Adaptive uncertainty quantification through adaptive and honest confidence sets

Alexandra Carpentier

Institut für Mathematik, Universität Potsdam
Funded by the DFG Emmy Noether grant CA 1488 1-1
MuSyAD

Van Dantzig Seminar, Amsterdam, Dec. 2th 2016

Scope of this talk

Talk :

- Adaptive and honest confidence sets : general presentation.
- Application in matrix completion.

Motivating example : Matrix completion

Problem :

Application : Recommendation system (e.g. Netflix).

Inference (estimation + uncertainty quantification) of the matrix?

Motivating example : Matrix completion

Problem :

Application : Recommendation system (e.g. Netflix).

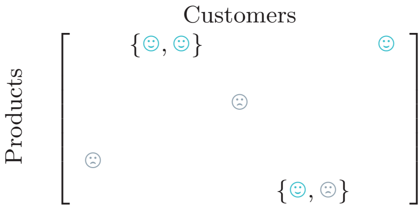
Inference (estimation + uncertainty quantification) of the matrix?

Trace Regression Model

f : matrix of dimension $d \times d$.
n observed data samples $(X_i, Y_i)_{i \leq n}$:

$$Y_i = f_{X_i} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $X_i \sim_{iid} \mathcal{U}_{\{1,\ldots,d\}^2}$ and ε is an indep. centered noise s. t. $|\varepsilon| \leq 1$.



Motivating example : Matrix completion

Problem :

Application : Recommendation system (e.g. Netflix).

Inference (estimation + uncertainty quantification) of the matrix?

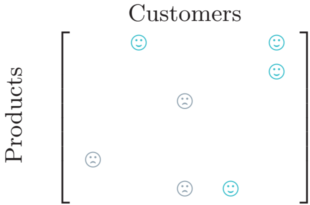
Bernoulli Model

f : matrix of dimension $d \times d$.

Data

$$Y_{i,j} = (f_{i,j}+\varepsilon_{i,j})B_{i,j}, \ \ (i,j) \in \{1,\ldots,d\}^2,$$

where $B_{i,j} \sim_{iid} \mathcal{B}(n/d^2)$ and ε is an indep. centered noise such that $|\varepsilon| \leq 1$.



Motivating example : Matrix completion

Problem :

Application : Recommendation system (e.g. Netflix).

Inference (estimation + uncertainty quantification) of the matrix?

High dimensional regime : $d^2 \geq n$.

Motivating example : Matrix completion

Problem :

Application : Recommendation system (e.g. Netflix).

Inference (estimation + uncertainty quantification) of the matrix?

High dimensional regime : $d^2 \geq n$.

Let for $1 \leq k \leq d$

$$\mathcal{C}(k) = \{f : \text{rank}(f) \leq k, \|f\|_\infty \leq 1\}.$$

Motivating example : Matrix completion

Problem :

Application : Recommendation system (e.g. Netflix).

Inference (estimation + uncertainty quantification) of the matrix?

High dimensional regime : $d^2 \geq n$.

Let for $1 \leq k \leq d$

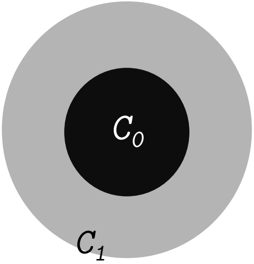
$$\mathcal{C}(k) = \{f : \text{rank}(f) \leq k, \|f\|_\infty \leq 1\}.$$

Question : If $f \in \mathcal{C}(k)$, then the "optimal" precision of inference should depend on k. Inference adaptive to k?

Adaptive estimation and confidence statements : See [Bickel, 1982], [Lepski and Spokoiny, 1997], [Cai and Low, 2004, 2006], [Robins and van der Vaart (2006)], [Massart, 2007], [Giné and Nickl, 2010], etc.

- "Large" sets $\mathcal{C}_0 \subset \mathcal{C}_1$ Low rank sets, smoothness classes, etc.
- Associated probability distributions $\mathbb{P}_f$ for $f \in \mathcal{C}_1$
- Receive a dataset of n i.i.d. entries according to $\mathbb{P}_f$

Adaptive inference : Adaptation to the set $\mathcal{C}_h$ when $f \in \mathcal{C}_h$, $h \in \{0, 1\}$.



