Unsupervised learning: symmetric low-rank matrix estimation, community detection and triplet loss.

Statistical physics and machine learning back together

Cargèse, August, 2018

Marc Lelarge

Unsupervised learning

Unsupervised machine learning is the machine learning task of inferring a function that describes the structure of "unlabeled" data (i.e. data that has not been classified or categorized). Since the examples given to the learning algorithm are unlabeled, there is no straightforward way to evaluate the accuracy of the structure that is produced by the algorithm.
Source:

Overview

- **Theoretical approach**
- **Data driven approach**

Overview

- **Theoretical approach**
  - Make assumptions, i.e. take a model for the data, and prove theorems.
- **Data driven approach**
  - Take a dataset and experiment!

Overview

- **Theoretical approach**
  - Make assumptions, i.e. take a model for the data, and prove theorems.
- **Data driven approach**
  - Take a dataset and experiment!
- **Common themes: graphs**

Overview

- **Theoretical approach**
  - Make assumptions, i.e. take a model for the data, and prove theorems.
    - Low-rank matrix estimation
    - Community detection
- **Data driven approach**
  - Take a dataset and experiment!
    - Unsupervised feature learning with deep neural network
- **Common themes: graphs**

Low-rank matrix estimation

"Spiked Wigner" model

$$\underbrace{\mathbf{Y}}_{\text{observations}} = \sqrt{\frac{\lambda}{n}}\,\underbrace{\mathbf{X}\mathbf{X}^\intercal}_{\text{signal}} + \underbrace{\mathbf{Z}}_{\text{noise}}$$

- $\mathbf{X}$: vector of dimension n with entries $X_i \overset{\text{i.i.d.}}{\sim} P_0$. $\mathbb{E}X_1 = 0$, $\mathbb{E}X_1^2 = 1$.
- $Z_{i,j} = Z_{j,i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$.
- λ: signal-to-noise ratio.
- λ and P_0 are known by the statistician.

Goal: recover the low-rank matrix $\mathbf{X}\mathbf{X}^\intercal$ from $\mathbf{Y}$.

Principal component analysis (PCA)

B.B.P. phase transition[1]

Spectral estimator:
Estimate $\mathbf{X}$ using the eigenvector $\hat{\mathbf{x}}_n$ associated with the largest eigenvalue μ_n of $\mathbf{Y}/\sqrt{n}$.

[1] SF Edwards and Raymund C Jones (1976). "The eigenvalue spectrum of a large symmetric random matrix". In: *Journal of Physics A: Mathematical and General* 9.10, p. 1595; TLH Watkin and J-P Nadal (1994). "Optimal unsupervised learning". In: *Journal of Physics A: Mathematical and General* 27.6, p. 1899.

Principal component analysis (PCA)

B.B.P. phase transition[1]

Spectral estimator:
Estimate $\mathbf{X}$ using the eigenvector $\hat{\mathbf{x}}_n$ associated with the largest eigenvalue μ_n of $\mathbf{Y}/\sqrt{n}$.

B.B.P. phase transition

- if $\lambda \leq 1$ $\begin{cases} \mu_n & \xrightarrow[n\to\infty]{\text{a.s.}} 2 \\ \mathbf{X} \cdot \hat{\mathbf{x}}_n & \xrightarrow[n\to\infty]{\text{a.s.}} 0 \end{cases}$
- if $\lambda > 1$ $\begin{cases} \mu_n & \xrightarrow[n\to\infty]{\text{a.s.}} \sqrt{\lambda} + \frac{1}{\sqrt{\lambda}} > 2 \\ |\mathbf{X} \cdot \hat{\mathbf{x}}_n| & \xrightarrow[n\to\infty]{\text{a.s.}} \sqrt{1 - 1/\lambda} > 0 \end{cases}$

Baik et al., 2005; Benaych-Georges and Nadakuditi, 2011

[1] SF Edwards and Raymund C Jones (1976). "The eigenvalue spectrum of a large symmetric random matrix". In: *Journal of Physics A: Mathematical and General* 9.10, p. 1595; TLH Watkin and J-P Nadal (1994). "Optimal unsupervised learning". In: *Journal of Physics A: Mathematical and General* 27.6, p. 1899.

Questions

- PCA fails when $\lambda \leq 1$, but is it still possible to recover the signal?

Questions

- PCA fails when $\lambda \leq 1$, but is it still possible to recover the signal?
- When $\lambda > 1$, is PCA optimal?

Questions

- PCA fails when $\lambda \leq 1$, but is it still possible to recover the signal?
- When $\lambda > 1$, is PCA optimal?
- More generally, what is the best achievable estimation performance in both regimes?

MMSE and information-theoretic threshold

Definitions

"MMSE" = Minimal Mean Square Error

$$\begin{aligned}\mathrm{MMSE}_n &= \min_{\hat{\theta}} \frac{1}{n^2}\mathbb{E}\left\|\mathbf{X}\mathbf{X}^\intercal - \hat{\theta}(\mathbf{Y})\right\|^2 \\ &= \frac{1}{n^2}\sum_{1\le i,j\le n}(X_iX_j - \mathbb{E}[X_iX_j|\mathbf{Y}])^2 \le \underbrace{\mathbb{E}_{P_0}[X_1^2]^2}_{\text{Dummy MSE}}\end{aligned}$$

The information-theoretic threshold is the critical value λ_c such that

- if $\lambda > \lambda_c$, $\lim_{n\to\infty} \mathrm{MMSE}_n <$ Dummy MSE
- if $\lambda < \lambda_c$, $\lim_{n\to\infty} \mathrm{MMSE}_n =$ Dummy MSE

Related work

A short overview

- Approximate Message Passing (AMP) algorithms: Rangan and Fletcher, 2012, Deshpande and Montanari, 2014; Lesieur et al., 2015b allows to derive the MMSE when AMP is optimal.
- In presence of a "hard phase", Barbier et al., 2016 uses AMP and spatial coupling techniques to compute the MMSE under some additional assumptions.
- Banks et al., 2016; Perry et al., 2016 obtained bounds on the information-theoretic threshold by second moment computations and contiguity.

Main result

Limiting formula for the MMSE

Theorem

$$\mathrm{MMSE}_n \xrightarrow[n\to\infty]{} \underbrace{\mathbb{E}_{P_0}[X^2]^2}_{\text{Dummy MSE}} - q^*(\lambda)^2$$

where $q^*(\lambda)$ is the maximizer of

$$q \geq 0 \mapsto \mathbb{E}_{\substack{X_0\sim P_0 \\ Z_0\sim\mathcal{N}}}\left[\log \int_{x_0} dP_0(x_0) e^{\sqrt{\lambda q} Z_0 x_0 + \lambda q X_0 x_0 - \frac{\lambda q}{2} x_0^2}\right] - \frac{\lambda}{4} q^2$$

Lelarge and Miolane, 2016 joint work with Léo Miolane.

This is the same formula as in Jean Barbier's talk, i.e. Replica Symmetric formula!

