Let $(x_1, x_2, \ldots, x_n)$ be a sequence of vectors:

$$x_i = \begin{bmatrix} x_i^1 \\ \vdots \\ x_i^m \end{bmatrix}, \; i = 1, \ldots, n.$$

In statistics one often has to compute the *sample mean* vector

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

and the *sample covariance* (or *variance-covariance*) matrix

$$\overline{V} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(x_i - \overline{x})^T$$

where x^T is the transpose of x.

1. What *canonical* form of information would you suggest to represent the sequence $(x_1, x_2, \ldots, x_n)$ in order to compute the sample mean vector and the sample covariance matrix?

Verify that all of the "desirable" properties of canonical information are satisfied.

Solution

Consider the scalar-vector-matrix triple (n, s, T) where

$$n = \sum_{i=1}^{n} 1, \quad s = \sum_{i=1}^{n} x_i, \quad \text{and} \quad T = \sum_{i=1}^{n} x_i x_i^T.$$

(a) **Uniqueness:** Note that (n, s, T) is uniquely determined as each is a function of well-defined vector operations.

  (i) **Elementary** canonical information: A single observation has the representation

$$x \mapsto (1, x, xx^T).$$

  (ii) **Empty** canonical information: Empty information has the representation

$$\{\} \mapsto (0, 0, 0),$$

  where each 0 is the respective scalar, vector, and matrix additive identity.

(b) **Composition** operation: Let $(n, s, T) \oplus (n', s', T') := (n + n', s + s', T + T')$. Commutativity and associativity are inherited directly from the respective additions for scalars, vectors, and matrices. Moreover, the neutral element consists of the triple of additive identities $(0, 0, 0)$.

(c) **Update** observation: Given x_{n+1}, we can update via the following schematic:

$$\begin{tikzcd} (S, T, n) \arrow[r] & \oplus \arrow[r] & (n+1, s + x_{n+1}, T + x_{n+1}{x_{n+1}}^T) \\ x_{n+1} \arrow[ur] & & \end{tikzcd}$$

Note that this is exactly the same operation one obtains by composing with the elementary element.

(d) **Completeness**: Recovering $\overline{x}$ is immediately given by $\overline{x} = \frac{s}{n}$. Some matrix algebra reveals

$$\begin{aligned}
\overline{V} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})^T \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n} x_i x_i^T - \overline{x}\left(\sum_{i=1}^{n} x_i\right)^T - \left(\sum_{i=1}^{n} x_i\right)\overline{x}^T + n\overline{x}\overline{x}^T\right\} \\
&= \frac{1}{n-1}\left\{T - \frac{s}{n}s^T - s\left(\frac{s}{n}\right)^T + n\frac{s}{n}\left(\frac{s}{n}\right)^T\right\} \\
&= \frac{1}{n-1}\left\{T - \frac{1}{n}ss^T\right\}.
\end{aligned}$$

Hence, we can recover each statistic using only the canonical information.

(e) Note that we require $n \geq 1$ to compute $\overline{x}$ and $n \geq 2$ in order to compute $\overline{V}$. Thus the minimum number of observations to compute $(\overline{x}, \overline{V})$ is 2.

□

2. * What *explicit* form of information would you suggest to represent the sequence $(x_1, x_2, \ldots, x_n)$? It should contain $\overline{x}$ and $\overline{V}$ and, perhaps, something else.

Solution

In the spirit of minimizing the number of quantities to keep track of, we can use the explicit variables $(\overline{x}, \overline{V}, n)$ to form an information system.

(a) Uniqueness follows from the fact that both computations are unique with respect to any representation (in particular, permutation of the coordinates). However, we lack both a well-defined **(i) Elementary** and **(ii) Empty** element as both $\overline{x}$ and $\overline{V}$ are undefined when $n = 0$ and $\overline{V}$ is undefined when $n = 1$.

To define the composition operation, let us first denote the map that takes the canonical information to the data in **Problem 1(d) Completeness**

$$\tau(n, s, T) = (n, \overline{x}, \overline{V}).$$

Observe

$$\begin{aligned}
&\overline{V} = \frac{1}{n-1}\left(T - \frac{ss^T}{n}\right) \\
\iff\ & T = (n-1)\overline{V} + \frac{1}{n}(n\overline{x})(n\overline{x})^T = (n-1)\overline{V} + n\overline{x}\overline{x}^T
\end{aligned}$$

Hence, τ is invertible by

$$\tau^{-1}(n, \overline{x}, \overline{V}) = \Big(n,\ n\overline{x},\ (n-1)\overline{V} + n\overline{x}\overline{x}^T\Big).$$

Schematically, we can construct the **(b) Composition** operation

$$
\begin{tikzcd}
& (n,\overline{x},\overline{V}) \arrow[r, "\tau^{-1}"] & (n,s,T) \arrow[dr] & & & \\
\widetilde{\oplus}: & & & \oplus \arrow[r] & (\widetilde{n},\widetilde{s},\widetilde{T}) \arrow[r, "\tau"] & (\widetilde{n},\widetilde{\overline{x}},\widetilde{\overline{V}}) \\
& (n',\overline{x}',\overline{V}') \arrow[r, "\tau^{-1}"] & (n',s',T') \arrow[ur] & & &
\end{tikzcd}
$$

The commutative monoid properties are inherited from $\oplus$ in Problem 1. I.e.

