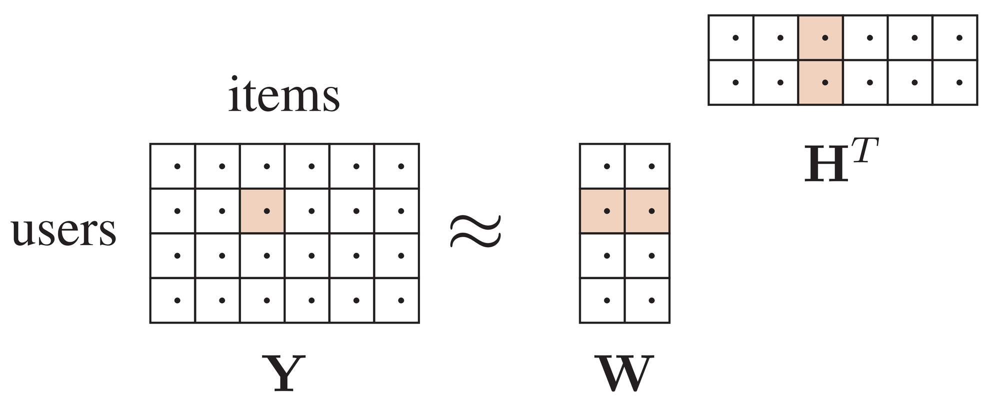
RECOMMENDATION FROM RAW DATA WITH ADAPTIVE COMPOUND POISSON FACTORIZATION

{ OLIVIER.GOUVERT, THOMAS.OBERLIN, CEDRIC.FEVOTTE }@IRIT.FR

INTRODUCTION

Collaborative filtering based on a matrix $\mathbf{Y}$: listening counts of U users on I items.



Poisson factorization (PF) (Gopalan et al., 2015): each entry is drawn from $y_{ui} \sim \text{Poisson}([\mathbf{WH}^T]_{ui})$.
- $\mathbf{W}$ of size $U \times K$: preferences of users
- $\mathbf{H}$ of size $I \times K$: attributes of items

Limitation: $var(y_{ui}|\mathbf{W}, \mathbf{H}) = \mathbb{E}[y_{ui}|\mathbf{W}, \mathbf{H}]$.

DCPF MODEL

Discrete compound Poisson factorization (dcPF) models over-dispersion through the *self-excitation* concept.

Generative model: (Basbug et al., 2016)

$$w_{uk} \sim \mathcal{G}(\alpha^W, \beta_u^W),\ h_{ik} \sim \mathcal{G}(\alpha^H, \beta_i^H),$$
$$n_{ui} \sim \text{Poisson}\left([\mathbf{WH}^T]_{ui}\right),$$
$$x_{l,ui} \sim \text{ED}(\theta, \kappa),\ \forall l \in \{1, \dots, n_{ui}\},$$
$$y_{ui} = \sum_{l=1}^{n_{ui}} x_{l,ui}.$$

where ED is the exponential dispersion family, defined by: $p(x; \theta, \kappa) = \exp(x\theta - \kappa^T \psi(\theta))h(x, \kappa)$, where θ is the *natural parameter* and κ is the *dispersion parameter*.

Assumption: $x_{l,ui} \in \{1, 2, \dots, +\infty\}$.

Interpretation of variable N: Number of listening sessions. During each session, a user can listen to an item at least one time.

Joint log-likelihood:

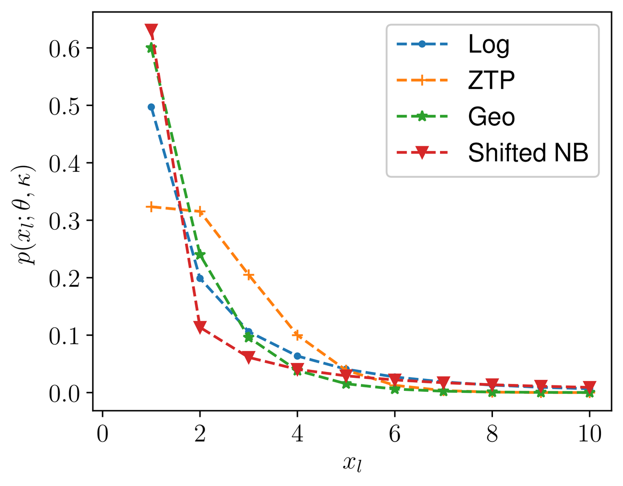
$$\log p(\mathbf{Y}, \mathbf{N}, \mathbf{W}, \mathbf{H}) = \underbrace{\log p(\mathbf{Y}|\mathbf{N}; \theta, \kappa)}_{\text{Mapping}}$$
$$+ \underbrace{\log p(\mathbf{N}|[\mathbf{WH}^T])}_{\text{PF structure}} + \underbrace{\log p(\mathbf{W}, \mathbf{H})}_{\text{Regularization}}.$$

