Linear Estimation Problem in Big Data Systems

Peter Golubtsov

Linear Experiment

As our next example, we will consider a widely used linear estimation process. Consider a linear measurement scheme of the form

$$y = Ax + v,$$

where $x \in \mathcal{D} = \mathbb{R}^m$ is the unknown vector of the m-dimensional Euclidean space – the measurement object, $y \in \mathcal{R} = \mathbb{R}^k$ is the measurement result, $A: \mathcal{D} \to \mathcal{R}$ is the linear mapping, represented by an $k \times m$-matrix, describing the distortions of the measuring system, the vector $v \in \mathcal{R}$ is the random noise vector with zero mean $\mathsf{E}v = 0$ and the given covariance matrix $S = \mathrm{cov}\,(v)$, which can be considered as a linear operator $S: \mathcal{R} \to \mathcal{R}$ – the covariance operator of the random vector $\in \mathcal{R}$.

It is easy to see that the covariance matrix of some random vector is symmetric and non-negative definite. We will consider only the measurements in which the matrix S is positive definite, $S > 0$, and hence is invertible. In essence, this means that the noise v is "possible in all directions", that is, there is no proper subspace $\tilde{\mathcal{R}} \subset \mathcal{R}$ such that $v \in \tilde{\mathcal{R}}$ with probability one.

Thus, the measurement data is represented by the triple (y, A, S) and includes the measurement result y and the measurement model described by the pair (A, S).

Optimal Estimation Problem

The problem of linear estimation of an unknown vector x consists in constructing a linear mapping $R: \mathcal{R} \to \mathcal{D}$ such that the estimate $\hat{x} = Ry$ is maximally close to x. Formally, let us consider the average estimation error

$$\begin{aligned}\mathsf{E}\|Ry - x\|^2 = \mathsf{E}\|R(Ax + v) - x\|^2 = \|(RA - I)x\|^2 + 2\mathsf{E}\langle (RA - I)x, Rv\rangle + \mathsf{E}\|Rv\|^2 \\ = \|(RA - I)x\|^2 + \mathrm{tr}RSR^*.\end{aligned}$$

Since there is an unknown vector x in the expression for $\mathsf{E}\|Ry - x\|^2$, we define the estimation error provided by the operator R as

$$H(R) = \sup_{x \in \mathcal{D}} \mathsf{E}\|Ry - x\|^2.$$

It is easy to see that if $RA \neq I$ then $\|(RA - I)x\|^2$ can take arbitrarily large values and, consequently,

$$H(R) = \begin{cases} +\infty, & \text{if } RA \neq I, \\ \mathrm{tr}RSR^*, & \text{if } RA = I. \end{cases}$$

Thus, the linear mapping R provides a finite estimation error $H(R)$ if and only if $RA = I$. It is easy to see that the last equation is equivalent to the requirement that the estimate $\hat{x} = Ry$ is unbiased, i.e., $\mathsf{E}Ry = x$. Thus, the problem of linear estimation can be regarded as the problem of conditional minimization:

$$\min_{R:\mathcal{R}\to\mathcal{D}} \{\mathrm{tr} RSR^* \mid RA = I\}.$$

It has a solution if and only if $A^*S^{-1}A: \mathcal{D} \to \mathcal{D}$ is nonsingular. In this case, the optimal estimate, known as the best linear unbiased estimate (BLUE), and the corresponding estimation error are given by the expressions:

$$\hat{x} = Ry = (A^*S^{-1}A)^{-1}A^*S^{-1}y,$$

$$\mathsf{E}\|\hat{x} - x\|^2 = \mathrm{tr}(A^*S^{-1}A)^{-1}.$$

Thus, the processing procedure $\boldsymbol{P}$ consists in converting the original data, represented by (y, A, S) into the processing result, the optimal estimate $\hat{x}$ of the vector x:

$$(y, A, S) \xrightarrow{\boldsymbol{P}} \hat{x} = (A^*S^{-1}A)^{-1}A^*S^{-1}y$$

Fig. 1. Optimal linear estimation processing.

