Estimation of the Conditional Variance in Paired Experiments

Alberto ABADIE & Guido W. IMBENS
Harvard University and NBER*

ABSTRACT. – In paired randomized experiments units are grouped in pairs, often based on covariate information, with random assignment within the pairs. Average treatment effects are then estimated by averaging the within-pair differences in outcomes. Typically the variance of the average treatment effect estimator is estimated using the sample variance of the within-pair differences. However, conditional on the covariates the variance of the average treatment effect estimator may be substantially smaller. Here we propose a simple way of estimating the conditional variance of the average treatment effect estimator by forming pairs-of-pairs with similar covariate values and estimating the variances within these pairs-of-pairs. Even though these within-pairs-of-pairs variance estimators are not consistent, their average is consistent for the conditional variance of the average treatment effect estimator and leads to asymptotically valid confidence intervals.

Estimation de la variance conditionnelle dans des expériences par paires

RÉSUMÉ. – Dans les expériences aléatoires d'appariement les unités sont regroupées par paires, souvent basées sur des caractéristiques explicatives, et avec appariement aléatoire. Les effets de traitement moyens sont alors estimés en faisant la moyenne des différences intra-paires dans les résultats. Typiquement, la variance de l'estimateur de l'effet de traitement moyen est estimée en utilisant la variance des différences intra-paires dans l'échantillon. Cependant, conditionnellement aux variables explicatives, l'estimateur de l'effet de traitement moyen peut être substantiellement plus petit. Nous proposons ici une manière simple d'estimer la variance conditionnelle de l'estimateur de l'effet de traitement moyen en formant des paires de paires avec des valeurs de variables explicatives similaires et en estimant les variances entre ces paires de paires. Même si ces estimateurs fondés sur les paires de paires ne sont pas convergents, leur moyenne est convergente pour la variance conditionnelle de l'estimateur de l'effet de traitement moyen et conduit à des intervalles de confiance asymptotiquement valides.

MOTS-CLÉS : Expériences d'appariement, variance conditionnelle, effets de traitement.

JEL CLASSIFICATION: C13, C14, C21

---

* A. ABADIE: John F. Kennedy School of Government, 79 John F. Kennedy Street, Cambridge, MA 02138 (alberto_abadie@harvard.edu). G. Imbens: Department of Economics, 1830 Cambridge Street, Cambridge, MA 02138 (imbens@fas.harvard.edu). We would like to thank John Lindsey and two anonymous referees for helpful suggestions.

1 Introduction

In paired randomized experiments units are grouped in pairs with randomized assignment within the pairs. Average treatment effects are then estimated by averaging the within-pair differences in outcomes. Typically, the variance of the average treatment effect estimator is estimated using the sample variance of the within-pair differences (*e.g.*, SNEDECOR and COCHRAN [1989]). However, often the pairing is based on covariate information at least partially available to the researcher (*e.g.*, ROSENBAUM [1995]). Conditional on such information the variance may be substantially smaller.

The conditional variance of the average treatment effect estimator can be expressed in terms of the conditional outcome variances. Consistent estimation of these conditional outcome variances is a difficult task which requires nonparametric estimation involving sample-size-dependent smoothing parameter choices (see, *e.g.*, STONE [1977]). Here we propose a simple way of estimating the conditional variance of the average treatment effect estimator by forming pairs-of-pairs with similar covariate values. These pairs-of-pairs allow us to obtain close-to-unbiased estimators of the conditional outcome variances. Even though these estimators are not consistent for the conditional outcome variances, their average is consistent for the conditional variance of the average treatment effect estimator and allows us to obtain asymptotically valid confidence intervals. A Monte Carlo simulation suggests that our estimator is accurate even in fairly small samples. The results for paired randomized experiments in this article complement previous result on the variance of matching estimators in observational settings (ABADIE and IMBENS [2006, 2008]).

2 Paired Experiments with Covariates

Consider a setup in which N pairs of units are matched on the basis of a vector of covariates. The covariates will be denoted by X_i, $i = 1, ..., N$. Let $\mathbf{X}' = (X_1, X_2, ..., X_N)$. For each i, two units are drawn from the subpopulation with $X = x_i$. One of the two units is randomly selected to receive the active treatment, and for this unit we record the response $Y_i(1)$. The second unit receives the control treatment and for this unit we record the response $Y_i(0)$. Let $\Delta(x)$ be the population average treatment effect for the subpopulation with $X = x$. Under standard conditions randomization implies:

$$\Delta(x) = E[Y_i(1) - Y_i(0)|X_i = x].$$

Let ε_i be the difference between the within-pair difference in outcomes and its population expectation conditional on X_i:

$$\varepsilon_i = Y_i(1) - Y_i(0) - \Delta(X_i).$$

Conditional on X_i, ε_i has mean zero and variance $\sigma_\varepsilon^2(X_i)$.

