Peter Golubtsov

Theoretical Basics of Big Data Analythics

Homework Assignments

Moscow
Faculty of Physics · Lomonosov Moscow State University
2024

Peter Golubtsov

Theoretical Basics of Big Data Analythics
Homework Assignments

The Homework Assignments book is intended for 5th-year students of the Physics Department of Moscow State University — students of the course “Theoretical Basics of Big Data Analytics”. This course involves completing homework assignments, which, in essence, are theoretical (research) or practical (software) projects. The manual contains conditions and detailed guidelines for completing the main stages of these assignments. The material corresponds to the course of lectures on the special course “Theoretical Foundations of Big Data Analytics”.

Contents

1 Canonical Information for Sample Mean and Covariance Matrix

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

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

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

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

and the **sample covariance** matrix

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

where x^T is the transpose of x.

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 the "desirable" properties of canonical information are satisfied:

a) **Existence and Uniqueness**. Can any sequence of raw data $(x_1, x_2, \ldots, x_n)$ be represented by canonical information in a unique way? Does this representation depend on the order of vectors x_i in the data sequence?

b) **Completeness**. Canonical information should retain ALL the information which was present in the original raw data. Specifically, an algorithm applied to canonical information (deployment phase) should produce the same results as the original algorithm applied to the original raw data.

Does your canonical representation of data conform this requirement? How would you compute X and V using only collected canonical information?

c) **Elementary** canonical information. Does canonical information exist for a single observation?

d) **Empty** canonical information. Does canonical information exist for an empty sequence of observations?

e) **Combination** (or composition) operation. How would you define composition of pieces of canonical information? Does it satisfy axioms for a commutative monoid? (commutativity, associativity, neutral element)

f) **Update** operation. How is canonical information updated when a new observation vector x arrives?

g) **Compactness and Efficiency.** What can you say about compactness (or minimality) and efficiency of your canonical form of information in terms of storage requirements and complexity of combination, update, and deployment operations?

h) What is the **minimum number** of observations n for which X and V are defined?

2 MapReduce Distributed Computing Model

Write a program simulating the MapReduce distributed computing model for constructing a sample mean and sample covariance matrix from homework 1.

Specifically, let $(x_1, x_2, \ldots, x_n)$ be a sequence of column vectors:

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

Write a MapReduce-style program which computes the **sample mean** vector

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

and the **sample covariance** matrix

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

where x^T is the transpose of the column vector x.

The key part is to design three functions:

a) The function which extracts canonical information from a dataset. This function will be applied to the list of datasets in the Map phase.

b) The function which combines two pieces of canonical information into one. It will be used to combine all the pieces of canonical information into one in the Reduce phase.

c) The function which comutes the result from combined cananical information.

Suggestions:

- Use the canonical information representation designed in the homework 1.
- As a guiding example, you can use the code “Mean_MR.py”, which illustrates the calculation of the arithmetic mean for a set of numbers.
- To generate a set of random columns with a given average and covariance matrix, you can use the code from “MV_ArGen.py”.

3 Linear Regression

Write a program which illustrates simple linear regression (or a more general variant of linear regression) and implements accumulation of canonical information.

a) For some fixed parameters a and b (or, in a more general case, $a_1, \ldots, a_m$) generate a sequence of “observations” (x_i, y_i):

$$y_i = f(x_i) + \varepsilon_i,$$

where

$$f(x) = a + bx \quad \text{or} \quad f(x) = a_1 + a_2 x + a_3 x^2 + \cdots + a_m x^{m-1}$$

ε_i are i.i.d. with zero mean and $\mathsf{E}\varepsilon_i^2 = \sigma^2$. Values x_i can be generated randomly with some mean and variance.

b) Accumulate canonical information, i.e., at each step, when a new observation (x_i, y_i) is produced, update canonical information.

c) Illustrate the real function $f(x)$ and its estimate $\widehat{f(x)}$.

d) Illustrate $\mathrm{Var}(\widehat{f(x)})$, assuming that σ^2 is known.

e) Illustrate $\widehat{\mathrm{Var}(\widehat{f(x)})}$, assuming that σ^2 is NOT known.

