

# UStat Package Manual

July 22, 2025

## 1 Overview

This document provides more detailed documentation for the functions to compute the  $U$ -statistic estimators of the variance-covariance and its sampling covariance proposed by [Rose, Schellenberg, and Shem-Tov \(2022\)](#). This package is available on PyPI here (link forthcoming).

The original paper was in the context of estimating teachers' value-added on student outcomes along several dimensions. We maintain the 'teacher-student' language when describing the empirical setup and estimators since it can help ground ideas about what the estimator inputs are. These estimators and functions would work in other settings with the empirical setup described below.

### 1.1 Empirical setup

The package assumes the researcher observes for each teacher  $j = 1, 2, \dots, J$  and outcome  $k = 1, 2, \dots, K$ :

$$y_j^k = (y_{j1}^k, \dots, y_{jT_j}^k)'$$

where  $y_{jt}^k = \alpha_j^k + e_{jt}^k$ . The parameter  $\alpha_j^k$  represents teacher  $j$ 's effect on outcome  $k$ . Different outcomes could refer to separate measures (e.g., math test scores and reading test scores) or separate sub-populations (e.g., male and female students). The term  $e_{jt}^k$  represents estimation error. The key assumptions are that:  $E[e_{jt}^k | \alpha_j^k] = 0$  for all  $j, k, t$ , and that  $E[e_{jt}^k e_{jt'}^l] = 0$  for  $t \neq t'$  and all  $j, k, l$ . Using the above, we can compute the teacher-level residuals:

$$\bar{Y}_{jt}^k = \alpha_j^k + \bar{v}_{jt}^k$$

We assume that the  $\bar{v}_{jt}^k$  are uncorrelated across years, i.e.  $E[\bar{v}_{jt}^k \bar{v}_{jt'}^l] = 0$  when  $t \neq t'$ , and that  $E[\bar{v}_{jt}^k] = 0$ . In this context, our main object of interest is  $\text{Cov}(\alpha_j^k, \alpha_j^l)$  and its sampling variance. Suppose we have two outcomes  $A$  and  $C$  from the possible set of outcomes, we can use the following covariance identity

$$\begin{aligned} \text{Cov}(\alpha_j^A, \alpha_j^C) &= \frac{1}{J} \sum_{j=1}^J \alpha_j^A \alpha_j^C - \left( \frac{1}{J} \sum_{j=1}^J \alpha_j^A \right) \left( \frac{1}{J} \sum_{j=1}^J \alpha_j^C \right) \\ &= \frac{J-1}{J^2} \sum_{j=1}^J \alpha_j^A \alpha_j^C - \frac{2}{J^2} \sum_{j=1}^{J-1} \sum_{k>j}^J \alpha_j^A \alpha_k^C \end{aligned}$$

We can estimate the above as the average of all pairs of products teacher-level residuals  $\bar{Y}_{jt}^A \bar{Y}_{jt'}^C$ . To eliminate the bias, we remove the products from the same year (i.e.  $\bar{Y}_{jt}^A \bar{Y}_{jt}^C$ ). That is,

$$\hat{\text{Cov}}(a_j^A, a_j^C) = \left( \frac{J-1}{J} \right) \frac{1}{J} \sum_{j=1}^J \binom{T_j}{2}^{-1} \sum_{t=1}^{T_j-1} \sum_{k=t+1}^{T_j} \bar{Y}_{jt}^A \bar{Y}_{jk}^C - \frac{2}{J^2} \sum_{j=1}^{J-1} \sum_{k>j}^J \bar{Y}_j^A \bar{Y}_k^C$$

We can obtain the sampling variance using a 2nd order  $U$ -statistic representation of the above (briefly outlined below in the Sampling Variance section).

The package produces estimates of  $\text{Var}(a_j^k)$  and  $\text{Cov}(a_j^k, a_j^l)$ , as well as estimates of their sampling variance. There are options to equally weight each these variance/covariance parameters as well as to

apply user-given weights. The package can also accomodate heavily unbalanced data, where  $T_j$  differs across teachers and/or across outcomes within teacher. Given the definition for the estimators, all the inputs referenced as ‘data’ or ‘arrays’ represent the residuals  $\bar{Y}_{jt}$  for each teacher  $j$  across time  $t = 1, 2, \dots, T_j$ .

