Tensor completion and tensor estimation

Andrea Montanari and Nike Sun

Stanford University and UC Berkeley

June 25, 2017

When did I first meet David?

2005: MSRI?

2007? 'There are some interesting mathematical things happening'

When did I first meet David?

2005: MSRI?

2007? ‘There are some interesting mathematical things happening’

When did I first meet David?

2005: MSRI?

2007? ‘There are some interesting mathematical things happening’

Tensor estimation: General question

Unknown tensor

$$\boldsymbol{X} \in \mathbb{R}^{d_1} \otimes \cdots \otimes \mathbb{R}^{d_k}$$
$$\boldsymbol{X} = (X_{i_1,\dots,i_k})_{i_1 \le d_1,\dots,i_k \le d_k}$$

Estimate $\boldsymbol{X}$ from noisy/incomplete observations

To symplify notations...

- $d_1 = d_2 = \cdots = d_k \equiv d$
- Tensors will be symmetric, e.g.:

$$X_{i_1.i_2,i_3} = X_{i_2,i_1,i_3} = X_{i_1,i_3,i_2} = \dots$$

- Results generalize.

Two concrete models:

Model #1: Spiked tensors

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{X} + \boldsymbol{W} \\ &= \lambda \boldsymbol{v}_0^{\otimes k} + \boldsymbol{W} \end{aligned}$$

Signal: $\boldsymbol{v}_0 \in \mathsf{S}^{d-1} \equiv \{\boldsymbol{x} \in \mathbb{R}^d : \ \|\boldsymbol{x}\|_2 = 1\}$.

Noise: $(W_{i_1,i_2,\dots,i_k})_{i_1<i_2<\dots<i_k} \sim_{iid} \mathsf{N}(0, 1/n)$

SNR: λ

Given $\boldsymbol{Y}$, estimate $\boldsymbol{v}_0$

[Montanari, Richard, 2015]

Model #1: Spiked tensors

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{X} + \boldsymbol{W} \\ &= \lambda \boldsymbol{v}_0^{\otimes k} + \boldsymbol{W} \end{aligned}$$

Signal: $\boldsymbol{v}_0 \in \mathsf{S}^{d-1} \equiv \{\boldsymbol{x} \in \mathbb{R}^d : \ \|\boldsymbol{x}\|_2 = 1\}$.

Noise: $(W_{i_1,i_2,\dots,i_k})_{i_1<i_2<\dots<i_k} \sim_{iid} \mathsf{N}(0, 1/n)$

SNR: λ

Given $\boldsymbol{Y}$, estimate $\boldsymbol{v}_0$

[Montanari, Richard, 2015]

Model #1: Spiked tensors

$k = 1$ (sequence model):

$$y_i = \lambda\, v_{0,i} + w_i$$

$k = 2$ (spiked matrix model):

$$Y_{ij} = \lambda\, v_{0,i} v_{0,j} + W_{ij}$$

$k = 3$ (spiked matrix model):

$$Y_{ijl} = \lambda\, v_{0,i} v_{0,i} v_{0,l} + W_{ijl}$$

Model #2: Tensor completion

$$\boldsymbol{X} = \sum_{\ell=1}^{r} \boldsymbol{v}_\ell^{\otimes k}$$

Observed entries $E \subseteq [d]^k \equiv \{1, \dots, d\}^k$

$$\Pi_E(\boldsymbol{X})_{i_1,\dots,i_k} = \begin{cases} X_{i_1,\dots,i_k} & \text{if } (i_1, \dots, i_k) \in E, \\ 0 & \text{otherwise} \end{cases}$$

Given $\boldsymbol{Y} = \Pi_E(\boldsymbol{X})$, estimate $\boldsymbol{X}$.

Model #2: Tensor completion

$$\boldsymbol{X} = \sum_{\ell=1}^{r} \boldsymbol{v}_\ell^{\otimes k}$$

Observed entries $E \subseteq [d]^k \equiv \{1, \dots, d\}^k$

$$\Pi_E(\boldsymbol{X})_{i_1,\dots,i_k} = \begin{cases} X_{i_1,\dots,i_k} & \text{if } (i_1, \dots, i_k) \in E, \\ 0 & \text{otherwise} \end{cases}$$

Given $\boldsymbol{Y} = \Pi_E(\boldsymbol{X})$, estimate $\boldsymbol{X}$.

Outline

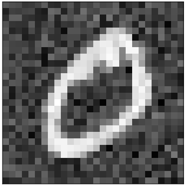
1. Why?
2. Tensor completion
3. Overcomplete tensors
4. Spiked tensor model (a.k.a. tensor PCA)
5. Conclusion

arXiv:1612.07866, CPAM
arXiv:1411.1076, NIPS

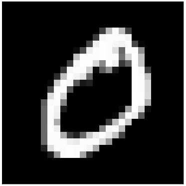
Why?

A cartoon application (not realistic!)

Image denoising



$=$



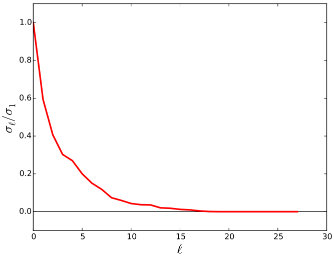
$+$ noise

$$\boldsymbol{Y} \quad = \quad \boldsymbol{X} \quad + \quad \boldsymbol{W}$$

Image denoising

Idea (not necessarily a good one): View $\boldsymbol{X}$ as a 28×28 matrix

Singular values:



[Dozens of references]

Singular value thresholding

Noisy image $\boldsymbol{Y} \in \mathbb{R}^{28\times 28}$

$$\boldsymbol{Y} = \sum_{\ell=1}^{28} \sigma_\ell \boldsymbol{u}_\ell \boldsymbol{v}_\ell^{\mathsf{T}}$$

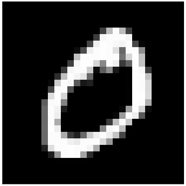
Denoised image

$$\boldsymbol{Y} = \sum_{\ell=1}^{r_*} \sigma_\ell \boldsymbol{u}_\ell \boldsymbol{v}_\ell^{\mathsf{T}}$$

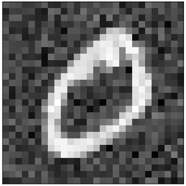
[. . . ; Candés, Sing-Long, Trzasko 2013; Donoho, Gavish 2014; Chatterjee 2015 . . .]