In your report present the source code and a few (around 3) nice graphs showing estimations for “small”, “intermediate”, and “large” number of observations.

Formulas & example

$$y_i = f_a(x_i) + \varepsilon_i = a_1 f_1(x_i) + \cdots + a_m f_m(x_i) + \varepsilon_i$$

or

$$y_i = F_{x_i} a + \varepsilon_i,$$

where

$$F_x = \begin{bmatrix} f_1(x) & f_2(x) & \cdots & f_m(x) \end{bmatrix}.$$

Function used in the demo — polynomial:

$$y_i = 1 + 1 \cdot x_i - 1 \cdot x_i^2 + 0.2 \cdot x_i^3 + \varepsilon_i$$

,

$$F_x = \begin{bmatrix} 1 & x & x^2 & x^3 \end{bmatrix}, \qquad m = 4, \qquad a = \begin{bmatrix} 1 \\ 1 \\ -1 \\ 0.2 \end{bmatrix}$$

Data: $(x_i, y_i), \qquad i = 1, \ldots, n$

Canonical information: (T, v, V, n)

Elementary information: (T_i, v_i, V_i, n_i)

$$n_i = 1, \qquad V_i = y_i^2, \qquad v_i = F_{x_i}^T \cdot y_i = \begin{bmatrix} f_1(x_i)\, y_i \\ \vdots \\ f_4(x_i)\, y_i \end{bmatrix},$$

$$T_i = F_{x_i}^T \cdot F_{x_i} = \begin{bmatrix} f_1(x_i)^2 & f_1(x_i)\, f_2(x_i) & \cdots & f_1(x_i)\, f_4(x_i) \\ \vdots & \vdots & \ddots & \vdots \\ f_4(x_i)\, f_1(x_i) & f_4(x_i)\, f_2(x_i) & \cdots & f_4(x_i)^2 \end{bmatrix}$$

Update:

$$(T, v, V, n) + (T_i, v_i, V_i, n_i) = (T + T_i, v + v_i, V + V_i, n + n_i)$$

Estimate $f(x)$:

$$(T, v, V, n) * x \mapsto$$

$$\widehat{f(x)} = F_x T^{-1} v,$$

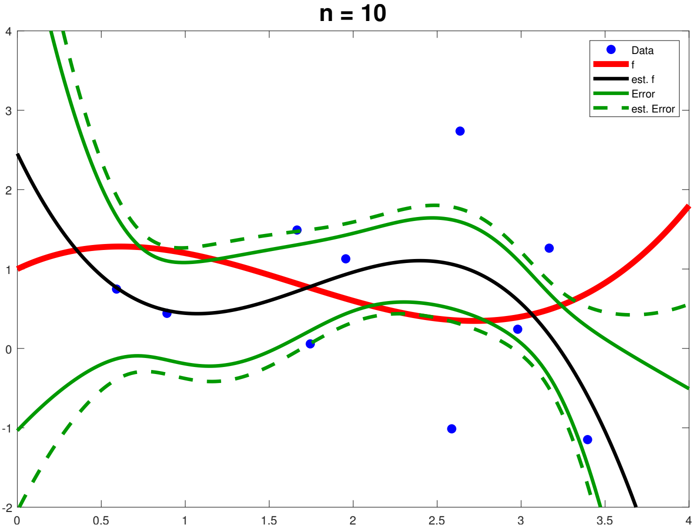
$$\mathrm{Var}(\widehat{f(x)}) = \sigma^2 F_x T^{-1} F_x^T,$$

$$\widehat{\mathrm{Var}(\widehat{f(x)})} = \frac{V - v^T T^{-1} v}{n - m} \cdot F_x T^{-1} F_x^T.$$

Part of the code in MatLab

```
in = in + Info([x,y]);    % Update: Elem. Info & Combine
est = in * xv;            % Apply Info
```