Properties:
- Scalability: $y_{ui} = 0 \Leftrightarrow n_{ui} = 0$. $\mathbf{N}$ is partially known and has the same zeros as $\mathbf{Y}$.
- Closed-form updates for the variable $\mathbf{N}$: $n_{ui} \leq y_{ui}$, $\mathbf{N}$ can take a finite number of values.

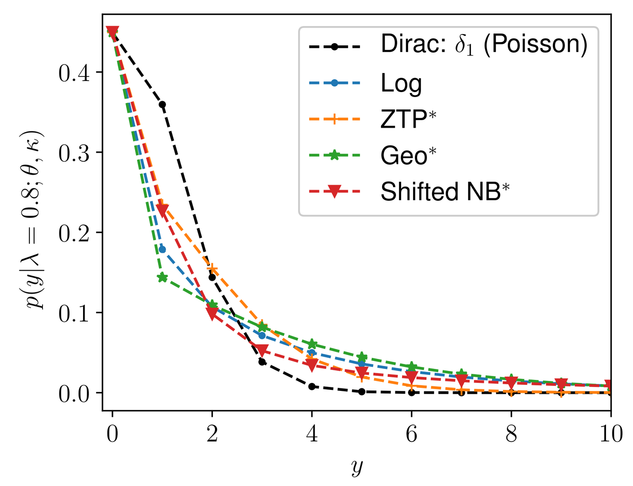
EXAMPLES OF DISTRIBUTIONS

Distribution	θ	Θ	θ^{raw}	θ^{bin}	κ	$\psi(\theta)$	$h(x,\kappa)$
$x_l \sim \text{Log}(p)$	$\log(p)$	$\mathbb{R}_-^*$	$-\infty$	0	1	$\log(-\log(1-e^\theta))$	$\frac{x!}{\kappa!} St_1(x,\kappa)$
$x_l \sim \text{ZTP}(p)$	$\log(p)$	$\mathbb{R}$	$-\infty$	$+\infty$	1	$\log(e^{e^\theta} - 1)$	$\frac{x!}{\kappa!} St_2(x,\kappa)$
$x_l \sim \text{shGeo}(1-p)$	$\log(p)$	$\mathbb{R}_-^*$	$-\infty$	0	1	$\log(\frac{e^\theta}{1-e^\theta})$	$\frac{x!}{\kappa!} St_3(x,\kappa)$
$x_l - 1 \sim \text{NB}(a,p)$	$\log(p)$	$\mathbb{R}_-^*$	$-\infty$	0	$(1,a)^T$	$(\theta, -\log(1-e^\theta))^T$	$\frac{\Gamma(x-\kappa_1+\kappa_2)}{\Gamma(x-\kappa_1+1)\Gamma(\kappa_2)}$

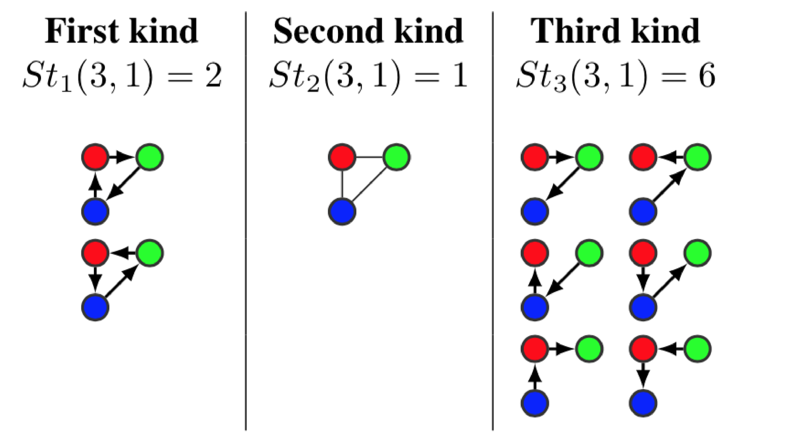
Distribution of x_l



Marginal distribution



Stirling numbers



A TRADE-OFF BETWEEN PF APPLIED TO RAW AND BINARIZED DATA

Proposition 1. *If there exists θ^{raw} such that $\lim_{\theta \to \theta^{raw}} \kappa^T \psi(\theta) = -\infty$, then the posterior of dcPF tends to the posterior of PF as θ goes to θ^{raw}.*