How well does it work? $r_0 = 12$, $\sigma = 30$

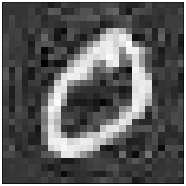
Original



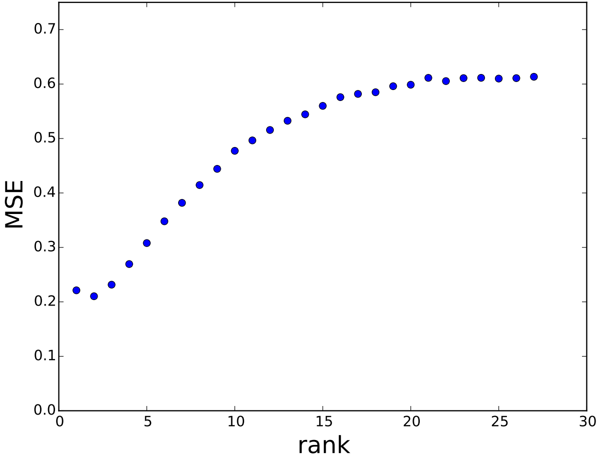
Noisy



Denoised

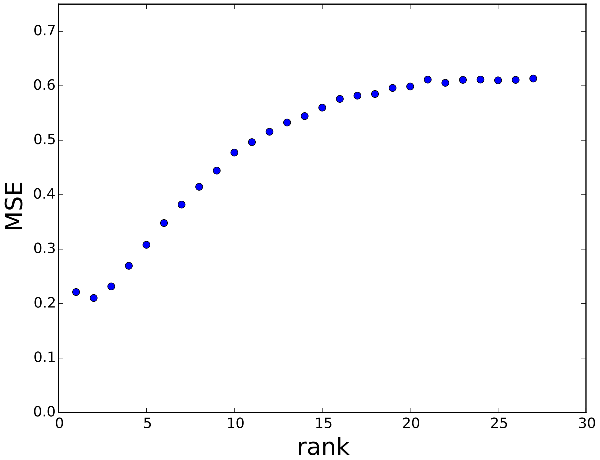


How well does it work? $\sigma = 100$



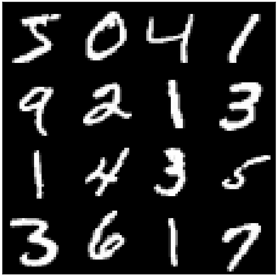
:(

How well does it work? $\sigma = 100$



:(

In reality we have many similar images



- $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_n \in \mathbb{R}^{d \times d}$
- Can we leverage similarities between images?

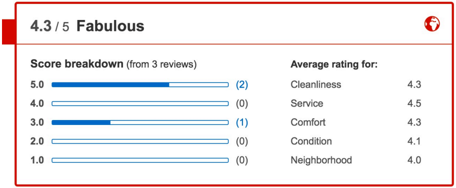
Idea

1. Stack images in a tensor:

$$\boldsymbol{X} = \Big[\boldsymbol{X}_1 \Big| \boldsymbol{X}_2 \Big| \cdots \Big| \boldsymbol{X}_n\Big] \in \mathbb{R}^d \otimes \mathbb{R}^d \otimes \mathbb{R}^n$$

2. Do tensor denoising

A better application



Collaborative filtering:

$$\text{Users} \times \text{Hotel} \times \text{Keyword}$$

Tensor completion

Model

- Unknown tensor $\boldsymbol{X} \in (\mathbb{R}^d)^{\otimes k}$:

$$\boldsymbol{X} = \sum_{\ell=1}^{r} \boldsymbol{v}_\ell^{\otimes k}$$

- Entries $E \subseteq [d]^k$, $|E| = n$ uniformly random

- Observations

$$\boldsymbol{Y} = \Pi_E(\boldsymbol{X})$$

Number of parameters $\approx r\, d$

Model

- Unknown tensor $\boldsymbol{X} \in (\mathbb{R}^d)^{\otimes k}$:

$$\boldsymbol{X} = \sum_{\ell=1}^{r} \boldsymbol{v}_\ell^{\otimes k}$$

- Entries $E \subseteq [d]^k$, $|E| = n$ uniformly random

- Observations

$$\boldsymbol{Y} = \Pi_E(\boldsymbol{X})$$

Number of parameters $\approx r\, d$

What do we know about matrices? ($k = 2$)

Theorem (Gross 2011)

If $X \in \mathbb{R}^{d_1 \times d_2}$, and $\text{rank}(X) \le r$ with factors factors are 'μ-incoherent', then we can reconstruct X exactly from

$$n \ge C(\mu)\, r(d_1 \vee d_2)\big(\log(d_1 \vee d_2)\big)^2$$

random entries. This is achieved by nuclear norm minimization.

[Candés, Recht, 2009; Candés, Tao, 2010; Keshavan, Montanari, Oh, 2010; Recht 2011; ...]

What do we know about matrices? ($k = 2$)

Theorem (Gross 2011)

If $\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}$*, and* $\operatorname{rank}(\boldsymbol{X}) \le r$ *with factors factors are '*μ*-incoherent', then we can reconstruct* $\boldsymbol{X}$ *exactly from*

$$n \ge C(\mu)\, r(d_1 \vee d_2)\big(\log(d_1 \vee d_2)\big)^2$$

random entries. This is achieved by nuclear norm minimization.

[Candés, Recht, 2009; Candés, Tao, 2010; Keshavan, Montanari, Oh, 2010; Recht 2011; . . .]

Let us try to redo the same with tensors!

- Number of unknown parameters $\approx rd$
- For $r = O(1)$ and nice factors, maximum likelihood succeds for

$$n \geq Crd \log d$$

[Do not know a reference]

- Generalizations of nuclear norm

$$n \geq C\big(r_{\boxplus}^{k-1} \vee r_{\boxplus}^{(k-1)/2} d^{1/2}\big)\, d(\log d)^2$$

[Yuan, Zhang, 2015; 2016]

Evaluating tensor nuclear norm is NP-hard!

Let us try to redo the same with tensors!

- Number of unknown parameters $\approx rd$
- For $r = O(1)$ and nice factors, maximum likelihood succeds for

$$n \geq Crd \log d$$

[Do not know a reference]

- Generalizations of nuclear norm

$$n \geq C\left(r_{\boxplus}^{k-1} \vee r_{\boxplus}^{(k-1)/2} d^{1/2}\right) d(\log d)^2$$

[Yuan, Zhang, 2015; 2016]

Evaluating tensor nuclear norm is NP-hard!

Let us try to redo the same with tensors!