Adaptive inference

Adaptive estimation and confidence statements : See [Bickel, 1982], [Lepski and Spokoiny, 1997], [Cai and Low, 2004, 2006], [Robins and van der Vaart (2006)], [Massart, 2007], [Giné and Nickl, 2010], etc.

- "Large" sets $\mathcal{C}_0 \subset \mathcal{C}_1$ Low rank sets, smoothness classes, etc.
- Associated probability distributions $\mathbb{P}_f$ for $f \in \mathcal{C}_1$
- Receive a dataset of n i.i.d. entries according to $\mathbb{P}_f$

Estimation :

- Minimax-optimal estimation errors r_0 (over $\mathcal{C}_0$) and r_1 (over $\mathcal{C}_1$) in $\|.\|$ norm

Minimax-opt. est. error

$$r_h = \inf_{\tilde{f} \text{ est.}} \sup_{f \in \mathcal{C}_h} \mathbb{E}_f \|\tilde{f} - f\|, \;\; h \in \{0, 1\}.$$

Minimax-optimal est. error in matrix completion over $\mathcal{C}(k)$:

$$\square d\sqrt{\frac{kd}{n}}.$$

See [Keshavan et al., 2009, Cai et al., 2010, Kolchinskii et al., 2011, Klopp and Gaiffas, 2015].

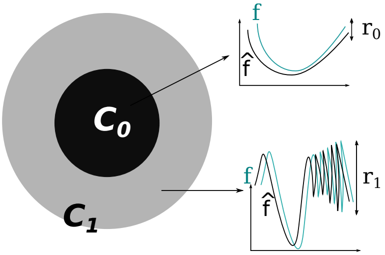
Adaptive inference

Adaptive estimation and confidence statements : See [Bickel, 1982], [Lepski and Spokoiny, 1997], [Cai and Low, 2004, 2006], [Robins and van der Vaart (2006)], [Massart, 2007], [Giné and Nickl, 2010], etc.

- "Large" sets $\mathcal{C}_0 \subset \mathcal{C}_1$ Low rank sets, smoothness classes, etc.
- Associated probability distributions $\mathbb{P}_f$ for $f \in \mathcal{C}_1$
- Receive a dataset of n i.i.d. entries according to $\mathbb{P}_f$

Adaptive estimation :

- Minimax-optimal estimation errors r_0 (over $\mathcal{C}_0$) and r_1 (over $\mathcal{C}_1$) in $\|.\|$ norm



Adaptive inference

Adaptive estimation and confidence statements : See [Bickel, 1982], [Lepski and Spokoiny, 1997], [Cai and Low, 2004, 2006], [Robins and van der Vaart (2006)], [Massart, 2007], [Giné and Nickl, 2010], etc.

- "Large" sets $\mathcal{C}_0 \subset \mathcal{C}_1$ Low rank sets, smoothness classes, etc.
- Associated probability distributions $\mathbb{P}_f$ for $f \in \mathcal{C}_1$
- Receive a dataset of n i.i.d. entries according to $\mathbb{P}_f$

Adaptive estimation :

- Minimax-optimal estimation errors r_0 (over $\mathcal{C}_0$) and r_1 (over $\mathcal{C}_1$) in $\|.\|$ norm
- In many models : adaptive estimator $\hat{f}$ exists

Adaptive estimation

$$\sup_{f \in \mathcal{C}_h} \mathbb{E}_f \|\hat{f} - f\| \leq \Box r_h, \quad \forall h \in \{0, 1\}.$$

Adaptive estimators exist in matrix completion See [Keshavan et al., 2009, Cai et al., 2010, Kolchinskii et al., 2011, Klopp and Gaiffas, 2015].

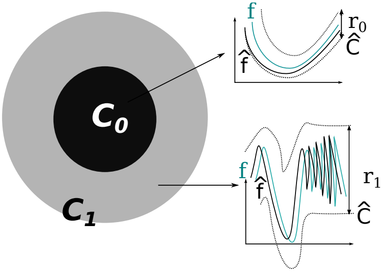
Adaptive inference

Adaptive estimation and confidence statements : See [Bickel, 1982], [Lepski and Spokoiny, 1997], [Cai and Low, 2004, 2006], [Robins and van der Vaart (2006)], [Massart, 2007], [Giné and Nickl, 2010], etc.

