Hierarchical Methods of Moments

Matteo Ruffini [1] Guillaume Rabusseau [2] Borja Balle [3]

[1]Universitat Politècnica de Catalunya, [2]McGill University, [3]Amazon Research, Cambridge

Methods of Moments for Latent Variable Models

Task: to learn **latent variable models**. Take a model with parameters

$$M = [\mu_1, \dots, \mu_k] \in \mathbb{R}^{d \times k}, \omega \in \Delta^{k-1}$$

From an iid sample, estimate the *moments:*

$$M_1 := \textstyle\sum_{i=1}^{k} \omega_i \, \mu_i \in \mathbb{R}^{d} \quad (1)$$

$$M_2 := \textstyle\sum_{i=1}^{k} \omega_i \, \mu_i \otimes \mu_i \in \mathbb{R}^{d \times d} \quad (2)$$

$$M_3 := \textstyle\sum_{i=1}^{k} \omega_i \, \mu_i \otimes \mu_i \otimes \mu_i \in \mathbb{R}^{d \times d \times d} \quad (3)$$

Parameters with tensor decomposition:

$$\mathcal{TD}(M_1, M_2, M_3, k) \to (M, \omega)$$

Advantages:

- Single pass through the data.
- Always run in polynomial time.
- Provable guarantees of optimality.

Example: single topic model

- μ_i are the topics, ω the topic proportions.
- Let X_s be the $s-th$ word of a document (one-hot encoded); we have:

$$M_1 = \mathbb{E}[X_s], \quad M_2 = \mathbb{E}[X_s \otimes X_t]$$
$$M_3 = \mathbb{E}[X_r \otimes X_s \otimes X_t]$$

State of the art

Tensor decomposition for methods of moments: *Tensor Power Method* [1], *SVD-based methods* [2], *Random-projections* [3]...
Provably recover a model if structure and number of latent states k are known.
No theory for data out of the model.

This paper

SIDIWO: first method of moments with guarantees when number of latent states is unknown.

- In the **realizable setting**, recovers the model generating the data.
- In the **misspecified setting**, recovers a model that optimally synthesizes the one generating the data.

Application: hierarchical method of moments.

Algorithm 1 SIDIWO

Require: An iid dataset $\mathcal{X} = (x^{(1)}, \dots, x^{(n)})$, and the number of latent states l

Estimate M_1, M_2 and M_3.

l-components SVD: $M_2 \approx U_l S_l U_l^{\top}$.

Get the whitening matrix: $E_l = U_l S_l^{1/2}$.

Define the set of feasible joint-diagonalizers:

$$\mathcal{D}_l = \{D : D = (E_l O_l)^{\dagger} \text{ for } O_l \text{ s.t. } O_l O_l^{\top} = \mathbb{I}_l\}$$

Find the matrix $D \in \mathcal{D}_l$ optimizing

$$\min_{D \in \mathcal{D}_l} \sum_{i \neq j} \left(\sum_{r=1}^{d} (D M_{3,r} D^{\top})_{i,j}^2 \right)^{1/2} \quad (4)$$

Find $(\tilde{M}, \tilde{\omega})$ solving $\begin{cases} \tilde{M}\tilde{\Omega}^{1/2} = D^{\dagger} \\ \tilde{M}\tilde{\omega}^{\top} = M_1 \end{cases}$

where $\tilde{\Omega} = diag(\tilde{\omega})$

return $(\tilde{M}, \tilde{\omega})$

SIDIWO: interpretation

SIDIWO: **S**imultaneous **Di**agonalization based on **W**hitening and **O**ptimization.

- Use the **whitening** matrix to reduce the dimension of the slices of M_3:

$$H_r = E_l^{\dagger} M_{3,r} (E_l^{\dagger})^{\top} \in \mathbb{R}^{l \times l}$$

- Find an orthogonal matrix O that tries to **simultaneously diagonalize** all the H_r.
- Return $(\tilde{M}, \tilde{\omega})$ such that $\tilde{M}\tilde{\Omega}^{1/2} = E_l O$.

Assume that data is generated by a model with k states. For $l \leq k$, we have:

- $(\tilde{M}, \tilde{\omega}) \in \mathbb{R}^{(d \times l) \times l}$ the output of SIDIWO.
- $(M, \omega) \in \mathbb{R}^{(d \times k) \times k}$ the parameters of the model generating the data.

Question: how are $(\tilde{M}, \tilde{\omega})$ and (M, ω) related?

Realizable setting

If $l = k$, then $(\tilde{M}, \tilde{\omega}) = (M, \omega)$ and SIDIWO provably recovers (asymptotically) the parameters of the model.