- Number of unknown parameters $\approx rd$
- For $r = O(1)$ and nice factors, maximum likelihood succeds for

$$n \geq Crd \log d$$

[Do not know a reference]

- Generalizations of nuclear norm

$$n \geq C\big(r_{\boxplus}^{k-1} \vee r_{\boxplus}^{(k-1)/2} d^{1/2}\big)\, d(\log d)^2$$

[Yuan, Zhang, 2015; 2016]

Evaluating tensor nuclear norm is NP-hard!

Let us try to redo the same with tensors!

- Number of unknown parameters $\approx rd$
- For $r = O(1)$ and nice factors, maximum likelihood succeds for

$$n \geq Crd \log d$$

[Do not know a reference]

- Generalizations of nuclear norm

$$n \geq C\left(r_{\boxplus}^{k-1} \vee r_{\boxplus}^{(k-1)/2} d^{1/2}\right) d(\log d)^2$$

[Yuan, Zhang, 2015; 2016]

Evaluating tensor nuclear norm is NP-hard!

Let us try to redo the same with tensors!

- Number of unknown parameters $\approx rd$

- For $r = O(1)$ and nice factors, maximum likelihood succeds for

$$n \geq Crd \log d$$

[Do not know a reference]

- Generalizations of nuclear norm

$$n \geq C\big(r_{\boxplus}^{k-1} \vee r_{\boxplus}^{(k-1)/2} d^{1/2}\big)\, d(\log d)^2$$

[Yuan, Zhang, 2015; 2016]

Evaluating tensor nuclear norm is NP-hard!

Idea #1: What about reducing to $k = 2$?

$k = a + b$

$$\mathsf{unfold}^{a\times b} : (\mathbb{R}^d)^{\otimes k} \to \mathbb{R}^{d^a} \otimes \mathbb{R}^{d^b}$$
$$\boldsymbol{Y} \to \mathsf{unfold}^{a\times b}(\boldsymbol{Y})$$

$$\mathsf{unfold}^{a\times b}(\boldsymbol{Y})_{i_1\cdots i_a;\, j_1\cdots j_b} \equiv Y_{i_1\cdots i_a j_1\cdots j_b}$$

- $\boldsymbol{M} = \mathsf{unfold}^{a\times b}(\boldsymbol{Y})$
- Do matrix completion (e.g. nuclear norm minimization) of $\boldsymbol{M}$
- Fold back

[Tomioka, Hayashi, Kashima, 2010; Tomioka, Suzuki, Hayashi, Kashima, 2011; Liu, Musialski, Wonka, Ye, 2013; Gandy, Recht, Yamada, 2011]

Idea #1: What about reducing to $k = 2$?

$k = a + b$

$$\mathsf{unfold}^{a\times b} : (\mathbb{R}^d)^{\otimes k} \to \mathbb{R}^{d^a} \otimes \mathbb{R}^{d^b}$$
$$\boldsymbol{Y} \to \mathsf{unfold}^{a\times b}(\boldsymbol{Y})$$

$$\mathsf{unfold}^{a\times b}(\boldsymbol{Y})_{i_1\cdots i_a;\, j_1\cdots j_b} \equiv Y_{i_1\cdots i_a j_1\cdots j_b}$$

- $\boldsymbol{M} = \mathsf{unfold}^{a\times b}(\boldsymbol{Y})$
- Do matrix completion (e.g. nuclear norm minimization) of $\boldsymbol{M}$
- Fold back

[Tomioka, Hayashi, Kashima, 2010; Tomioka, Suzuki, Hayashi, Kashima, 2011; Liu, Musialski, Wonka, Ye, 2013; Gandy, Recht, Yamada, 2011]

Idea #1: What about reducing to $k = 2$?

$k = a + b$

$$\mathsf{unfold}^{a\times b} : \ (\mathbb{R}^d)^{\otimes k} \to \mathbb{R}^{d^a} \otimes \mathbb{R}^{d^b}$$
$$\boldsymbol{Y} \to \mathsf{unfold}^{a\times b}(\boldsymbol{Y})$$

$$\mathsf{unfold}^{a\times b}(\boldsymbol{Y})_{i_1\cdots i_a;\, j_1\cdots j_b} \equiv Y_{i_1\cdots i_a j_1\cdots j_b}$$

- $\boldsymbol{M} = \mathsf{unfold}^{a\times b}(\boldsymbol{Y})$
- Do matrix completion (e.g. nuclear norm minimization) of $\boldsymbol{M}$
- Fold back

[Tomioka, Hayashi, Kashima, 2010; Tomioka, Suzuki, Hayashi, Kashima, 2011; Liu, Musialski, Wonka, Ye, 2013; Gandy, Recht, Yamada, 2011]

Idea #1: What about reducing to $k = 2$?

Corollary

If $\boldsymbol{X} \in (\mathbb{R}^d)^{\otimes k}$, $\mathrm{rank}(\boldsymbol{X}) \le r$, *is such that* $\mathrm{unfold}^{a \times b}(\boldsymbol{X})$ *satisfies incoherence, then it can be reconstructed whp from* n *random entries, provided*

$$n \ge C\, r\, d^{a \vee b}\, (\log d)^2$$

Insights (?)

- Optimal choice $a = \lfloor k/2 \rfloor$, $b = \lceil k/2 \rceil$.
- Gap: $rd \ll n \ll rd^{\lceil k/2 \rceil}$.
- Unfolding cannot beat the barrier $n \gtrsim rd^{\lceil k/2 \rceil}$.

Idea #1: What about reducing to $k = 2$?

Corollary

If $\boldsymbol{X} \in (\mathbb{R}^d)^{\otimes k}$, $\mathrm{rank}(\boldsymbol{X}) \le r$, *is such that* $\mathsf{unfold}^{a\times b}(\boldsymbol{X})$ *satisfies incoherence, then it can be reconstructed whp from* n *random entries, provided*

$$n \ge C\, r\, d^{a\vee b}\, (\log d)^2$$

Insights (?)

- Optimal choice $a = \lfloor k/2 \rfloor$, $b = \lceil k/2 \rceil$.
- Gap: $rd \ll n \ll rd^{\lceil k/2 \rceil}$.
- Unfolding cannot beat the barrier $n \gtrsim rd^{\lceil k/2 \rceil}$.

