Random Matrix Improved Covariance Estimation for a Large Class of Metrics

Malik TIOMOKO, Florent BOUCHARD,
Guillaume GINOLHAC and Romain COUILLET

GSTATS IDEX DataScience Chair, GIPSA-lab, University Grenoble–Alpes, France.
Laboratoire des Signaux et Systèmes (L2S), University Paris-Sud.
LISTIC, University Savoie Mont-Blanc, France

June 10, 2019

Context

Observations:

- $X = [x_1, \ldots, x_n]$, $x_i \in \mathbb{R}^p$ with $\mathbb{E}[x_i] = 0$, $\mathbb{E}[x_i x_i^\mathsf{T}] = C$.

Context

Observations:

- $X = [x_1, \ldots, x_n]$, $x_i \in \mathbb{R}^p$ with $\mathbb{E}[x_i] = 0$, $\mathbb{E}[x_i x_i^\mathsf{T}] = C$.

Objective:

- From the data x_i, estimate C.

Context

Observations:

- $X = [x_1, \ldots, x_n]$, $x_i \in \mathbb{R}^p$ with $\mathbb{E}[x_i] = 0$, $\mathbb{E}[x_i x_i^\mathsf{T}] = C$.

Objective:

- From the data x_i, estimate C.

State of the Art:

- Sample Covariance Matrix (SCM):

$$\hat{C} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\mathsf{T} = \frac{1}{n} X X^\mathsf{T}.$$

Context

Observations:

- $X = [x_1, \ldots, x_n]$, $x_i \in \mathbb{R}^p$ with $\mathbb{E}[x_i] = 0$, $\mathbb{E}[x_i x_i^\mathsf{T}] = C$.

Objective:

- From the data x_i, estimate C.

State of the Art:

- Sample Covariance Matrix (SCM):

$$\hat{C} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\mathsf{T} = \frac{1}{n} X X^\mathsf{T}.$$

$\longrightarrow$ **Often justified by Law of Large Numbers**: $n \to \infty$.

Context

Observations:

- $X = [x_1, \ldots, x_n]$, $x_i \in \mathbb{R}^p$ with $\mathbb{E}[x_i] = 0$, $\mathbb{E}[x_i x_i^\mathsf{T}] = C$.

Objective:

- From the data x_i, estimate C.

State of the Art:

- Sample Covariance Matrix (SCM):

$$\hat{C} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\mathsf{T} = \frac{1}{n} X X^\mathsf{T}.$$

$\longrightarrow$ **Often justified by Law of Large Numbers**: $n \to \infty$.

- Numerical inversion of asymptotic spectrum (QuEST).
  1. Bai-Silverstein equation: Estimate $\lambda(\hat{C})$ from $\lambda(C)$ in "large p, n" regime.
  2. Need for non trivial inversion of the equation.

Key Idea

- Elementary idea
$$C \equiv \operatorname{argmin}_{M \succ 0} \delta(M, C)$$
where $\delta(M, C)$ can be the Fisher, Bhattacharyya, KL, Rényi divergence.

Key Idea

- Elementary idea

$$C \equiv \operatorname{argmin}_{M \succ 0} \delta(M, C)$$

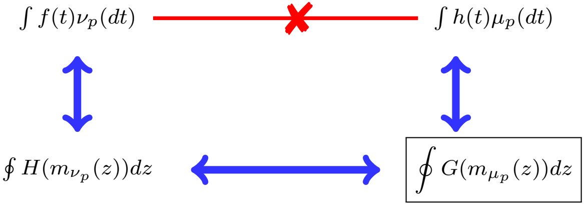
  where $\delta(M,C)$ can be the Fisher, Bhattacharyya, KL, Rényi divergence.
- Divergence $\delta(M,C) = \int f(t) d\nu_p(t)$ inaccessible, $\nu_p \equiv \frac{1}{p} \sum_{i=1}^{p} \delta_{\lambda_i(M^{-1}C)}$.

Key Idea

- Elementary idea

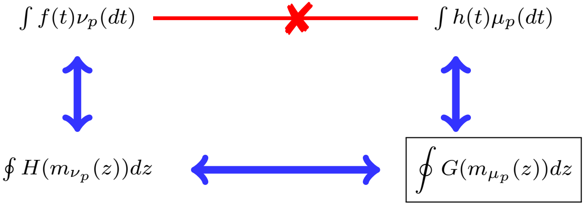
$$C \equiv \operatorname{argmin}_{M \succ 0} \delta(M, C)$$

  where $\delta(M, C)$ can be the Fisher, Bhattacharyya, KL, Rényi divergence.
- Divergence $\delta(M, C) = \int f(t) d\nu_p(t)$ inaccessible, $\nu_p \equiv \frac{1}{p} \sum_{i=1}^{p} \delta_{\lambda_i(M^{-1}C)}$.
- Random Matrix improved estimate $\boxed{\hat{\delta}(M, X)}$ of $\delta(M, C)$ using $\mu_p \equiv \frac{1}{p} \sum_{i=1}^{p} \delta_{\lambda_i(M^{-1}\hat{C})}$.