Misspecified setting

If $l < k$, we prove, under mild requirements:

- The columns of $\tilde{M}$ are **non-trivial** linear combination of those of M; we call them **pseudocenters**.
- Problem (4) is equivalent to

$$\min_{D \in \mathcal{D}_l} \sum_{i \neq j} \sup_{v \in \mathcal{V}_M} \sum_{h=1}^{k} \langle d_i, \mu_h \rangle \langle d_j, \mu_h \rangle \omega_h v_h \quad (5)$$

with $d_1, \dots, d_l$ rows of a feasible D and

$$\mathcal{V}_M = \{v \in \mathbb{R}^k : v = \alpha^{\top} M, \text{where } \|\alpha\|_2 \leq 1\}$$

maximizing the disjoint support of vectors $u_1, \dots, u_l$, where

$$u_i = [\langle d_i, \mu_1 \sqrt{\omega_1} \rangle, \dots, \langle d_i, \mu_k \sqrt{\omega_k} \rangle]$$

Interpretation: Each pseudocenter tries to be aligned with some of the true centers and orthogonal to the others.

SIDIWO: optimization

We can rewrite problem (4) as:

$$\min_{O_l^{\top} O_l = \mathbb{I}_l} \sum_{i \neq j} \left(\sum_{r=1}^{d} (O_l^{\top} E_l^{\dagger} M_{3,r} (E_l^{\dagger})^{\top} O_l)_{i,j}^2 \right)^{1/2}$$

where O_l are orthogonal matrices.

- If $2 < l \leq k$ SIDIWO can be optimized with Jacobi's method [4].
- If $l = 2$ use the fact that the orthogonal matrix O_2 has the form

$$O_2(a) = \begin{bmatrix} \sqrt{1-a^2} & a \\ -a & \sqrt{1-a^2} \end{bmatrix}, \quad a \in [-1, 1]$$

and optimize w.r.t. a by griding on $[-1, 1]$.

Hierarchical Method of Moments

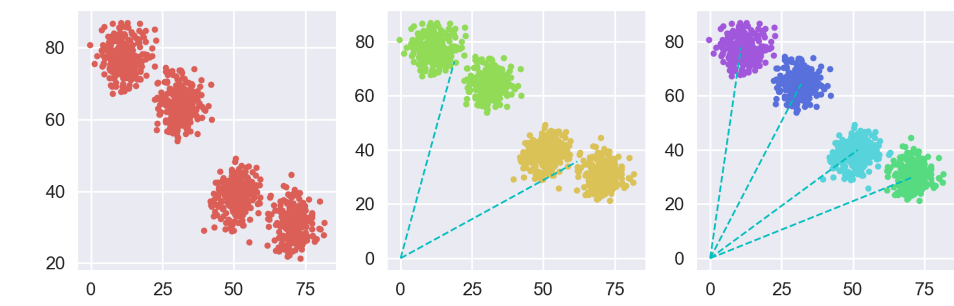
When $l = 2$, SIDIWO returns $[\tilde{\mu}_1, \tilde{\mu}_2]$; each pseudocenter synthesizes some of the true centers.
Define the set of centers approximated by $\tilde{\mu}_j$:

$$\mathcal{C}_j = \{\mu_i : \mu_i \text{ is approximated by } \tilde{\mu}_j\}$$

Idea: use the pseudocenters to bipartite a dataset, via MAP assignment on each sample $x^{(i)}$:

$$Cluster(i) = \arg\max_j \mathbb{P}[X = x^{(i)} | \tilde{\omega}, \tilde{\mu}_j]$$

Ideally, points generated by centers in $\mathcal{C}_j$ will belong to the cluster with center $\tilde{\mu}_j$, for $j = 1, 2$.
Recursively iterating: a *divisive hierarchical clustering algorithm*, with a *hierarchical representation* of our latent variable model.

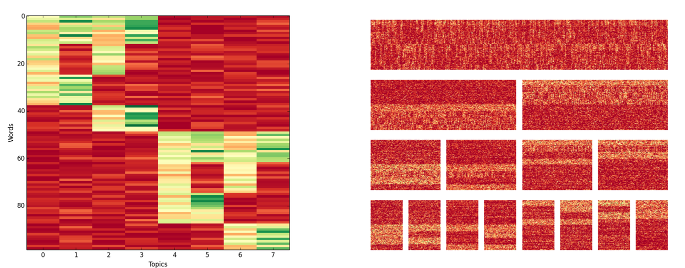


Hierarchical Topic Modeling

Latent variable model: **Single Topic Model**.

