hat{\boldsymbol{\epsilon}}) = \hat{\sigma}^2 \mathbf{u}_i^\top \mathbf{I}_n \mathbf{u}_i = \hat{\sigma}^2,
$$

where $\hat{\sigma}^2$ is the estimated variance of ϵ and

$$
E(\mathbf{v}_i^\top \mathbf{v}_i) = E(\hat{\boldsymbol{\epsilon}}^\top \mathbf{u}_i \mathbf{u}_i^\top \hat{\boldsymbol{\epsilon}}) = Var(\mathbf{u}_i^\top \hat{\boldsymbol{\epsilon}}) + [E(\mathbf{u}_i^\top \hat{\boldsymbol{\epsilon}})]^2 = \hat{\sigma}^2.
$$

Consequently,

$$
E(\hat{\epsilon}_n^\top H_p \hat{\epsilon}_n) = E(\sum_{i=1}^n \lambda_i \mathbf{v}_i^\top \mathbf{v}_i) = \sum_{i=1}^n \lambda_i E(\mathbf{v}_i^\top \mathbf{v}_i) = \hat{\sigma}^2 \sum_{i=1}^n \lambda_i = trace(H_p) \cdot \hat{\sigma}^2 = o_p(\sqrt{n}) \cdot \hat{\sigma}^2,
$$

so $\hat{\epsilon}_n^\top H_p \hat{\epsilon}_n = o_p(\sqrt{n})$. On the other hand, $2(\hat{f}_n - f_0)^\top H_p f_0 = o_p(\sqrt{n})$ because $\hat{f}_n$ is a consistent estimator of f_0. Moreover, since $y_i - f_{0,i} = o_p(1)$, we know $\frac{2}{n}(y - f_0)^\top H_p f_0 = o_p(\sqrt{n})$. So,

$$
\hat{\epsilon}_n^\top H_p \hat{\epsilon}_n - \frac{2}{n}(y - f_0)^\top H_p f_0 = o_p(\sqrt{n}) \qquad (4)
$$

Therefore, by **Lemma 1**, we have

$$\sqrt{n}(\psi_p(\hat{f}_n) - \psi_p(f_0)) \xrightarrow{d} N(0, 4\sigma^2 \|H_p f_0\|_n^2).$$

$\square$

Theorem 2. Asymptotic Distribution of Variable Importance (multivariate). *Suppose* $y_i = f_0(\mathbf{x}_i) + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2),\ \mathbf{x}_i \in \mathcal{R}^P,\ i = 1, \dots, n$. *Denote* $\boldsymbol{\psi} = [\psi_1, \dots, \psi_P]$ *for* ψ_p *as defined in* (1), *and if* $H_p = o_p(\sqrt{n})$, *then* $\boldsymbol{\psi}(\hat{f}_n)$ *asymptotically converges toward a multivariate normal distribution surrounding* $\boldsymbol{\psi}(f_0)$, *i.e.,*

$$\sqrt{n}(\boldsymbol{\psi}(\hat{f}_n) - \boldsymbol{\psi}(f_0)) \xrightarrow{d} MVN(\mathbf{0}, \mathbf{V}_0),$$

where $\mathbf{V}_0$ *is a* $P \times P$ *matrix such that* $[\mathbf{V}_0]_{p1,p2} = 4\sigma^2 \langle H_{p1} f_0, H_{p2} f_0 \rangle_n$.

Proof. Define a mean function $m : F \to E(F)$, where F is the distribution. Then in our case, $f_0 = E(F) = m(F)$. According to **Lemma 2**, we have

$$[\boldsymbol{\psi}(f_0)]_p = \psi_p(E(F)) = \psi_p(m(F)) = [\boldsymbol{\phi}(F)]_p,$$

i.e., $\boldsymbol{\phi}(\cdot) = \boldsymbol{\psi}(m(\cdot))$ and $[\boldsymbol{\phi}(\cdot)]_p = \psi_p(m(\cdot))$, where $\boldsymbol{\phi} : \mathcal{R} \to \mathcal{R}^P$. Therefore,

$$\begin{aligned}
[\boldsymbol{\phi}'_F(\delta_y - F)]_p &= \psi'_p(m(\delta_y - F)) \\
&= \frac{d}{dt}\,|_{t=0}\ \psi_p(m((1-t)F + t\delta_y)) \\
&= \frac{d}{dt}\,|_{t=0}\ \psi_p((1-t)f_0 + ty) \\
&= \frac{d}{dt}\,|_{t=0}\ \frac{1}{n}[(1-t)f_0 + ty]^\top H_p [(1-t)f_0 + ty] \\
&= \frac{2}{n}(y - f_0)^\top H_p f_0 \\
&= [IF_{\boldsymbol{\phi},F}(y)]_p.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
[\mathbf{V}_0]_{p1,p2} &= \int \langle [IF_{\boldsymbol{\phi},F}(y)]_{p1}, [IF_{\boldsymbol{\phi},F}(y)]_{p2} \rangle dF \\
&= 4 \int \frac{1}{n} \cdot f_0^\top H_{p1}(y - f_0)(y - f_0)^\top H_{p2} f_0 \cdot \frac{1}{n} dF \\
&= 4\sigma^2 \langle H_{p1} f_0, H_{p2} f_0 \rangle_n.
\end{aligned}$$

Moreover, we have

$$[\int \boldsymbol{\phi}'_F(\delta_y - F) dF]_p = \frac{2}{n} \int (y - f_0)^\top H_p f_0 dF = 0,$$

and

$$
\begin{aligned}
[\sqrt{n}\mathbf{R}_n]_p &= \sqrt{n}[[\boldsymbol{\phi}(\mathcal{F}_n)]_p - [\boldsymbol{\phi}(F)]_p - \frac{1}{n}\sum_i [\boldsymbol{\phi}'_F(\delta_{y_i} - F)]_p] \\
&= \frac{1}{\sqrt{n}}[\hat{f}_n^\top H_p \hat{f}_n - f_0^\top H_p f_0 - \frac{2}{n}(y - f_0)^\top H_p f_0] \\
&= \frac{1}{\sqrt{n}}[(\hat{f}_n - f_0)^\top H_p (\hat{f}_n - f_0) + 2(\hat{f}_n - f_0)^\top H_p f_0 - \frac{2}{n}(y - f_0)^\top H_p f_0] \\
&= \frac{1}{\sqrt{n}}[\hat{\epsilon}_n^\top H_p \hat{\epsilon}_n + 2(\hat{f}_n - f_0)^\top H_p f_0 - \frac{2}{n}(y - f_0)^\top H_p f_0] && (5) \\
&= \frac{1}{\sqrt{n}} o_p(\sqrt{n}) && (6) \\
&= o_p(1) \xrightarrow{P} 0,
\end{aligned}
$$

where $\hat{\epsilon}_n = \hat{f}_n - f_0$ and the reason from (5) to (6) is because of (4). Therefore, by **Lemma 2**, we have

$$\sqrt{n}(\boldsymbol{\psi}(\hat{f}_n) - \boldsymbol{\psi}(f_0)) \xrightarrow{d} MVN(\mathbf{0}, \mathbf{V}_0),$$

where $\mathbf{V}_0$ is a $P \times P$ matrix such that $[\mathbf{V}_0]_{p1,p2} = 4\sigma^2 \langle H_{p1} f_0, H_{p2} f_0 \rangle_n$. □