Idea #2: Non-convex optimization

$\boldsymbol{X} = v_0^{\otimes k}$, $\|v_0\|_2 = 1$

Maximum likelihood

$$\begin{aligned} \text{maximize} \quad & \mathcal{L}_n(\boldsymbol{\theta}) = \frac{1}{2}\big\|\Pi_E(\boldsymbol{Y} - \boldsymbol{\theta}^{\otimes k})\big\|_F^2\,, \\ \text{subject to} \quad & \|\boldsymbol{\theta}\|_2 = 1\,. \end{aligned}$$

Heuristic analysis

$$\nabla\mathcal{L}_n(\boldsymbol{\theta}) = -k\,\Pi_E(\boldsymbol{Y})\{\boldsymbol{\theta}^{\otimes(k-1)}\} + \text{smaller terms}$$

Idea #2: Non-convex optimization

$\boldsymbol{X} = v_0^{\otimes k}$, $\|v_0\|_2 = 1$
Maximum likelihood

$$\begin{aligned} \text{maximize} \quad & \mathcal{L}_n(\boldsymbol{\theta}) = \frac{1}{2}\big\|\Pi_E(\boldsymbol{Y} - \boldsymbol{\theta}^{\otimes k})\big\|_F^2 \,, \\ \text{subject to} \quad & \|\boldsymbol{\theta}\|_2 = 1 \,. \end{aligned}$$

Heuristic analysis

$$\nabla \mathcal{L}_n(\boldsymbol{\theta}) = -k\, \Pi_E(\boldsymbol{Y})\{\boldsymbol{\theta}^{\otimes(k-1)}\} + \text{smaller terms}$$

Idea #2: Non-convex optimization

$\boldsymbol{X} = v_0^{\otimes k}$, $\|v_0\|_2 = 1$

Maximum likelihood

$$\begin{aligned} \text{maximize} \quad & \mathcal{L}_n(\boldsymbol{\theta}) = \frac{1}{2}\big\|\Pi_E(\boldsymbol{Y} - \boldsymbol{\theta}^{\otimes k})\big\|_F^2\,, \\ \text{subject to} \quad & \|\boldsymbol{\theta}\|_2 = 1\,. \end{aligned}$$

Heuristic analysis

$$\nabla\mathcal{L}_n(\boldsymbol{\theta}) = -k\,\Pi_E(\boldsymbol{Y})\{\boldsymbol{\theta}^{\otimes(k-1)}\} + \text{smaller terms}$$

Idea #2: Non-convex optimization

$$\begin{aligned}\Pi_E(\boldsymbol{Y}) &= \mathbb{E}\Pi_E(\boldsymbol{Y}) + \{\Pi_E(\boldsymbol{Y}) - \mathbb{E}\Pi_E(\boldsymbol{Y})\} \\ &= \frac{n}{d^k}\, v_0^{\otimes k} + \boldsymbol{W}\end{aligned}$$

Assume random initialization $\langle \boldsymbol{\theta}, v_0\rangle = c d^{-1/2}$, $c = O(1)$:

$$\begin{aligned}\langle v_0, \nabla \mathcal{L}_n(\boldsymbol{\theta})\rangle &= -k\,\langle \Pi_E(\boldsymbol{Y}), v_0 \otimes \boldsymbol{\theta}^{\otimes(k-1)}\rangle + \text{smaller terms} \\ &= -\frac{kn}{d^k}\langle v_0, \boldsymbol{\theta}\rangle^{k-1} - k\langle \boldsymbol{W}, v_0 \otimes \boldsymbol{\theta}^{\otimes(k-1)}\rangle + \dots \\ &= -ck\frac{n}{d^{(3k-1)/2}} \pm \frac{n^{1/2}}{d^k}\end{aligned}$$

Gradient points in a random direction unless $n \gtrsim d^{k-1}$

Is there a practical estimator for $n \leq rd^{1.1}$?

Barak, Moitra, 2014, $k = 3$:

- Under 'Feige's hypothesis,' no polynomil algorithm exists for

$$n \ll d^{3/2}$$

- Degree-6 Sum-Of-Squares works (approximate reconstruction) if

$$n \geq Cr_*^2 d^{3/2} (\log d)^4$$

- Complexity $O(d^{15})$

Practical algorithms? Better rank dependency?

Is there a practical estimator for $n \leq rd^{1.1}$?

Barak, Moitra, 2014, $k = 3$:

- Under 'Feige's hypothesis,' no polynomil algorithm exists for

$$n \ll d^{3/2}$$

- Degree-6 Sum-Of-Squares works (approximate reconstruction) if

$$n \geq Cr_*^2 d^{3/2} (\log d)^4$$

- Complexity $O(d^{15})$

Practical algorithms? Better rank dependency?

Is there a practical estimator for $n \leq rd^{1.1}$?

Barak, Moitra, 2014, $k = 3$:

- Under 'Feige's hypothesis,' no polynomil algorithm exists for

$$n \ll d^{3/2}$$

- Degree-6 Sum-Of-Squares works (approximate reconstruction) if

$$n \geq Cr_*^2 d^{3/2} (\log d)^4$$

- Complexity $O(d^{15})$

Practical algorithms? Better rank dependency?

Is there a practical estimator for $n \leq rd^{1.1}$?

Barak, Moitra, 2014, $k = 3$:

- Under 'Feige's hypothesis,' no polynomil algorithm exists for

$$n \ll d^{3/2}$$

- Degree-6 Sum-Of-Squares works (approximate reconstruction) if

$$n \geq Cr_*^2 d^{3/2} (\log d)^4$$

- Complexity $O(d^{15})$

Practical algorithms? Better rank dependency?

Is there a practical estimator for $n \le rd^{1.1}$?

Barak, Moitra, 2014, $k = 3$:

- Under 'Feige's hypothesis,' no polynomil algorithm exists for

$$n \ll d^{3/2}$$

- Degree-6 Sum-Of-Squares works (approximate reconstruction) if

$$n \ge C r_*^2 d^{3/2} (\log d)^4$$

- Complexity $O(d^{15})$

Practical algorithms? Better rank dependency?

Is there a practical estimator for $n \le rd^{1.1}$?

Barak, Moitra, 2014, $k = 3$:

- Under 'Feige's hypothesis,' no polynomil algorithm exists for

$$n \ll d^{3/2}$$

- Degree-6 Sum-Of-Squares works (approximate reconstruction) if

$$n \ge Cr_*^2 d^{3/2} (\log d)^4$$

- Complexity $O(d^{15})$

Practical algorithms? Better rank dependency?