- "Large" sets $\mathcal{C}_0 \subset \mathcal{C}_1$ Low rank sets, smoothness classes, etc.
- Associated probability distributions $\mathbb{P}_f$ for $f \in \mathcal{C}_1$
- Receive a dataset of n i.i.d. entries according to $\mathbb{P}_f$

Adaptive and honest confidence sets :

- Minimax-optimal estimation errors r_0, r_1 in $\|.\|$ norm
- Confidence set $\hat{C}$: contains f and has adaptive diameter



Adaptive inference

Adaptive estimation and confidence statements : See [Bickel, 1982], [Lepski and Spokoiny, 1997], [Cai and Low, 2004, 2006], [Robins and van der Vaart (2006)], [Massart, 2007], [Giné and Nickl, 2010], etc.

- "Large" sets $\mathcal{C}_0 \subset \mathcal{C}_1$ Low rank sets, smoothness classes, etc.
- Associated probability distributions $\mathbb{P}_f$ for $f \in \mathcal{C}_1$
- Receive a dataset of n i.i.d. entries according to $\mathbb{P}_f$

Adaptive and honest confidence sets :

- Minimax-optimal estimation errors r_0, r_1 in $\|.\|$ norm
- Confidence set $\hat{C}$: contains f and has adaptive diameter

α-adapt. and honest conf. set

Honesty :

$$\sup_{f \in \mathcal{C}_1} \mathbb{P}_f(f \in \hat{C}) \geq 1 - \alpha.$$

Adaptivity :

$$\sup_{f \in \mathcal{C}_h} \mathbb{E}_f \|\hat{C}\| \leq \square r_h, \quad \forall h \in \{0, 1\}.$$

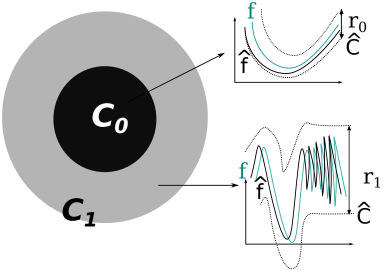
Uncertainty quantification

Adaptive and honest confidence statements : See [Ingster (1987, 1993)], [Spokoiny (1996)], [Cai and Low (2004, 2006)], [Robins and van der Vaart (2006)], [Giné and Nickl (2010)], etc.

- "Large" sets $\mathcal{C}_0 \subset \mathcal{C}_1$
- Associated probability distributions $\mathbb{P}_f$ for $f \in \mathcal{C}_1$
- Receive a dataset of n i.i.d. entries according to $\mathbb{P}_f$

Adaptive and honest confidence sets :

- Minimax-optimal estimation errors r_0, r_1 in $\|.\|$ norm
- Confidence set $\hat{C}$: contains f and has adaptive diameter



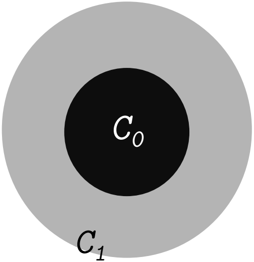
Uncertainty quantification

Adaptive and honest confidence statements : See [Ingster (1987, 1993)], [Spokoiny (1996)], [Cai and Low (2004, 2006)], [Robins and van der Vaart (2006)], [Giné and Nickl (2010)], etc.

- "Large" sets $\mathcal{C}_0 \subset \mathcal{C}_1$
- Associated probability distributions $\mathbb{P}_f$ for $f \in \mathcal{C}_1$
- Receive a dataset of n i.i.d. entries according to $\mathbb{P}_f$

A related testing problem :

$$H_0 : f \in \mathcal{C}_0 \quad \text{vs}$$
$$H_1 : f \in \mathcal{C}_1 \setminus \mathcal{C}_0.$$



Uncertainty quantification

Adaptive and honest confidence statements : See [Ingster (1987, 1993)], [Spokoiny (1996)], [Cai and Low (2004, 2006)], [Robins and van der Vaart (2006)], [Giné and Nickl (2010)], etc.

