SPARSE BOOLEAN MATRIX FACTORIZATIONS

Pauli Miettinen

15.12.2010

max planck institut
informatik

BOOLEAN FACTORIZATIONS

- Input: a 0/1 (i.e. Boolean) n-by-m matrix $\mathbf{A}$ and integer k (i.e. the rank)
- Output: 0/1 n-by-k matrix $\mathbf{B}$ and 0/1 k-by-m matrix $\mathbf{C}$
- Goal: minimize $\sum_{i,j}|\mathbf{A}_{ij} - (\mathbf{B}\circ\mathbf{C})_{ij}|$
  - Boolean matrix multiplication: $(\mathbf{B}\circ\mathbf{C})_{ij} = \vee_p \mathbf{B}_{ip}\mathbf{C}_{pj}$
  - Like normal, but addition defined as 1+1=1

SOME EXITING PROPERTIES

- Easy to interpret
- Generalizes many data mining techniques
- Boolean rank can be exponentially smaller than normal rank
  - Boolean factorizations can have less error than SVD
- Computations become combinatorial

SOME BAD NEWS

- Computations become combinatorial
- Finding optimal Boolean factorizations is computationally hard
- Hard inapproximability results for:
  - best Boolean rank-*k* factorization of a given matrix
  - Boolean rank of a given matrix
    - As hard as finding graph's minimum chromatic number

GOOD NEWS

- Sparsity helps!

SPARSE FACTORIZATIONS

- Ideally, sparse matrices have sparse factors
    - Not true with many factorization methods
- Sparse Boolean matrices have sparse decompositions

SPARSE FACTORIZATIONS

- Ideally, sparse matrices have sparse factors
  - Not true with many factorization methods
- Sparse Boolean matrices have sparse decompositions

Theorem 1. For any n-by-m 0/1 matrix $\mathbf{A}$ of Boolean rank k, there exist n-by-k and k-by-m 0/1 matrices $\mathbf{B}$ and $\mathbf{C}$ such that $\mathbf{A}=\mathbf{B}\circ\mathbf{C}$ and
$|\mathbf{B}|+|\mathbf{C}|\leq 2|\mathbf{A}|$.

APPROXIMATING THE BOOLEAN RANK

- Sparsity is not enough; we need some structure in it
- An *n*-by-*m* 0/1 matrix **A** is *f(n)*-uniformly sparse, if all of its columns have at most *f(n)* 1s

Theorem 2. The Boolean rank of log(*n*)-uniformly sparse matrix can be approximated to within $O(\log(m))$ in time $\tilde{O}(m^2 n)$.

max planck institut
informatik

NON-UNIFORMLY SPARSE MATRICES

- Uniform sparsity is very restricted; what can we do
  - Trade non-uniformity with approximation accuracy