A theorem

A1. $$\max_{i_1\dots i_k} |X_{i_1\dots i_k}|^2 \le \frac{\alpha}{d^k}\|\boldsymbol{X}\|_F^2$$

A2. $$\|\mathsf{unfold}^{a\times b}(\boldsymbol{X})\|_{\rm op}^2 \le \frac{\mu}{r}\|\mathsf{unfold}^{a\times b}(\boldsymbol{X})\|_F^2$$

Theorem (Montanari, Sun, 2017)

Under the above assumptions there exists a spectral algorithm achieving $\|\widehat{\boldsymbol{X}}(\boldsymbol{Y}) - \boldsymbol{X}\|_F \le \varepsilon\|\boldsymbol{X}\|_F$ *whp, provided*[a] $r \le d^{c(k)}$ *and*

$$n \ge C(\mu,\alpha,\varepsilon)\, r d^{k/2} (\log d)^9 .$$

[a] $c(3) = 3/4$, $c(k) = k/2$ (k even), $c(k) = (k/2) - 1$ ($k \ge 5$ odd).

A theorem

$$\textbf{A1.} \qquad \max_{i_1 \dots i_k} |X_{i_1 \dots i_k}|^2 \le \frac{\alpha}{d^k} \|\boldsymbol{X}\|_F^2$$

$$\textbf{A2.} \qquad \|\mathsf{unfold}^{a\times b}(\boldsymbol{X})\|_{\rm op}^2 \le \frac{\mu}{r} \|\mathsf{unfold}^{a\times b}(\boldsymbol{X})\|_F^2$$

Theorem (Montanari, Sun, 2017)

Under the above assumptions there exists a spectral algorithm achieving $\|\widehat{\boldsymbol{X}}(\boldsymbol{Y}) - \boldsymbol{X}\|_F \le \varepsilon \|\boldsymbol{X}\|_F$ *whp, provided*[a] $r \le d^{c(k)}$ *and*

$$n \ge C(\mu, \alpha, \varepsilon)\, r d^{k/2} (\log d)^9 \,.$$

[a] $c(3) = 3/4$, $c(k) = k/2$ (k even), $c(k) = (k/2) - 1$ ($k \ge 5$ odd).

A theorem

Theorem (Montanari, Sun, 2017)

Under the above assumptions there exists a spectral algorithm achieving $\|\widehat{\boldsymbol{X}}(\boldsymbol{Y}) - \boldsymbol{X}\|_F \le \varepsilon \|\boldsymbol{X}\|_F$ *whp, provided*[a] $r \le d^{c(k)}$ *and*

$$n \ge C(\mu, \alpha, \varepsilon)\, r d^{k/2} (\log d)^9 \,.$$

[a] $c(3) = 3/4$, $c(k) = k/2$ (k even), $c(k) = (k/2) - 1$ ($k \ge 5$ odd).

- **FALSE:** Optimal choice $a = \lfloor k/2 \rfloor$, $b = \lceil k/2 \rceil$.
- **FALSE:** Gap: $rd \ll n \ll rd^{\lceil k/2 \rceil}$.
- **FALSE:** Unfolding cannot beat the barrier $n \gtrsim rd^{\lceil k/2 \rceil}$.

A theorem

Theorem (Montanari, Sun, 2017)

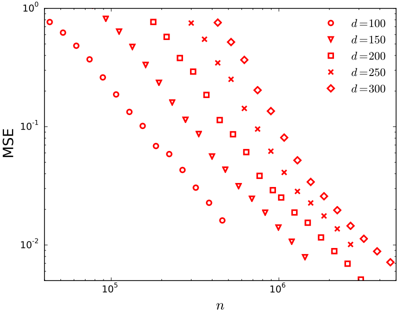
Under the above assumptions there exists a spectral algorithm achieving $\|\widehat{\boldsymbol{X}}(\boldsymbol{Y}) - \boldsymbol{X}\|_F \le \varepsilon \|\boldsymbol{X}\|_F$ *whp, provided*[a] $r \le d^{c(k)}$ *and*

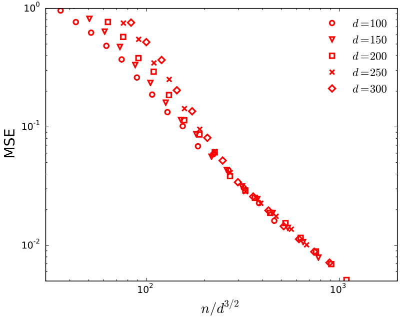
$$n \ge C(\mu, \alpha, \varepsilon)\, r d^{k/2} (\log d)^9 \,.$$

[a] $c(3) = 3/4$, $c(k) = k/2$ (k even), $c(k) = (k/2) - 1$ ($k \ge 5$ odd).

- **FALSE:** Optimal choice $a = \lfloor k/2 \rfloor$, $b = \lceil k/2 \rceil$.
- **FALSE:** Gap: $rd \ll n \ll rd^{\lceil k/2 \rceil}$.
- **FALSE:** Unfolding cannot beat the barrier $n \gtrsim rd^{\lceil k/2 \rceil}$.

Simulations: $k = 3$, $r = 4$





Algorithm idea

$k = a + b,\ a < b$

$$M = \mathsf{unfold}^{a \times b}(Y)$$

$d^a \ll n \ll d^{k/2}$:

- Cannot complete M
- Can estimate the top left singular space!

Algorithm, for $k = 3 \quad (\delta = nd^3)$

1. Compute $M = \text{unfold}^{1\times 2}(X)$ and

$$A = \frac{1}{\delta^2}\Pi^{\perp}_{\text{diag}}(MM^{\mathsf{T}}) + \frac{1}{\delta}\Pi_{\text{diag}}(MM^{\mathsf{T}})$$

2. Eigenvalue decomposition:

$$A = \sum_{i=1}^{d} \lambda_i u_i u_i^{\mathsf{T}}$$

3. Projector

$$Q = \sum_{i=1}^{d} \mathbf{1}_{\lambda_i \geq \lambda_*} u_i u_i^{\mathsf{T}}$$

4. Let $\mathcal{Q} \equiv Q \otimes Q \otimes Q$ and return

$$\widehat{X} = \frac{1}{\delta}\mathcal{Q}(X)$$

Algorithm, for $k = 3 \quad (\delta = nd^3)$

1. Compute $\boldsymbol{M} = \text{unfold}^{1\times 2}(\boldsymbol{X})$ and

$$\boldsymbol{A} = \frac{1}{\delta^2}\Pi_{\text{diag}}^{\perp}(\boldsymbol{M}\boldsymbol{M}^{\mathsf{T}}) + \frac{1}{\delta}\Pi_{\text{diag}}(\boldsymbol{M}\boldsymbol{M}^{\mathsf{T}})$$

2. Eigenvalue decomposition:

$$\boldsymbol{A} = \sum_{i=1}^{d} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^{\mathsf{T}}$$