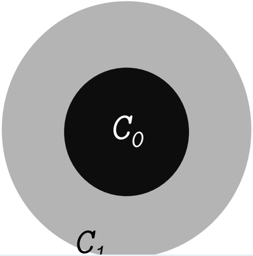
- "Large" sets $\mathcal{C}_0 \subset \mathcal{C}_1$
- Associated probability distributions $\mathbb{P}_f$ for $f \in \mathcal{C}_1$
- Receive a dataset of n i.i.d. entries according to $\mathbb{P}_f$

Uniform consistency :

$$H_0 : f \in \mathcal{C}_0 \quad \text{vs}$$
$$H_1 : f \in \mathcal{C}_1 \setminus \mathcal{C}_0.$$

α-unif. consistent test T

$$\sup_{f \in H_0} \mathbb{E}_f T + \sup_{f \in H_1} \mathbb{E}_f [1 - T] \leq 2\alpha.$$



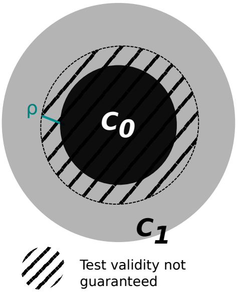
Uncertainty quantification

Adaptive and honest confidence statements : See [Ingster (1987, 1993)], [Spokoiny (1996)], [Cai and Low (2004, 2006)], [Robins and van der Vaart (2006)], [Giné and Nickl (2010)], etc.

- "Large" sets $\mathcal{C}_0 \subset \mathcal{C}_1$
- Associated probability distributions $\mathbb{P}_f$ for $f \in \mathcal{C}_1$
- Receive a dataset of n i.i.d. entries according to $\mathbb{P}_f$

Uniform consistency:

$$H_0 : f \in \mathcal{C}_0 \quad \text{vs}$$
$$H_\rho : f \in \mathcal{C}_1, \|f - \mathcal{C}_0\| \geq \rho.$$



Uncertainty quantification

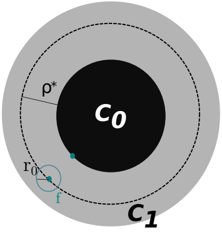
Adaptive and honest confidence statements : See [Ingster (1987, 1993)], [Spokoiny (1996)], [Cai and Low (2004, 2006)], [Robins and van der Vaart (2006)], [Giné and Nickl (2010)], etc.

- "Large" sets $\mathcal{C}_0 \subset \mathcal{C}_1$
- Associated probability distributions $\mathbb{P}_f$ for $f \in \mathcal{C}_1$
- Receive a dataset of n i.i.d. entries according to $\mathbb{P}_f$

Non-existence result :

Theorem (Hoffmann and Nickl, 2010)

If for $\rho \geq \square r_0$ there exists no α-uniformly consistent test, then α-adaptive and honest confidence sets do not exist.



NO

Uncertainty quantification

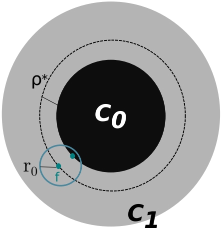
Adaptive and honest confidence statements : See [Ingster (1987, 1993)], [Spokoiny (1996)], [Cai and Low (2004, 2006)], [Robins and van der Vaart (2006)], [Giné and Nickl (2010)], etc.

- "Large" sets $\mathcal{C}_0 \subset \mathcal{C}_1$
- Associated probability distributions $\mathbb{P}_f$ for $f \in \mathcal{C}_1$
- Receive a dataset of n i.i.d. entries according to $\mathbb{P}_f$

Existence result :

Theorem (C., 2015, C., Klopp, Löffler and Nickl, 2016)

If for $\rho \leq \Box r_0$ there exists an α uniformly consistent test and if an oracle estimator exists, then α-adaptive and honest confidence sets exist.



OK

Results in some models

Setting	Sub-models	Norm	Existence iif...
Non-param. regression	Hölder smooth. $s_0 > s_1$	L_∞	Never [Baraud, 2004], [Cai and Low, 2004]
Density est. and non-param. reg.	Hölder smooth. $s_0 > s_1$	L_∞	Never [Robins and van der Vart, 2006]
Density estimation	Sobolev smooth. $s_0 > s_1$	L_2	$s_0 \geq s_1/2$ [Bull and Nickl, 2012]
Non-param. regression	Besov smooth. $s_0 > s_1$	L_p	$s_0 \geq s_1(1-1/p)$ [C., 2013]
Sparse reg. in dim. p, n samples	Sparsity $k_0 < k_1$	L_2	$k_0 \geq \square \min(\sqrt{p}, \sqrt{n})$ [Nickl and van de Geer, 2013]
Extreme value index : 1st order coeff	2nd order Pareto $\beta_0 < \beta_1$	$\|.\|$	Never [C. and Kim, 2015]