The average treatment effect for the sample conditional on the covariates is:

$$\tau(\mathbf{X}) = \frac{1}{N}\sum_{i=1}^{N} \Delta(X_i).$$

The average treatment effect for the population is the expected value of $\Delta(X)$ (which is the same as the expected value of $\tau(\mathbf{X})$ over the distribution of X in the population, $f_X(x)$:

$$\tau = E[\tau(\mathbf{X})] = E[\Delta(X)] = \int \Delta(x) F_X(x) dx.$$

The estimator we consider is the average over the sample of the within-pair differences:

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N} (Y_i(1) - Y_i(0)).$$

Conditional on $\mathbf{X}$, $\hat{\tau}$ is unbiased for the average treatment effect $\tau(\mathbf{X})$. Hence, if the elements of $\mathbf{X}$ are chosen as random draws from the distribution of the covariates, $\hat{\tau}$ is unconditionally unbiased for τ.

The variance of $\hat{\tau}$ conditional on $X_1, \ldots, X_N$ is

$$V(\hat{\tau} \mid \mathbf{X}) = E[(\hat{\tau} - \tau(\mathbf{X}))^2 \mid \mathbf{X}] = \frac{1}{N^2}\sum_{i=1}^{N} \sigma_\varepsilon^2(X_i).$$

The unconditional variance of $\hat{\tau}$ is $V(\hat{\tau}) = E[(\hat{\tau} - E[\hat{\tau}])^2]$. If the vector X is chosen at random from $f_X(x)$ the unconditional variance is

$$\begin{aligned} V(\hat{\tau}) &= E[(\hat{\tau} - \tau)^2] = E[V(\hat{\tau} \mid \mathbf{X})] + V(E[\hat{\tau} \mid \mathbf{X}]) \\ &= E\left[\frac{1}{N^2}\sum_{i=1}^{N} \sigma_\varepsilon^2(X_i)\right] + V\left(\frac{1}{N}\sum_{i=1}^{N} \Delta(X_i)\right). \end{aligned}$$

This last equation shows that the marginal variance $V(\hat{\tau})$ is larger than the average of the conditional variance $V(\hat{\tau} \mid \mathbf{X})$ by the variance of the treatment effect $\Delta(X)$. Therefore, if the average effect of the treatment varies substantially with the covariates, the difference between the marginal variance and the average conditional variance will be large.

It is straightforward to estimate the normalized unconditional variance using the sample variance of the within-pair differences:

$$N \cdot \hat{V}(\hat{\tau}) = \frac{1}{N-1}\sum_{i=1}^{N} (Y_i(1) - Y_i(0) - \hat{\tau})^2, \tag{1}$$

which has expectation equal to $N \cdot V(\hat{\tau})$. See for example SNEDECOR and COCHRAN [1989]. Estimating the normalized conditional variance $N \cdot V(\hat{\tau} \mid \mathbf{X})$ is more dif-

ficult because it involves the unknown function $\sigma_\varepsilon^2(x)$. In this article, we propose a simple matching estimator of the conditional variance of $\hat{\tau}$.

The choice of conditional or unconditional variance corresponds to a focus on the sample average treatment effect $\tau(X)$ versus the population average treatment effect τ. There are two reasons for our interest in the former. The first is that in many paired experiments the sample is not chosen at random from a well-defined population and there is therefore no well-defined population average effect. The second reason is that we view it as useful to separate the uncertainty about the treatment effects for the sample (as captured by the conditional variance) from the uncertainty coming from the uncertainty stemming from the extrapolation from the sample to the population (which is combined with the former in the unconditional variance).

