Kernel Tuning for Compressive Clustering

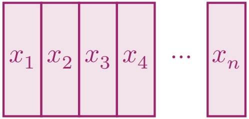
Antoine Chatalic (antoine.chatalic@irisa.fr), Rémi Gribonval

PANAMA research group – IRISA

December 2, 2020 – iTWIST

Context: large-scale (unsupervised) machine learning

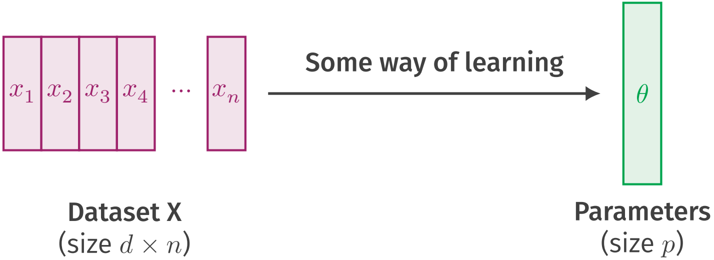
Goal: learn from the dataset about the underlying distribution!



Dataset X
(size $d \times n$)

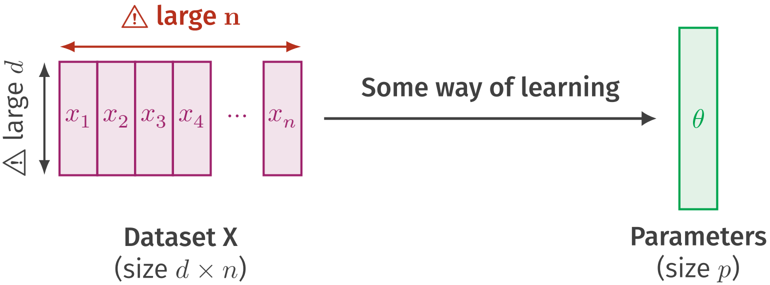
Context: large-scale (unsupervised) machine learning

Goal: learn from the dataset about the underlying distribution!



Context: large-scale (unsupervised) machine learning

Goal: learn from the dataset about the underlying distribution!

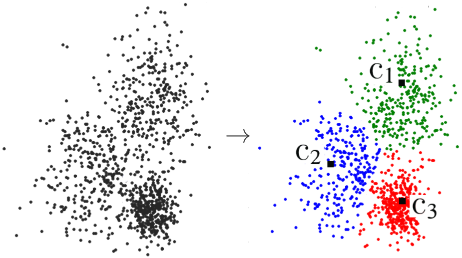


What do we learn? What is θ?

For today:

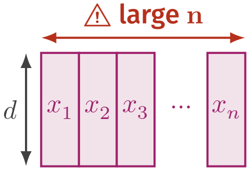
- **k-means** clustering.
  $\theta = \{c_1, \ldots, c_k\}$, k centroids in $\mathbb{R}^d$ minimizing the sum of squared errors:

$$\mathrm{SSE}((c_j)_{1 \le j \le k}, X) = \sum_{i=1}^{n} \min_j \|x_i - c_j\|^2.$$



Each sample x_i is assigned to the closest centroid.

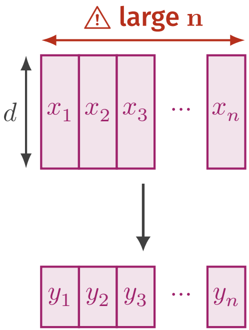
Handling large datasets: several approaches



Use the whole dataset
e.g. empirical risk minimization
⚠ Requires storage, RAM, time, GPUs.

“Compressive” approaches?

Handling large datasets: several approaches

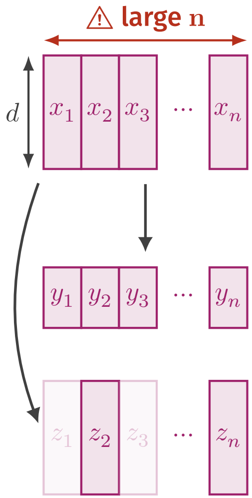


Use the whole dataset
e.g. empirical risk minimization
⚠ Requires storage, RAM, time, GPUs.

"Compressive" approaches?

Dimensionality reduction
New dimension $d' \ll d$.