NON-UNIFORMLY SPARSE MATRICES

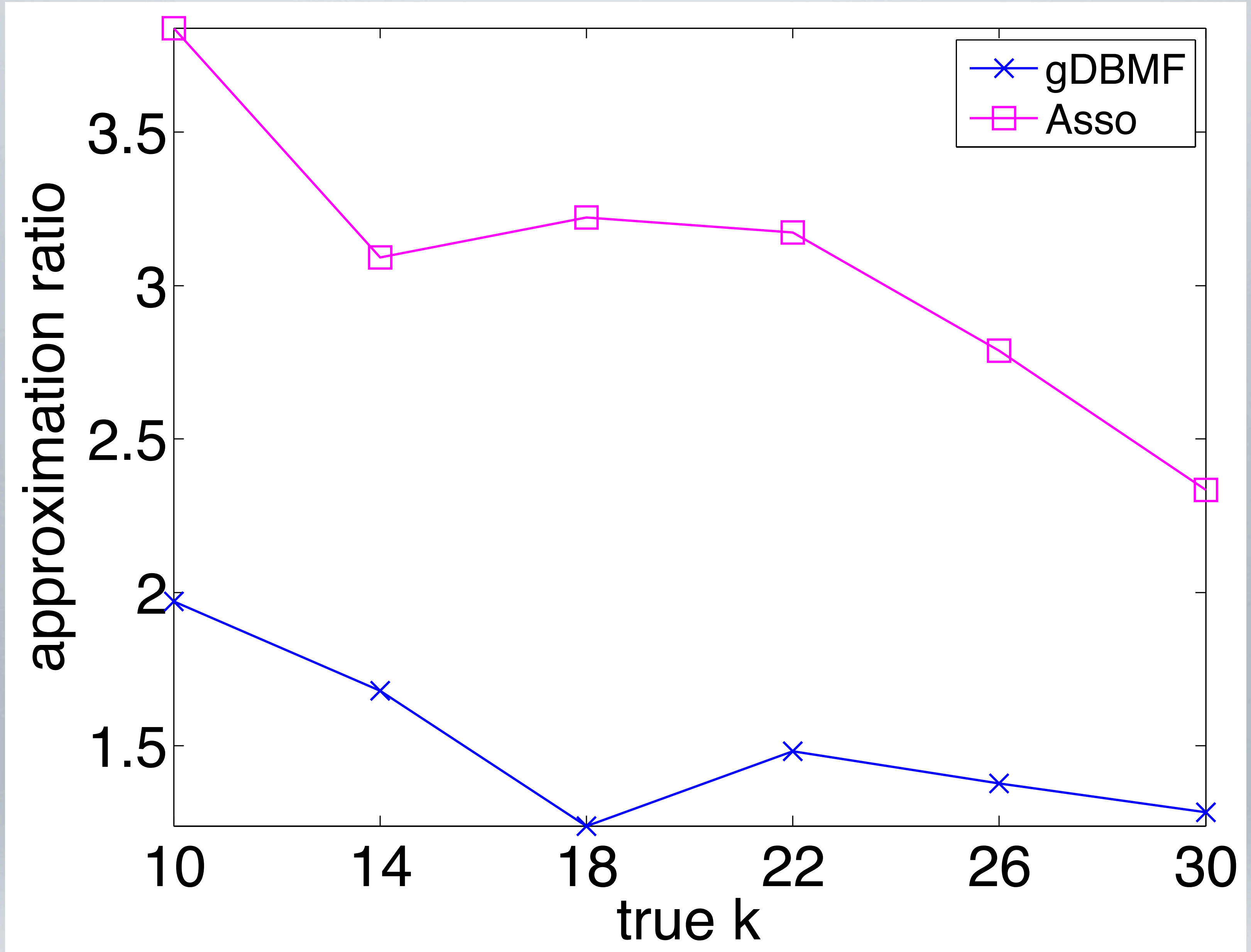
- Uniform sparsity is very restricted; what can we do
  - Trade non-uniformity with approximation accuracy

Theorem 3. If there are at most $\log(m)$ columns with more than $\log(n)$ 1s, then we can approximate the Boolean rank in polynomial time to within $O(\log^2(m))$.

