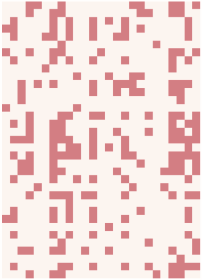
technische universität
dortmund

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

The Trustworthy Pal: Controlling the False Discovery Rate in Boolean Matrix Factorization

Sibylle Hess, Katharina Morik and Nico Piatkowski

Given a Binary Data Matrix..

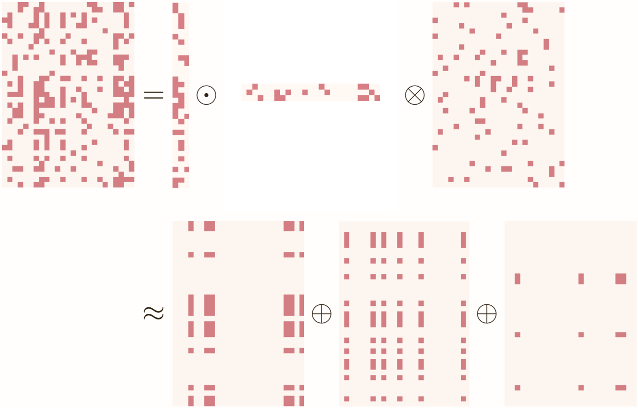


Depending on the data domain, one could ask for

- groups of users which like the same set of movies,
- groups of patients having a similar set of gene mutations,
- groups of costumers buying a similar set of items.

Finding a Factorization

Solve $\min_{X,Y} |D - Y \odot X^\top|$ for binary matrices X and Y



The False Discovery Rate (FDR)

> Boy called wolf once – Type 1 error
> Boy called wolf twice – Type 2 error

The ***FDR*** is ***controlled*** at level q if

$$\mathbb{E}\left(\frac{v}{r}\right) \leq q$$

v : false alarms
r : alarms in total

False Discoveries and BMF

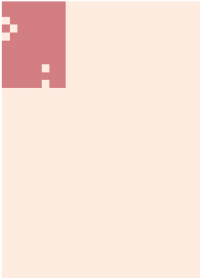
Given a factorization of rank r, define

$$Z_s = \begin{cases} 1 \text{ if } Y_{\cdot s} X_{\cdot s}^\top \text{ covers (mostly) noise} \\ 0 \text{ if } Y_{\cdot s} X_{\cdot s}^\top \text{ is a part of the model} \end{cases}$$

The ***FDR*** is ***controlled*** at level q if $\mathbb{P}(Z_s = 1) < q$:

$$\mathbb{E}\left(\frac{v}{r}\right) = \frac{1}{r} \sum_{s=1}^{r} \mathbb{P}(Z_s = 1)$$

Properties of Outer Products



If the binary matrices solve

$$(X, Y) \in \arg\min |D - Y \odot X^\top|$$

then any outer product $A = D \circ Y_{\cdot s} X_{\cdot s}^\top$ has a ***high density***

$$\delta = \frac{Y_{\cdot s}^\top D X_{\cdot s}}{|X_{\cdot s}||Y_{\cdot s}|} \geq \frac{1}{2}$$

and a ***high coherence***

$$\eta = \max_{1 \leq i \neq k \leq n} \langle A_{\cdot i}, A_{\cdot k} \rangle > \delta |Y_{\cdot s}| \frac{\delta |X_{\cdot s}| - 1}{|X_{\cdot s}| - 1}$$

Theorem (Density Bound)

Suppose N is an $m \times n$ Bernoulli matrix with parameter p. The probability that a δ-dense tile of size $|x| \geq a$ and $|y| \geq b$ exists is no larger than

$$\binom{n}{a}\binom{m}{b}\exp(-2ab(\delta - p)^2) \; . \tag{1}$$

Proof (sketch): Hoeffding's inequality yields

$$\begin{aligned} \mathbb{P}\left(\frac{y^\top N x}{|x||y|} \geq \delta\right) &= \mathbb{P}\left(\left(\frac{1}{ab}\sum_{i,j} x_i y_j N_{ji}\right) - p \geq \delta - p\right) \\ &\leq \exp(-2ab(\delta - p)^2), \end{aligned}$$

The union bound yields the final result.

