Big Data Analytics. Homework #1

Let $(x_1, x_2, \dots, x_n)$ be a sequence of vectors: $x_i = \begin{bmatrix} x_i^1 \\ \vdots \\ x_i^m \end{bmatrix}$, $i = 1, \dots, n$.

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

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

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

$$\bar{V} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})',$$

where x' is the transpose of x.

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

Verify that all the “desirable” properties of canonical information are satisfied:

a) **Uniqueness**: each set of raw data $(x_1, x_2, \dots, x_n)$ can be represented by canonical information in a unique way.
b) **Elementary** canonical information: does canonical information exist for a single observation?
c) **Empty** canonical information: does it exist?
d) **Composition** operation: is it a commutative monoid? (commutativity, associativity, identity).
e) **Update** operation: how is canonical information updated when a new observation x arrives.
f) **Completeness**: canonical information should retain ALL the information which was present in the original raw data. Specifically, an algorithm applied to canonical information should produce the same results as the original algorithm applied to the original raw data.
g) What is the **minimum number of observations** (n) for which $\bar{x}$ and $\bar{V}$ are defined?

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

Analyze all the questions of problem 1 with regard to explicit information.