Proof ideas

A planted spin system

$$\mathbb{P}(\mathbf{X} = \mathbf{x} \,|\, \mathbf{Y}) = \frac{1}{\mathcal{Z}_n} P_0(\mathbf{x}) e^{H_n(\mathbf{x})} \text{ where}$$

$$H_n(\mathbf{x}) = \sum_{i<j} \sqrt{\frac{\lambda}{n}} Y_{i,j} x_i x_j - \frac{\lambda}{2n} x_i^2 x_j^2$$

$$= \sum_{i<j} \underbrace{\sqrt{\frac{\lambda}{n}} Z_{i,j} x_i x_j}_{\text{SK}} + \underbrace{\frac{\lambda}{n} X_i X_j x_i x_j - \frac{\lambda}{2n} x_i^2 x_j^2}_{\text{planting}}$$

Proof ideas

A planted spin system

$$\mathbb{P}(\mathbf{X} = \mathbf{x} \,|\, \mathbf{Y}) = \frac{1}{\mathcal{Z}_n} P_0(\mathbf{x}) e^{H_n(\mathbf{x})} \text{ where}$$

$$H_n(\mathbf{x}) = \sum_{i<j} \sqrt{\frac{\lambda}{n}} Y_{i,j} x_i x_j - \frac{\lambda}{2n} x_i^2 x_j^2$$

$$= \sum_{i<j} \underbrace{\sqrt{\frac{\lambda}{n}} Z_{i,j} x_i x_j}_{\text{SK}} + \underbrace{\frac{\lambda}{n} X_i X_j x_i x_j - \frac{\lambda}{2n} x_i^2 x_j^2}_{\text{planting}}$$

Lower bound: Guerra's interpolation technique. Adapted in Korada and Macris, 2009; Krzakala et al., 2016.

$$\begin{cases} \mathbf{Y} &= \sqrt{t} \quad \sqrt{\lambda/n} \quad \mathbf{X}\mathbf{X}^\intercal + \mathbf{Z} \\ \mathbf{Y}' &= \sqrt{1-t} \quad \sqrt{\lambda} \quad \mathbf{X} + \mathbf{Z}' \end{cases}$$

Proof ideas

Upper bound: cavity computations and the scalar channel

Cavity computations (Mézard et al., 1987) **in physics** $=$ **in mathematics**

Aizenman-Sims-Starr scheme: Aizenman et al., 2003; Talagrand, 2010 to compute the limit of the free energy $F_n = \frac{1}{n}\mathbb{E}\log \mathcal{Z}_n$ because

$$\text{Constant} - F_n = \frac{1}{n} I(\mathbf{X}; \mathbf{Y}) \xrightarrow[\text{I-MMSE theorem}]{\partial \lambda} \text{MMSE}$$

Proof ideas

Upper bound: cavity computations and the scalar channel

Cavity computations (Mézard et al., 1987) **in physics =in mathematics**

Aizenman-Sims-Starr scheme: Aizenman et al., 2003; Talagrand, 2010 to compute the limit of the free energy $F_n = \frac{1}{n}\mathbb{E}\log \mathcal{Z}_n$ because

$$\text{Constant} - F_n = \frac{1}{n} I(\mathbf{X};\mathbf{Y}) \xrightarrow[\text{I-MMSE theorem}]{\partial\lambda} \text{MMSE}$$

Lesieur et al., 2015a conjectured that the problem is characterized par the scalar channel:

$$Y_0 = \sqrt{\gamma} X_0 + Z_0$$

and the scalar free energy: $\mathcal{F}(\gamma) = \mathbb{E}\left[\log \sum_{x_0} P_0(x_0) e^{\sqrt{\gamma} Y_0 x_0 - \frac{\gamma}{2} x_0^2}\right]$

Proof ideas

Upper bound: cavity computations and the scalar channel

Cavity computations (Mézard et al., 1987) **in physics =in mathematics**

Aizenman-Sims-Starr scheme: Aizenman et al., 2003; Talagrand, 2010 to compute the limit of the free energy $F_n = \frac{1}{n}\mathbb{E}\log \mathcal{Z}_n$ because

$$\text{Constant} - F_n = \frac{1}{n} I(\mathbf{X};\mathbf{Y}) \xrightarrow[\text{I-MMSE theorem}]{\partial\lambda} \text{MMSE}$$

Lesieur et al., 2015a conjectured that the problem is characterized par the scalar channel:

$$Y_0 = \sqrt{\gamma} X_0 + Z_0$$

and the scalar free energy: $\mathcal{F}(\gamma) = \mathbb{E}\left[\log \sum_{x_0} P_0(x_0) e^{\sqrt{\gamma} Y_0 x_0 - \frac{\gamma}{2} x_0^2}\right]$

Replica symmetric formula

$$F_n \xrightarrow[n\to\infty]{} \sup_{q\geq 0} \mathcal{F}(\lambda q) - \frac{\lambda}{4} q^2$$

Some curves

Recall $\mathbf{Y} = \sqrt{\lambda/n}\mathbf{X}\mathbf{X}^\intercal + \mathbf{Z}$, where $(X_i)_{1\le i\le n} \overset{\text{i.i.d.}}{\sim} P_0$.

- If $P_0 = \mathcal{N}(0,1)$, PCA is optimal.

Some curves

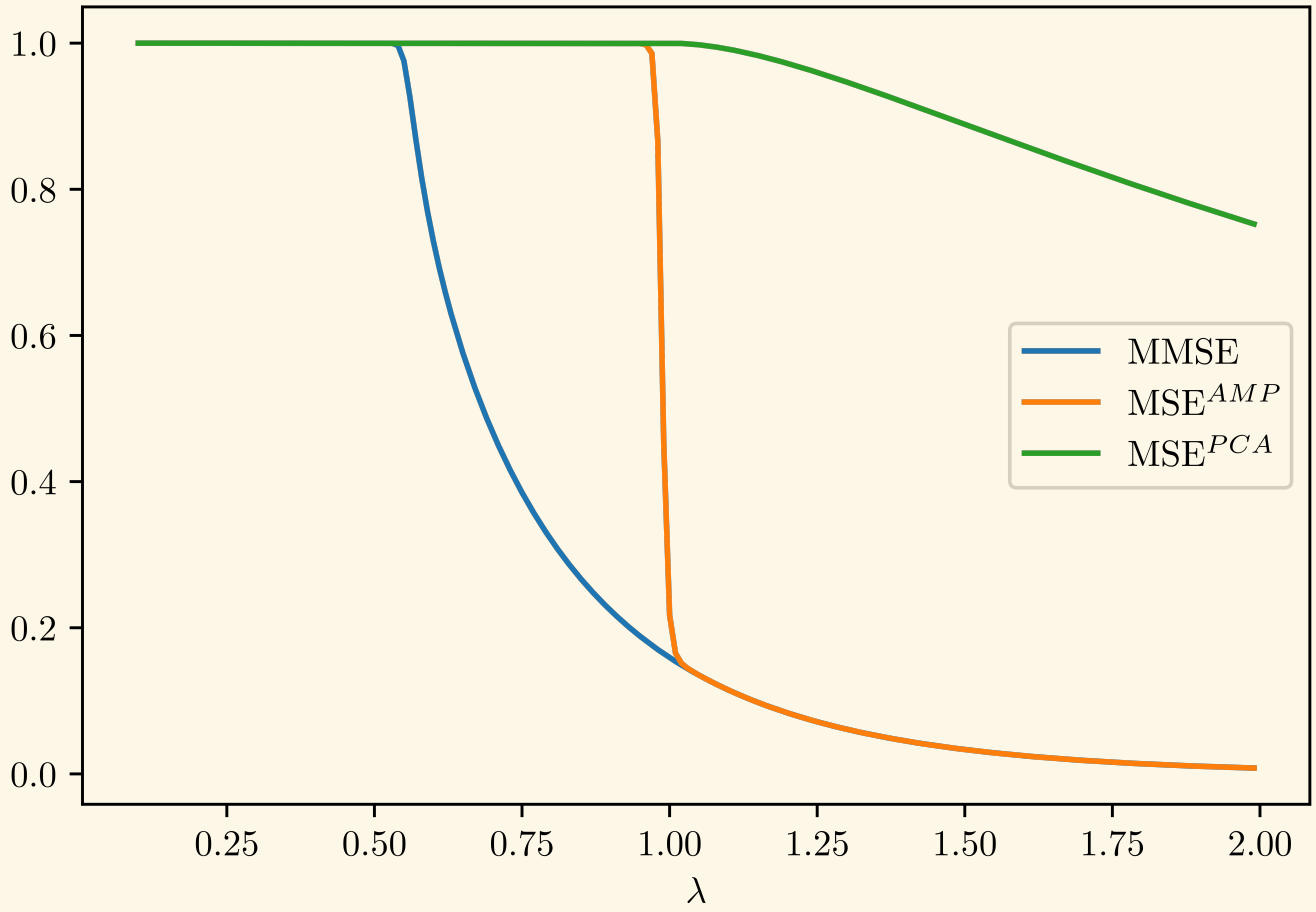
Recall $\mathbf{Y} = \sqrt{\lambda/n}\mathbf{X}\mathbf{X}^\intercal + \mathbf{Z}$, where $(X_i)_{1\le i\le n} \overset{\text{i.i.d.}}{\sim} P_0$.

