Lecture Notes 13
The Bootstrap

1 Introduction

The bootstrap is a method for estimating the variance of an estimator and for finding approximate confidence intervals for parameters. Although the method is nonparametric, it can also be used for inference about parameters in parametric and nonparametric models.

2 Empirical Distribution

Let $X_1, \ldots, X_n \sim P$. Recall that the empirical distribution P_n is defined by

$$P_n(A) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \in A).$$

In other words, P_n puts mass $1/n$ at each X_i. Recall also that a parameter of the form $\theta = T(P)$ is called a statistical functionl and that the plug-in estimator is $\widehat{\theta}_n = T(P_n)$.

An iid sample of size n drawn from P_n is called a *bootstrap* sample, denoted by

$$X_1^*, \ldots, X_n^* \sim P_n.$$

Bootstrap samples play an important role in what follows. Note that drawing an iid sample $X_1^*, \ldots, X_n^*$ from P_n is equivalent to drawing n observations, with replacement, from the original data $\{X_1, \ldots, X_n\}$. Thus, bootstrap sampling is often described as "resampling the data."

3 The Bootstrap

Now we give the bootstrap algorithms for estimating the variance of $\widehat{\theta}_n$ and for constructing confidence intervals. The explanation of why (and when) the bootstrap works is mainly deferred until Section 5. Let $\widehat{\theta}_n = g(X_1, \ldots, X_n)$ denote some estimator.

We would like to find the variance of $\widehat{\theta}_n$. Let

$$\mathrm{Var}_P(\widehat{\theta}_n) = \mathrm{Var}_P(g(X_1, \ldots, X_n)) \equiv S_n(P).$$

Note that $\text{Var}_P(\widehat{\theta}_n)$ is some function pf P (and n) so I have written $\text{Var}_P(\widehat{\theta}_n) = S_n(P)$. If we knew P, we could approximate $S_n(P)$ by simulation as follows:

$$\begin{aligned}
\text{draw} \quad & X_1, \ldots, X_n \sim P \\
\text{compute} \quad & \widehat{\theta}_n^{(1)} = g(X_1, \ldots, X_n) \\
\text{draw} \quad & X_1, \ldots, X_n \sim P \\
\text{compute} \quad & \widehat{\theta}_n^{(2)} = g(X_1, \ldots, X_n) \\
& \vdots \\
\text{draw} \quad & X_1, \ldots, X_n \sim P \\
\text{compute} \quad & \widehat{\theta}_n^{(B)} = g(X_1, \ldots, X_n).
\end{aligned}$$

Let s^2 be the sample variance of $\widehat{\theta}_n^{(1)}, \ldots, \widehat{\theta}_n^{(B)}$. So

$$s^2 = \frac{1}{B}\sum_{j=1}^{B}(\widehat{\theta}_n^{(j)})^2 - \left(\frac{1}{B}\sum_{j=1}^{B}\widehat{\theta}_n^{(j)}\right)^2.$$

By the law of large numbers,

$$s^2 \xrightarrow{P} \mathbb{E}[\widehat{\theta}_n^2] - (\mathbb{E}[\widehat{\theta}_n])^2 = \text{Var}_P(\widehat{\theta}_n) = S_n(P).$$

Since we can take B as large as we want, we have that $s^2 \approx \text{Var}_P(\widehat{\theta}_n)$. In other words, **we can approximate $S_n(P)$ by repeatedly simulating n observations from P.**

But we don't know P. So we estimate $S_n(P)$ with $S_n(P_n)$ where P_n is the empirical distribution. Since P_n is a consistent estimator, we expect that $S_n(P_n) \approx S_n(P)$. In other words:

$$\text{Bootstrap approximation of the variance}: \quad \text{estimate } S_n(P) \quad \text{with} \quad S_n(P_n)$$

or in other words

$$\widehat{\text{Var}_P(\widehat{\theta}_n)} = \text{Var}_{P_n}(\widehat{\theta}_n).$$

But how do we compute $S_n(P_n)$? We use the simulation method above, except that we simulate from P_n instead of P. This leads to the following algorithm:

Bootstrap Variance Estimator

1. Draw a bootstrap sample $X_1^*, \ldots, X_n^* \sim P_n$. Compute $\widehat{\theta}_n^* = g(X_1^*, \ldots, X_n^*)$.
2. Repeat the previous step, B times, yielding estimators $\widehat{\theta}_{n,1}^*, \ldots, \widehat{\theta}_{n,B}^*$.
3. Compute:
$$\widehat{s} = \sqrt{\frac{1}{B}\sum_{j=1}^{B}(\widehat{\theta}_{n,j}^* - \overline{\theta})^2}$$
where $\overline{\theta} = \frac{1}{B}\sum_{j=1}^{B}\widehat{\theta}_{n,j}^*$.
4. Output $\widehat{s}$.