Results in some models

Setting	Sub-models	Norm	Existence iif...
Non-param. regression	Hölder smooth. $s_0 > s_1$	L_∞	Never [Baraud, 2004], [Cai and Low, 2004]
Density est. and non-param. reg.	Hölder smooth. $s_0 > s_1$	L_∞	Never [Robins and van der Vart, 2006]
Density estimation	Sobolev smooth. $s_0 > s_1$	L_2	$s_0 \geq s_1/2$ [Bull and Nickl, 2012]
Non-param. regression	Besov smooth. $s_0 > s_1$	L_p	$s_0 \geq s_1(1 - 1/p)$ [C., 2013]
Sparse reg. in dim. p, n samples	Sparsity $k_0 < k_1$	L_2	$k_0 \geq \square \min(\sqrt{p}, \sqrt{n})$ [Nickl and van de Geer, 2013]
Extreme value index : 1st order coeff	2nd order Pareto $\beta_0 < \beta_1$	$\lvert . \rvert$	Never [C. and Kim, 2015]

In all these models, adaptive and honest confidence sets typically do not exist over the entire parameter range.

Results in some models

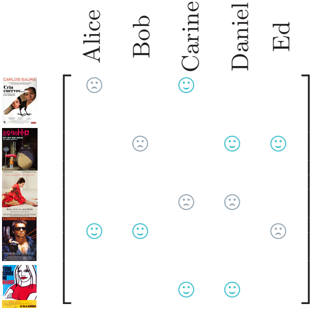
Setting	Sub-models	Norm	Existence iif...
Non-param. regression	Hölder smooth. $s_0 > s_1$	L_∞	Never [Baraud, 2004], [Cai and Low, 2004]
Density est. and non-param. reg.	Hölder smooth. $s_0 > s_1$	L_∞	Never [Robins and van der Vart, 2006]
Density estimation	Sobolev smooth. $s_0 > s_1$	L_2	$s_0 \geq s_1/2$ [Bull and Nickl, 2012]
Non-param. regression	Besov smooth. $s_0 > s_1$	L_p	$s_0 \geq s_1(1 - 1/p)$ [C., 2013]
Sparse reg. in dim. p, n samples	Sparsity $k_0 < k_1$	L_2	$k_0 \geq \Box \min(\sqrt{p}, \sqrt{n})$ [Nickl and van de Geer, 2013]
Extreme value index : 1st order coeff	2nd order Pareto $\beta_0 < \beta_1$	$\|.\|$	Never [C. and Kim, 2015]

Interesting research direction : restrict these spaces to large spaces where such confidence sets exist [Szabó, van der Vaart, van Zanten (2015)], [Ray, 2015], [Nickl and Szabó (2016)].

Matrix completion : Trace regression

Problem :

Application : Recommendation system (e.g. Netflix).



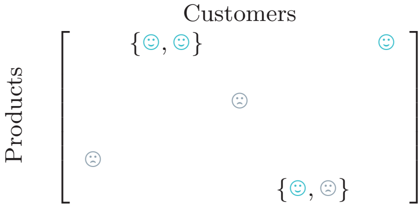
Inference (estimation + uncertainty quantification) of the matrix?

Trace Regression Model

f : matrix of dimension $d \times d$.
n observed data samples $(X_i, Y_i)_{i \le n}$:

$$Y_i = f_{X_i} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $X_i \sim_{iid} \mathcal{U}_{\{1,\ldots,d\}^2}$ and ε is an indep. centered noise s. t. $|\varepsilon| \le 1$.



Confidence sets : Trace Regression Model

Theorem (C., Klopp, Löffler and Nickl, 2016)

In the matrix completion "trace regression" model, α-adaptive and honest confidence sets exist over the entire range of parameters (i.e. for any $1 \leq k_0 < k_1 \leq d$).

Dimension reduction in the smaller model not too radical.

Idea of the proof : known variance

Let $\hat{f}$ be a minimax estimator s.t. $\mathrm{rank}(\hat{f}) \leq \mathrm{rank}(f)$ whp. Set

$T_n = \dfrac{\|Y - \mathcal{X}\hat{f}\|^2}{n} - \mathbb{V}(\epsilon)$, where $\mathcal{X}$ is the sampling operator.