3. Projector

$$\boldsymbol{Q} = \sum_{i=1}^{d} \mathbf{1}_{\lambda_i \geq \lambda_*} \boldsymbol{u}_i \boldsymbol{u}_i^{\mathsf{T}}$$

4. Let $\mathcal{Q} \equiv \boldsymbol{Q} \otimes \boldsymbol{Q} \otimes \boldsymbol{Q}$ and return

$$\widehat{\boldsymbol{X}} = \frac{1}{\delta}\mathcal{Q}(\boldsymbol{X})$$

Algorithm, for $k = 3 \quad (\delta = nd^3)$

1. Compute $\boldsymbol{M} = \text{unfold}^{1\times 2}(\boldsymbol{X})$ and

$$\boldsymbol{A} = \frac{1}{\delta^2}\Pi^{\perp}_{\text{diag}}(\boldsymbol{M}\boldsymbol{M}^{\mathsf{T}}) + \frac{1}{\delta}\Pi_{\text{diag}}(\boldsymbol{M}\boldsymbol{M}^{\mathsf{T}})$$

2. Eigenvalue decomposition:

$$\boldsymbol{A} = \sum_{i=1}^{d} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^{\mathsf{T}}$$

3. Projector

$$Q = \sum_{i=1}^{d} \mathbf{1}_{\lambda_i \geq \lambda_*} \boldsymbol{u}_i \boldsymbol{u}_i^{\mathsf{T}}$$

4. Let $\mathcal{Q} = Q \otimes Q \otimes Q$ and return

$$\widehat{\boldsymbol{X}} = \frac{1}{\delta}\mathcal{Q}(\boldsymbol{X})$$

Algorithm, for $k = 3 \quad (\delta = nd^3)$

1. Compute $M = \text{unfold}^{1\times 2}(X)$ and

$$A = \frac{1}{\delta^2}\Pi^{\perp}_{\text{diag}}(MM^{\mathsf{T}}) + \frac{1}{\delta}\Pi_{\text{diag}}(MM^{\mathsf{T}})$$

2. Eigenvalue decomposition:

$$A = \sum_{i=1}^{d} \lambda_i u_i u_i^{\mathsf{T}}$$

3. Projector

$$Q = \sum_{i=1}^{d} \mathbf{1}_{\lambda_i \geq \lambda_*} u_i u_i^{\mathsf{T}}$$

4. Let $\mathcal{Q} \equiv Q \otimes Q \otimes Q$ and return

$$\widehat{X} = \frac{1}{\delta}\mathcal{Q}(X)$$

Algorithm, for $k = 3$ $\quad (\delta = nd^3)$

1. Compute $\boldsymbol{M} = \text{unfold}^{1\times 2}(\boldsymbol{X})$ and

$$\boldsymbol{A} = \frac{1}{\delta^2}\Pi^{\perp}_{\text{diag}}(\boldsymbol{M}\boldsymbol{M}^{\mathsf{T}}) + \frac{1}{\delta}\Pi_{\text{diag}}(\boldsymbol{M}\boldsymbol{M}^{\mathsf{T}})$$

2. Eigenvalue decomposition:

$$\boldsymbol{A} = \sum_{i=1}^{d} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^{\mathsf{T}}$$

3. Projector

$$\boldsymbol{Q} = \sum_{i=1}^{d} \mathbf{1}_{\lambda_i \geq \lambda_*} \boldsymbol{u}_i \boldsymbol{u}_i^{\mathsf{T}}$$

4. Let $\mathcal{Q} \equiv \boldsymbol{Q} \otimes \boldsymbol{Q} \otimes \boldsymbol{Q}$ and return

$$\widehat{\boldsymbol{X}} = \frac{1}{\delta}\mathcal{Q}(\boldsymbol{X})$$

Similar method developed independently by Yuan, Xia, 2017

Back to image denoising

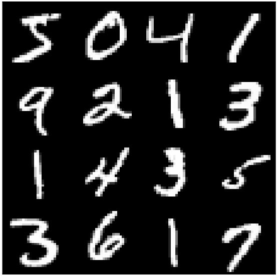
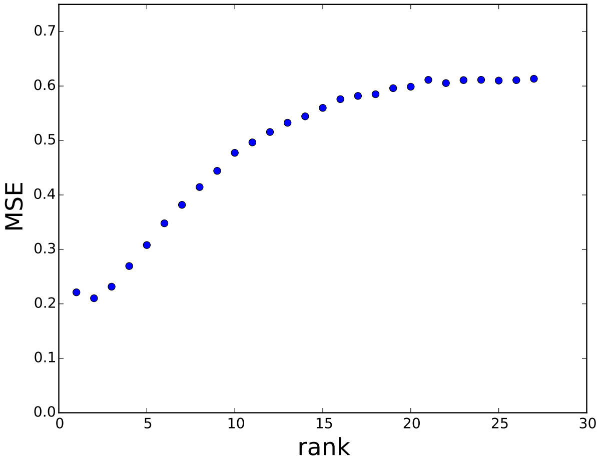
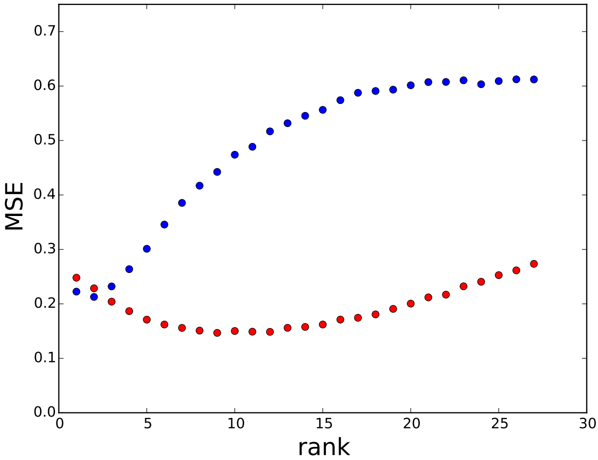


Image-by-image matrix denoising



:(

Tensor denoising



:)

Overcomplete tensors

$k = 3$

$$\boldsymbol{X} = \sum_{\ell=1}^{r} \boldsymbol{v}_\ell \otimes \boldsymbol{v}_\ell \otimes \boldsymbol{v}_\ell$$

$\boldsymbol{X}$ can have rank larger than d!!!

Can we use spectral methods?

$$Y = \Pi_E(X)$$
$$M = \mathrm{unfold}^{1\times 2}(Y)$$

Seems a lost cause:

$$\mathrm{rank}(X) \geq d \quad \Rightarrow \quad \mathrm{rank}(M) = d$$

Structure is lost!