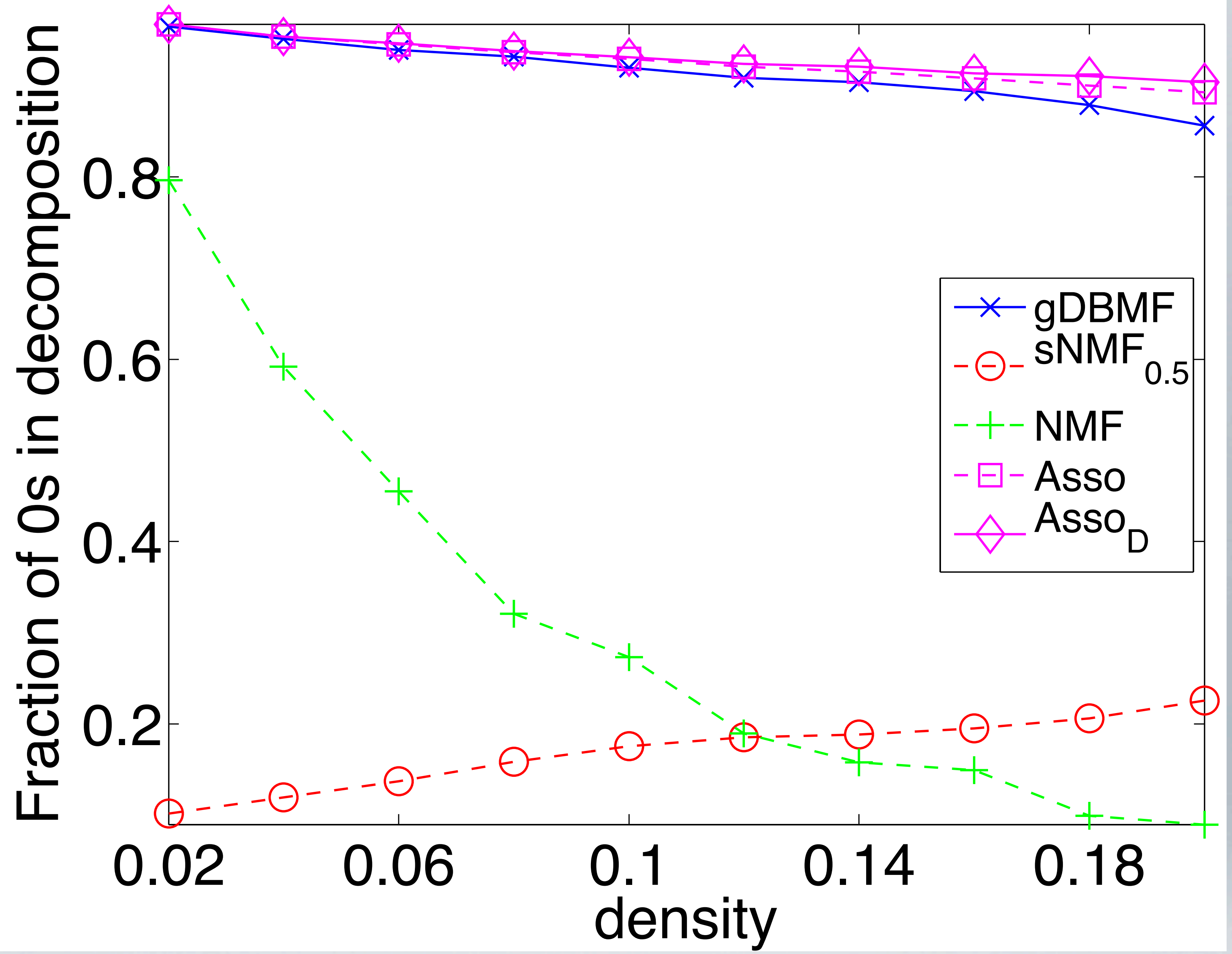
APPROXIMATING DOMINATED COVERS

Theorem 4. If n-by-m 0/1 matrix A is O(log n)-uniformly sparse, we can approximate the best dominated *k*-cover of A by e/(e-1) in polynomial time.