3 A Matching Estimator for the Variance

The conditional variance of the average treatment effect estimator depends on the conditional variance of the outcome differences, $\sigma_\varepsilon^2(x)$. Estimating these consistently requires non-parametric estimation of conditional expectations, which in turn requires choices of smoothing parameters that depend on sample size (see, *e.g.*, STONE [1977]). However, we are not interested in the conditional variances at every x, only in the average of these conditional variances in the form of the normalized conditional variance:

$$N \cdot \hat{V}(\hat{\tau} \mid \mathbf{X}) = \frac{1}{N} \sum_{i=1}^{N} \sigma_\varepsilon^2(X_i). \tag{2}$$

To estimate (2) we propose a nearest neighbor or matching approach. This matching approach produces an approximately unbiased estimator of $\sigma_\varepsilon^2(x)$ at every x, but not a consistent one. However, the average of these inconsistent variance estimators is consistent for the average of the variances in the same way that, although the unit-level difference $Y_i(1) - Y_i(0)$ are unbiased but not consistent for $\Delta(X_i)$, the average difference $\sum_i (Y_i(1) - Y_i(0))/N$ is consistent for τ.

Suppose we have two pairs i and j with the same covariates, $X_i = X_j = x$. The average of the squared difference between the two within-pair differences is:

$$E[((Y_i(1) - Y_i(0)) - (Y_j(1) - Y_j(0)))^2 \mid X_i = X_j = x] = 2 \cdot \sigma_\varepsilon^2(x).$$

Therefore,

$$\frac{1}{2}((Y_i(1) - Y_i(0)) - (Y_j(1) - Y_j(0)))^2,$$

is unbiased for $\sigma_\varepsilon^2(x)$. In practice, it may not be possible to find different pairs with the same value of the covariates. Hence let us consider the nearest pair to pair i by solving

$$j(i) = \text{argmin}_{j:j\neq i} \|X_i - X_j\|,$$

where $\|a\|$ is the standard vector norm, $\|a\| = \left(\sum_{n=1}^{k} a_n^2\right)^{1/2}$. Let

$$s_\varepsilon^2(X_i) = \frac{1}{2}((Y_i(1) - Y_i(0)) - (Y_{j(i)}(1) - Y_{j(i)}(0)))^2.$$

Consider the conditional expectation of this variance estimator:

$$E[s_\varepsilon^2(X_i) \mid \mathbf{X}] = \frac{1}{2}(\sigma_\varepsilon^2(X_i) + \sigma_\varepsilon^2(X_{j(i)}) + (\Delta(X_i) - \Delta(X_{j(i)}))^2),$$

which differs from $\sigma_\varepsilon^2(x)$ by a bias term

$$B_i = E[s_\varepsilon^2(X_i) \mid \mathbf{X}] - \sigma_\varepsilon^2(X_i) = \frac{1}{2}(\sigma_\varepsilon^2(X_{j(i)}) - \sigma_\varepsilon^2(X_i) + (\Delta(X_i) - \Delta(X_{j(i)}))^2).$$

At an intuitive level, if the conditional moments of $Y_i(1) - Y_i(0)$ given X_i are sufficiently smooth in the covariates, the bias of the pair-level variance estimates, B_i will vanish asymptotically if the matching discrepancies, $\|X_i - X_{j(i)}\|$, vanish as N increases. However, even if the bias goes to zero as the sample size increases, the pair-level variance estimators, $s_\varepsilon^2(X_i)$, do not estimate the variances $\sigma_\varepsilon^2(X_i)$ consistently because the variances of $s_\varepsilon^2(X_i)$ do not vanish. In fact, as N increases the variance of $s_\varepsilon^2(X_i)$, conditional on X_i, converges to $V(\varepsilon_i^2 \mid X_i)/2$.

We use these pair-level variance estimates to estimate the normalized conditional variance as:

$$N \cdot \widehat{V(\hat{\tau} \mid \mathbf{X})} = \frac{1}{N}\sum_{i=1}^{N} s_\varepsilon^2(X_i). \tag{3}$$

Now, if the bias of $N \cdot \widehat{V(\hat{\tau} \mid \mathbf{X})}$ converges to zero, then it is sufficient to show that the variance also converges to zero in order to prove consistency. Notice that the pair-level variance estimates $s_\varepsilon^2(X_i)$ are not all independent as the same pair may get used as a match more than once. Nevertheless, we will show that because the maximum number of times that a pair can be used as a match is bounded (depending only on the dimension of X, see MILLER *et al.* [1997]) the average, $N \cdot \widehat{V(\hat{\tau} \mid \mathbf{X})}$, is consistent.

The next two assumptions contain sufficient regularity conditions for consistency of our estimator of the conditional variance in paired experiments, defined in equation (3).

