Gaussian Mean Testing under Truncation

Clément L. Canonne*, Themis Gouleakis, Yuhao Wang, Joy Qiping Yang*

*University of Sydney, Nanyang Technological University, National University of Singapore

Introduction

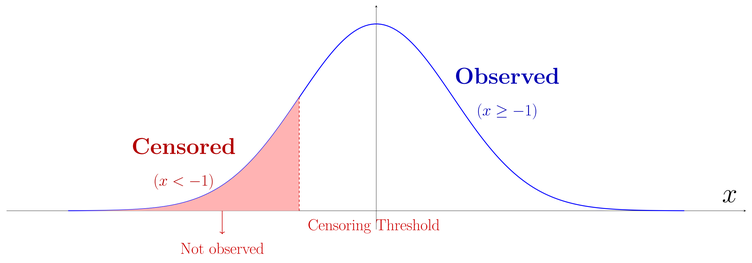
Problem: Learning a high-dimensional Gaussian distribution when data is **missing not at random (MNAR)**.

- **Self-Censoring:** Each coordinate is missing depending on its own.
- **Linear Thresholding:** Missingness depends linearly on the full data vector.

Self-Censoring Model

Model:

- Observations are censored versions of samples from $\mathcal{N}(\mu^*, \Sigma^*)$.
- Coordinate i is observed if $y_i \in S_i$.



Main Theorem (Recovery under Self-Censoring)

- If any two coordinates are observed together with probability at least α.
- Algorithm learns μ^* and Σ^* up to TV distance ε.
- Sample complexity: $\widetilde{O}\left(d^2(\lambda_{\max}/\lambda_{\min})^2/\alpha\varepsilon^2\right)$.

Algorithm: `Truncation_PSGD`

Algorithm 1: Recovery of Gaussian parameters under self-censoring via truncated PSGD.

Input: Data $\mathbf{x} \in \mathbb{R}^{n\times d}$, where $n = \frac{1}{\alpha\varepsilon'^2}$

for $i \leftarrow 1$ ***to*** d **do**
  $\hat{\mu}_i,\ \hat{\Sigma}_{ii} \leftarrow$ `Uni_SGD_trunc`$(\mathbf{x}_i, S_i)$
for $i \leftarrow 1$ ***to*** $d-1$ **do**
  for $j \leftarrow i+1$ ***to*** d **do**
    $\hat{\Sigma}^{ij}_{12} \leftarrow$ `Biv_SGD_trunc`$(\mathbf{x}_i, \mathbf{x}_j, S_i \times S_j)$
$\hat{\boldsymbol{\mu}} \leftarrow [\hat{\mu}_1, \ldots, \hat{\mu}_d]$
for $i \leftarrow 1$ ***to*** $d-1$ **do**
  for $j \leftarrow i+1$ ***to*** d **do**
    $\hat{\Sigma}_{ij},\ \hat{\Sigma}_{ji} \leftarrow \hat{\Sigma}^{ij}_{12}$
return $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$

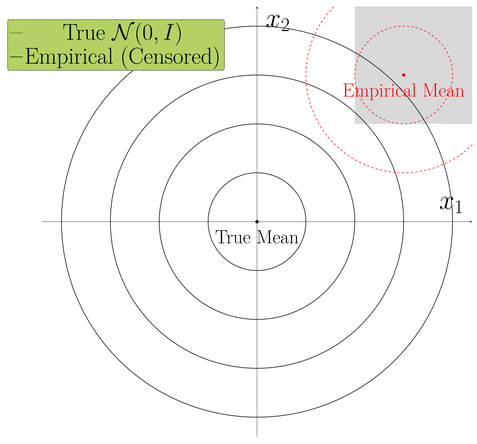
Folklore: Learning a Gaussian Without Missingness

Given n i.i.d. samples $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)} \sim \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$:

- Estimate the mean: $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^n \mathbf{y}^{(i)}$
- Estimate the covariance: $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n (\mathbf{y}^{(i)} - \hat{\boldsymbol{\mu}})(\mathbf{y}^{(i)} - \hat{\boldsymbol{\mu}})^\top$
- With $n = \tilde{\mathcal{O}}\left(d^2/\varepsilon^2\right)$ samples, we recover:

$$\|\boldsymbol{\mu}^* - \hat{\boldsymbol{\mu}}\| = \mathcal{O}(\varepsilon), \quad \left\|\boldsymbol{\Sigma}^* - \hat{\boldsymbol{\Sigma}}\right\|_F = \mathcal{O}(\varepsilon)$$