We have whp and knowing $\tilde{\theta}$

$$|T_n - \|f - \hat{f}\|^2| \lesssim d^2 \frac{(\mathrm{rank}(\hat{f}) + 1)d}{n}.$$

If $\theta \in \mathcal{C}(k_0)$:

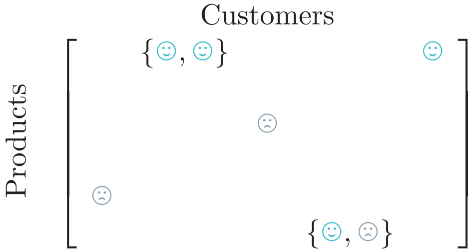
$$\hat{f} \in \mathcal{C}(k_0) \quad \text{and} \quad T_n \lesssim d^2 \frac{(k_0 + 1)d}{n} \simeq r_{k_0}^2.$$

If $r_{k_0} \lesssim \|f - \mathcal{C}(k_0)\|$:

$$\hat{f} \notin \mathcal{C}(k_0) \quad \text{or} \quad r_{k_0}^2 - d^2 \frac{(k_0 + 1)d}{n} \lesssim r_{k_0}^2 \lesssim T_n.$$

This concludes the proof accepting the test if either $\hat{f} \in \mathcal{C}(k_0)$ or if $T_n \lesssim r_{k_0}^2$: $\rho^2 \leq r_{k_0}^2$.

Idea of the proof : unknown variance



More than $\frac{n^2}{d^2} \wedge n$ entries sampled twices, i.e. N. Set $(\tilde{X}_i, Z_i, Z_i')_{i \leq N}$

Replace T_n by

$$T_n = \frac{1}{N} \sum_{i \leq N} (Z_i - \tilde{\mathcal{X}} \hat{f})(Z_i' - \tilde{\mathcal{X}} \hat{f}).$$

We have as before whp and knowing $\tilde{\theta}$

$$|T_n - \|f - \hat{f}\|^2| \lesssim d^2 \frac{(\mathrm{rank}(\hat{f}) + 1)d}{n}.$$

This conclude the proof : $\rho^2 \leq r_{k_0}^2$.

Matrix completion : Bernoulli Model

Problem :

Application : Recommendation system (e.g. Netflix).

Inference (estimation + uncertainty quantification) of the matrix?

Bernoulli Model

f : matrix of dimension $d \times d$.
Data

$$Y_{i,j} = (f_{i,j} + \varepsilon_{i,j}) B_{i,j}, \ \ (i,j) \in \{1, \ldots, d\}^2,$$

where $B_{i,j} \sim_{iid} \mathcal{B}(n/d^2)$ and ε is an indep. centered noise such that $|\varepsilon| \leq 1$.

Confidence sets : Bernoulli Model

Theorem (C., Klopp, Löffler and Nickl, 2016)

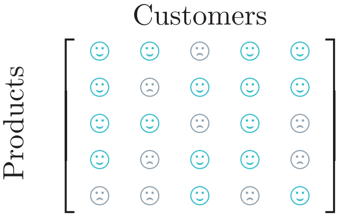
- **Bernoulli Model with known noise variance** : *Adaptive and honest confidence sets exist.*
- **Bernoulli Model with unknown noise variance** : *Adaptive and honest confidence sets do not exist (unless maybe $k_0^2 \geq k_1$? Open question).*

The two models are not equivalent in this case!

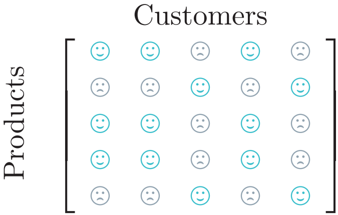
(Simplified) Idea of the proof : Unknown variance

No entries sampled twice! First example : rank one

H_0 : Random opinions!



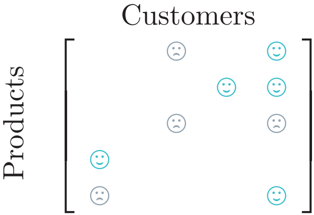
H_1 : Rank one opinions.



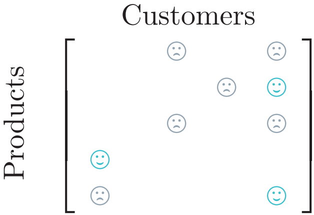
(Simplified) Idea of the proof : Unknown variance

No entries sampled twice! First example : rank one

H_0 : Random opinions!