Theorem (Coherence Bound)

Let N be an $m \times n$ Bernoulli matrix with parameter p and let $\mu > p^2$. The function value of η satisfies $\eta\left((1/\sqrt{m})N\right) \geq \mu$ with probability no larger than

$$\frac{n(n-1)}{2} \exp\left(-\frac{3}{2} m \frac{(\mu - p^2)^2}{2p^2 + \mu}\right). \tag{2}$$

Proof (sketch): The Bernstein inequality yields

$$\mathbb{P}(\langle N_{\cdot i}, N_{\cdot k}\rangle \geq m\mu) \leq \exp\left(-\frac{3}{2} m \frac{(\mu - p^2)^2}{2p^2 + \mu}\right) \ ,$$

The union bound yields the final result.

Applying the Bounds

Given $Y \odot X^\top \approx D$, calculate for $1 \leq s \leq r$ the ***density*** δ_s and ***coherence*** η_s.

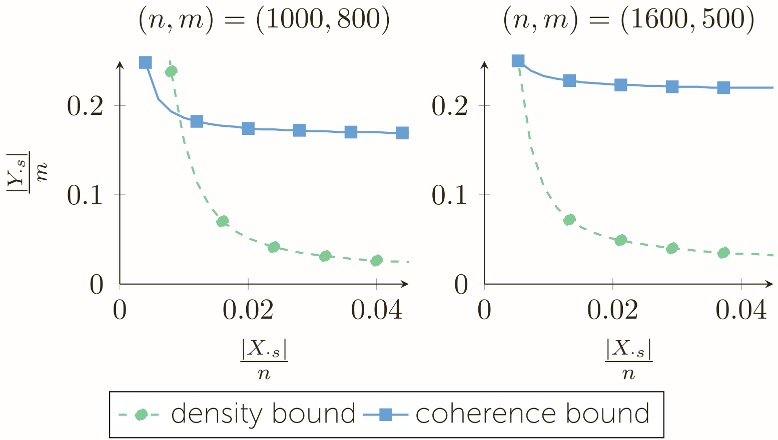
Toss $(X_{\cdot s}, Y_{\cdot s})$ if all of the following bounds are ***larger than*** q.

Corollary

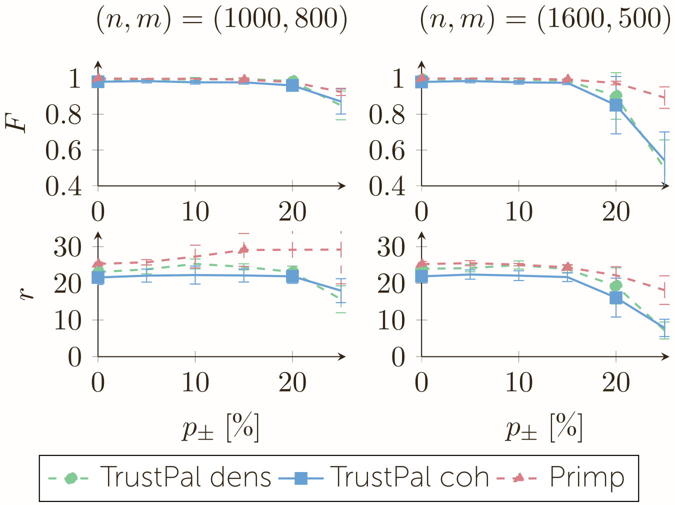
$$\mathbb{P}(Z_s = 1) \leq \binom{n}{|X_{\cdot s}|}\binom{m}{|Y_{\cdot s}|} \exp(-2|X_{\cdot s}||Y_{\cdot s}|(\delta_s - p)^2)$$

$$\mathbb{P}(Z_s = 1) \leq \exp\left(-\frac{3}{2}m\frac{(\eta_s/m - p^2)^2}{2p^2 + \eta_s/m}\right)$$

How Good are These Bounds?



Synthetic Experiments in PalTiling



Take Home

- Establishing ***quality guarantees*** for ***unsupervised*** tasks is pretty interesting
- ***Concentration inequalities*** (e.g., Hoeffding bound) are powerful stuff,
- the ***binary*** nature of BMF enables satisfying ***solutions*** to problems which are much harder in fuzzier tasks like NMF