You can think about it like this:

$$\frac{1}{B}\sum_{j=1}^{B}(\widehat{\theta}_{n,j}^* - \overline{\theta})^2 \underbrace{\approx}_{\text{simulation error}} S_n(P_n) \underbrace{\approx}_{\text{estimation error}} S_n(P)$$

The are two sources of error in this apprixmation. The first is due to the fact that n is finite and the second is due to the fact that B is finite. However, we can make B as large as we like. (In practice, it usually suffices to take $B = 10,000$.) So we ignore the error due to finite B.

Theorem 1 *Under appropriate regularity conditions,* $\frac{s^2}{\text{Var}(\widehat{\theta}_n)} \xrightarrow{P} 1$ *as* $n \to \infty$.

Now we describe the confidence interval algorithm. This will look less intuitive than the variance estimator; I'll eplain it in Section 5.

Bootstrap Confidence Interval

1. Draw a bootstrap sample $X_1^*, \ldots, X_n^* \sim P_n$. Compute $\widehat{\theta}_n^* = g(X_1^*, \ldots, X_n^*)$.
2. Repeat the previous step, B times, yielding estimators $\widehat{\theta}_{n,1}^*, \ldots, \widehat{\theta}_{n,B}^*$.
3. Let
$$\widehat{F}(t) = \frac{1}{B}\sum_{j=1}^{B} I\Big(\sqrt{n}(\widehat{\theta}_{n,j}^* - \widehat{\theta}_n\Big) \leq t).$$
4. Let
$$C_n = \left[\widehat{\theta}_n - \frac{t_{1-\alpha/2}}{\sqrt{n}},\ \widehat{\theta}_n - \frac{t_{\alpha/2}}{\sqrt{n}}\right]$$
where $t_{\alpha/2} = \widehat{F}^{-1}(\alpha/2)$ and $t_{1-\alpha/2} = \widehat{F}^{-1}(1-\alpha/2)$.
5. Output C_n.

Theorem 2 *Under appropriate regularity conditions,*

$$\mathbb{P}(\theta \in C_n) = 1 - \alpha - O\left(\frac{1}{\sqrt{n}}\right).$$

as $n \to \infty$.

See the appendix for a discussion of the regularity conditions.

4 Examples

Example 3 *Consider the polynomial regression model* $Y = g(X) + \epsilon$ *where* $X, Y \in \mathbb{R}$ *and* $g(x) = \beta_0 + \beta_1 x + \beta_2 x^2$. *Given data* $(X_1, Y_1), \ldots, (X_n, Y_n)$ *we can estimate* $\beta = (\beta_0, \beta_1, \beta_2)$ *with the least squares estimator* $\widehat{\beta}$. *Suppose that* $g(x)$ *is concave and we are interested in the location at which* $g(x)$ *is maximized. It is easy to see that the maximum occurs at* $x = \theta$ *where* $\theta = -(1/2)\beta_1/\beta_2$. *A point estimate of* θ *is* $\widehat{\theta} = -(1/2)\widehat{\beta}_1/\widehat{\beta}_2$. *Now we use the bootstrap to get a confidence interval for* θ. *Figure 1 shows 50 points drawn from the above model with* $\beta_0 = -1$, $\beta_1 = 2$, $\beta_2 = -1$. *The* X_i*'s were sample uniformly on* $[0, 2]$ *and we took* $\epsilon_i \sim N(0, .2^2)$. *In this case,* $\theta = 1$. *The true and estimated curves are shown in the figure. At the bottom of the plot we show the 95 percent boostrap confidence interval based on* $B = 1,000$.