Key Idea

- Elementary idea
$$C \equiv \operatorname{argmin}_{M \succ 0} \delta(M, C)$$
where $\delta(M, C)$ can be the Fisher, Bhattacharyya, KL, Rényi divergence.
- Divergence $\delta(M, C) = \int f(t) d\nu_p(t)$ inaccessible, $\nu_p \equiv \frac{1}{p} \sum_{i=1}^{p} \delta_{\lambda_i(M^{-1}C)}$.
- Random Matrix improved estimate $\boxed{\hat{\delta}(M, X)}$ of $\delta(M, C)$ using $\mu_p \equiv \frac{1}{p} \sum_{i=1}^{p} \delta_{\lambda_i(M^{-1}\hat{C})}$.



- $\hat{\delta}(M, X) < 0$ with non zero probability.

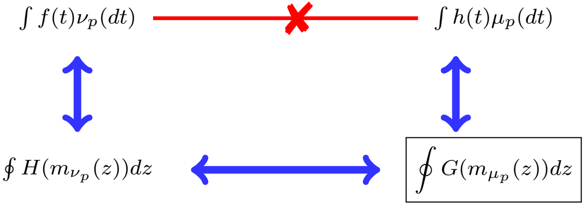
Key Idea

- Elementary idea

$$C \equiv \operatorname{argmin}_{M \succ 0} \delta(M, C)$$

where $\delta(M, C)$ can be the Fisher, Bhattacharyya, KL, Rényi divergence.

- Divergence $\delta(M, C) = \int f(t) d\nu_p(t)$ inaccessible, $\nu_p \equiv \frac{1}{p} \sum_{i=1}^{p} \delta_{\lambda_i(M^{-1}C)}$.
- Random Matrix improved estimate $\boxed{\hat{\delta}(M, X)}$ of $\delta(M, C)$ using $\mu_p \equiv \frac{1}{p} \sum_{i=1}^{p} \delta_{\lambda_i(M^{-1}\hat{C})}$.



- $\hat{\delta}(M, X) < 0$ with non zero probability.
- Proposed estimation

$$\boxed{\check{C} \equiv \operatorname{argmin}_{M \succ 0} h(M), \quad h(M) = \hat{\delta}(M, X)^2}$$

- Gradient descent over the Positive Definite manifold.

Algorithm 1 Proposed estimation algorithm.

Require $M_0 \in C_n^{++}$.

Repeat $M \leftarrow M^{\frac{1}{2}} \exp\left(-tM^{-\frac{1}{2}} \nabla h_X(M) M^{-\frac{1}{2}}\right) M^{\frac{1}{2}}$.

Until Convergence.

Return $\check{C} = M$.

Experiments

- 2 Data classes $x_1^{(1)}, \ldots, x_{n_1}^{(1)} \sim N(\mu_1, C_1)$ and $x_1^{(2)}, \ldots, x_{n_2}^{(2)} \sim N(\mu_2, C_2)$.
- Classify point x using Linear Discriminant Analysis based on the sign of

$$\delta_x^{\mathrm{LDA}} = (\hat{\mu}_1 - \hat{\mu}_2)^\mathsf{T} \check{C}^{-1} x + \frac{1}{2} \hat{\mu}_2^\mathsf{T} \check{C}^{-1} \hat{\mu}_2 - \frac{1}{2} \hat{\mu}_1^\mathsf{T} \check{C}^{-1} \hat{\mu}_1.$$

- Estimate $\check{C} \equiv \frac{n_1}{n_1+n_2} \check{C}_1 + \frac{n_2}{n_1+n_2} \check{C}_2$.

Experiments

- 2 Data classes $x_1^{(1)}, \ldots, x_{n_1}^{(1)} \sim N(\mu_1, C_1)$ and $x_1^{(2)}, \ldots, x_{n_2}^{(2)} \sim N(\mu_2, C_2)$.
- Classify point x using Linear Discriminant Analysis based on the sign of

$$\delta_x^{\mathrm{LDA}} = (\hat{\mu}_1 - \hat{\mu}_2)^\mathsf{T} \check{C}^{-1} x + \frac{1}{2}\hat{\mu}_2^\mathsf{T} \check{C}^{-1} \hat{\mu}_2 - \frac{1}{2}\hat{\mu}_1^\mathsf{T} \check{C}^{-1} \hat{\mu}_1.$$

- Estimate $\check{C} \equiv \frac{n_1}{n_1+n_2}\check{C}_1 + \frac{n_2}{n_1+n_2}\check{C}_2$.

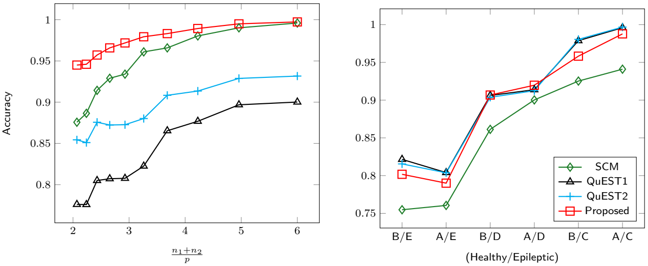


Figure: Mean accuracy obtained over 10 realizations of LDA classification. (Left) C_1 and C_2 Toeplitz-0.2/Toeplitz-0.4, and (Right) real EEG data.