- If $P_0 = \mathcal{N}(0,1)$, PCA is optimal.
- Next, we plot the MMSE and MSE^{PCA} curves for priors of the form

$$X_i = \begin{cases} \sqrt{\frac{1-p}{p}} & \text{with probability } p \\ -\sqrt{\frac{p}{1-p}} & \text{with probability } 1-p \end{cases}$$

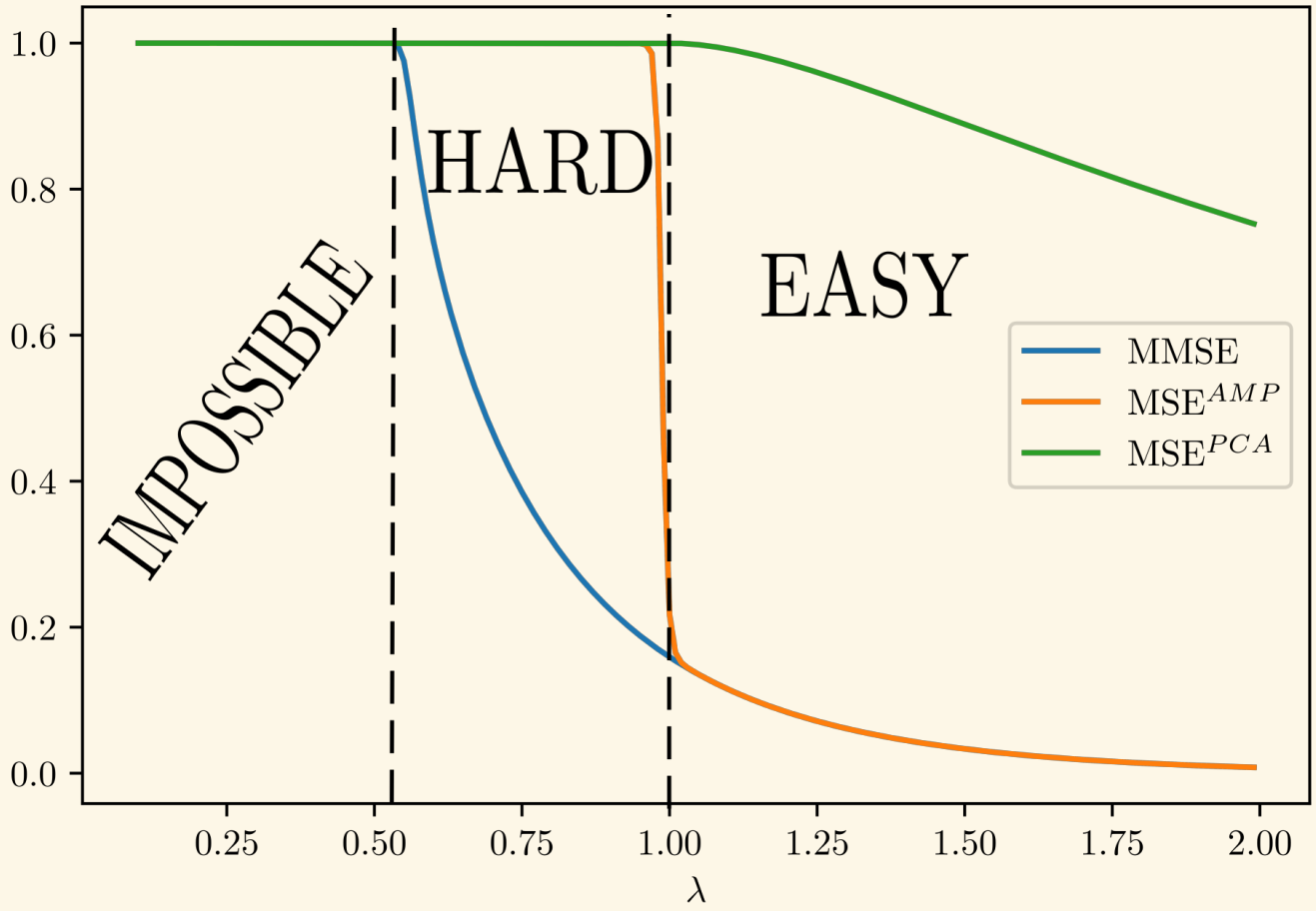
  for some $p \in (0,1)$.

Plot of MSEs



MMSE, MSE$^{\text{PCA}}$ and MSE$^{\text{AMP}}$, $p = 0.05$.

Plot of MMSE



MMSE, MSE$^{\text{PCA}}$ and MSE$^{\text{AMP}}$, $p = 0.05$.

Community detection

From Bernoulli to Gaussian noise

$$A_{i,j} \sim \text{Ber}\left(\frac{d}{n} + \frac{\sqrt{d}\sqrt{\lambda}}{n}\tilde{X}_i\tilde{X}_j\right) \tag{1}$$

where $$\tilde{X}_k = \begin{cases} \sqrt{(1-p)/p} & \text{with probability } p \\ -\sqrt{p/(1-p)} & \text{with probability } 1-p \end{cases}.$$

[2] Yash Deshpande et al. (2017). "Asymptotic mutual information for the balanced binary stochastic block model". In: *Information and Inference: A Journal of the IMA* 6.2, pp. 125–170.

Community detection

From Bernoulli to Gaussian noise

$$A_{i,j} \sim \text{Ber}\left(\frac{d}{n} + \frac{\sqrt{d}\sqrt{\lambda}}{n}\tilde{X}_i\tilde{X}_j\right) \tag{1}$$

where $\tilde{X}_k = \begin{cases} \sqrt{(1-p)/p} & \text{with probability } p \\ -\sqrt{p/(1-p)} & \text{with probability } 1-p \end{cases}$.

The Bernoulli noise model (1) is "equivalent" to the Gaussian noise model (when $n, d \to \infty$)[2]:

$$A'_{i,j} = \frac{d}{n} + \frac{\sqrt{d}\sqrt{\lambda}}{n}\tilde{X}_i\tilde{X}_j + \sqrt{\frac{d}{n}}Z_{i,j} \tag{2}$$

where $Z_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$,

[2] Yash Deshpande et al. (2017). "Asymptotic mutual information for the balanced binary stochastic block model". In: *Information and Inference: A Journal of the IMA* 6.2, pp. 125–170.