ASSUMPTION 1: $(1/N)\sum_{i=1}^{N} \left\| X_i - X_{j(i)} \right\|^2 \to 0$.

ASSUMPTION 2: *(i)* $\Delta(x)$ *and* $\sigma_\varepsilon^2(x)$ *are Lipschitz in* $\mathbb{X}$, *with constants* C_Δ *and* $C_{\sigma_\varepsilon^2}$, *respectively, and (ii) the fourth moments of the conditional distribution of* ε_i *given* $X = x$ *exist and are uniformly bounded by some constant,* $\bar{\mu}_4$.

Assumption 1 is not primitive. However, the next lemma shows that boundedness of the set $\mathbb{X}$, from which the elements of $\mathbf{X}$ are chosen, is enough for Assumption 1 to hold.

LEMMA 1: *If the components of the vector* X_i *are chosen from a bounded set,* $\mathbb{X}$, *then Assumption 1 holds.*

Assumption 2 contains regularity conditions. The following theorem shows that, under assumptions 1 and 2 the estimator of the conditional variance described in equation (3) is consistent.

THEOREM 1: *Suppose that assumptions 1 and 2 hold. Then, conditional on X:*

$$N(\widehat{V(\hat{\tau} \mid \mathbf{X})} - V(\hat{\tau} \mid \mathbf{X})) = \frac{1}{N}\sum_{i=1}^{N} (s_\varepsilon^2(X_i) - \sigma_\varepsilon^2(X_i)) \to 0.$$

4 Discussion

The pair-level variance estimators, $s_\varepsilon^2(X_i)$, can be interpreted as nearest-neighbor nonparametric estimators of $\sigma_\varepsilon^2(X_i)$. For consistency of such estimators it is typically required that the number of neighbors increases with the sample size. However, increasing the number of neighbors is not required for consistency of the normalized conditional treatment effect variance estimator, $N\widehat{V(\hat{\tau} \mid \mathbf{X})} = \sum_i s_\varepsilon^2(X_i)/N$. This is somewhat similar to EICKER [1967] estimation of the variance in regression models with heteroskedasticity. Although his estimators do not estimate the form of the heteroskedasticity consistently, they do consistently estimate the average variance.

Although it is not required for consistency, one can use more than one neighbor to estimate the variances, $\sigma_\varepsilon^2(X_i)$. Let $D_i = Y_i(1) - Y_i(0)$ and let $j_m(i)$ be the index of the m-th closest match to pair i in term of the covariates. An estimator of $\sigma_\varepsilon^2(X_i)$ that uses M neighbors is:

$$\hat{\sigma}_\varepsilon^2(X_i) = \frac{1}{M}\sum_{m=0}^{M} (D_{j_m(i)} - \bar{D}_{M(i)})^2,$$

where $D_{j_0(i)} = D_i$ and $\bar{D}_{M(i)} = (M+1)^{-1} \sum_{m=0}^{M} D_{j_m(i)}$. If $\sigma_\varepsilon^2(X_i)$ does not vary much with the covariates, using multiple neighbors may result in a more precise estimator for the variance and thus in improved confidence intervals in small samples. Notice that for the limiting case of $M = N - 1$, the estimators in (1) and (3) are identical.

Notice also that alternative estimators of $N \cdot V(\hat{\tau} \mid \mathbf{X})$ could be constructed as:

$$\frac{1}{N} \sum_{i=1}^{N} \hat{\sigma}_\varepsilon^2(X_i),$$

where $\hat{\sigma}_\varepsilon^2(x)$ is a consistent estimator of $\sigma_\varepsilon^2(x)$ given by non-parametric smoothing techniques (*e.g.*, series or kernel regression). The advantage of our matching estimator of $N \cdot V(\hat{\tau} \mid \mathbf{X})$ is that it does not require consistent estimation of the function $\sigma_\varepsilon^2(x)$, and therefore it does not force researchers to choose smoothing parameters as functions of the sample size.

If the vector X is chosen at random from the distribution, $f_X(x)$, then the standard variance estimator in (1) provides asymptotically conservative confidence intervals for the conditional average treatment effect $\tau(\mathbf{X})$, as well as valid confidence intervals for the population average treatment effect τ. The variance estimator proposed here in (3) provides asymptotically tighter, but valid, confidence intervals for the conditional average treatment effect $\tau(\mathbf{X})$ regardless of how $\mathbf{X}$ is chosen, but not for the population average treatment effect τ. Conditioning on the covariates can therefore be interpreted as changing the estimand. Which estimand is of interest may differ in applications, although often interest is in the specific sample at hand, and thus in $\tau(\mathbf{X})$, especially when the sample is not representative of the population of interest in their covariate distribution.