Can we use spectral methods?

$$\boldsymbol{Y} = \Pi_E(\boldsymbol{X})$$
$$\boldsymbol{M} = \mathsf{unfold}^{1\times 2}(\boldsymbol{Y})$$

Seems a lost cause:

$$\text{rank}(\boldsymbol{X}) \geq d \quad \Rightarrow \quad \text{rank}(\boldsymbol{M}) = d$$

Structure is lost!

Can we use spectral methods?

$$\boldsymbol{Y} = \Pi_E(\boldsymbol{X})$$
$$\boldsymbol{M} = \mathsf{unfold}^{1\times 2}(\boldsymbol{Y})$$

Seems a lost cause:

$$\mathrm{rank}(\boldsymbol{X}) \geq d \quad \Rightarrow \quad \mathrm{rank}(\boldsymbol{M}) = d$$

Structure is lost!

Idea: Do something before spectral analysis

$$A = (A_{i_1 i_2\,;\,j_1 j_2})_{i_1,i_2,j_1,j_2 \le d} \in \mathbb{R}^{d^2 \times d^2},$$

$$A_{i_1 i_2\,;\,j_1 j_2} = \sum_{\ell=1}^{d} Y_{i_1 j_1 \ell} Y_{i_2 j_2 \ell}\,.$$

NOT EQUAL TO:

$$M = \mathsf{unfold}^{1\times 2}(Y) \in \mathbb{R}^{d \times d^2},$$

$$B = M^{\mathsf{T}} M$$

Claim: $\mathrm{rank}_*(A) \approx r, \quad \mathrm{rank}(B) \le d$

Idea: Do something before spectral analysis

$$\boldsymbol{A} = (A_{i_1 i_2\,;\,j_1 j_2})_{i_1,i_2,j_1,j_2 \le d} \in \mathbb{R}^{d^2 \times d^2}\,,$$

$$A_{i_1 i_2\,;\,j_1 j_2} \equiv \sum_{\ell=1}^{d} Y_{i_1 j_1 \ell}\, Y_{i_2 j_2 \ell}\,.$$

NOT EQUAL TO:

$$\boldsymbol{M} = \mathsf{unfold}^{1\times 2}(\boldsymbol{Y}) \in \mathbb{R}^{d \times d^2}\,,$$

$$\boldsymbol{B} = \boldsymbol{M}^{\mathsf{T}} \boldsymbol{M}$$

Claim: $\text{rank}_*(\boldsymbol{A}) \approx r, \quad \text{rank}(\boldsymbol{B}) \le d$

Idea: Do something before spectral analysis

$$\boldsymbol{A} = (A_{i_1 i_2\,;\,j_1 j_2})_{i_1,i_2,j_1,j_2 \le d} \in \mathbb{R}^{d^2 \times d^2}\,,$$
$$A_{i_1 i_2\,;\,j_1 j_2} \equiv \sum_{\ell=1}^{d} Y_{i_1 j_1 \ell} Y_{i_2 j_2 \ell}\,.$$

NOT EQUAL TO:

$$\boldsymbol{M} = \mathsf{unfold}^{1\times 2}(\boldsymbol{Y}) \in \mathbb{R}^{d\times d^2}\,,$$
$$\boldsymbol{B} = \boldsymbol{M}^{\mathsf{T}}\boldsymbol{M}$$

Claim: $\mathrm{rank}_*(\boldsymbol{A}) \approx r, \quad \mathrm{rank}(\boldsymbol{B}) \le d$

$\boldsymbol{X} = \sum_{\ell=1}^{r} \boldsymbol{v}_\ell^{\otimes 3}$

Theorem (Montanari, Sun, 2017)

Assume $(v_\ell)_{\ell \le r} \sim_{iid} \mathsf{N}(0, \mathrm{I}_d)$*, and* $r \le d^2$*. Then there exists a spectral algorithm achieving* $\|\widehat{\boldsymbol{X}} - \boldsymbol{X}\|_F \le \varepsilon \|\boldsymbol{X}\|_F$ *whp, for* n *random entries, provided*

$$n \ge C(\varepsilon)(d \vee r) d^{3/2} (\log d)^{c_0}$$

Similar ideas: Hopkins, Schramm, Shi, Steurer 2016; Raghavendra, Schramm 2016

$\boldsymbol{X} = \sum_{\ell=1}^{r} \boldsymbol{v}_\ell^{\otimes 3}$

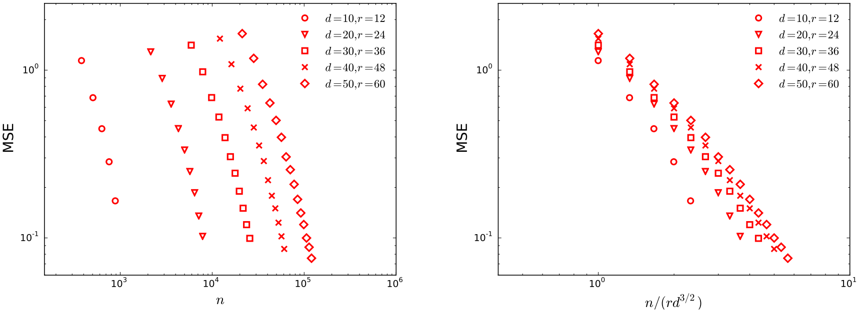
Theorem (Montanari, Sun, 2017)

Assume $(\boldsymbol{v}_\ell)_{\ell \le r} \sim_{iid} \mathsf{N}(0, \mathrm{I}_d)$, *and* $r \le d^2$. *Then there exists a spectral algorithm achieving* $\|\widehat{\boldsymbol{X}} - \boldsymbol{X}\|_F \le \varepsilon \|\boldsymbol{X}\|_F$ *whp, for* n *random entries, provided*

$$n \ge C(\varepsilon)(d \vee r) d^{3/2} (\log d)^{c_0}$$

Similar ideas: Hopkins, Schramm, Shi, Steurer 2016; Raghavendra, Schramm 2016

Simulations: $k = 3$, $r = 4$



Spiked tensor model (a.k.a. tensor PCA)

Reminder: A much simpler model

$$\boldsymbol{Y} = \lambda \boldsymbol{v}_0^{\otimes k} + \boldsymbol{W}$$

Signal: $\boldsymbol{v}_0 \in \mathsf{S}^{d-1} \equiv \{\boldsymbol{x} \in \mathbb{R}^d : \ \|\boldsymbol{x}\|_2 = 1\}$.