Community detection

From Bernoulli to Gaussian noise

$$A_{i,j} \sim \text{Ber}\left(\frac{d}{n} + \frac{\sqrt{d}\sqrt{\lambda}}{n}\tilde{X}_i\tilde{X}_j\right) \tag{1}$$

where $\tilde{X}_k = \begin{cases} \sqrt{(1-p)/p} & \text{with probability } p \\ -\sqrt{p/(1-p)} & \text{with probability } 1-p \end{cases}$.

The Bernoulli noise model (1) is "equivalent" to the Gaussian noise model (when $n, d \to \infty$)[2]:

$$A'_{i,j} = \frac{d}{n} + \frac{\sqrt{d}\sqrt{\lambda}}{n}\tilde{X}_i\tilde{X}_j + \sqrt{\frac{d}{n}}Z_{i,j} \tag{2}$$

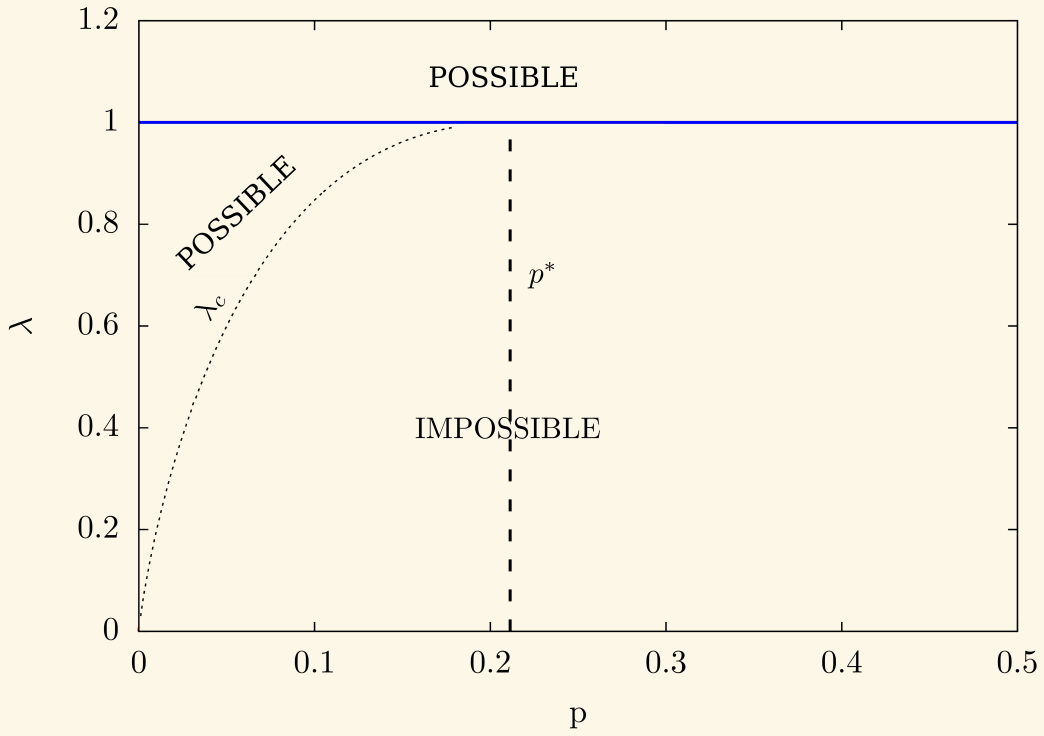
where $Z_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, and thus to

$$\sqrt{\frac{n}{d}}A'_{i,j} - \sqrt{\frac{d}{n}} = Y_{i,j} = \sqrt{\frac{\lambda}{n}}\tilde{X}_i\tilde{X}_j + Z_{i,j}$$

[2] Yash Deshpande et al. (2017). "Asymptotic mutual information for the balanced binary stochastic block model". In: *Information and Inference: A Journal of the IMA* 6.2, pp. 125–170.

Phase diagram for asymmetric community detection

Large degrees asymptotic



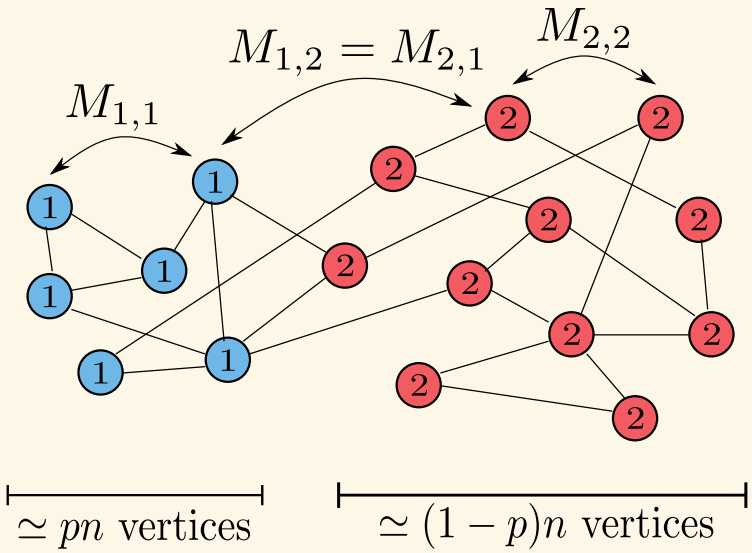
with $p^* = \frac{1}{2} - \frac{1}{2\sqrt{3}}$ as in Guilhem Semerjian's talk!
Phase diagram from Caltagirone et al., 2017 joint work with Francesco Caltagirone and Léo Miolane.

Community detection (the sparse case)

The Stochastic Block Model (SBM)

$\mathbf{G}$ is generated as follows:

- n vertices: $1, \dots, n$.
- Each vertex i has a label $X_i \in \{1, 2\}$ where $(X_k)_k \overset{\text{i.i.d.}}{\sim} 1 + \mathsf{Ber}(1-p)$.
- Two vertices i, j are then connected with probability M_{X_i, X_j}.

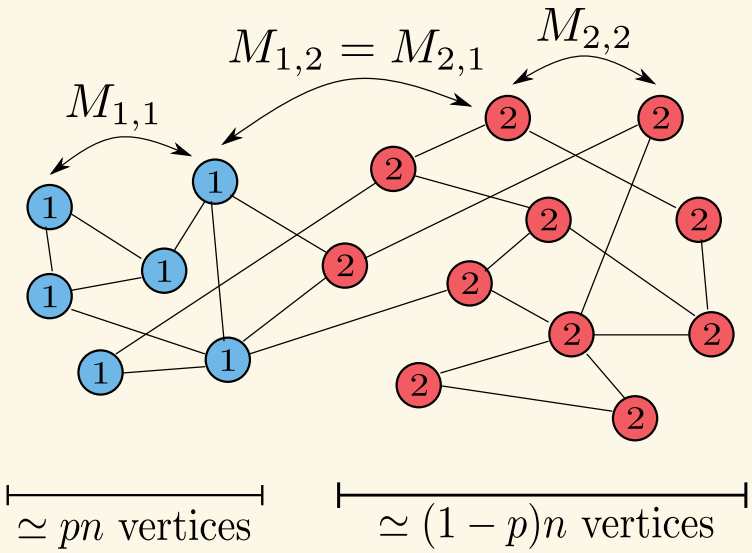


Community detection (the sparse case)

The Stochastic Block Model (SBM)

$\mathbf{G}$ is generated as follows:

- n vertices: $1, \dots, n$.
- Each vertex i has a label $X_i \in \{1, 2\}$ where $(X_k)_k \overset{\text{i.i.d.}}{\sim} 1 + \mathsf{Ber}(1-p)$.
- Two vertices i, j are then connected with probability M_{X_i, X_j}.