Example of an illustration



4 Simple Linear Estimation Problem and Canonical Information

Consider the following series of measurements of the unknown value x:

$$y_i = a_i x + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where y_i are measurement results, a_i are known coefficients, and ε_i represent random error of measurement and are independent identically distributed (i.i.d.) with zero mean and variance σ^2:

$$\mathsf{E}\varepsilon_i = 0, \qquad \mathsf{E}\varepsilon_i^2 = \sigma^2, \qquad i = 1, \ldots, n,$$

a) Construct an optimal, linear in $y_1, \ldots, y_n$, estimate $\widehat{x}$ for x? $\widehat{x} = ?$

b) Is it a biased or an unbiased estimate?

c) What is its variance (expressed through σ^2)? $\mathrm{Var}(\widehat{x}) = ?$

d) How would you estimate σ^2 if it is unknown? $\widehat{\sigma^2} = ?$

e) What would you use as an estimate for $\mathrm{Var}(\widehat{x})$ if σ^2 is unknown? $\widehat{\mathrm{Var}(\widehat{x})} = ?$

f) Suppose that the variance σ^2 is known. What "canonical information" would be sufficient to extract from the series of records

$$(y_1, a_1), \ldots, (y_n, a_n), \qquad i = 1, \ldots, n$$

in order to compute the estimate $\widehat{x}$, and its variance $\mathrm{Var}(\widehat{x})$?

g) Suppose that the variance σ^2 is NOT known. What "canonical information" would be sufficient to extract from the series of observations in order to compute $\widehat{x}$, $\widehat{\sigma^2}$, and $\widehat{\mathrm{Var}(\widehat{x})}$?

h) How should we update such “information” when a new record (y_{n+1}, a_{n+1}) arrives?

i) How should we “combine” (merge) two pieces of “canonical information”?

Please do not try to use general formulas, but develop as much as possible from scratch.

5 Optimal Linear Estimation Examples

1. (This problem is a particular case of Problem 3. So, if you feel confident you can skip it and then just extract answers from Problem 3).

Consider the following set of measurements of the unknown variables x_1 and x_2:

$$y_1 = x_1 + x_2 + \nu_1,$$

$$y_2 = x_1 - x_2 + \nu_2,$$

$$y_3 = -x_1 + x_2 + \nu_3,$$

where y_i are measurement results, and ν_i represent random error of measurement and are independent identically distributed (i.i.d.) with zero mean and variance σ^2:

$$\mathsf{E}\nu_i = 0, \qquad \mathsf{E}\varepsilon_i^2 = \sigma^2, \qquad i = 1, 2, 3.$$

(a) Write it in matrix form

$$y = Ax + \nu$$

and write the matrices A and $S = \mathrm{Var}(\nu)$.

(b) Find the variance matrix $\mathrm{Var}(\widehat{x})$ for the optimal linear estimate of x and variances of $\widehat{x_1}$ and $\widehat{x_2}$.

2. Consider two measurements of one unknown variable x with the correlated noise. Specifically, suppose that

$$y_1 = x + \nu_1,$$

$$y_2 = x + \nu_2,$$

where

$$\nu_1 = \varepsilon_1 + \varepsilon_0,$$

$$\nu_2 = \varepsilon_2 + \varepsilon_0,$$

$$\varepsilon_1, \varepsilon_2 \sim (0, \sigma_1^2), \qquad \varepsilon_0 \sim (0, \sigma_0^2),$$

$$\sigma_0^2 + \sigma_1^2 = \sigma^2, \qquad r = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_1^2}.$$

(a) Same as in 1.

(b) Same as in 1.

(c) Analyze how the variance of $\widehat{x}$ depends on the correlation parameter r for $0 \le r \le 1$. Is higher correlation good or bad for estimation in this example? A graph might be helpful. How would you explain such behavior?

3. Consider the same measurement scheme as in Problem 1, but with the correlated noise. Specifically, suppose that

$$\nu_1 = \varepsilon_1 + \varepsilon_0,$$

$$\nu_2 = \varepsilon_2 + \varepsilon_0,$$

$$\nu_3 = \varepsilon_3 + \varepsilon_0,$$

$$\varepsilon_1, \varepsilon_2, \varepsilon_3 \sim (0, \sigma_1^2), \qquad \varepsilon_0 \sim (0, \sigma_0^2),$$

$$\sigma_0^2 + \sigma_1^2 = \sigma^2, \qquad r = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_1^2}.$$

(a) Same as in 1.