5 A Small Simulation Study

In this section we carry out a small simulation study to investigate the small sample properties of the proposed variance estimator and the associated confidence intervals. We draw samples of $N = 50$ and $N = 200$ pairs. In each replication the scalar covariate is drawn from a uniform distribution on [0, 4]. In our initial Monte Carlo specification, conditional on $X_i = x$, $Y_i(0)$ has a normal distribution with mean x and variance equal to 1, and $Y_i(1)$ has a normal distribution with mean zero and variance equal to 0.5. In Table 1, we report the average standard error based on the standard formula (1), the average standard error based on the proposed formula (3), and the coverage rates of the associated 95% and 90% confidence intervals of the conditional treatment effect $\tau(\mathbf{X})$ (which differs between replications because the covariates are re-drawn each time). We use 50.000 Monte Carlo repetitions. In the simulations in Table 1, for both sample sizes ($N = 50$ and $N = 200$) the average standard error is considerably smaller for the matching variance estimator. In addition, the confidence intervals based on this variance estimator have approximately the right coverage, whereas the standard variance estimator leads to substantial over-coverage. A sample size of 50 pairs seems sufficiently large to lead to fairly accurate estimates of the variance.

TABLE 1

Monte-Carlo Simulation: Basic Results

$X \sim U[0, 4]$
$Y(0)|X = x \sim N(x, 1)$, $Y(1)|X = x \sim N(0, 1/2)$
50.000 replications, one match

	$N = 50$			$N = 200$		
	average s.e.	95% conf. interv.	90% conf. interv.	average s.e.	95% conf. interv.	90% conf. interv.
Standard variance estimator	.2370	.9915	.9742	.1189	.9918	.9743
Matching variance estimator	.1716	.9410	.8892	.0864	.9463	.8963

Table 2, reports Monte Carlo results for the cases of multiple matches (Panel A) and heteroskedasticity (Panel B). In Panel A of Table 2, we repeat the analysis of Table 1 using one, five, and twenty-five matches for the calculation of the conditional variance estimator. For $N = 50$ coverage rates improve by using five matches, relative to just one. For $N = 50$, coverage rates deteriorate when we go from five to twenty-five matches, but they stay reasonably close to nominal levels. For $N = 200$, coverage rates improve as we increase the number of matches from one to five, and from five to twenty-five. In Panel B, we repeat the analysis of Panel A, this time allowing for heteroskedasticity, and we obtain the same results.

TABLE 2

Monte-Carlo Simulation: Multiple Matches and Heteroskedasticity

Panel A: Homoskedasticity
$X \sim U[0, 4]$
$Y(0)|X = x \sim N(x, 1)$, $Y(1)|X = x \sim N(0, 1/2)$
50.000 replications, one and multiple matches

	$N = 50$			$N = 200$		
	average s.e.	95% conf. interv.	90% conf. interv.	average s.e.	95% conf. interv.	90% conf. interv.
Standard variance estimator	.2370	.9915	.9742	.1189	.9918	.9743
Matching variance estimator						
1 match	.1716	.9410	.8892	.0864	.9463	.8963
5 matches	.1732	.9472	.8961	.0865	.9474	.8971
25 matches	.1920	.9688	.9296	.0871	.9488	.9003

Panel B: Heteroskedasticity
$X \sim U[0, 4]$
$Y(0)|X = x \sim N(x, 1)$, $Y(1)|X = x \sim N(0, 1 - x + x^2/4)$
50.000 replications, one and multiple matches

	$N = 50$			$N = 200$		
	average s.e.	95% conf. interv.	90% conf. interv.	average s.e.	95% conf. interv.	90% conf. interv.
Standard variance estimator	.2297	.9926	.9775	.1153	.9940	.9784
Matching variance estimator						
1 match	.1616	.9403	.8887	.0814	.9463	.8965
5 matches	.1629	.9456	.8940	.0815	.9478	.8970
25 matches	.1787	.9659	.9259	.0819	.9491	.8985