- True centers $\mu_1, \dots, \mu_k$: set of topics.
- Pseudocenters $\tilde{\mu}_1, \tilde{\mu}_2$: generic topics, summing up the concepts of the topics they synthesize.
- More specialized if deeper in the hierarchy.

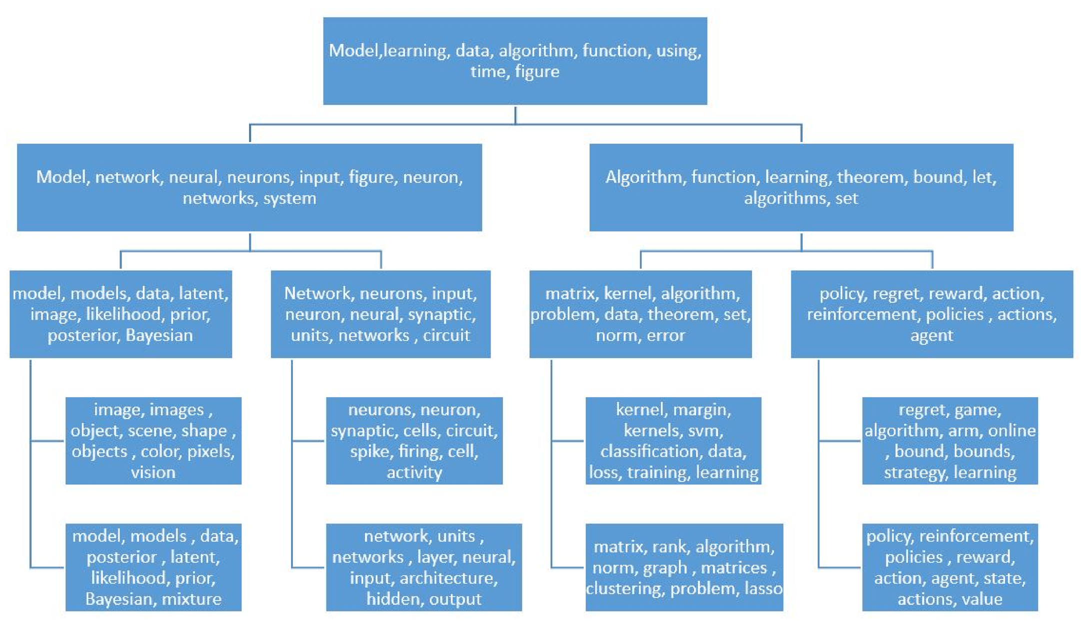
Synthetic Data



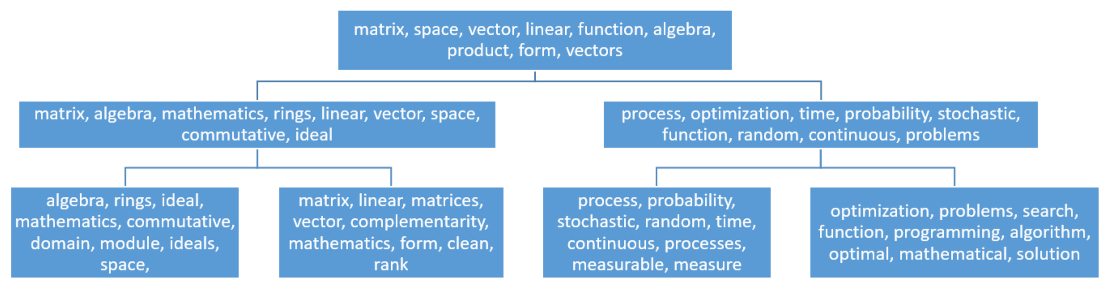
Method	Adj. Rand Idx Mean	St. dev.	Run. Time
TPM	0.93	0.06	1.2 sec.
SVD	0.52	0.13	0.1 sec.
Rand. Proj.	0.72	0.06	16 min.
SIDIWO	0.98	0.01	0.4 sec.

Generate single topic model data, do hierarchical clustering and study accuracy. Comparison with existing flat methods of moments.

Nips Papers 1987-2015



Wikipedia Mathematical Pages



References

[1] A. Anandkumar et al, (2014), Tensor decompositions for learning latent variable models.
[2] A. Anandkumar et al, (2012), A method of moments for mixture models and HMM.
[3] V. Kuleshov et al, (2016), Tensor factorization via matrix factorization.
[4] J. Cardoso et al, (1996), Jacobi angles for simultaneous diagonalization.