Rose, Schellenberg, and Shem-Tov (2022) offer complete details about deriving the sampling variance, as well as discussion of the assumptions required for the estimators to estimate the variance of causal teacher effects.

## 2 Functions

### 2.1 Variance-covariance: varcovar

The ‘ustat.varcovar( $A, C$ )’ function computes the unbiased covariance between two datasets  $A$  and  $C$  which contain the residuals  $\bar{Y}_{jt}^A$  and  $\bar{Y}_{jt}^C$ . The function also supports weighted variance calculations (where each weight corresponds to a row of  $A$  and  $C$ ) and weighting by year. Specifically, the function calculates any of the following:

(1) Unweighted:

$$\hat{C}_{unweighted} = \left(\frac{J-1}{J}\right) \frac{1}{J} \sum_{j=1}^J \binom{T_j}{2}^{-1} \sum_{t=1}^{T_j-1} \sum_{k=t+1}^{T_j} \bar{Y}_{jt}^A \bar{Y}_{jk}^C - \frac{2}{J^2} \sum_{j=1}^{J-1} \sum_{k>j}^J \bar{Y}_j^A \bar{Y}_k^C \quad (1)$$

(2) Weighting each agent

$$\hat{C}_w = \sum_{j=1}^J \binom{T_j}{2}^{-1} \tilde{w}_j (1 - \tilde{w}_j) \sum_{t=1}^{T_j-1} \sum_{k=t+1}^{T_j} \bar{Y}_{jt}^A \bar{Y}_{jk}^C - 2 \sum_{j=1}^{J-1} \sum_{k>j}^J \tilde{w}_j \bar{Y}_j^A \tilde{w}_k \bar{Y}_k^C \quad (2)$$

where  $\tilde{w}_j = w_j / \sum_{j=1}^J w_j$ , and  $|T_j^A|$  represents the number of time periods individual  $j$  is observed for outcome  $A$ .

**Notes:**

1. due to the debiasing procedure of removing the  $t = t$  product terms, this function can yield negative variance estimates. Negative variance estimates occur when the variance of teacher means is close to 0.

#### 2.1.1 Arguments

ustat\_var.varcovar( $A, C, w, \text{quiet}=\text{True}$ )

1.  $A, C$  = two  $J$ -by- $\max(T_j)$  arrays containing the residuals for each teacher/individual across time for outcome  $A$  and  $C$ . These arrays can contain missing values (in the form of a Nan), and each row of  $A$  and  $C$  can have missings in different spots.
2.  $w$  = a  $J$ -by-1 array containing weights for the rows of  $A, C$ . Used to compute a weighted variance-covariance. Empty/none by default, so that default behaviour is to estimate an unweighted covariance.
3. quiet = True/false on whether to report to user what type of variance was calculated and whether the panels were balanced/unbalanced. Reporting messages suppressed by default.

#### 2.1.2 Usage

---

```
import ustat_var as ustat
import numpy as np

# Data and weights
np.random.seed(48912)
```

```

n_teachers, n_time = 50, 10
X, Y = ustat.generate_test_data.generate_unique_nan_arrays(n_rows=n_teachers, n_cols=n_time,
    n_arrays=2, min_int=1, max_int=9, nan_prob=0.25, seed = 48912, balanced = False)

weights = np.random.exponential(size = n_teachers)

# Variance-covariance
ustat.varcovar(X, X) # Var(X)
ustat.varcovar(X, Y) # Cov(X, Y)

ustat.varcovar(X, X, w = weights) # weighted Var(X)
ustat.varcovar(X, Y, w = weights) # weighted Cov(X, Y)

ustat.varcovar(X, X, yearWeighted = True) # year weighted Var(X)
ustat.varcovar(X, Y, yearWeighted = True) # year weighted Cov(X, Y)