- The connectivity matrix is of the form:

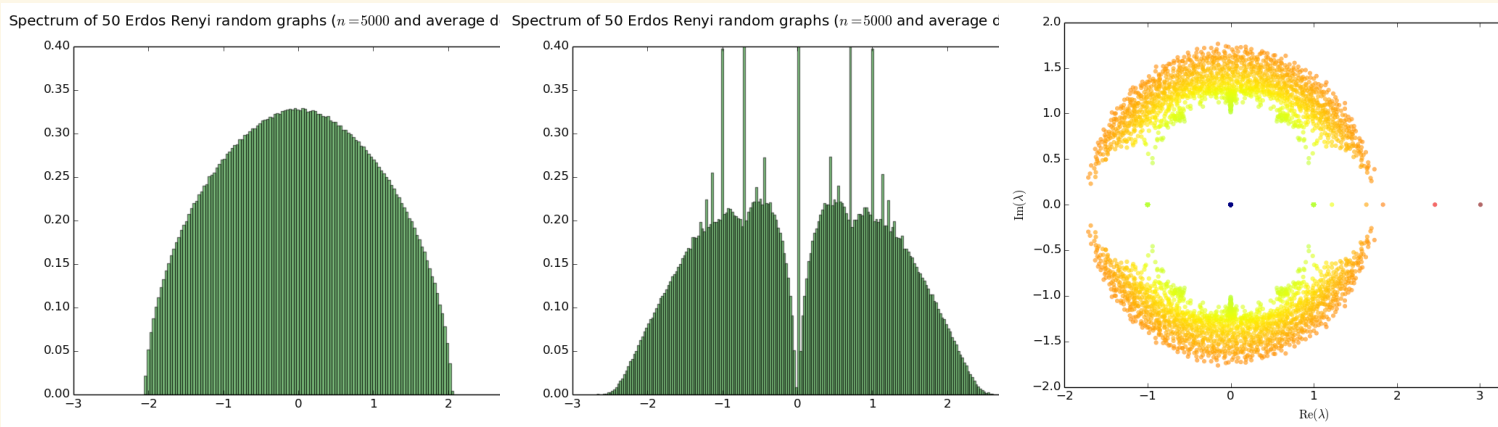
$$\mathbf{M} = \frac{d}{n} \begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

  $a, c > b$ and $pa + (1-p)b = pb + (1-p)c = 1$.
- **Important quantity:** the signal-to-noise ratio

$$\lambda = d(1-b)^2$$

The Non-Backtracking Matrix

The spectral redemption[3]



The problem: if $d \to \infty$, then Wigner's semi-circle law + BBP phase transition but if $d < \infty$ as $n \to \infty$, then Lifshitz tails.

The solution: the non-backtracking matrix on directed edges of the graph: $B_{u \to v, v \to w} = 1(\{u, v\} \in E)1(\{v, w\} \in E)1(u \neq w)$ achieves weak reconstruction on the SBM as soon as $\lambda > 1$.

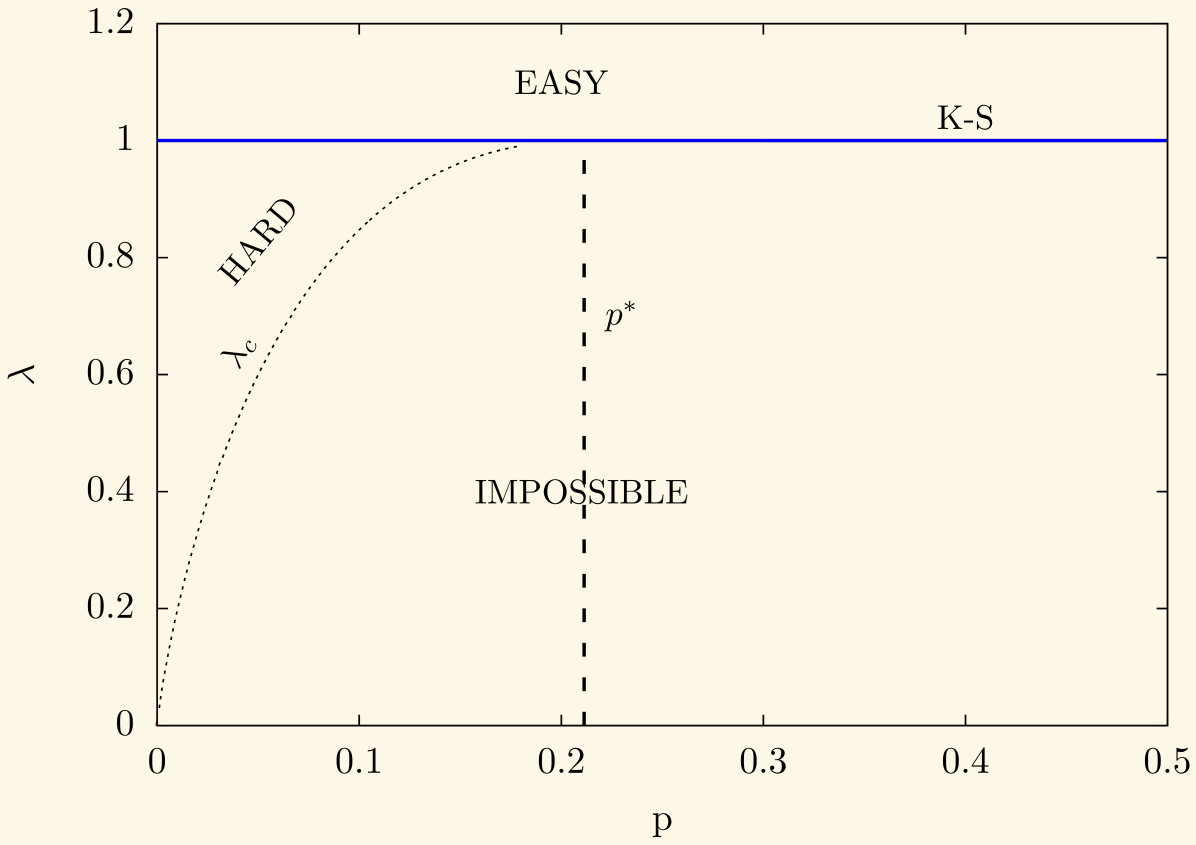
Bordenave et al., 2015 joint work with Charles Bordenave and Laurent Massoulié.

[3] Florent Krzakala et al. (2013). "Spectral redemption in clustering sparse networks". In: *Proceedings of the National Academy of Sciences* 110.52, pp. 20935–20940.

Asymmetric communities

The main picture

- if $\lambda > 1$, recovery is possible and easy for all $p < 1/2$
- No proof for the curve λ_c for sparse graphs.



The degree-corrected SBM

Impossibility result

The non-backtracking matrix is also working for the symetric degree-corrected SBM.

$$P(u \sim v) = \frac{\phi_u \phi_v}{n} \begin{cases} a & \text{if } X_u = X_v \\ b & \text{if } X_u \neq X_v \end{cases}$$

Theorem

Reconstruction is impossible if:

$$(a-b)^2 \Phi^{(2)} \leq 2(a+b),$$

with $\Phi^{(2)}$ the second moment of the weights.

Gulikers et al., 2015 joint work with Lennart Gulikers and Laurent Massoulié.

Robustness of the non-backtracking matrix

Weak recovery

Let $\rho = \Phi^{(2)} \frac{a+b}{2}$ and $\mu = \Phi^{(2)} \frac{a-b}{2}$.