Censoring



Linear Thresholding Model

Model: Each coordinate is observed if a random linear inequality is satisfied. For coordinate $i \in [d]$, we observe y_i only if:

$$v_i^\top \mathbf{y} \leq b_i$$

where $v_i \in \mathbb{R}^d$ and $b_i \in \mathbb{R}$ are (possibly random) threshold parameters.

Goal: Recover the true mean $\boldsymbol{\mu}^*$ given covariance matrix $\boldsymbol{\Sigma}^*$ of a Gaussian $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ from censored samples.

Assumptions:

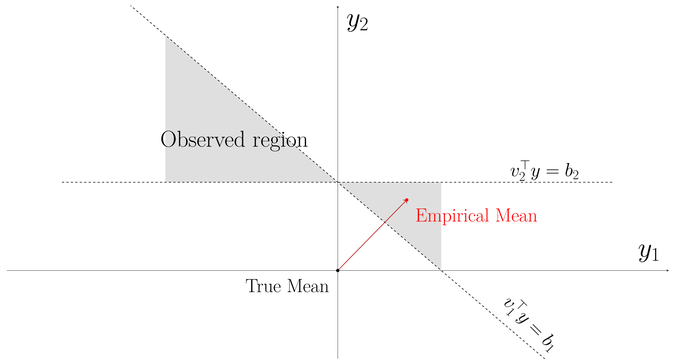
- Small subsets (up to βd coordinates) are observed together with probability at least α.
- A stable anchoring subset C is always observed (parameter γ).

Main Result:

- If $\boldsymbol{\Sigma}$ is known, there is a polynomial-time algorithm to estimate $\boldsymbol{\mu}^*$.
- Sample complexity:

$$\mathrm{poly}\left(d, 1/\alpha, 1/\beta, 1/\gamma, \lambda_{\max}/\lambda_{\min}, 1/\varepsilon, \log\left(1/\delta\right)\right)$$

Linear Thresholding Missingness Model



Algorithm: `MissingDescent`

Algorithm 2: Mean recovery algorithm given oracle access to incomplete data samples.

Input: Access to data generator $\mathcal{O}$; parameters $\beta, \lambda_{\text{sgd}}, \eta_{\text{lmc}}, R_{\text{lmc}}, r_{\text{proj}}, M_{\text{init}}, M_{\text{sgd}}, M_{\text{grad}}$

$\boldsymbol{\mu}^{(0)} \leftarrow$ `Initialize`$(\mathcal{O}, \beta, M_{\text{init}})$;
for $i \leftarrow 1$ ***to*** M_{sgd} **do**
  Sample $(A^{(i)}, \mathbf{x}^{(i)})$ from $\mathcal{O}$;
  $\eta_i \leftarrow \frac{1}{\lambda_{\text{sgd}} \cdot i}$;
  $\mathbf{g}^{(i)} \leftarrow$ `SampleGradient`$\left((A^{(i)}, \mathbf{x}^{(i)}), \boldsymbol{\mu}^{(i-1)}, \eta_{\text{lmc}}, R_{\text{lmc}}, M_{\text{grad}}\right)$;
  $\mathbf{v}^{(i)} \leftarrow \boldsymbol{\mu}^{(i-1)} - \eta_i \mathbf{g}^{(i)}$;
  $\boldsymbol{\mu}^{(i)} \leftarrow$ `ProjectToDomain`$(\boldsymbol{\mu}^{(0)}, \mathbf{v}^{(i)}, r_{\text{proj}})$;
$\bar{\boldsymbol{\mu}} \leftarrow \frac{1}{M_{\text{sgd}}}\sum_{i=1}^{M_{\text{sgd}}} \boldsymbol{\mu}^{(i)}$;
return $\bar{\boldsymbol{\mu}}$

Future Direction

- Extend recovery guarantees to unknown covariance settings.
- Study more general non-linear missingness patterns.
- Explore robust estimation under adversarial censoring.

Paper Link