Note, that the mapping $\boldsymbol{P}$ is not everywhere defined. It is defined only when the operator $A^*S^{-1}A$ is invertible.

Linear Estimation for Multiple Independent Measurements

Now suppose that there are many independent measurements of the same unknown vector $x \in \mathcal{D}$:

$$y_i = A_i x + v_i, \qquad i = 1, \ldots, n,$$

where $y_i \in \mathcal{R}_i$ are measurement results, $A_i: \mathcal{D} \to \mathcal{R}_i$ are linear mappings, and $v_i \in \mathcal{R}_i$ are independent random vectors with zero means $\mathsf{E}v_i = 0$ and covariance operators $S_i: \mathcal{R}_i \to \mathcal{R}_i$. In general, the measurement spaces $\mathcal{R}_i = \mathbb{R}^{k_i}$ can be different.

To process such n measurements, one have to collect all the pieces of data (y_i, A_i, S_i) in one place, reorganize them in the form of block matrices, possibly very large dimensions, and apply the transformation $\boldsymbol{P}$ to the combined data (Fig. 2).

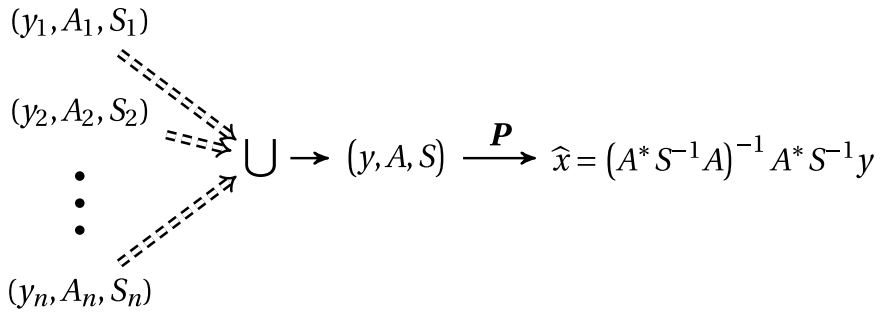

Fig. 2. The standard scheme of linear estimation for a large number of measurements.

Here

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in \mathcal{R}, \qquad A = \begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_n \end{pmatrix}: \mathcal{D} \to \mathcal{R},$$

$$S = \begin{pmatrix} S_1 & 0 & \cdots & 0 \\ 0 & S_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & S_1 \end{pmatrix} : \mathcal{R} \to \mathcal{R},$$

$$\mathcal{R} = \mathcal{R}_1 \times \mathcal{R}_2 \times \cdots \times \mathcal{R}_n, \qquad \dim \mathcal{R} = \sum_{i=1}^{n} \dim \mathcal{R}_i = \sum_{i=1}^{n} k_i.$$

For a large number of measurements, the dimension of the combined data can become extremely large, which makes this approach unfeasible. Besides, the addition of new data would lead to the increase in the dimensions of the merged data, which in turn would require increasing resources for their storage and processing (application of the transformation $\boldsymbol{P}$).

Parallelizing Processing by Extracting the Intermediate Information

Let us show that the data processing in the linear estimation problem can be divided into two phases $\boldsymbol{P} = \boldsymbol{P}_2 \circ \boldsymbol{P}_1$, where the first phase $\boldsymbol{P}_1$ extracts some compact intermediate information from the initial data, and the second $\boldsymbol{P}_2$ calculates the estimation result based on this intermediate information. Moreover, our goal will be to find such a factorization that the application of the transformation $\boldsymbol{P}_1$ to the combined data set can be replaced by the parallel application of $\boldsymbol{P}_1$ to individual data and the subsequent "addition" of the extracted information fragments.