Leading eigenvalue of the non-backtracking matrix B is asymptotic to ρ.

Second eigenvalue is asymptotic to μ when $\mu^2 > \rho$, but asymptotically bounded by $\sqrt{\rho}$ when $\mu^2 \leq \rho$.

All the remaining eigenvalues are asymptotically bounded by $\sqrt{\rho}$.

Consequently, a clustering positively-correlated with the true communities can be obtained based on the second eigenvector of B in the regime where $\mu^2 > \rho$ i.e. when $(a-b)^2\Phi^{(2)} > 2(a+b)$.

Gulikers et al., 2016, joint work with Lennart Gulikers and Laurent Massoulié.

Overview

- **Theoretical approach**
  - Make assumptions, i.e. take a model for the data, and prove theorems.
    - Low-rank matrix estimation
    - Community detection
- **Data driven approach**
  - Take a dataset and experiment!
    - Unsupervised feature learning with deep neural network
- **Common themes: graphs**

Unsupervised learning in practice

- If you have acces to an expert.
  a) Ask him to do feature engineering.
  b) Apply your favorite clustering algorithm on the feature vectors.
  c) Ask your expert if your algorithm was correct.

Unsupervised learning in practice

- If you have acces to an expert.
  a) Ask him to do feature engineering.
  b) Apply your favorite clustering algorithm on the feature vectors.
  c) Ask your expert if your algorithm was correct.
- In the first part of this talk, we replaced the expert by a model.

Unsupervised learning in practice

- If you have acces to an expert.
  a) Ask him to do feature engineering.
  b) Apply your favorite clustering algorithm on the feature vectors.
  c) Ask your expert if your algorithm was correct.
- In the first part of this talk, we replaced the expert by a model.
- Task a) is labor-intensive. Can we leverage the success of deep learning for feature extraction in an unsupervised framework?

Unsupervised learning in practice

- If you have acces to an expert.
  a) Ask him to do feature engineering.
  b) Apply your favorite clustering algorithm on the feature vectors.
  c) Ask your expert if your algorithm was correct.
- In the first part of this talk, we replaced the expert by a model.
- Task a) is labor-intensive. Can we leverage the success of deep learning for feature extraction in an unsupervised framework?
- The motivation is not new! Representation learning with Boltzmann Machines, Auto-Encoders, Generative Adversarial Networks...

Unsupervised learning in practice

- If you have acces to an expert.
  a) Ask him to do feature engineering.
  b) Apply your favorite clustering algorithm on the feature vectors.
  c) Ask your expert if your algorithm was correct.
- In the first part of this talk, we replaced the expert by a model.
- Task a) is labor-intensive. Can we leverage the success of deep learning for feature extraction in an unsupervised framework?
- The motivation is not new! Representation learning with Boltzmann Machines, Auto-Encoders, Generative Adversarial Networks...

- We propose a new data-driven approach based on the triplet loss on graphs.

Let's agree on a dataset!

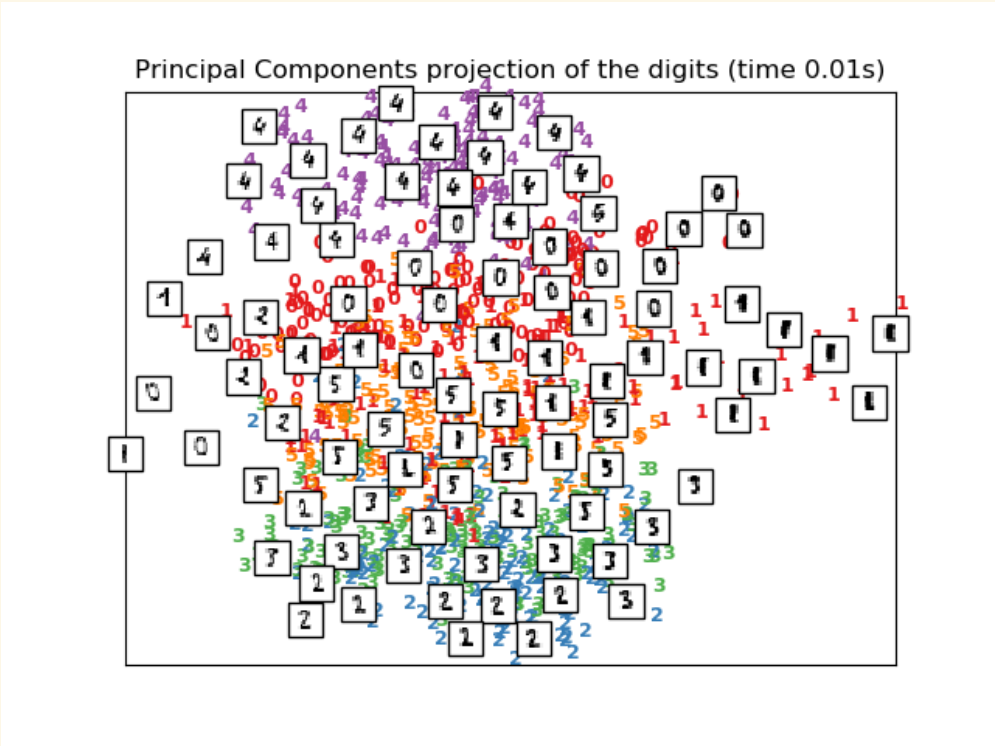
MNIST

We need to agree on the desired output of the algorithm.

Dimension reduction

PCA

Without an expert, it is safe to reduce the dimension before the clustering step, i.e. we replace feature engineering by dimension reduction.

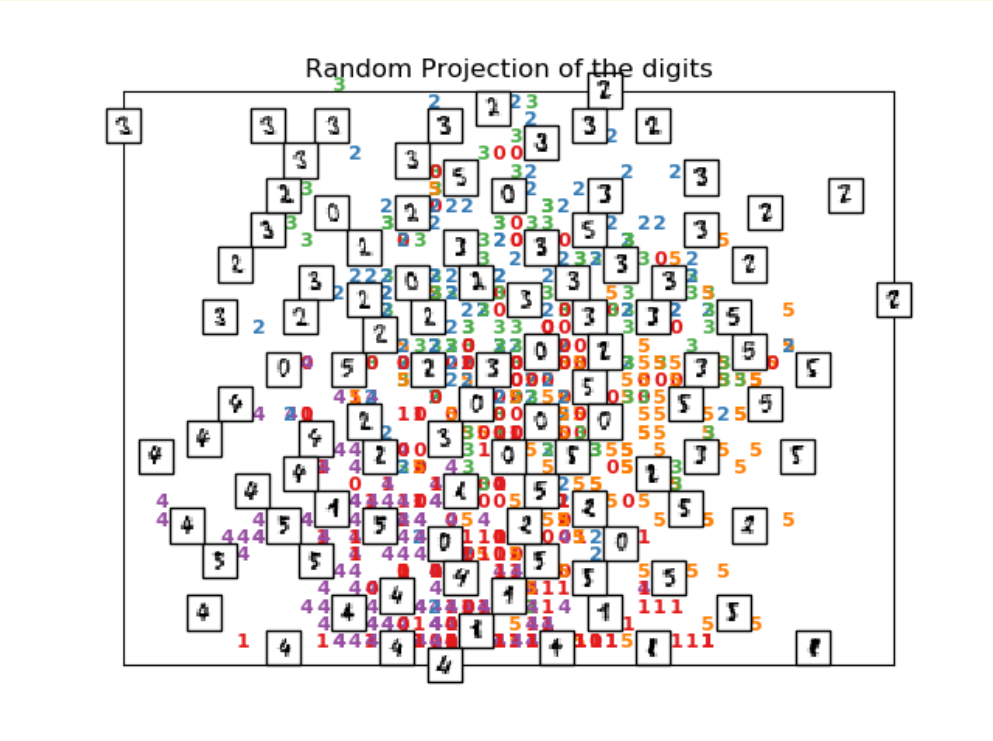


How to improve the feature engineering now?