Handling large datasets: several approaches



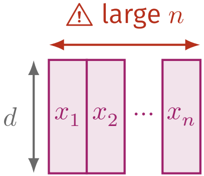
Use the whole dataset
e.g. empirical risk minimization
⚠ Requires storage, RAM, time, GPUs.

"Compressive" approaches?

Dimensionality reduction
New dimension $d' \ll d$.

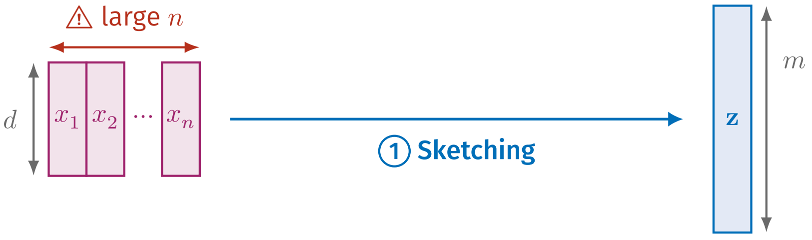
Subsampling (Coresets, Nyström methods)
$n' \ll n$ samples

Compressive learning: yet another idea



Dataset X
(size $d \times n$)

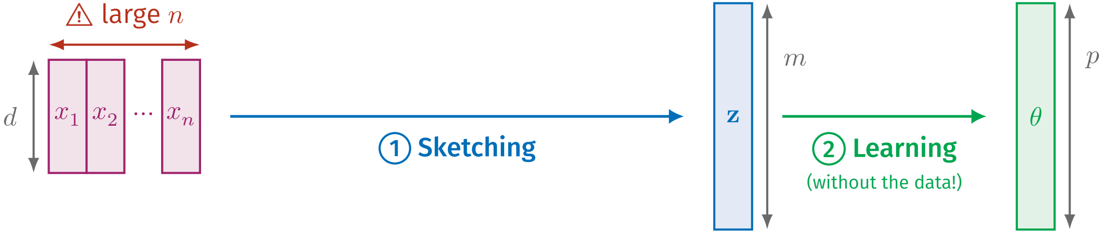
Compressive learning: yet another idea



Dataset X
(size $d \times n$)

Sketch
(size $m = \Theta(p) \ll nd$)

Compressive learning: yet another idea



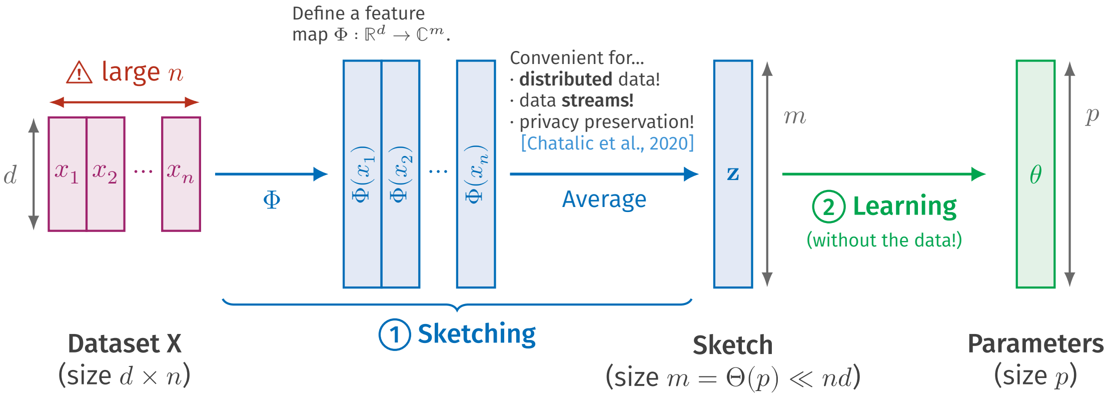
Dataset X
(size $d \times n$)

Sketch
(size $m = \Theta(p) \ll nd$)

Parameters
(size p)

[Bourrier, Gribonval, and Pérez, 2013. **"Compressive Gaussian Mixture Estimation"**]

Compressive learning: yet another idea



[Bourrier, Gribonval, and Pérez, 2013. **"Compressive Gaussian Mixture Estimation"**]

Which feature map Φ? How to learn?