As we have just seen, the vector y and matrices A, and S describing the combined data can become extremely large, which can make the application of the transformation $\boldsymbol{P}$ impossible. However, it can be shown that the main parts of the expression $\hat{x} = (A^* S^{-1} A)^{-1} A^* S^{-1} y$ can be decomposed into pieces:

$$A^* S^{-1} y = A_1^* S_1^{-1} y_1 + \cdots + A_n^* S_n^{-1} y_n,$$

$$A^* S^{-1} A = A_1^* S_1^{-1} A_1 + \cdots + A_n^* S_n^{-1} A_n.$$

This implies that all the information needed for further processing related to the i-th piece of data (y_i, A_i, S_i) can be represented by a pair (v_i, T_i), where

$$v_i = A_i^* S_i^{-1} y_i \in \mathcal{D}, \qquad T_i = A_i^* S_i^{-1} A_i : \mathcal{D} \to \mathcal{D},$$

and T_i is a non-negative definite operator. Obviously, the pair (v, T) in which $v = v_1 + \cdots + v_n$ and $T = T_1 + \cdots + T_n$ will correspond to the combined data (y, A, S).

Canonical Information Space

We will call the pair $(v, T) = (A^* S^{-1} y,\ A^* S^{-1} A)$ the **canonical information** for the data (y, A, S), and the set $\Im$ of all such pairs - canonical **information space** for the problem of linear estimation of a vector from the space $\mathcal{D}$. It can be shown that $\Im$ consist of all the pairs (v, T) in which $v \in \operatorname{im} T$. Thus,

$$\Im = \{(v, T) \mid T \in \mathbb{S}_{\mathcal{D}}^+, v \in \operatorname{im} T\},$$

where $\mathbb{S}_{\mathcal{D}}^+$ is the set of nonnegative definite operators on $\mathcal{D}$ – a convex cone in the linear space $\mathbb{S}_{\mathcal{D}}$ of selfadjoint operators on the space $\mathcal{D}$. If $\dim \mathcal{D} = m$, then $\dim \mathbb{S}_{\mathcal{D}} = \frac{m(m+1)}{2}$. Thus, $\Im \subset \mathcal{D} \times \mathbb{S}_{\mathcal{D}}^+$

is a convex cone in the $\frac{m(m+3)}{2}$-dimensional vector space $\mathcal{D} \times \mathbb{S}_{\mathcal{D}}$. It implies, in particular, that any element of the information space $\Im$ can be represented by $\frac{m(m+3)}{2}$ numbers.

Obviously, the process of linear estimation can be divided into two phases $\boldsymbol{P} = \boldsymbol{P}_2 \circ \boldsymbol{P}_1$, where the first phase $\boldsymbol{P}_1$ consists in constructing the canonical information:

$$(v, T) = \boldsymbol{P}_1(y, A, S) = (A^* S^{-1} y, A^* S^{-1} A),$$

and the second phase $\boldsymbol{P}_2$ calculates the estimation result based on this information (Fig. 3):

$$\hat{x} = \boldsymbol{P}_2(v, T) = T^{-1} v.$$

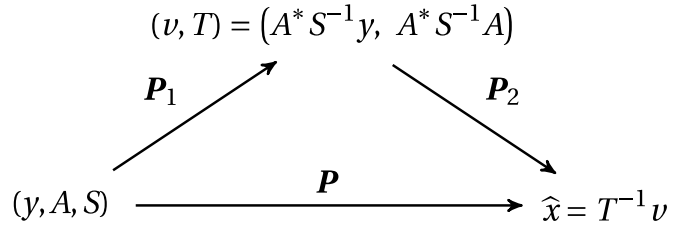

Fig. 3. Splitting data processing into two phases.

As was shown above, the combination of the initial data (y_1, A_1, S_1) and (y_2, A_2, S_2) can be represented by the composition of the corresponding pieces of canonical information (v_1, T_1) and (v_2, T_2), defined as

$$(v_1, T_1) \oplus (v_2, T_2) = (v_1 + v_2, T_1 + T_2).$$

This can be written as

$$\boldsymbol{P}_1(y_1, A_1, S_1) \otimes \boldsymbol{P}_1(y_2, A_2, S_2) = \boldsymbol{P}_1\big((y_1, A_1, S_1) \cup (y_2, A_2, S_2)\big),$$

where $(y_1, A_1, S_1) \cup (y_2, A_2, S_2)$ is combining two data sets into one, Fig. 4.

