Covariance/Variance Decomposition

Here we show that the population covariance can be expressed as a difference equation.

For $\Delta_k a_t \equiv a_{t+k} - a_t$ (the difference) and $\bar{a} \equiv \frac{1}{T}\sum_{t=1}^{T} a_t$ (the mean)

Given sequences $\{a_t\}_{t=1}^{T}$ and $\{b_t\}_{t=1}^{T}$, we have

$$\frac{1}{T}\sum_{t=1}^{T}(a_t - \bar{a})(b_t - \bar{b}) = \frac{1}{T^2}\sum_{k=1}^{T-1}\sum_{t=1}^{T-k}\Delta_k a_t \Delta_k b_t$$

Strategy

The population covariance can be written as follows:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}(a_t - \bar{a})(b_t - \bar{b}) &= \frac{1}{T}\sum_{t=1}^{T}\left(a_t b_t - \bar{a} b_t - a_t \bar{b} + \bar{a}\bar{b}\right) && (1)\\
&= \left(\frac{1}{T}\sum_{t=1}^{T} a_t b_t\right) - \bar{a}\bar{b} && (2)
\end{aligned}$$

so the proof strategy is to show that the difference equation can be reduced to the same form.

Proof

Since $\Delta_k a_t \equiv a_{t+k} - a_t$ we can expand the difference equation as:

$$\begin{aligned}
\frac{1}{T^2}\sum_{k=1}^{T-1}\sum_{t=1}^{T-k}\Delta_k a_t \Delta_k b_t &= \frac{1}{T^2}\sum_{k=1}^{T-1}\sum_{t=1}^{T-k}\left(a_{t+k} - a_t\right)\left(b_{t+k} - b_t\right) && (3)\\
&= \frac{1}{T^2}\sum_{k=1}^{T-1}\sum_{t=1}^{T-k}\left(a_{t+k} b_{t+k} + a_t b_t\right) && (4)\\
&- \frac{1}{T^2}\sum_{k=1}^{T-1}\sum_{t=1}^{T-k} a_t b_{t+k} - \frac{1}{T^2}\sum_{k=1}^{T-1}\sum_{t=1}^{T-k} a_{t+k} b_t && (5)
\end{aligned}$$

and we'll evaluate each of the 4 terms in turn.

term 1:

Making a change of variable: $s = t + k$ and noting that when t ranges from 1 to Tk, s ranges from $1 + k$ to T, we have:

$$\frac{1}{T^2}\sum_{k=1}^{T-1}\sum_{t=1}^{T-k} a_{t+k}b_{t+k} = \frac{1}{T^2}\sum_{k=1}^{T-1}\sum_{s=k+1}^{T} a_s b_s \tag{6}$$

$$= \frac{1}{T^2}\sum_{s=2}^{T}\sum_{k=1}^{s-1} a_s b_s \tag{7}$$

$$= \frac{1}{T^2}\sum_{s=2}^{T}(s-1)\, a_s b_s \tag{8}$$

where in the second-last line we note that $a_s b_s$ doesn't depend on k so we count each $a_s b_s$ exactly $s-1$ times for $s \in [2, T]$ (i.e. due to the inner sum). Note that this is

$$\frac{1}{T^2}\left(a_2b_2 + 2a_3b_3 + 3a_4b_4 + \cdots + + (T-2)\, a_{T-1}b_{T-1} + (T-1)\, a_T b_T\right)$$

term 2:

The second term is

$$\frac{1}{T^2}\sum_{k=1}^{T-1}\sum_{t=1}^{T-k} a_t b_t$$

and it can be decomposed directly to:

$$\frac{1}{T^2}\left((T-1)a_1b_1 + (T-2)a_2b_2 + (T-3)a_3b_3 + \cdots + a_{T-1}b_{T-1}\right)$$

i.e. $\frac{1}{T^2}$ times the following sum of $T-1$ values:

$$\begin{aligned}
&+ (a_1b_1 + a_2b_2 + \cdots + a_{T-3}b_{T-3} + a_{T-2}b_{T-2} + a_{T-1}b_{T-1}) \\
&+ (a_1b_1 + a_2b_2 + \cdots + a_{T-3}b_{T-3} + a_{T-2}b_{T-2}) \\
&\qquad\qquad\vdots \\
&+ (a_1b_1 + a_2b_2) \\
&+ (a_1b_1)
\end{aligned}$$

So combining the first and second terms we get $\frac{T-1}{T^2}\sum_{t=1}^{T} a_t b_t$, and we're on our way towards showing that

$$\frac{1}{T^2}\sum_{k=1}^{T-1}\sum_{t=1}^{T-k} \Delta_k a_t \Delta_k b_t = \left(\frac{1}{T}\sum_{t=1}^{T} a_t b_t\right) - \bar{a}\bar{b}$$

term 3:

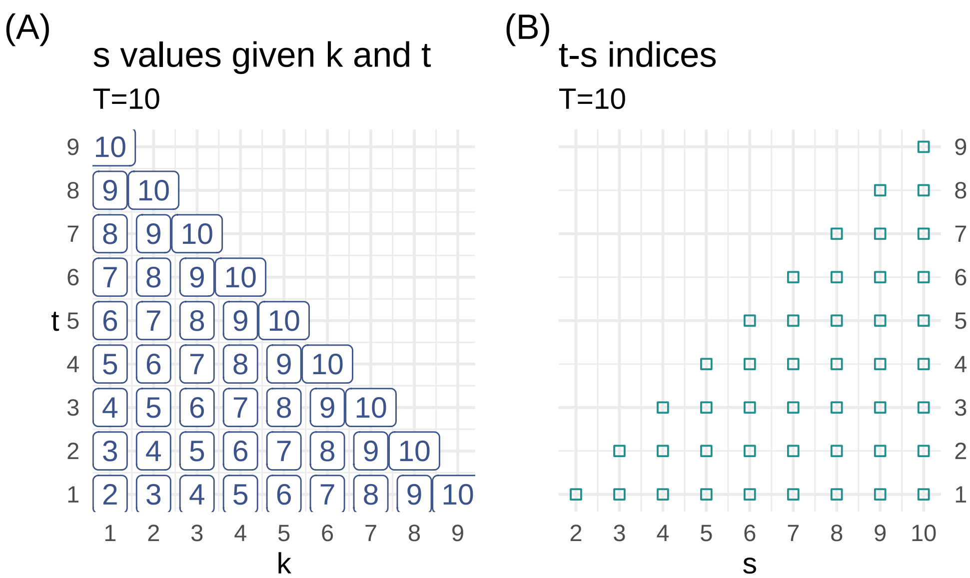
The third term is $-\frac{1}{T^2}\sum_{k=1}^{T-1}\sum_{t=1}^{T-k} a_{t+k}b_t$ and again we will use the change of variables $s = t + k$.

Before changing variables, the outer sum goes from $k = 1$ to $k = T - 1$ and the inner sum goes from $t = 1$ to $t = T - k$. The (k, t) indices used in $-\frac{1}{T^2}\sum_{k=1}^{T-1}\sum_{t=1}^{T-k} a_{t+k}b_t$ are shown in figure (A) below (for $T = 10$), along with the $s = t + k$ values.

With the change of variable, the outer sum goes from $s = t + k = 2$ to $s = t + k = T$ and the inner sum goes from $t = 1$ to $t = s - 1$. The (s, t) indices used to sum $-\frac{1}{T^2}\sum_{s=2}^{T}\sum_{t=1}^{s-1} a_s b_t$ are shown in figure (B) below (for $T = 10$).

The first index (s) is along the horizontal axis, while the second index (t) is along the vertical axis.

```
T <- 10
res <- 1:(T-1) |> # outer sum
  purrr::map(
    (\(k){ 1:(T-k) |> # inner sum
        purrr::map(\(t)data.frame("k"=k,"t"=t, "s"=t+k)) }
    )
  ) |> dplyr::bind_rows()
```



Note that the inner sum on t gives $s - 1$ pairs, so that the total number of pairs is

$$\sum_{s=2}^{T}(s-1) = 1 + 2 + 3 + \cdots (T-1) = \frac{T(T-1)}{2}$$

or half of all T^2 index pairs (t, s) less the diagonal (s, s). From figure (B) we see that it sums the lower diagonal of all index pair values, less the diagonal.

This sum also counts the number of unique pairs (s, t) where $2 \le s \le T$ and $1 \le t < s$ which is equivalent to counting the number of pairs where $1 \le t < s \le T$

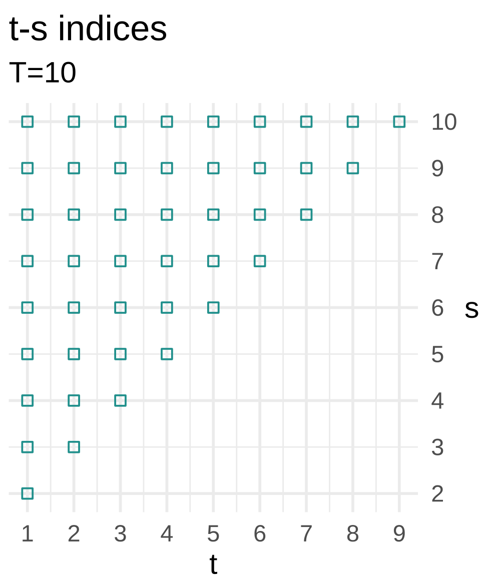
term 4:

The fourth term is $-\frac{1}{T^2}\sum_{k=1}^{T-1}\sum_{t=1}^{T-k} a_t b_{t+k}$ and again we will use the change of variables $s = t + k$.