Random projection

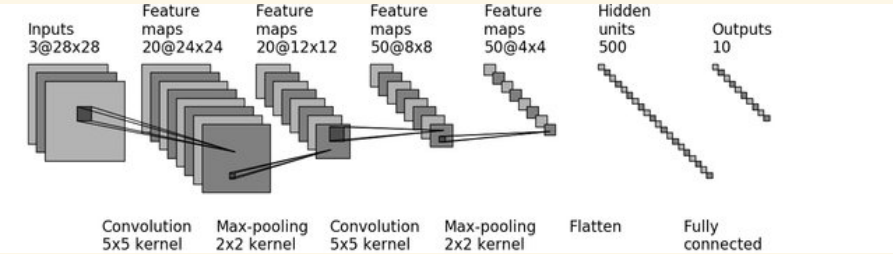
Johnson-Lindenstrauss dimension reduction lemma

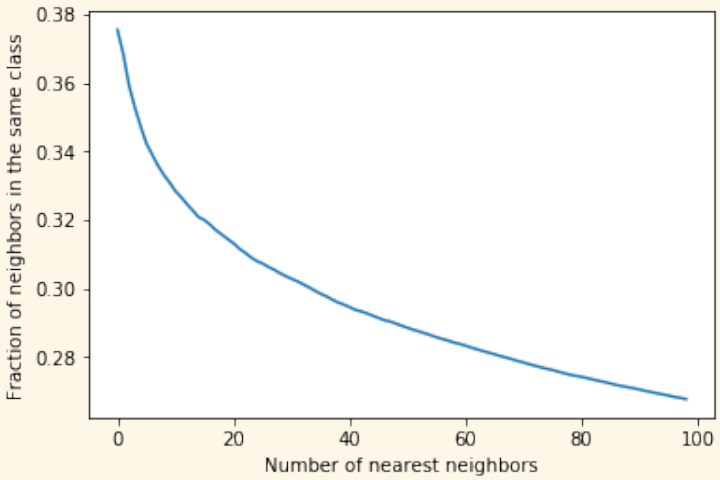
The probability that a random projection gives a distance-preserving dimension reduction is high.



Johnson-Lindenstrauss lemma with random LeNet

k-nearest neighbor graph

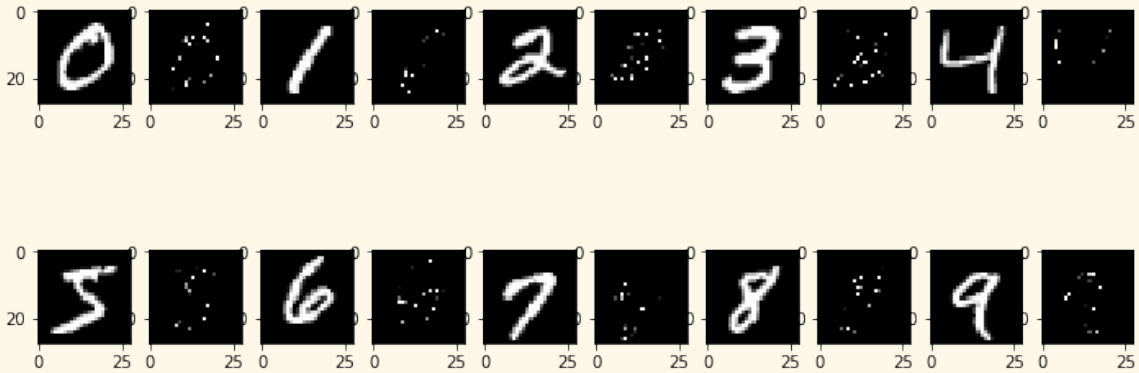




Side remark

MNIST is too clean!

We are using a noisy version of MNIST.



Learning the neural network

Summary so far:

- We used a neural network with random weights to produce noisy feature vectors f_i from the images.
- We built the k-nearest neighbor graph from the f_i.

We need to improve the feature vectors, i.e. learn the weights of the neural network.

Learning the neural network

Summary so far:

- We used a neural network with random weights to produce noisy feature vectors f_i from the images.
- We built the k-nearest neighbor graph from the f_i.

We need to improve the feature vectors, i.e. learn the weights of the neural network.

We do not have labels, but there is signal in the graph: for each edge there is $\approx 30\%$ of chances that both end-points belong to the same class, which is much better than the 10% given by a random pairing.

Learning the neural network

Summary so far:

- We used a neural network with random weights to produce noisy feature vectors f_i from the images.
- We built the k-nearest neighbor graph from the f_i.

We need to improve the feature vectors, i.e. learn the weights of the neural network.

We do not have labels, but there is signal in the graph: for each edge there is $\approx 30\%$ of chances that both end-points belong to the same class, which is much better than the 10% given by a random pairing.

Idea of the triplet loss: for each edge $i \to j$, pick a random vertex k in the graph and ensure that $\langle f_i, f_j \rangle > \langle f_i, f_k \rangle + \alpha$, by using the hinge loss with margin α.

Back to the Stochastic Block Model

Community detection with the triplet loss

Online algorithm: each incoming edge $i \to j$ induces a loss

$$\ell_{(i,j)} = (\langle f_i, f_K \rangle + \alpha - \langle f_i, f_j \rangle)^+ \text{ with a random } K;$$

once a batch arrived, update the embeddings f_i to minimize the loss.

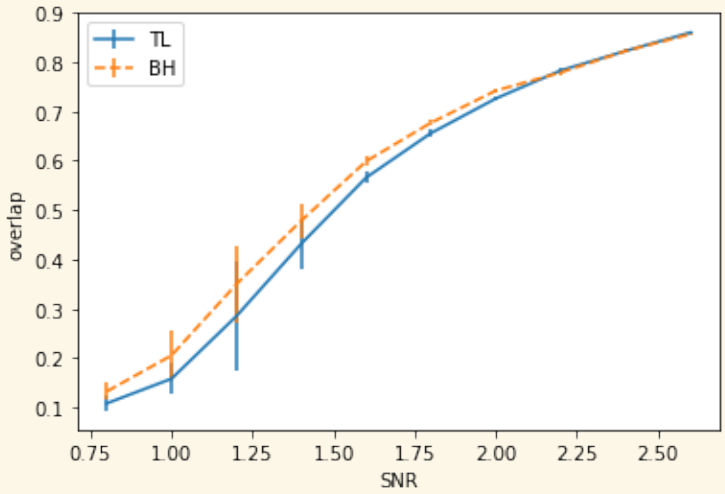
Back to the Stochastic Block Model

Community detection with the triplet loss

Online algorithm: each incoming edge $i \to j$ induces a loss

$$\ell_{(i,j)} = (\langle f_i, f_K \rangle + \alpha - \langle f_i, f_j \rangle)^+ \text{ with a random } K;$$

once a batch arrived, update the embeddings f_i to minimize the loss.