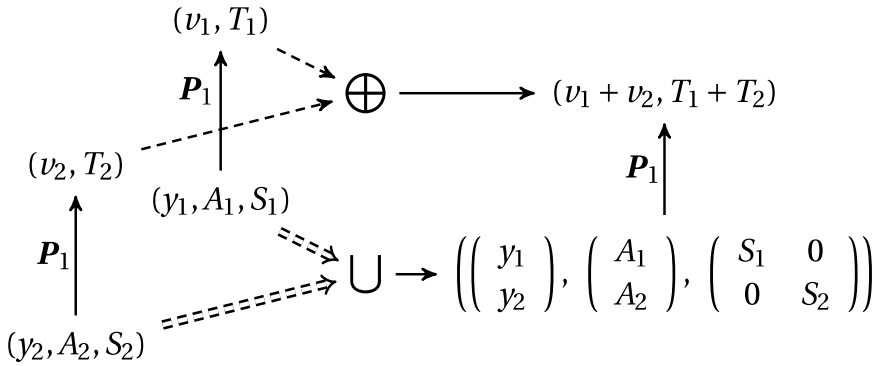

Fig. 4. The correspondence between the composition of fragments of canonical information and combining sets of input data.

Revised Processing Scheme

As a result of the introduction of canonical information and the factorization of algorithm *P* into two phases, the data processing scheme presented in Fig. 2 can be transformed to the one, shown on Fig. 6. From each individual fragment (y_i, A_i, S_i) of the data, the canonical information (v_i, T_i) is extracted, which is subsequently combined and used to calculate the estimation result.

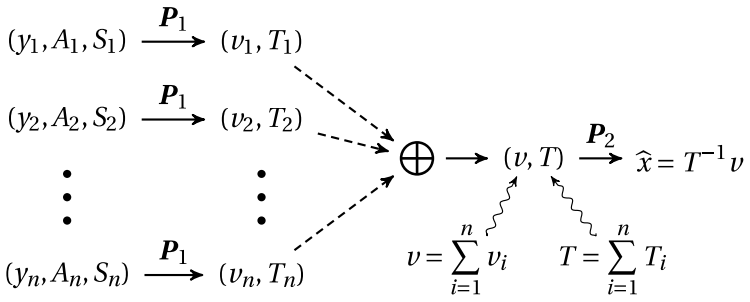

Fig. 6. Modified scheme for processing distributed data.

Let us outline the main features of such a modified scheme. The amount of memory required to store information in the canonical form does not depend on the volume of the represented original data and is $\frac{m(m+3)}{2}$ real numbers (m-dimensional vector and symmetric $m \times m$ matrix). Computing the canonical information (v_i, T_i) from the i-th set of data (transformation $\boldsymbol{P}_1$) can be performed on the computers, where the data is located, in parallel and independently. Only compact fragments of the canonical information of the same volume are transferred. The addition of the parts of canonical information is maximally simplified and is determined by the componentwise addition of the pairs (v_i, T_i). Resources requirements for the second phase P_2, consisting in constructing the result from the compact accumulated information (v, T), are determined only by the dimension m of the space of unknown x and do not depend on the volume of the original data. As a new data becomes available, it would only be necessary to extract from it the canonical information and "add" it to the accumulated information. In this case, the final processing $\boldsymbol{P}_2$ would have to be reapplied to the compact information of a fixed volume.

Quality of Information

In the problem of linear estimation, considered above, our goal was to construct an estimate $\hat{x}$, that is, $\boldsymbol{P}(y, A, S) = \hat{x}$. The corresponding estimation error is $\mathsf{E}\|\hat{x} - x\|^2 = \mathrm{tr}Q$, where $Q = (A^* S^{-1} A)^{-1} = T^{-1}$ represents the covariance matrix of $\hat{x}$, i.e., $Q = \mathrm{cov}(\hat{x})$. Matrix Q also allows to determine the estimation errors for the individual components of the vector x since $\mathsf{E}(\hat{x}_j - x_j)^2 = \mathrm{var}(\hat{x}_j) = Q_{jj}$.