In Table 3, we consider the case in which the variance of $\Delta(X_i)$ is equal to zero, so both the standard variance estimator in (1) and the matching variance estimator in (3) produce valid inference for $\tau(\mathbf{X})$. As expected, confidence intervals constructed using the standard variance estimator and the matching variance estimator produce coverage rates that are close to nominal levels. ■

TABLE 3

Monte-Carlo Simulation: $V(\Delta(X)) = 0$

$X \sim U[0, 4]$
$Y(0)|X = x \sim N(x, 1)$, $Y(1)|X = x \sim N(0, 1/2)$
50.000 replications, one and multiple matches

	$N = 50$			$N = 200$		
	average s.e.	95% conf. interv.	90% conf. interv.	average s.e.	95% conf. interv.	90% conf. interv.
Standard variance estimator	.1723	.9467	.8940	.0865	.9476	.8976
Matching variance estimator						
1 match	.1715	.9412	.8890	.0864	.9463	.8962
5 matches	.1721	.9456	.8936	.0865	.9473	.8970
25 matches	.1722	.9454	.8942	.0865	.9473	.8976

References

ABADIE A. and IMBENS G. (2006). – "Large Sample Properties of Matching Estimators for Average Treatment Effects", *Econometrica* 74, 235-267.

ABADIE A. and IMBENS G. (2008). – "On the Failure of the Bootstrap for Matching Estimators", *Econometrica* 76, 1537-1557.

EICKER F. (1967). – "Limit Theorems for Regression with Unequal and Dependent Errors", *Fifth Berkeley Symposium on Mathematical Statistics and Probability* eds., L. LeCam and J. Neyman, Berkeley, University of California, 59-82.

IMBENS G. (2004). – "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review", *Review of Economics and Statistics* 86, 4-29.

MILLER G.L., TENG S., THURSTON W. and VAVASIS S.A. (1997). – "Separators for Sphere-Packings and Nearest Neighbor Graphs", *Journal of the ACM* 44, 1-29.

ROSENBAUM P. (1995). – *Observational Studies*, Springer Verlag, New York.

SNEDECOR G. and COCHRAN W. (1989). – *Statistical Methods*, eighth edition, Iowa State University Press, Ames, Iowa.

STONE C. (1977). – "Consistent Nonparametric Regression" (with discussion), *Annals of Statistics* 5, 595-645.

Annex

A: *Proofs*

Proof of Lemma 1: Because the set $\mathbb{X}$ is bounded, it is enough to prove that the $(1/N)\sum_{i=1}^{N}\left\|X_i - X_{j(i)}\right\|$ converges to zero. (Because the matching discrepancies are bounded by the diameter of $\mathbb{X}$, convergence of $(1/N)\sum_{i=1}^{N}\left\|X_i - X_{j(i)}\right\|$ to zero implies that $(1/N)\sum_{i=1}^{N}\left\|X_i - X_{j(i)}\right\|^2$ converges to zero too.) Given that the set $\mathbb{X}$ is bounded, we can always embed $\mathbb{X}$ in a hypersphere of the same diameter. Without loss of generality, assume that the radius of such hypersphere is equal to one. Now, suppose that there are M matching discrepancies greater than 2ε, with $\varepsilon < 1/2$: $\left\|X_i - X_{j(i)}\right\| > 2\varepsilon$. Construct an open ball of radius ε around each such point. These M balls do not intersect because their radii are smaller than half of the distances between their centers. The volume of each such ball is equal to $(\varepsilon^k/k)S_k$, where S_k is the surface area of a unit hypersphere in k dimensions. Obviously, all such balls are embedded in a hypersphere of radius $(1+\varepsilon)$. Therefore: $M(\varepsilon^k/k)S_k < ((1+\varepsilon)^k/k)S_k$. As a result, $\forall\varepsilon < 1/2$:

$$M < \left(\frac{1+\varepsilon}{\varepsilon}\right)^k.$$

Then, because the diameter of $\mathbb{X}$ is bounded by 2, we obtain:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|X_i - X_{j(i)}\right\| < 2\frac{M}{N} + 2\varepsilon\frac{(N-M)}{N} < 2(M/N+\varepsilon) < 2\left(\frac{1}{N}\left(\frac{1+\varepsilon}{\varepsilon}\right)^k + \varepsilon\right).$$