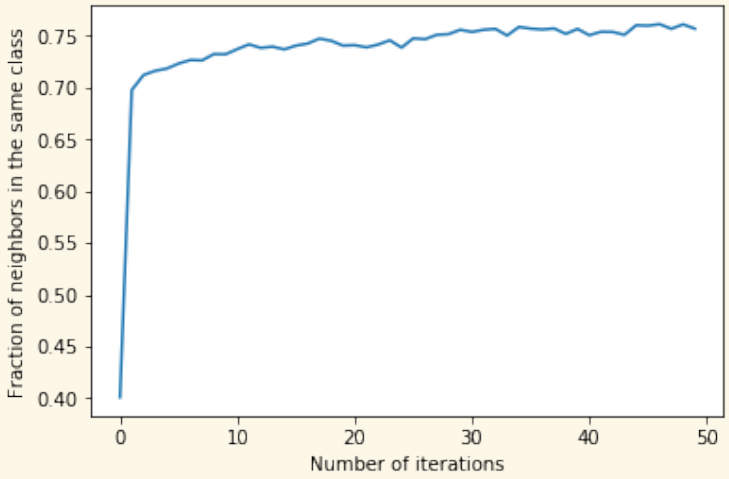


Comparison of Triplet Loss and Non-Backtracking.

Our algorithm learns a graph embedding.

Learning the neural network

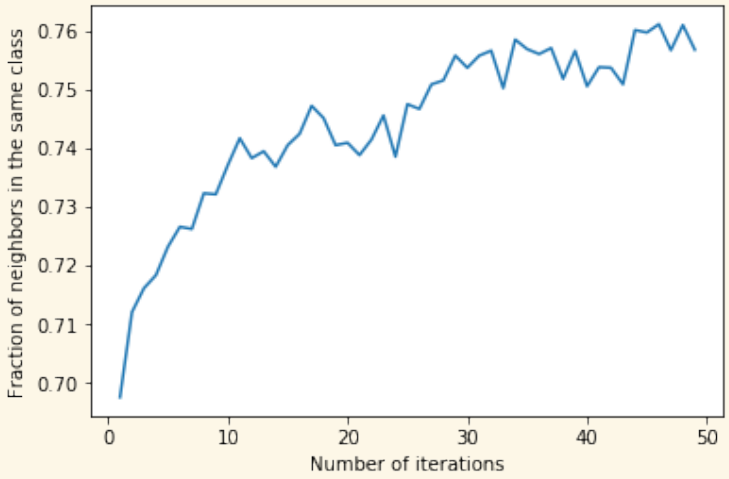
Empirical results



The accuracy of the clustering increases from $\approx 25\%$ for the random projection to more than 50%.

Learning the neural network

Empirical results



The accuracy of the clustering increases from $\approx 25\%$ for the random projection to more than 50%.

Summary

- **Theoretical approach**
  - Proof of the information theoretic threshold for the low-rank matrix estimation
  - Performance analysis of spectral clustering of the non-backtracking matrix for the community detection problem.
- **Data driven approach**
  - Deep learning algorithm based on the triplet loss and nearest neighbor graph for unsupervised feature extraction.
- **Common themes: graphs**

Thank you for your attention.

References I

- Aizenman, Michael, Robert Sims, and Shannon L Starr (2003). "Extended variational principle for the Sherrington-Kirkpatrick spin-glass model". In: *Physical Review B* 68.21, p. 214403.
- Baik, Jinho, Gérard Ben Arous, and Sandrine Péché (2005). "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices". In: *Annals of Probability*, pp. 1643–1697.
- Banks, Jess et al. (2016). "Information-theoretic bounds and phase transitions in clustering, sparse PCA, and submatrix localization". In: *arXiv preprint arXiv:1607.05222v2*.
- Barbier, Jean et al. (2016). "Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula". In: *Advances in Neural Information Processing Systems*, pp. 424–432.
- Benaych-Georges, Florent and Raj Rao Nadakuditi (2011). "The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices". In: *Advances in Mathematics* 227.1, pp. 494–521.

References II

- Bordenave, Charles, Marc Lelarge, and Laurent Massoulié (2015). "Non-backtracking spectrum of random graphs: community detection and non-regular Ramanujan graphs". In: *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*. IEEE, pp. 1347–1357.
- Caltagirone, Francesco, Marc Lelarge, and Léo Miolane (2017). "Recovering asymmetric communities in the stochastic block model". In: *IEEE Transactions on Network Science and Engineering*.
- Deshpande, Yash and Andrea Montanari (2014). "Information-theoretically optimal sparse PCA". In: *2014 IEEE International Symposium on Information Theory*. IEEE, pp. 2197–2201.
- Deshpande, Yash, Emmanuel Abbe, and Andrea Montanari (2017). "Asymptotic mutual information for the balanced binary stochastic block model". In: *Information and Inference: A Journal of the IMA* 6.2, pp. 125–170.
- Edwards, SF and Raymund C Jones (1976). "The eigenvalue spectrum of a large symmetric random matrix". In: *Journal of Physics A: Mathematical and General* 9.10, p. 1595.

References III

- Gulikers, Lennart, Marc Lelarge, and Laurent Massoulié (2015). "An Impossibility Result for Reconstruction in a Degree-Corrected Planted-Partition Model". In: *arXiv preprint arXiv:1511.00546*.
- – (2016). "Non-Backtracking Spectrum of Degree-Corrected Stochastic Block Models". In: *arXiv preprint arXiv:1609.02487*.
- Korada, Satish Babu and Nicolas Macris (2009). "Exact solution of the gauge symmetric p-spin glass model on a complete graph". In: *Journal of Statistical Physics* 136.2, pp. 205–230.
- Krzakala, Florent et al. (2013). "Spectral redemption in clustering sparse networks". In: *Proceedings of the National Academy of Sciences* 110.52, pp. 20935–20940.
- Krzakala, Florent, Jiaming Xu, and Lenka Zdeborová (2016). "Mutual information in rank-one matrix estimation". In: *Information Theory Workshop (ITW), 2016 IEEE*. IEEE, pp. 71–75.
- Lelarge, Marc and Léo Miolane (2016). "Fundamental limits of symmetric low-rank matrix estimation". In: *arXiv preprint arXiv:1611.03888*.

References IV

- Lesieur, Thibault, Florent Krzakala, and Lenka Zdeborová (2015a). "MMSE of probabilistic low-rank matrix estimation: Universality with respect to the output channel". In: *53rd Annual Allerton Conference on Communication, Control, and Computing, Allerton 2015, Allerton Park & Retreat Center, Monticello, IL, USA, September 29 - October 2, 2015*, pp. 680–687.
- – (2015b). "Phase transitions in sparse PCA". In: *IEEE International Symposium on Information Theory, ISIT 2015, Hong Kong, China, June 14-19, 2015*. IEEE, pp. 1635–1639. ISBN: 978-1-4673-7704-1. DOI: 10.1109/ISIT.2015.7282733. URL: http://dx.doi.org/10.1109/ISIT.2015.7282733.
- Mézard, Marc, Giorgio Parisi, and Miguel Virasoro (1987). *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*. Vol. 9. World Scientific Publishing Co Inc.
- Perry, Amelia et al. (2016). "Optimality and Sub-optimality of PCA for Spiked Random Matrices and Synchronization". In: *arXiv preprint arXiv:1609.05573*.

References V

- Rangan, Sundeep and Alyson K Fletcher (2012). "Iterative estimation of constrained rank-one matrices in noise". In: *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*. IEEE, pp. 1246–1250.
- Talagrand, Michel (2010). *Mean field models for spin glasses: Volume I: Basic examples*. Vol. 54. Springer Science & Business Media.
- Watkin, TLH and J-P Nadal (1994). "Optimal unsupervised learning". In: *Journal of Physics A: Mathematical and General* 27.6, p. 1899.