Noise: $(W_{i_1,i_2,\dots,i_k})_{i_1<i_2<\dots<i_k} \sim_{iid} \mathsf{N}(0, 1/d)$

SNR: λ

Given $\boldsymbol{Y}$, estimate $\boldsymbol{v}_0$

[Montanari, Richard, 2015]

A lot of information from statistical physics

Gibbs measure

$$\mu_{\beta,\lambda}(\mathrm{d}\boldsymbol{\theta}) = \frac{1}{Z(\beta,\lambda)} \exp\left\{\beta\langle \boldsymbol{Y}, \boldsymbol{\theta}^{\otimes k}\rangle\right\} \mu_0(\mathrm{d}\boldsymbol{\theta})$$

- $\beta = \infty$: Maximum Likelihood
- $\beta = \lambda/k!$: Bayes posterior
- $\mu_0(\,\cdot\,)$: Uniform measure on S^{d-1}

[Crisanti, Sommers, 1992, 1995; Auffinger, Ben Arous, Cerny, 2013; Chen, 2013; Subag, 2016; Krzakala, Lelarge, Miolane, Zdeborova, 2016;...]

Theorem

There exists $\lambda_{\rm Bayes}(k)$ (explicit!) such that:

$$\lim_{n\to\infty} \mathbb{E}|\langle v_0, v_{\rm Bayes}(\boldsymbol{Y})\rangle| = \begin{cases} 0 & \text{if } \lambda < \lambda_{\rm Bayes}(k), \\ > 0 & \text{if } \lambda > \lambda_{\rm Bayes}(k). \end{cases}$$

A lot of information from statistical physics

Gibbs measure

$$\mu_{\beta,\lambda}(\mathrm{d}\boldsymbol{\theta}) = \frac{1}{Z(\beta,\lambda)} \exp\left\{\beta\langle \boldsymbol{Y}, \boldsymbol{\theta}^{\otimes k}\rangle\right\} \mu_0(\mathrm{d}\boldsymbol{\theta})$$

- $\beta = \infty$: Maximum Likelihood
- $\beta = \lambda/k!$: Bayes posterior
- $\mu_0(\,\cdot\,)$: Uniform measure on S^{d-1}

[Crisanti, Sommers, 1992, 1995; Auffinger, Ben Arous, Cerny, 2013; Chen, 2013; Subag, 2016; Krzakala, Lelarge, Miolane, Zdeborova, 2016;...]

Theorem

There exists $\lambda_{\rm Bayes}(k)$ (explicit!) such that:

$$\lim_{n\to\infty} \mathbb{E}|\langle \boldsymbol{v}_0, \boldsymbol{v}_{\rm Bayes}(\boldsymbol{Y})\rangle| = \begin{cases} 0 & \text{if } \lambda < \lambda_{\rm Bayes}(k), \\ > 0 & \text{if } \lambda > \lambda_{\rm Bayes}(k). \end{cases}$$

What about polynomial-time estimators?

Theorem (Montanari, Richard, 2014; Hopkins, Shi, Steurer, 2015)

There exists poly-time estimator achieving $\mathbb{E}\{|\langle \hat{\boldsymbol{\theta}}^{\text{Poly}}, \boldsymbol{\theta}_0 \rangle|\} \geq 1 - \varepsilon$, *provided*

$$\lambda \geq C(\varepsilon) n^{(k-2)/4}.$$

No poly-time algorithm known for $1 \ll \lambda \ll n^{(k-2)/4}$.

What about polynomial-time estimators?

Theorem (Montanari, Richard, 2014; Hopkins, Shi, Steurer, 2015)

There exists poly-time estimator achieving $\mathbb{E}\{|\langle \hat{\boldsymbol{\theta}}^{\text{Poly}}, \boldsymbol{\theta}_0 \rangle|\} \geq 1 - \varepsilon$, *provided*

$$\lambda \geq C(\varepsilon) n^{(k-2)/4} .$$

No poly-time algorithm known for $1 \ll \lambda \ll n^{(k-2)/4}$.

What about polynomial-time estimators?

Theorem (Montanari, Richard, 2014; Hopkins, Shi, Steurer, 2015)

There exists poly-time estimator achieving $\mathbb{E}\{|\langle\hat{\boldsymbol{\theta}}^{\rm Poly},\boldsymbol{\theta}_0\rangle|\}\ge 1-\varepsilon$, *provided*

$$\lambda\ge C(\varepsilon)n^{(k-2)/4}\,.$$

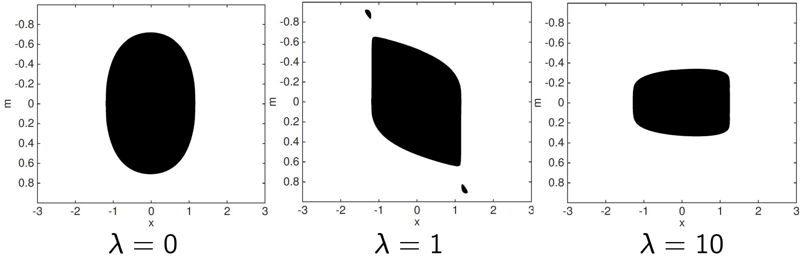
No poly-time algorithm known for $1\ll\lambda\ll n^{(k-2)/4}$.

A case study in failure: Maximum likelihood

$$\begin{aligned} \text{maximize} \quad & \langle \boldsymbol{Y}, \boldsymbol{\theta}^{\otimes k} \rangle \\ \text{subject to} \quad & \boldsymbol{\theta} \in \mathsf{S}^{d-1} \end{aligned}$$

$$N(x, m) \equiv \#\Big\{ \text{ critical points with } \langle \boldsymbol{Y}, \boldsymbol{\theta}^{\otimes k} \rangle \approx x,\ \langle v_0, \boldsymbol{\theta} \rangle \approx m \ \Big\}$$

A peek at complexity



$$\mathbb{E}N(x,m) = e^{d\,\Phi(x,m)+o(d)}\,,$$
$$\Phi(x,m) = \text{explicit expression}$$

[Ben Arous, Mei, Montanari, Nica, in progress]

Conclusion

Conclusion

- Tensors are useful for modeling multivariate data
- Estimation requires entirely new ideas
- Information-computation gap
- Many open problems

Thanks! Happy birthday Dave!

Conclusion

- Tensors are useful for modeling multivariate data
- Estimation requires entirely new ideas
- Information-computation gap
- Many open problems

Thanks! Happy birthday Dave!