For k-means clustering and GMM fitting, **random Fourier features** [Rahimi and Recht, 2008] :

$$\Phi(x) = \begin{bmatrix} e^{-i\omega_1^T x} \\ \vdots \\ e^{-i\omega_m^T x} \end{bmatrix} \in \mathbb{C}^m, \text{ with } \omega \overset{i.i.d.}{\sim} \mathcal{N}(0, \tfrac{1}{\sigma^2} I) \in \mathbb{R}^d.$$

Which feature map Φ? How to learn?

For k-means clustering and GMM fitting, **random Fourier features** [Rahimi and Recht, 2008] :

$$\Phi(x) = \begin{bmatrix} e^{-i\omega_1^T x} \\ \vdots \\ e^{-i\omega_m^T x} \end{bmatrix} \in \mathbb{C}^m, \text{ with } \omega \overset{i.i.d.}{\sim} \mathcal{N}(0, \tfrac{1}{\sigma^2} I) \in \mathbb{R}^d.$$

Learn from the empirical sketch = **moment-matching** problem (cf. [Hall, 2005]).

k-means clustering: looking for centroids $C = c_1, ..., c_k$ in $\mathbb{R}^d$ and weights α:

$$(C, \alpha) = \underset{\substack{C, \alpha \geq 0 \\ \alpha^T \mathbf{1} = 1}}{\arg\min} \Bigg\| \underbrace{\sum_{i=1}^{k} \alpha_i \Phi(c_i)}_{\text{sketch of the centroids } (c_i)_{1 \leq i \leq k}} - \underbrace{\mathbf{z}}_{\text{empirical sketch}} \Bigg\|_2 .$$

⚠ This is non-convex!
Heuristics exist (Continuous OMP [Keriven et al., 2017] , Message passing [Byrne et al., 2019] ...)

Interpretation of $\Phi(x) = \begin{bmatrix} e^{-i\omega_1^T x} \\ \vdots \\ e^{-i\omega_m^T x} \end{bmatrix} \in \mathbb{C}^m$

- **For probabilists:**
  - Sketch $=$ random samples of the empirical **characteristic function** φ.
  - Learning from the sketch $\approx$ generalized method of moments

Interpretation of $\Phi(x) = \begin{bmatrix} e^{-i\omega_1^T x} \\ \vdots \\ e^{-i\omega_m^T x} \end{bmatrix} \in \mathbb{C}^m$

- **For probabilists:**
  - Sketch $=$ random samples of the empirical **characteristic function** φ.
  - Learning from the sketch $\approx$ generalized method of moments
- **For sparsity addicts:**
  - Akin to **compressive sensing** [Foucart and Rauhut, 2013] .
  - Sketch = noisy linear measurements of the distribution via $\mathcal{A}(\pi) = \mathbf{E}_{x\sim\pi}\Phi(x)$.
  - Recovery possible with additional regularity assumptions (e.g. : recover mixture of diracs).

Interpretation of $\Phi(x) = \begin{bmatrix} e^{-i\omega_1^T x} \\ \vdots \\ e^{-i\omega_m^T x} \end{bmatrix} \in \mathbb{C}^m$

- **For probabilists:**
  - Sketch $=$ random samples of the empirical **characteristic function** φ.
  - Learning from the sketch $\approx$ generalized method of moments
- **For sparsity addicts:**
  - Akin to **compressive sensing** [Foucart and Rauhut, 2013] .
  - Sketch = noisy linear measurements of the distribution via $\mathcal{A}(\pi) = \mathbf{E}_{x\sim\pi}\Phi(x)$.
  - Recovery possible with additional regularity assumptions (e.g. : recover mixture of diracs).
- **Signal processing perspective:**
  - $\|\mathbf{z}_{\pi_1} - \mathbf{z}_{\pi_2}\|_2 \approx \|\kappa \star \pi_1 - \kappa \star \pi_2\|_{L^2(\mathbb{R}^d)}$ where $\kappa : u \mapsto \exp(-\frac{\|u\|^2}{2\sigma^2})$ (Remember $\omega \sim \mathcal{N}(0, \frac{1}{\sigma^2}I)$)
    i.e. sketching = low-pass filtering with a Gaussian kernel.

Statistical Learning Guarantees?

Yes! (Control of the excess risk)

Successful recovery provided that:

- Condition on the kernel

$$\sigma^2 \leq \varepsilon^2 / \log(k)$$

  where $\varepsilon =$ separation (minimum distance) between clusters.