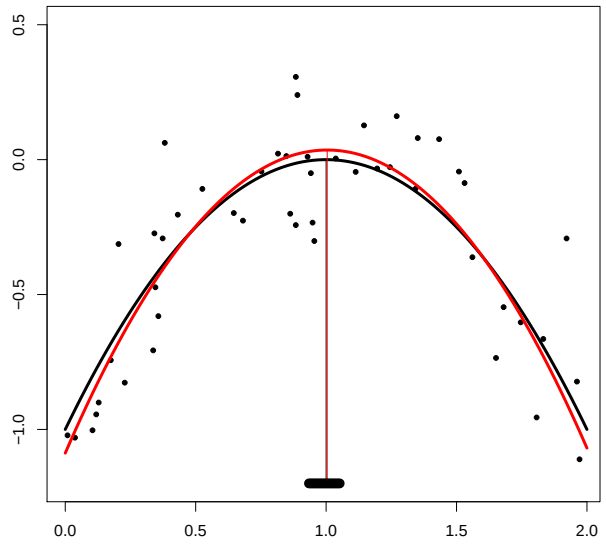

Figure 1: *50 points drawn from the model* $Y_i = -1 + 2X_i - X_i^2 + \epsilon_i$ *where* $X_i \sim \text{Uniform}(0, 2)$ *and* $\epsilon_i \sim N(0, .2^2)$. *In this case, the maximum of the polynomail occurs at* $\theta = 1$. *The true and estimated curves are shown in the figure. At the bottom of the plot we show the 95 percent boostrap confidence interval based on* $B = 1,000$.

Example 4 *Let* $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n) \sim P$ *where* $X_i \in \mathbb{R}, Y_i \in \mathbb{R}, Z_i \in \mathbb{R}^d$. *The partial correlation of* X *and* Y *given* Z *is*

$$\theta = -\frac{\Omega_{12}}{\sqrt{\Omega_{11}\Omega_{22}}}$$

where $\Omega = \Sigma^{-1}$ *and* Σ *is the covariance matrix of* $W = (X, Y, Z)^T$. *The partial correlation measures the linear dependence between* X *and* Y *after removing the effect of* Z. *For illustration, suppose we generate the data as follows: we take* $Z \sim N(0, 1)$, $X = 10Z + \epsilon$ *and* $Y = 10Z + \delta$ *where* $\epsilon, \delta \sim N(0, 1)$. *The correlation between* X *and* Y *is very large. But the partial correlation is 0. We generated* $n = 100$ *data points from this model. The sample correlation was 0.99. However, the estimate partial correaltion was -0.16 which is much closer to 0. The 95 percent bootstrap confidence interval is [-.33,.02] which includes the true value, namely, 0.*

5 Why Does the Bootstrap Work?

To explain why the bootstrap works, let us begin with a heuristic. Let

$$F_n(t) = \mathbb{P}(\sqrt{n}(\widehat{\theta} - \theta) \leq t).$$

If we knew F_n we could easily construct a confidence interval. Let

$$C_n = \left[\widehat{\theta} - \frac{t_{1-\alpha/2}}{\sqrt{n}},\ \ \widehat{\theta} - \frac{t_{\alpha/2}}{\sqrt{n}}\right]$$

where $t_\alpha = F_n^{-1}(\alpha)$. Then

$$\begin{aligned}\mathbb{P}(\theta \in C_n) &= \mathbb{P}\left(\widehat{\theta} - \frac{t_{1-\alpha/2}}{\sqrt{n}} \leq \theta \leq \widehat{\theta} - \frac{t_{\alpha/2}}{\sqrt{n}}\right) \\ &= \mathbb{P}(t_{\alpha/2} \leq \sqrt{n}(\widehat{\theta} - \theta) \leq t_{1-\alpha/2}) = F_n(t_{1-\alpha/2}) - F_n(t_{\alpha/2}) \\ &= F_n(F_n^{-1}(1-\alpha/2)) - F_n(F_n^{-1}(\alpha/2)) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1-\alpha.\end{aligned}$$

The problem is that we do not know F_n. The bootstrap estimates F_n with

$$\widehat{F}_n(t) = \mathbb{P}\Big(\sqrt{n}(\widehat{\theta}^* - \widehat{\theta}_n) \leq t \ \Big|\ X_1, \ldots, X_n\Big).$$

If $\widehat{F}_n \approx F_n$ then the bootstrap will work.