Consider $\varepsilon = N^{-1/(1+k)}$. As $N \to \infty$, $\varepsilon \to 0$, and

$$\frac{1}{N}\left(\frac{1+\varepsilon}{\varepsilon}\right)^k = \frac{N^{k/(1+k)}}{N}O(1) = o(1).$$

As a result,

$$\frac{1}{N}\sum_{i=1}^{N}\left\|X_i - X_{j(i)}\right\| \to 0,$$

which along with boundedness of $\mathbb{X}$ proves the statement of the lemma.
Proof of Theorem 1: Because,

$$E[s_\varepsilon^2(X_i)\,|\,\mathbf{X}] - \sigma_\varepsilon^2(X_i) = \frac{1}{2}(\sigma_\varepsilon^2(X_{j(i)}) - \sigma_\varepsilon^2(X_i) + (\Delta(X_i) - \Delta(X_{j(i)}))^2),$$

and Assumption 2, we obtain:

$$\left|E[s_\varepsilon^2(X_i)\,|\,\mathbf{X}]-\sigma_\varepsilon^2(X_i)\right| \le \frac{C_\Delta}{2}\left\|X_i - X_{j(i)}\right\| + \frac{C_{\sigma_\varepsilon^2}^2}{2}\left\|X_i - X_{j(i)}\right\|^2.$$

By Assumption 1:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|X_i - X_{j(i)}\right\|^2 \to 0.$$

Cauchy-Schwarz Inequality implies that:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|X_i - X_{j(i)}\right\| \le \left(\frac{1}{N}\sum_{i=1}^{N}\left\|X_i - X_{j(i)}\right\|^2\right)^{1/2} \to 0.$$

Therefore:

$$\frac{1}{N}\sum_{i=1}^{N}\left(E[s_\varepsilon^2(X_i)\,|\,\mathbf{X}]-\sigma_\varepsilon^2(X_i)\right) \to 0. \tag{A.1}$$

Notice that:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\left(s_\varepsilon^2(X_i)-E[s_\varepsilon^2(X_i)\,|\,\mathbf{X}]\right) = {} & \frac{1}{N}\sum_{i=1}^{N}\left(\varepsilon_i^2-\sigma_\varepsilon^2(X_i)\right)\\ & +\frac{1}{2N}\sum_{i=1}^{N}\left(\varepsilon_{j(i)}^2-\sigma_\varepsilon^2(X_{j(i)})\right)\\ & -\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i\varepsilon_{j(i)}\\ & +\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i\left(\Delta(X_i)-\Delta(X_{j(i)})\right)\\ & -\frac{1}{N}\sum_{i=1}^{N}\varepsilon_{j(i)}\left(\Delta(X_i)-\Delta(X_{j(i)})\right).\end{aligned} \tag{A.2}$$

Clearly, the conditional expectations of all the terms on the right-hand-side of equation (A.2) are equal to zero. Next, we show that the conditional variances of all the terms on the right-hand-side of equation (A.2) converge to zero. Because the expectations of all the terms, conditional on **X** are equal to zero, the variances are equal to the expectations of the squares. The maximum of number of times that an observation can be used as a match, given that the dimension of X is equal to k, is bounded by $\bar{L}(k)$. The integer $\bar{L}(k) < \infty$ (sometimes called the "kissing number" in k dimensions) is equal to the maximum number of non-overlapping unit balls in $\mathbb{R}^k$ that can be arranged to overlap with a unit ball. For $k = 1, 2, 3$, $\bar{L}(k) = 2, 6, 12,$

respectively. (See MILLER *et al.* [1997].) For each of the five terms on the right-hand-side of equation (A.2), the conditional second moments are:

$$E\left[\left(\frac{1}{2N}\sum_{i=1}^{N}(\varepsilon_i^2-\sigma_\varepsilon^2(X_i))\right)^2 \mid \mathbf{X}\right]=\frac{1}{4N^2}\sum_{i=1}^{N}V(\varepsilon_i^2 \mid \mathbf{X})\leq\frac{\bar{\mu}_4}{4N}.$$