Moreover, the smaller the covariance matrix, the less the estimation error: that is, if $Q \le \tilde{Q}$, then $Q_{jj} \le \tilde{Q}_{jj}$ and $\mathrm{tr}Q \le \mathrm{tr}\tilde{Q}$. It means that if Q and $\tilde{Q}$ are covariance matrices for two estimates of x, then the estimate with the smaller covariance matrix provides better precision in all respects. Define the partial order on the set of symmetric matrices of the same dimension as follows:

$$Q \ge \tilde{Q} \;\Leftrightarrow\; Q - \tilde{Q} \ge 0,$$

that is, Q is greater or equal than $\tilde{Q}$ if $Q - \tilde{Q}$ is nonnegative definite.

We will say that the information (v, T) is not less **accurate** (not worse) than $(\tilde{v}, \tilde{T})$ and write $(v, T) \succcurlyeq (\tilde{v}, \tilde{T})$ if $T \ge \tilde{T}$. If $(v, T) \succcurlyeq (\tilde{v}, \tilde{T})$ and $(\tilde{v}, \tilde{T}) \succcurlyeq (v, T)$, then we say that (v, T) and $(\tilde{v}, \tilde{T})$ have the same accuracy and denote this $(v, T) \approx (\tilde{v}, \tilde{T})$. Obviously, this is equivalent to the condition $T = \tilde{T}$. It is easy to see that more accurate information provides more accurate estimation. Indeed, let $T \ge \tilde{T}$ and the couples (v, T) and $(\tilde{v}, \tilde{T})$ allow to construct the

corresponding estimates, that is, T and $\tilde{T}$ are invertible. This implies that $T^{-1} \leq \tilde{T}^{-1}$ and hence $Q \leq \tilde{Q}$, where Q and $\tilde{Q}$ are the covariance matrices of the corresponding estimates.

Properties of Canonical Information

Let us summarize the properties of information spaces that we observed in the above examples. These properties not only represent an independent interest, but also can serve as an example of the general properties of information spaces that arise in the tasks of processing large volumes of distributed data.

Existence

Any source dataset must allow the presentation of information in the canonical form. Note that the calculation of the final result may not be possible for some data. Strictly speaking, the transformation $\boldsymbol{P}$ can be (and often is) not everywhere defined. At the same time, we require $\boldsymbol{P}_1$ to be defined everywhere.

For instance, as we have seen in the estimation problem, the information contained in the data (y, A, S) may not allow the construction of the estimation result. Namely, if $A^*S^{-1}A$ is singular, then the estimate of the unknown vector cannot be produced. In particular, if dimension of the observation y is less than the dimension of the unknown x, this matrix is singular and the estimate cannot be computed. Nevertheless, the canonical information (v, T) can be constructed even for such data. Moreover, even the complete lack of measurements (carrying zero information) can be represented in canonical form. Formally, any measurement (y, A, S), in which $A = 0: \mathcal{D} \to \mathcal{R}$ is a zero mapping, does not carry any information about the vector being measured. Any such measurement corresponds to the canonical information $\mathbf{0} = (0,0)$, i.e., $v = 0 \in \mathcal{D}$ and $T = 0: \mathcal{D} \to \mathcal{D}$.

Similarly, in the second example, to compute the sample variance one needs at least two values. However, even single-element "atomic" dataset (x) and an empty dataset () can be represented respectively by the elements $(1, x, x^2)$ and $\mathbf{0} = (0,0,0)$ of the corresponding information space.

Sufficiency

Representation in the canonical form should retains all the information contained in the original data, namely, it should lead to the same result as the original data from which it was derived. This property resembles the concept of sufficiency in mathematical statistics. Formally, it means that $\boldsymbol{P}(D) = \boldsymbol{P}_2(\boldsymbol{P}_1(D))$ for any data D from the domain of definition of the transformation $\boldsymbol{P}$.

Composition operation

The canonical information space $\mathfrak{I}$ should be equipped with a composition operation $\oplus$, representing the combination of the corresponding fragments of data, such that the following properties hold for any $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathfrak{I}$:

- $\boldsymbol{a} \oplus \boldsymbol{b} = \boldsymbol{b} \oplus \boldsymbol{a}$. *(Commutativity)* – Changing the order of the pieces of information does not change the result.
- $(\boldsymbol{a} \oplus \boldsymbol{b}) \oplus \boldsymbol{c} = \boldsymbol{a} \oplus (\boldsymbol{b} \oplus \boldsymbol{c})$. *(Associativity)* – Composition of pieces of information does not depend on the order of the pairwise compositions.