Usually, F_n will be close to some limiting distribution L. Similarly, $\widehat{F}_n$ will be close to some limiting distribution $\widehat{L}$. Moreover, L and $\widehat{L}$ will be close which implies that F_n and $\widehat{F}_n$ are close. In practice, we usually approximate $\widehat{F}_n$ by its Monte Carlo version

$$\overline{F}(t) = \frac{1}{B}\sum_{j=1}^{B} I(\sqrt{n}(\widehat{\theta}_j^* - \widehat{\theta}_j) \leq t).$$

But $\overline{F}$ is close to $\widehat{F}_n$ as long as we take B large. See Figure 2.

Now we will give more detail in a simple, special case. Suppose that $X_1, \ldots, X_n \sim P$ where X_i has mean μ and variance σ^2. Suppose we want to construct a confidence interval for μ.

Let $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and define

$$F_n(t) = \mathbb{P}(\sqrt{n}(\widehat{\mu}_n - \mu) \leq t). \tag{1}$$

We want to show that

$$\widehat{F}_n(t) = \mathbb{P}\Big(\sqrt{n}(\widehat{\mu}_n^* - \widehat{\mu}_n) \leq t \ \Big|\ X_1, \ldots, X_n\Big)$$

is close to F_n.

Theorem 5 (Bootstrap Theorem) *Suppose that $\mu_3 = \mathbb{E}|X_i|^3 < \infty$. Then,*

$$\sup_t |\widehat{F}_n(t) - F_n(t)| = O_P\left(\frac{1}{\sqrt{n}}\right).$$

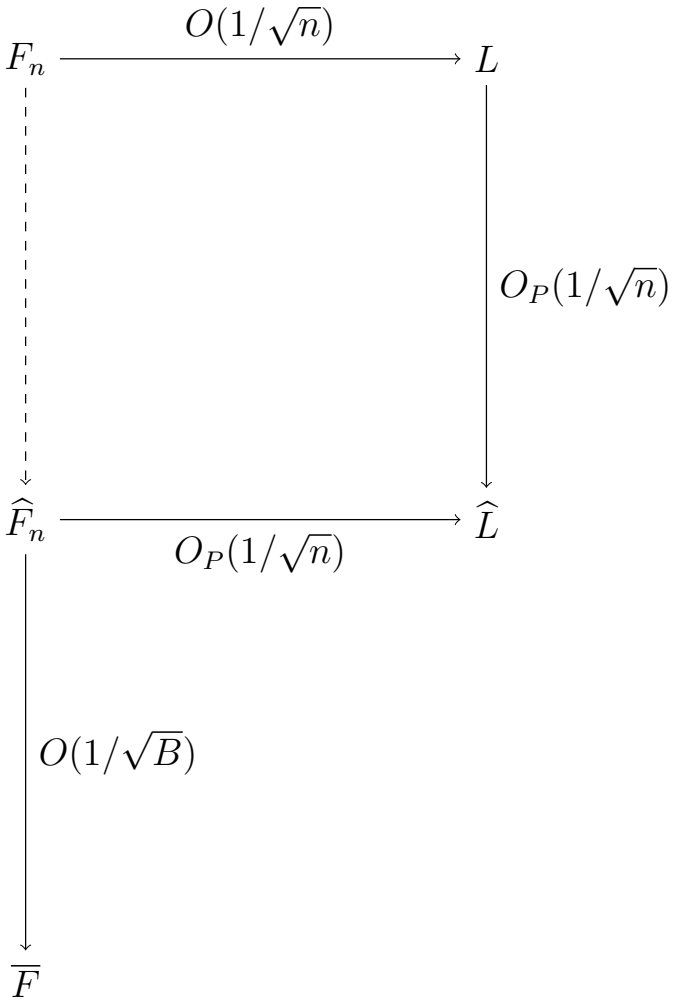

Figure 2: *The distribution* $F_n(t) = \mathbb{P}(\sqrt{n}(\widehat{\theta}_n - \theta) \leq t)$ *is close to some limit distribution* L*. Similarly, the bootstrap distribution* $\widehat{F}_n(t) = \mathbb{P}(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n) \leq t | X_1, \ldots, X_n)$ *is close to some limit distribution* $\widehat{L}$*. Since* $\widehat{L}$ *and* L *are close, it follows that* F_n *and* $\widehat{F}_n$ *are close. In practice, we approximate* $\widehat{F}_n$ *with its Monte Carlo version* $\overline{F}$ *which we can make as close to* $\widehat{F}_n$ *as we like by taking* B *large.*

To prove this result, let us recall that Berry-Esseen Theorem.