With the change of variable, and with s now the second index, the inner sum goes from $s = t+1$ to $s = T$ and the outer sum goes from $t = 1$ to $t = T-1$. The (t, s) indices used to sum $-\frac{1}{T^2}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T} a_t b_s$ are shown in figure (B) below (for $T = 10$).

In fact the $a_t b_{t+k}$ terms are the transpose of the terms $a_{t+k} b_t$ of figure (B) as can be seen in figure (C).

(C)



So per terms 3 & 4, each cross-product ($a_i b_j$, $i \neq j$) appears exactly once with a negative sign, for a total of $T^2 - T$ terms (since each of terms 3 & 4 has $\frac{T(T-1)}{2}$ terms.

So we are missing a set of T diagonal terms from the total T^2 terms.

but the sum of the first two terms is $\frac{T-1}{T^2}\sum_{t=1}^{T} a_t b_t$, which provides the missing diagonal (i.e. $-\frac{1}{T^2}\sum_{t=1}^{T} a_t b_t$), so we have:

$$\frac{T-1}{T^2}\sum_{t=1}^{T} a_t b_t - \frac{1}{T^2}\sum_{s=1}^{T}\sum_{t=1,\, t\neq s}^{T} a_t b_s = \tag{9}$$

$$\frac{1}{T}\sum_{t=1}^{T} a_t b_t - \frac{1}{T^2}\sum_{s=1}^{T}\sum_{t=1}^{T} a_t b_s = \tag{10}$$

$$\frac{1}{T}\sum_{t=1}^{T} a_t b_t - \frac{1}{T}\sum_{s=1}^{T} b_s \times \frac{1}{T}\sum_{t=1}^{T} a_t = \tag{11}$$

$$\frac{1}{T}\sum_{t=1}^{T} a_t b_t - \bar{b}\bar{a} \tag{12}$$

compute example for covariance:

```r
set.seed(8740); T = 20
a <- 1 + rnorm(n = T)
b <- a + 4 + rnorm(n = T, sd = 2)

# calculate using difference equation
cov_diff <-
  1:(T-1) |> purrr::map_vec(
    (\(k){ ((dplyr::lead(a,k) - a) * (dplyr::lead(b,k) - b)) |> sum(na.rm=TRUE) })
  ) |> sum(na.rm=TRUE) / (T*T)

# summarize results
tibble::tibble(x = a, y = b) |>
  dplyr::mutate(x = x - mean(x), y = y - mean(y), prod = x*y) |>
  dplyr::summarize(pop_cov = mean(prod) ) |>
  tibble::add_column(cov_diff = cov_diff) |>
  gt::gt() |>
  gt::tab_header(
    title = "Covariance Calculations"
    , subtitle = stringr::str_glue("T={T}")) |>
  gt::cols_width( everything() ~ gt::px(150) ) |>
  gt::cols_label(
    pop_cov = gt::md("using product formula")
    , cov_diff = "using diff formula"
  ) |>
  gt::tab_options(table.width = gt::pct(50), table.align = "center") |>
  gtExtras::gt_theme_espn() #|> gt::as_latex()
```

Covariance Calculations
T=20

using product formula	using diff formula
0.9179426	0.9179426

compute example for covariance:

The difference formula for (population) covariance is

$\frac{1}{T^2}\sum_{k=1}^{T-1}\sum_{t=1}^{T-k}\left(a_{t+k}-a_t\right)\left(b_{t+k}-b_t\right)$

And the same expression holds for variance calculations: the population variance is calculated as the mean of the square of centered values, can also be calculated using all the differences between values.

```r
a_var <- (a-mean(a))^2 |> mean()

# calculate using difference equation
var_diff <-
  1:(T-1) |> purrr::map_vec(
    (\(k){ ((dplyr::lead(a,k) - a)^2) |> sum(na.rm=TRUE) })
  ) |> sum(na.rm=TRUE) / (T*T)

# summarize results
tibble::tibble("using product formula" = a_var, "using difference formula" = var_diff) |>
  gt::gt() |>
  gt::tab_header(
    title = "Variance Calculations"
    , subtitle = stringr::str_glue("with variable a | T={T}")
  ) |>
  gt::tab_options(table.width = gt::pct(50), table.align = "center") |>
  gtExtras::gt_theme_espn()
```

Variance Calculations

with variable a | T=20

using product formula	using difference formula
0.6150733	0.6150733