$$
\begin{aligned}
(n,\overline{x},\overline{V})\widetilde{\oplus}(n',\overline{x}',\overline{V}') &= \tau\Big(\tau^{-1}(n,\overline{x},\overline{V}) \oplus \tau^{-1}(n',\overline{x}',\overline{V}')\Big) \\
&= \tau\Big(\tau^{-1}(n',\overline{x}',\overline{V}') \oplus \tau^{-1}(n,\overline{x},\overline{V})\Big) \\
&= (n',\overline{x}',\overline{V}')\widetilde{\oplus}(n,\overline{x},\overline{V})
\end{aligned}
$$

and

$$
\begin{aligned}
\Big((n,\overline{x},\overline{V})\widetilde{\oplus}(n',\overline{x}',\overline{V}')\Big)\widetilde{\oplus}(n'',\overline{x}'',\overline{V}'') &= \tau\Big(\tau^{-1}(n',\overline{x}',\overline{V}') \oplus \tau^{-1}(n,\overline{x},\overline{V})\Big)\widetilde{\oplus}(n'',\overline{x}'',\overline{V}'') \\
&= \tau\bigg(\tau^{-1}\tau\Big(\tau^{-1}(n',\overline{x}',\overline{V}') \oplus \tau^{-1}(n,\overline{x},\overline{V})\Big) \oplus \tau^{-1}(n'',\overline{x}'',\overline{V}'')\bigg) \\
&= \tau\bigg(\tau^{-1}(n',\overline{x}',\overline{V}') \oplus \tau^{-1}\tau\Big(\tau^{-1}(n,\overline{x},\overline{V}) \oplus \tau^{-1}(n'',\overline{x}'',\overline{V}'')\Big)\bigg) \\
&= \tau\bigg(\tau^{-1}(n',\overline{x}',\overline{V}') \oplus \tau^{-1}\Big((n,\overline{x},\overline{V})\widetilde{\oplus}(n'',\overline{x}'',\overline{V}'')\Big)\bigg) \\
&= (n,\overline{x},\overline{V})\widetilde{\oplus}\Big((n',\overline{x}',\overline{V}')\widetilde{\oplus}(n'',\overline{x}'',\overline{V}'')\Big)
\end{aligned}
$$

and

$$
\begin{aligned}
(n,\overline{x},\overline{V})\widetilde{\oplus}(0,0,0) &= \tau(\tau^{-1}(n,\overline{x},\overline{V}) \oplus (0, 0\cdot 0, (0-1)0+0)) \\
&= \tau\tau^{-1}(n,\overline{x},\overline{V}) \\
&= (n,\overline{x},\overline{V}).
\end{aligned}
$$

Explicitly, this results in the expression $(n,\overline{x},\overline{V})\widetilde{\oplus}(n',\overline{x}',\overline{V}') = (\widetilde{n},\widetilde{\overline{x}},\widetilde{\overline{V}})$ where

$$
\widetilde{n} = n + n', \quad \widetilde{x} = s + s' = n\overline{x} + n'\overline{x}',
$$

and

$$
\begin{aligned}
\widetilde{\overline{V}} &= (T+T') + \frac{1}{n+n'}(s+s')(s+s')^T \\
&= \Big((n-1)\overline{V} + n\overline{x}\overline{x}^T + (n'-1)\overline{V}' + n'\overline{x}'\overline{x}'^T\Big) + \frac{1}{n+n'}(n\overline{x}+n'\overline{x}')(n\overline{x}+n'\overline{x}')^T \\
&= \Big((n-1)\overline{V} + n\overline{x}\overline{x}^T + (n'-1)\overline{V}' + n'\overline{x}'\overline{x}'^T\Big) + \frac{1}{n+n'}\big(n\overline{x}\overline{x}^T + n'n(\overline{x}'\overline{x} + \overline{x}\overline{x}'^T) + n'^2\overline{x}'\overline{x}'^T\big).
\end{aligned}
$$

This suggests that if we wish to save on computation time, we could add to the canonical information $W = \overline{x}\overline{x}'$, and the expression above simplifies to

$$\ldots = \Big((n-1)\overline{V} + nW + (n'-1)\overline{V}' + n'W'\Big) + \frac{1}{n+n'}\big(nW + n'n(\overline{x}'\overline{x} + \overline{x}\overline{x}'^T) + n'^2W'\big).$$

The **(c) Update** map is very similar to the one derived for scalar mean and variance:

$$\begin{aligned}
\overline{x}_{n+1} &= \frac{1}{n+1}\sum_{i=1}^{n+1} \\
&= \frac{n}{n+1}\overline{x}_n + \frac{1}{n+1}x_{n+1} \\
&= \overline{x}_n + \frac{1}{n+1}(x_{n+1} - \overline{x}_n),
\end{aligned}$$

and

$$\begin{aligned}
\overline{V}_{n+1} &= \frac{1}{n}\sum_{i=1}^{n+1}(x_i - \overline{x}_i)(x_{-}\overline{x}_i)^T \\
&= \frac{n-1}{n}\overline{V}_n + \frac{1}{n}(x_{n+1} - \overline{x}_{n+1})(x_{n+1} - \overline{x}_{n+1})^T \\
&= \overline{V}_n + \frac{1}{n}\Big((x_{n+1} - \overline{x}_{n+1})(x_{n+1} - \overline{x}_{n+1}) - \overline{V}_n\Big).
\end{aligned}$$

Note that $\overline{V}_{n+1}$ is expressed in terms of $\overline{x}_{n+1}$, so the computation for $\overline{x}_{n+1}$ should precede $\overline{V}_{n+1}$. Also, this suggests that $(0, 0, 0)$ and $(1, x_1, 0)$ could be used for the **(ii) Empty** and **(i) Elementary** elements respectively.

Being an explicit representation, this is clearly **(e) Complete**, and as before, meaningful $\overline{x}$ and $\overline{V}$ are obtained only for $n \geq 2$.

$\square$