Theorem 6 (Berry-Esseen Theorem) *Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2. Let $\mu_3 = \mathbb{E}[|X_i - \mu|^3] < \infty$. Let $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$ be the sample mean and let Φ be the cdf of a $N(0,1)$ random variable. Let $Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$. Then*

$$\sup_z \Big| \mathbb{P}(Z_n \leq z) - \Phi(z) \Big| \leq \frac{33}{4}\frac{\mu_3}{\sigma^3\sqrt{n}}. \tag{2}$$

Proof of the Bootstrap Theorem. Let $\Phi_\sigma(t)$ denote the cdf of a Normal with mean 0 and variance σ^2. Let $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \widehat{\mu}_n)^2$. Thus, $\widehat{\sigma}^2 = \mathrm{Var}(\sqrt{n}(\widehat{\mu}_n^* - \widehat{\mu}_n)|X_1, \ldots, X_n)$. Now, by the triangle inequality,

$$\begin{aligned}\sup_t |\widehat{F}_n(t) - F_n(t)| &\leq \sup_t |F_n(t) - \Phi_\sigma(t)| + \sup_t |\Phi_\sigma(t) - \Phi_{\widehat{\sigma}}(t)| + \sup_t |\widehat{F}_n(t) - \Phi_{\widehat{\sigma}}(t)| \\ &= \mathrm{I} + \mathrm{II} + \mathrm{III}.\end{aligned}$$

Let $Z \sim N(0,1)$. Then, $\sigma Z \sim N(0, \sigma^2)$ and from the Berry-Esseen theorem,

$$\begin{aligned}\mathrm{I} &= \sup_t |F_n(t) - \Phi_\sigma(t)| = \sup_t \left| \mathbb{P}\left(\sqrt{n}(\widehat{\mu}_n - \mu) \leq t\right) - \mathbb{P}\left(\sigma Z \leq t\right) \right| \\ &= \sup_t \left| \mathbb{P}\left(\frac{\sqrt{n}(\widehat{\mu}_n - \mu)}{\sigma} \leq \frac{t}{\sigma}\right) - \mathbb{P}\left(Z \leq \frac{t}{\sigma}\right) \right| \leq \frac{33}{4}\frac{\mu_3}{\sigma^3\sqrt{n}}.\end{aligned}$$

Using the same argument on the third term, we have that

$$\mathrm{III} = \sup_t |\widehat{F}_n(t) - \Phi_{\widehat{\sigma}}(t)| \leq \frac{33}{4}\frac{\widehat{\mu}_3}{\widehat{\sigma}^3\sqrt{n}}$$

where $\widehat{\mu}_3 = \frac{1}{n}\sum_{i=1} |X_i - \widehat{\mu}_n|^3$ is the empirical third moment. By the strong law of large numbers, $\widehat{\mu}_3$ converges almost surely to μ_3 and $\widehat{\sigma}$ converges almost surely to σ. So, almost surely, for all large n, $\widehat{\mu}_3 \leq 2\mu_3$ and $\widehat{\sigma} \geq (1/2)\sigma$ and $\mathrm{III} \leq \frac{33}{4}\frac{4\mu_3}{\sqrt{n}}$. From the fact that $\widehat{\sigma} - \sigma = O_P(\sqrt{1/n})$ it may be shown that $\mathrm{II} = \sup_t |\Phi_\sigma(t) - \Phi_{\widehat{\sigma}}(t)| = O_P(\sqrt{1/n})$. (This may be seen by Taylor expanding $\Phi_{\widehat{\sigma}}(t)$ around σ.) This completes the proof. □

We have shown that $\sup_t |\widehat{F}_n(t) - F_n(t)| = O_P\left(\frac{1}{\sqrt{n}}\right)$. From this, it may be shown that, for each $0 < \beta < 1$, $t_\beta - z_\beta = O_P\left(\frac{1}{\sqrt{n}}\right)$. From this, one can prove Theorem 2.