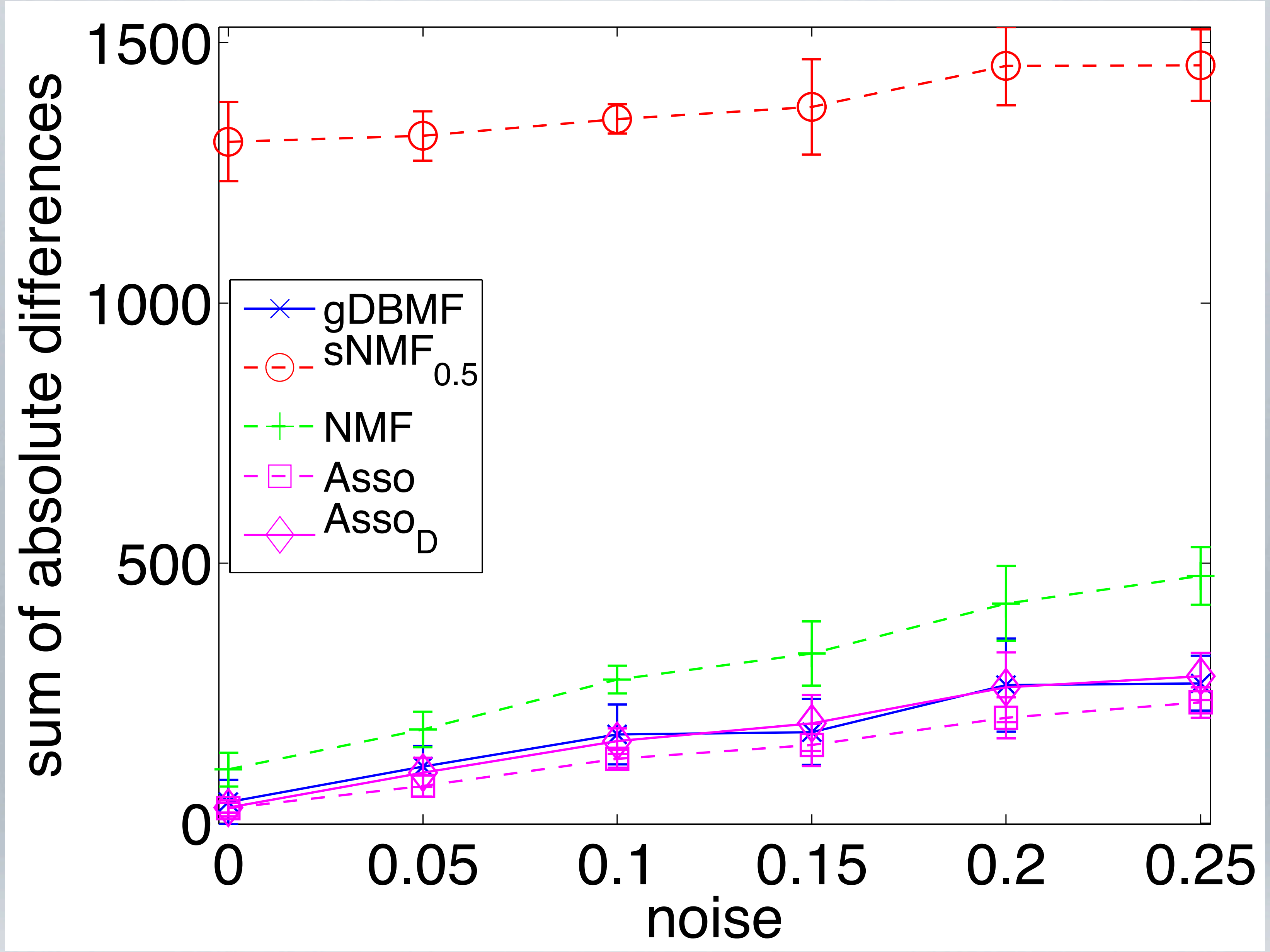
- Dominated *k*-cover: The rank is *k* and if $(\mathbf{B} \circ \mathbf{C})_{ij} = 1$, then $\mathbf{A}_{ij} = 1$
  - Has applications e.g. in role mining



APPROXIMATING THE RANK


Fraction of 0s in decomposition
0.8
0.6
0.4
0.2
0.02
0.06
0.1
0.14
0.18
density
gDBMF
sNMF0.5
NMF
Asso
AssoD

SPARSITY



APPROXIMATION ERROR

max planck institut
informatik

CONCLUSIONS

- Sparse Boolean matrices have sparse decompositions
  - Not true with “normal” decompositions
- Sparsity helps with computational complexity
  - Requires some regularity in sparsity
- Initial work; better results to be expected.

CONCLUSIONS

- Sparse Boolean matrices have sparse decompositions
  - Not true with "normal" decompositions
- Sparsity helps with computational complexity
  - Requires some regularity in sparsity
- Initial work; better results to be expected.

Thank You!