(b) Same as in 1.

(c) Analyze how the variancees of $\widehat{x_1}$ and $\widehat{x_2}$ depend on the correlation parameter r for $0 \leq r \leq 1$ when $\sigma^2 = \text{const.}$ Is higher correlation good or bad for estimation in this example? A graph might be helpful. How would you explain such behavior?

Feel free to use symbolic packages (e.g., Maple, Symbolic Toolbox in MatLab,...).

6 Canonical Information from Multiple Observations & Prior Information

In this homework you are asked to write a program which would implement and demonstrate various aspects of optimal linear estimation. The underlying process imitates a simple signal measurement experiment.

Items 1 and 2 below describe in more details certain suggestions for your program. The problem itself (in fact, four closely related ones) is formulated in item 3.

1. Measurement Simulation (see sample code).

   (a) Choose some profile x or generate it randomly:

      i. Generate random "white noise" $\mu \sim (0, I)$, i.e., μ_i are i.i.d. with $\mathsf{E}\mu_i = 0$ and $\mathrm{Var}(\mu)_i = 1$.

      ii. "Smooth" it with some matrix B.

      iii. Then $x = B\mu \sim (0, F)$, where $F = BB^T$. Use this x (one and the same) in all your numeric simulations.

   (b) Create matrix A (see sample code).

   (c) Simulate a measurement $y = Ax + \nu$.

2. Estimation: Construct an optimal linear estimate $\widehat{x}$ and the variance matrix $Q = \mathrm{Var}(\widehat{x} - x)$. Show on the same graph:

   (a) The original signal x (a curve with components x_i),

   (b) Its estimate $\widehat{x}$ (a curve with components $\widehat{x}_i$),

   (c) Standard deviations for the estimates $\widehat{x}_i$ ($= \sqrt{\mathrm{Var}(\widehat{x}_i - x_i)} = \sqrt{Q_{ii}}$). It can be illustrated by showing the corresponding "corridor" around $\widehat{x}_i$).

3. Illustrate estimation (see item 2) in different settings:

(a) Single measurement (y, A, S).

i. Transform (y, A, S) to canonical form (T, v).

ii. Construct the estimate, based on the canonical information.

(b) Single measurement (y, A, S) with the prior information $x \sim (0, F)$:

i. Transform the prior information to canonical form.

ii. Transform the measurement to canonical form.

iii. Combine the pieces of canonical information.

iv. Construct the estimate, based on the combined canonical information.

(c) Many measurements (y_j, A_j, S_j), no the prior information.

i. Simulate a sequence of measurements of the same signal x, but with differing matrices A_j (and, possibly, S_j).

ii. Extract canonical information from each measurement.

iii. Combine pieces of canonical information.

iv. Construct the estimate, based on the combined canonical information.

(d) Many measurements (y_j, A_j, S_j), now with the prior information $x \sim (0, F)$. Same steps as in item (c).

i. Extract canonical information from the prior information.

ii. Simulate a sequence of measurements of the same signal x, but with differing matrices A_j (and, possibly, S_j).

iii. Extract canonical information from each measurement.

iv. Combine the pieces of canonical information.

v. Construct the estimate, based on the combined canonical information.

7 Calibration Problem

In this assignment you will implement and demonstrate optimal calibration for a linear estimation. The underlying process imitates a simple signal measurement experiment and is heavily based on Homework 6. Here are suggested phases of the project.

1. Measurement Simulation:

(a) Randomly generate some "unknown" profile $x \sim (0, F)$.

(b) Create a matrix A.

(c) Simulate a measurement

$$y = Ax + \nu, \quad \nu \sim (0, \sigma^2 I).$$

2. Calibration (assuming that A is unknown):

(a) Randomly generate calibration signals $\varphi_k,\ k = 1, \ldots, K$ in the same way as you generated x.