So far we have focused on the mean. Similar theorems may be proved for more general parameters. The details are complex so we will not discuss them here. More information is in the appendix. See also Chapter 23 of van der Vaart (1998).

6 The Parametric Bootstrap

The bootstrap can also be used for parametric inference. Suppose that $X_1, \ldots, X_n \sim p(x; \theta)$. Let $\widehat{\theta}$ be the mle. Let $\psi = g(\theta)$ and $\widehat{\psi} = g(\widehat{\theta})$. To estimate the standard error of $\widehat{\psi}$ we could find the Fishher information followed by the delta method.

Alternatively, we simply compute the standard deviation of the bootstrap replications $\widehat{\psi}_1^*, \ldots, \widehat{\psi}_B$. The only difference is that now we draw from the bootstrap sampples from $p(x; \widehat{\theta})$. In other words:

$$X_1^*, \ldots, X_n^* \sim p(x; \widehat{\theta}).$$

7 A Few Remarks About the Bootstrap

Here are some random remarks about the bootstrap:

1. The bootstrap is nonparametric but it does require some assumptions. You can't assume it is always valid. (See the appendix.)

2. The bootstrap is an asymptotic method. Thus the coverage of the confidence interval is $1 - \alpha + r_n$ where, typically, $r_n = C/\sqrt{n}$.

3. There is a related method called the jackknife where the standard error is estimated by leaving out one observation at a time. However, the bootstrap is valid under weaker conditions than the jackknife. See Shao and Tu (1995).

4. Another way to construct a bootstrap confidence interval is to set $C = [a, b]$ where a is the $\alpha/2$ quantile of $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_B^*$ and b is the $1 - \alpha/2$ quantile. This is called the percentile interval. This interval seems very intuitive but does not have the theoretical support for the interval B_n. However, in practice, the percentile interval and B_n are often quite similar.

5. There are many cases where the bootstrap is not formally justified. This is especially true with discrete structures like trees and graphs. Nonethless, the bootstrap can be used in an informal way to get some intuition of the variability of the procedure. But keep in mind that the formal guarantees may not apply in these cases. For example, see Holmes (2003) for a discussion of the bootstrap applied to phylogenetic tres.

6. There is a method related to the bootstrap called subsampling. In this case, we draw samples of size $m < n$ without replacement. Subsampling produces valid confidence intervals under weaker conditions than the bootstrap. See Politis, Romano and Wolf (1999).

7. There are many modifications of the bootstrap that lead to more accurate confidence intervals; see Efron (1996).

8. There is a version of the bootstrap that works in high dimensions. We discuss this in 10/36-702.

8 Summary

The bootstrap provides nonparametric standard errors and confidence intervals. To draw a bootstrap sample we draw n observations $X_1^*, \ldots, X_n^*$ from the empirical distribution P_n. This is equivalent to drawing n observations with replacement from the original daa $X_1, \ldots, X_n$. We then compute the estimator $\widehat{\theta}^* = g(X_1^*, \ldots, X_n^*)$. If we repeat this whole process B times we get $\widehat{\theta}_1^*, \ldots, \theta_B^*$. The standard deviation of these values approximates the stanard error of $\widehat{\theta}_n = g(X_1, \ldots, X_n)$.

9 References

Efron, Bradley and Tibshirani, Robert. (1994). *An introduction to the bootstrap.* CRC press.

van der Vaart, A. (1998). *Asymptotic Statistics.* Chapter 23.

Appendix

Hadamard Differentiability. The key condition needed for the bootstrap is *Hadamard differentiability*. Let D and E be normed spaces and let $T : D \to E$. We say that T is Hadamard differentiable at $P \in D$ if there exists a continuous linear map $T_p' : D \to E$ such that

$$\left|\left| \frac{T(P + tQ_t) - T(P)}{t} - T_P'(Q) \right|\right|_R \to 0$$

whenever $t \downarrow 0$ and $Q_t \to Q$.