[Gribonval et al., 2017. ***Compressive Statistical Learning with Random Feature Moments***]

Statistical Learning Guarantees?

Yes! (Control of the excess risk)

Successful recovery provided that:

- Condition on the kernel

$$\sigma^2 \leq \varepsilon^2 / \log(k)$$

  where $\varepsilon =$ separation (minimum distance) between clusters.

- Minimum sketch size

$$m \gtrsim k^2 d.$$

  In practice: $m = \Theta(kd)$ is sufficient.

  (Remember: we are learning kd parameters.)

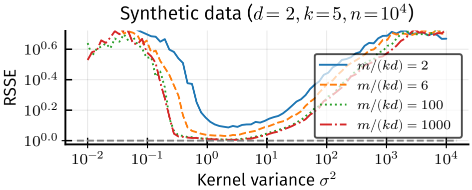
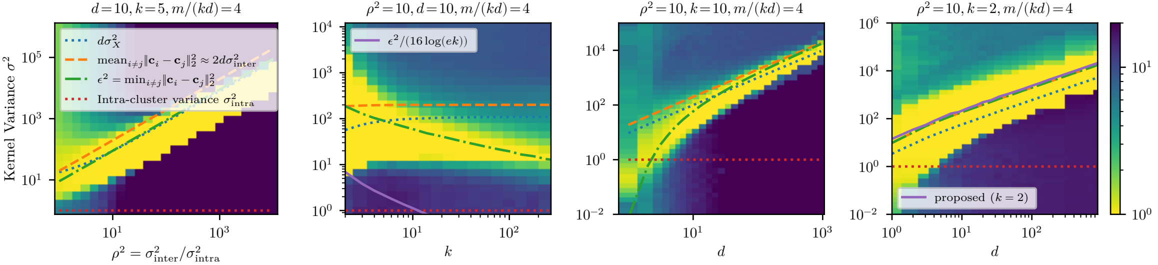


Figure: Impact of the sketch size m on clustering error (RSSE = SSE error w.r.t. standard k-means).

[Gribonval et al., 2017. ***Compressive Statistical Learning with Random Feature Moments***]

How to choose σ^2? [Chatalic and Gribonval, 2020]

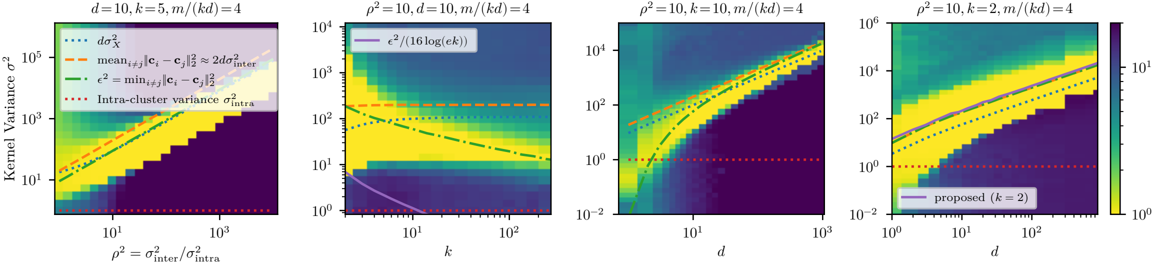
Simulations with data drawn as $x \overset{i.i.d.}{\sim} \sum_{i=1}^{k} \frac{1}{k}\mathcal{N}(\mathbf{c}_i, \sigma^2_{\text{intra}} I)$ with $\mathbf{c}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2_{\text{inter}} I)$.



- Yellow = good (RSSE ≈ 1), blue = bad.

How to choose σ^2? [Chatalic and Gribonval, 2020]