$$\begin{aligned}E\left[\left(\frac{1}{2N}\sum_{i=1}^{N}(\varepsilon_{j(i)}^2-\sigma_\varepsilon^2(X_{j(i)}))\right)^2 \mid \mathbf{X}\right]&=E\left[\frac{1}{4N^2}\sum_{i=1}^{N}(\varepsilon_{j(i)}^2-\sigma_\varepsilon^2(X_{j(i)}))^2 \mid \mathbf{X}\right]\\&\quad+E\left[\frac{1}{2N^2}\sum_{i=1}^{N}\sum_{t>i}(\varepsilon_i^2-\sigma_\varepsilon^2(X_i))(\varepsilon_{j(i)}^2-\sigma_\varepsilon^2(X_{j(i)}))\mid \mathbf{X}\right]\\&\leq\frac{\bar{\mu}_4}{4N}+\frac{\bar{L}(k)\bar{\mu}_4}{2N},\end{aligned}$$

$$\begin{aligned}E\left[\left(\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i\varepsilon_{j(i)}\right)^2 \mid \mathbf{X}\right]&=E\left[\frac{1}{N^2}\sum_{i=1}^{N}\varepsilon_i^2\varepsilon_{j(i)}^2 \mid \mathbf{X}\right]+E\left[\frac{2}{N^2}\sum_{i=1}^{N}\sum_{t>i}\varepsilon_i\varepsilon_{j(i)}\varepsilon_t\varepsilon_{j(t)} \mid \mathbf{X}\right]\\&\leq\frac{\bar{\mu}_4}{N}+\frac{2\bar{L}(k)\bar{\mu}_4}{N},\end{aligned}$$

$$\begin{aligned}E\left[\left(\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i(\Delta(X_i)-\Delta(X_{j(i)}))\right)^2 \mid \mathbf{X}\right]&=E\left[\frac{1}{N^2}\sum_{i=1}^{N}\varepsilon_i^2(\Delta(X_i)-\Delta(X_{j(i)}))^2 \mid \mathbf{X}\right]\\&\leq\frac{\bar{\mu}_4^{1/2}}{N}E\left[\frac{1}{N}\sum_{i=1}^{N}(\Delta(X_i)-\Delta(X_{j(i)}))^2 \mid \mathbf{X}\right]\\&\leq\frac{\bar{\mu}_4^{1/2}}{N}C_\Delta^2\frac{1}{N}\sum_{i=1}^{N}\|X_i-X_{j(i)}\|^2,\end{aligned}$$

$$\begin{aligned}E\left[\left(\frac{1}{N}\sum_{i=1}^{N}\varepsilon_{j(i)}(\Delta(X_i)-\Delta(X_{j(i)}))\right)^2 \mid \mathbf{X}\right]&=E\left[\frac{1}{N^2}\sum_{i=1}^{N}\varepsilon_{j(i)}^2(\Delta(X_i)-\Delta(X_{j(i)}))^2 \mid \mathbf{X}\right]\\&\quad+E\left[\frac{1}{N^2}\sum_{i=1}^{N}\sum_{t\neq i}\varepsilon_{j(i)}(\Delta(X_i)-\Delta(X_{j(i)}))\varepsilon_{j(t)}(\Delta(X_t)-\Delta(X_{j(t)}))\mid \mathbf{X}\right]\\&\leq\frac{\bar{\mu}_4^{1/2}}{N}C_\Delta^2\frac{1}{N}\sum_{i=1}^{N}\|X_i-X_{j(i)}\|^2\\&\quad+\bar{L}(k)\bar{\mu}_4^{1/2}C_\Delta^2\frac{1}{N^2}\sum_{i=1}^{N}\|X_i-X_{j(i)}\|\sum_{j(t)=j(i)}\|X_t-X_{j(t)}\|\\&\leq o(1)+\bar{L}(k)\bar{\mu}_4^{1/2}C_\Delta^2\frac{1}{N}\sum_{i=1}^{N}\|X_i-X_{j(i)}\|\left(\frac{1}{N}\sum_{i=1}^{N}\|X_t-X_{j(t)}\|\right).\end{aligned}$$

As a result, the conditional second moments of each of the five terms on the right-hand-side of equation (A.2) converge to zero. This implies that

$$V\left(\frac{1}{N}\sum_{i=1}^{N}(s_\varepsilon^2(X_i)-E[s_\varepsilon^2(X_i)\,|\,\mathbf{X}])\,|\,\mathbf{X}\right)=o(1),$$

and therefore, conditional on X:

$$\frac{1}{N}\sum_{i=1}^{N}(s_\varepsilon^2(X_i)-E[s_\varepsilon^2(X_i)\,|\,\mathbf{X}])=o_p(1). \tag{A.3}$$

Now, the result of the theorem follows from equations (A.1) and (A.3).