(b) Simulate calibration measurements

$$\psi_k = A\varphi_k + \nu_k.$$

(c) Collect canonical calibration information (G, H).

(d) Compute A_0 (an estimate of A) and J.

3. Using your simulated observation y construct an optimal linear estimate $\widehat{x}$ and the variance matrix $\mathrm{Var}(\widehat{x} - x)$. Show on the same graph:

(a) The original signal x (a curve with components x_i),

(b) Its estimate $\widehat{x}$ (a curve with components $\widehat{x}_i$),

(c) Standard deviations for the estimates $\widehat{x}_i$

$$\sqrt{\mathsf{E}(\widehat{x}_i - x_i)^2} = \sqrt{\mathrm{Var}(\widehat{x} - x)_{ii}} = \sqrt{Q_{ii}}$$

can be illustrated by showing the corresponding "corridor" around $\widehat{x}_i$).

4. Illustrate estimation (Phase 3) when you not only increase the number of calibration measurements K but also measure the original signal x N times:

(a) Simulate N measurements $y_n = Ax + \nu_n$ for $n = 1, \ldots, N$ and collect the appropriate information.

(b) Simulate K calibration measurements, collect canonical calibration information, and compute A_0 and J.

(c) Using these two types of information construct and show (as in Phase 3) an optimal linear estimate $\widehat{x}$ and its precision.

(d) Using these two types of information construct and show (as in Phase 3) an optimal linear estimate $\widehat{x}$ and its precision.

(e) Do the above for several cases with numbers N and K "small" "medium", and "large". Reminder: N and K are "balanced" when $K \approx N \cdot M$, where M is the dimension of the unknown x.

(f) (Optional) Show how estimation precision depends on N and K. To do that you could show total estimation error

$$\mathsf{E}||\widehat{x} - x||^2 = \mathrm{tr}Q$$

as a function of N and K. To illustrate a function of two variables you can show it, e.g., as a surface or as a pseudocolor image. It might be interesting to indicate contour lines (curves along which the function has constant values).

8 Multicriteria Optimization: Trade-off between Systematic and Random Errors

In this homework you are asked to write a program which would implement and demonstrate separate treatment of systematic and random errors in linear estimation. The underlying process imitates a simple signal measurement experiment and is heavily based on Homework 6. Here are suggested phases of the project.

1. Constuct Pareto sets (curves) for different information scenarios:

   (a) Very little information (small and even singular T). Single measurement with a short (truncated) matrix A (the number of observations is less than the number of unknowns).

   (b) Moderate information. Around 4 measurements with normal A.

   (c) More information. Around 16 measurements with normal A.

   and show them on the same graph.

2. Illustrate estimation results for several (6) interesting cases:

   (a) No random noise suppression ($\lambda \approx 0$).

   (b) Moderate or strong noise suppression.

   for different information scenarios.

3. For each estimation illustration show on the same graph:

(a) The original signal x (a curve with components x_i),

(b) Its estimate $\widehat{x}$ (a curve with components $\widehat{x}_i$),

(c) Standard deviations for the random (additive) error $\sqrt{H_{ii}}$. It can be illustrated by showing the corresponding "corridor" around $\widehat{x}_i$).

(d) Illustration of systematic (multiplicative) error $\sqrt{G_{ii}}$. Perhaps on a separate graph.

Instead of the original parametrization by $\lambda \in [0, +\infty]$ it might be convenient to use $\lambda = \frac{t}{1-t}$ where $t \in [0, 1]$.

9 The Smallest Information Space

Let $\{x_1, ..., x_n\} \in \mathbb{R}^+$ be a multiset of reals. For the processing $p : \mathbb{R}^+ \to \mathbb{R} \cup \{\frownie\}$ which computes the spread of points:

$$p(\{x_1, ..., x_n\}) = \begin{cases} \max\limits_{0 \le i \le n} x_i - \min\limits_{0 \le i \le n} x_i & n > 0 \\ \frownie & n = 0 \end{cases}$$

find a candidate for the smallest information space and prove that it really is the smallest.