- $\boldsymbol{a} \oplus \mathbf{0} = \boldsymbol{a}$. *(Neutral property of zero element)* - Adding zero information to any information does not change it.

It means that $(\mathfrak{I}, \oplus, \mathbf{0})$ is a commutative monoid.

In addition, the monoid $(\mathfrak{I}, \oplus, \mathbf{0})$ also has the *cancellation property*:

- $\boldsymbol{a} \oplus \boldsymbol{b} = \boldsymbol{a} \oplus \boldsymbol{c} \;\Rightarrow\; \boldsymbol{b} = \boldsymbol{c}$,

but does not have invertible elements other than $\mathbf{0}$, i.e. there is no "negative" information.

It is easy to see that these properties are satisfied in all the considered examples.

Preorder relation

For a certain important class of processing problems, it is possible to define the preorder relation $\succcurlyeq$ on the information space, reflecting *accuracy* of information. Such relation naturally appears in "optimal" processing problems, where the processing in a certain sense optimizes the quality of the result (e.g., estimation precision). For example, in the linear estimation problem, such preorder relation is intrinsically related to the estimation precision.

A preorder relation $\succcurlyeq$ is a binary relation, which satisfies the following properties:

- $\boldsymbol{a} \succcurlyeq \boldsymbol{a}$ (*Reflexivity*)
- $\boldsymbol{a} \succcurlyeq \boldsymbol{b}$ & $\boldsymbol{b} \succcurlyeq \boldsymbol{c} \;\Rightarrow\; \boldsymbol{a} \succcurlyeq \boldsymbol{c}$ (*Transitivity*)

It can be easily verified, that the accuracy relation defined above in the context of the linear estimation problem does indeed satisfy these properties.

Besides, the order structure of the information space should be consistent with the algebraic structure:

- $\boldsymbol{a} \succcurlyeq \mathbf{0}$. Any information is more accurate than the lack of information.
- $\boldsymbol{a} \oplus \boldsymbol{b} \succcurlyeq \boldsymbol{a}, \boldsymbol{b}$. The composition of two fragments of information is more precise than each of them individually.
- $\boldsymbol{a} \succcurlyeq \boldsymbol{b}$ & $\boldsymbol{c} \succcurlyeq \boldsymbol{e} \Rightarrow \boldsymbol{a} \oplus \boldsymbol{c} \succcurlyeq \boldsymbol{b} \oplus \boldsymbol{e}$. Composition of a more accurate pieces of information gives a more accurate result.

Uniqueness

For any original data, there should be a unique representation in the information space $\mathfrak{I}$, consistent with the composition operation. In fact, this property means that there is no redundancy in canonical information.

In particular, since the estimation result does not depend on the order of the data in the source set, the canonical information should not depend on the order of the data.

Finally, we will mention two "practical" desirable properties of such special form of representing intermediate information. They are more of a technical nature, related to the implementation of the corresponding algorithms.

Compactness

The information presented in the canonical form should occupy a small (preferably minimal) volume, if possible, independent of the amount of data presented. In the estimation example, the canonical form occupies a fixed volume of $\frac{m(m+3)}{2}$ numbers.

Efficiency

The presentation of the intermediate information in canonical form should ensure the efficient implementation of all stages of data processing. Specifically, in the estimation example:

- *Extracting* canonical information from the original data requires several matrix multiplications for the matrices determined by the individual data fragments. Besides, extraction of canonical information from individual fragments can be performed in parallel.
- *Combining and accumulating* canonical information reduces to the addition of vectors and matrices of fixed dimension and requires insignificant computational resources.
- *Computing the result* based on the accumulated canonical information requires solving a system of linear equations of fixed size $m \times m$ (or inversion of the corresponding matrix.) Even with the constant arrival of new data, updating of the estimate can be carried out only from time to time.