```

---

## 2.2 Sampling variance: ustat\_samp\_covar

The ‘`ustat.ustat_samp_covar(A, B, C, D)`’ function computes the sampling covariance of  $\text{Cov}(A, B)$  and  $\text{Cov}(C, D)$ . Note that we do not impose any logical cap on the sampling variance, meaning this function can yield sampling covariances-variances which imply correlations exceeding 1. Specifically, the function computes an estimator for:

$$\begin{aligned}
& \text{Cov} \left( \hat{\text{Cov}}(a_j^A, a_j^B) - \text{Cov}(a_j^A, a_j^B), \hat{\text{Cov}}(a_j^C, a_j^D) - \text{Cov}(a_j^C, a_j^D) \right) = \\
& \sum_i \sigma_i^{AC} \left( \sum_{k \neq i} C_{ik}^{AB} a_{j(k)}^B \right) \left( \sum_{k \neq i} C_{ik}^{CD} a_{j(k)}^D \right) + \sum_i \sigma_i^{AD} \left( \sum_{k \neq i} C_{ik}^{AB} a_{j(k)}^B \right) \left( \sum_{k \neq i} C_{ik}^{DC} a_{j(k)}^C \right) \\
& + \sum_i \sigma_i^{BC} \left( \sum_{k \neq i} C_{ik}^{BA} a_{j(k)}^A \right) \left( \sum_{k \neq i} C_{ik}^{CD} a_{j(k)}^D \right) + \sum_i \sigma_i^{BD} \left( \sum_{k \neq i} C_{ik}^{BA} a_{j(k)}^A \right) \left( \sum_{k \neq i} C_{ik}^{DC} a_{j(k)}^C \right) + \\
& \sum_i \sigma_i^{AD} \sum_{k \neq i} C_{ik}^{AB} C_{ik}^{DC} \sigma_k^{BC} + \sum_i \sigma_i^{AC} \sum_{k \neq i} C_{ik}^{AB} C_{ik}^{CD} \sigma_k^{BD} \quad (3)
\end{aligned}$$

where  $\sigma_i^{AC}$  represents the covariance between  $A$  and  $C$  and

$$C_{ik}^{AC} = \begin{cases} \frac{J-1}{J^2} \frac{1}{|T_j^A| |T_j^C| - |T_j^A \cap T_j^C|} & \text{if } j(i) = j(k) \\ \frac{1}{J^2} \frac{-1}{|T_{j(i)}^A| |T_{j(k)}^C|} & \text{if } j(i) \neq j(k) \end{cases}$$

### Notes:

1. the function computes *unbiased* estimators of the product-sums  $\left( \sum_{k \neq i} C_{ik}^{AB} a_{j(k)}^B \right) \left( \sum_{k \neq i} C_{ik}^{CD} a_{j(k)}^D \right)$ . As with the variance-covariance estimator embodied in `varcovar()`, this means estimated sampling variances *can* be negative, though this does not happen often.
2. the function accepts row-weights to compute teacher/individual level weighted sampling covariances. This enters the function through the  $C_{ik}^{AC}$  coefficients. Instead of being pre-multiplied by  $(J-1)/J^2$  and  $1/J^2$ , they are pre-multiplied instead by  $\tilde{w}_j(1-\tilde{w}_j)$  and  $\tilde{w}_j^2$ , where  $\tilde{w}_j = w_j / \sum_{j=1}^T w_j$  and  $w_j$  represents the weight given to row/individual/teacher  $j$ .

### 2.2.1 Arguments

`ustat_var.ustat_samp_covar(A, B, C, D, w)`

1.  $A, B, C, D =$  four  $J$ -by- $\max(T_j)$  arrays containing the residuals for each teacher/individual across time for outcome  $A, B, C$  and  $D$ . Each can contain missing values (in the form of a Nan), and each row of each array can contain missing values in different spots.

2.  $w$  = a  $J$ -by-1 array containing the weights to be applied to each row/individual/teacher of  $A, B, C, D$ . Used to compute a weighted variance-covariance. Empty/none by default, so that default behaviour is to estimate an unweighted covariance.

### 2.2.2 Usage

---

```
import ustat_var as ustat
import numpy as np

# Data and weights
np.random.seed(48912)
n_teachers, n_time = 50, 10
A, B, C, D = ustat.generate_test_data.generate_unique_nan_arrays(n_rows=n_teachers,
    n_cols=n_time, n_arrays=4, min_int=1, max_int=9, nan_prob=0.25, seed = 48912, balanced =
    False)

# Compute
ustat.ustat_samp_covar(A, A, A, A) # Var(Var(A))
ustat.ustat_samp_covar(A, B, A, B) # Var(Cov(A, B))
ustat.ustat_samp_covar(A, B, C, D) # Cov(Cov(A, B), Cov(C, D))
```

---