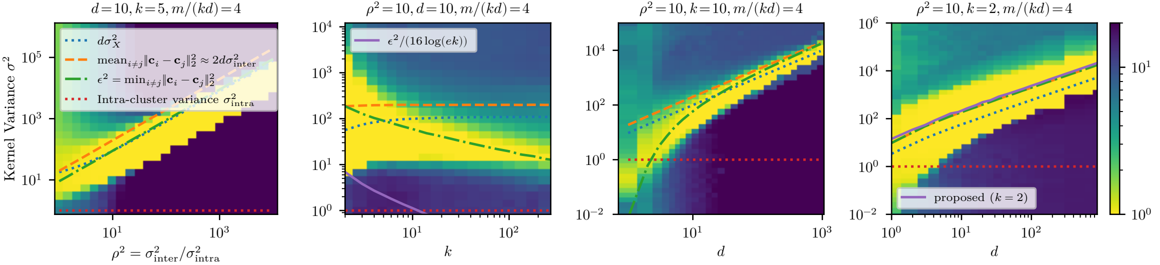
Simulations with data drawn as $x \overset{i.i.d.}{\sim} \sum_{i=1}^{k} \frac{1}{k}\mathcal{N}(\mathbf{c}_i, \sigma^2_{\text{intra}} I)$ with $\mathbf{c}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2_{\text{inter}} I)$.



- Yellow = good (RSSE ≈ 1), blue = bad.
- Dot-dashed green curve: squared separation
- Other curves: heuristics proposed so far

How to choose σ^2? [Chatalic and Gribonval, 2020]

Simulations with data drawn as $x \overset{i.i.d.}{\sim} \sum_{i=1}^{k} \frac{1}{k} \mathcal{N}(\mathbf{c}_i, \sigma^2_{\text{intra}} I)$ with $\mathbf{c}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2_{\text{inter}} I)$.



- Yellow = good (RSSE ≈ 1), blue = bad.
- Dot-dashed green curve: squared separation
- Other curves: heuristics proposed so far

Conclusion:

- Use the clusters separation
- The $\log(k)^{-1}$ dependency might be a proof artifact

How to learn the separation ε from the dataset?

Consider a mixture of dirac $P_C = \frac{1}{k}\sum_{1\leq i\leq k}\delta_{c_i}$.
Define $\mathbf{d}_{ij} \triangleq \frac{1}{2}(\mathbf{c}_i - \mathbf{c}_j)$ for any i, j.

When $\mathbf{k} = \mathbf{2}$, we have

$$1 - |\varphi(\boldsymbol{\omega})|^2 = \sin^2(\boldsymbol{\omega}^\top \mathbf{d}_{12}) \approx |\boldsymbol{\omega}^\top \mathbf{d}_{12}|^2 \text{ provided } \sigma_\omega \ll 1/\|\mathbf{d}_{12}\|$$

$\rightarrow$ we can get an approximation of $\varepsilon = \|\mathbf{d}_{12}\|$

When $\mathbf{k} > \mathbf{2}$,

$$\frac{1}{2}(k^2|\varphi(t\boldsymbol{\omega})|^2 - k) = \sum_{i<j}\cos(2\pi f_{ij} t) \text{ where } f_{ij} \triangleq \frac{1}{\pi}|\boldsymbol{\omega}^T \mathbf{d}_{ij}|.$$

$\rightarrow$ ... to be continued!

Bibliography I

Bourrier, Anthony, Rémi Gribonval, and Patrick Pérez (2013). "Compressive Gaussian Mixture Estimation". In: ICASSP-38th International Conference on Acoustics, Speech, and Signal Processing, pp. 6024–6028.

Byrne, Evan et al. (Sept. 2019). "Sketched Clustering via Hybrid Approximate Message Passing". In: *IEEE Transactions on Signal Processing* 67.17, pp. 4556–4569.

Chatalic, Antoine and Rémi Gribonval (June 2020). "Learning to Sketch for Compressive Clustering". In: International Traveling Workshop on Interactions between Low-Complexity Data Models and Sensing Techniques (iTWIST).

Chatalic, Antoine et al. (Mar. 3, 2020). "Compressive Learning with Privacy Guarantees". Submitted to Information and Inference (under review). Submitted to Information and Inference (under review).

Foucart, Simon and Holger Rauhut (2013). *A Mathematical Introduction to Compressive Sensing*. Vol. 1. 3. Birkhäuser Basel.

Gribonval, Rémi et al. (2017). *Compressive Statistical Learning with Random Feature Moments*. arXiv: 1706.07180.

Hall, Alastair R. (2005). *Generalized Method of Moments*. Oxford university press.

Keriven, Nicolas et al. (Mar. 5, 2017). "Compressive K-Means". In: International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Bibliography II

Rahimi, Ali and Benjamin Recht (2008). “Random Features for Large-Scale Kernel Machines”. In: *Advances in Neural Information Processing Systems*, pp. 1177–1184.