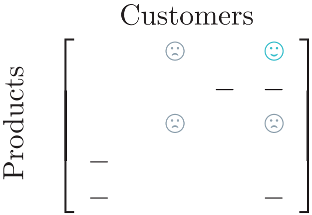
H_1 : Rank one opinions.



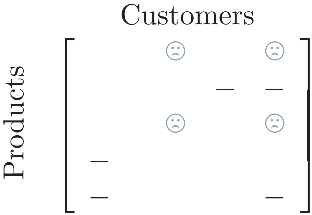
(Simplified) Idea of the proof : Unknown variance

No entries sampled twice! First example : rank one

H_0 : Random opinions!



H_1 : Rank one opinions.



(Simplified) Idea of the proof : Unknown variance

No entries sampled twice! First example : rank one

H_0 : Random opinions!

Customers

Products
$$\begin{bmatrix} & \text{☹} & & \text{☺} \\ & & - & - \\ & \text{☹} & & \text{☹} \\ - & & & \\ - & & & - \end{bmatrix}$$

H_1 : Rank one opinions.

Customers

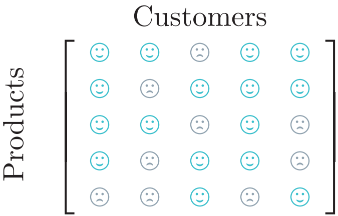
Products
$$\begin{bmatrix} & \text{☹} & & \text{☹} \\ & & - & - \\ & \text{☹} & & \text{☹} \\ - & & & \\ - & & & - \end{bmatrix}$$

Less than $\frac{n^4}{d^4}$ such cycles whp $\rightarrow$ distinguishability only if $n \gg d$.

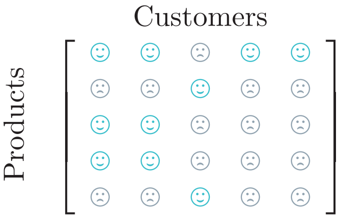
(Simplified) Idea of the proof : Unknown variance

No entries sampled twice! General case : rank k

H_0 : Random opinions!



H_1 : Rank one opinions.

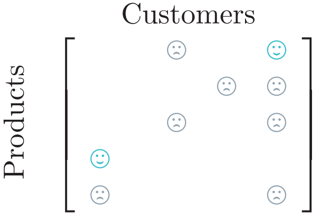


(Simplified) Idea of the proof : Unknown variance

No entries sampled twice! General case : rank k

H_0 : Random opinions!

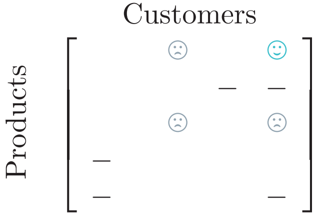
H_1 : Rank one opinions.



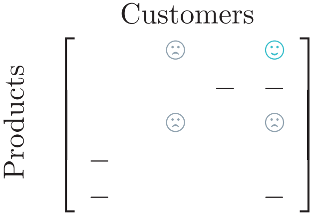
(Simplified) Idea of the proof : Unknown variance

No entries sampled twice! General case : rank k

H_0 : Random opinions!



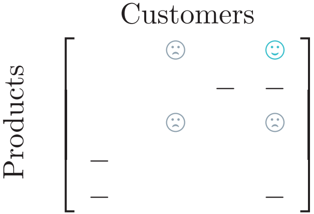
H_1 : Rank one opinions.



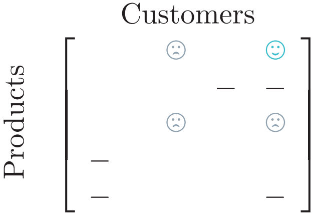
(Simplified) Idea of the proof : Unknown variance

No entries sampled twice! General case : rank k

H_0 : Random opinions!



H_1 : Rank one opinions.



Less than $\frac{n^4}{d^4 k^3}$ *correct* cycles (taking rank groups into account)
$\rightarrow$ distinguishability only if $n \gg k^{3/4} d$.

Conclusion on uncertainty quantification

Conclusions :

- Strong link between model testing and adaptive uncertainty quantification.
- Adaptive estimators and and honest confidence sets : while adaptive estimation is often possible in standard models, adaptive uncertainty quantification is often impossible.
- Matrix completion, trace regression model : adaptive and honest confidence sets exist in some models! Sensitivity of adaptive and honest confidence sets on the model assumptions.

Relevant open question : exact model testing. Complexity of the null hypothesis?

Thank you!