Proposition 2. *If there exists θ^{bin} such that $\lim_{\theta \to \theta^{bin}} \kappa^T \psi(\theta) = +\infty$, then the posterior of dcPF tends to the posterior of PF applied to binarized data as θ goes to θ^{bin}, i.e.: $\lim_{\theta \to \theta^{bin}} p(\mathbf{W}, \mathbf{H}|\mathbf{Y}) = p(\mathbf{W}, \mathbf{H}|\mathbf{N} = \mathbf{Y}^b)$, where $y_{ui}^b = \mathbb{1}[y_{ui} > 0]$.*

Adaptivity of dcPF to over-dispersion: the natural parameter $\theta = \log p$ is strongly correlated to the variance-mean ratio.

Dataset	NIPS	TP	Last.fm
mean of non-zeros	2.7	2.7	3.9
var of non-zeros	20.9	25.9	65.7
ratio var/mean	7.6	9.8	17.0
Log - p	0.74	0.80	0.90
ZTP - p	1.40	1.95	2.35
Geo - p	0.51	0.60	0.69
sh. NB - p	0.86	0.87	0.90
sh. NB - κ_2	0.17	0.21	0.27

EXPERIMENTS

Dataset: Taste Profile (TP) dataset, subset size: $16k$ users and $12k$ items

Splitting: 80% of non-zeros for training; 20% for testing.

Evaluation score: Normalized discount cumulative gain (NDCG) with threshold s.

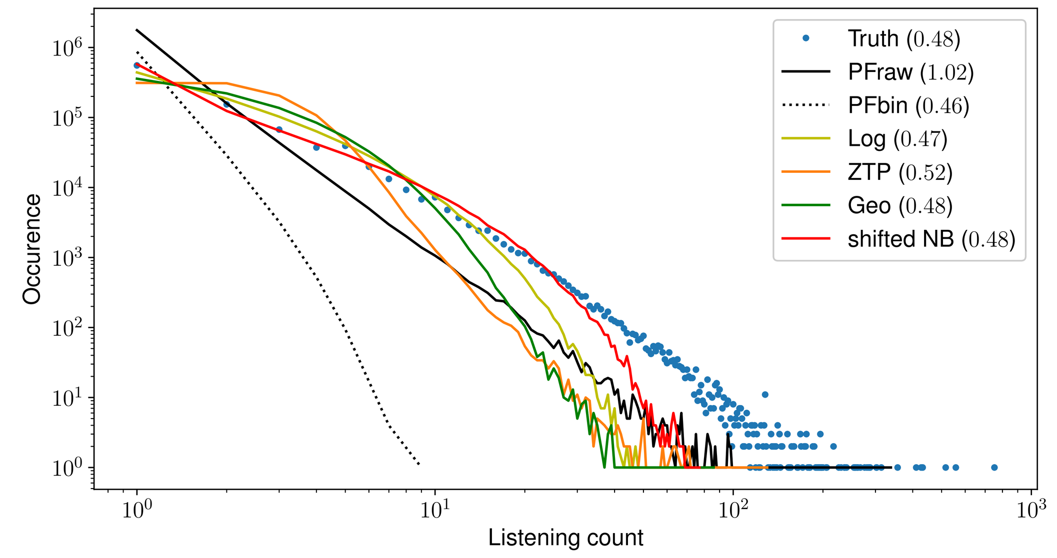
Algorithms:
- Variational Bayes algorithm with maximum likelihood estimation of the parameters κ and θ (VBEM).
- Variational Bayes algorithm with a grid search on θ.

Compared methods: limit cases PFraw and PFbin.

Model	Est.	$p = e^\theta$	κ	NDCG0	NDCG1	NDCG2	NDCG5
Log	VBEM	0.803	1	**0.200** ($\pm$**3.0** $\mathbf{10^{-3}}$)	**0.182** ($\pm$**2.3** $\mathbf{10^{-3}}$)	**0.166** ($\pm$**2.0** $\mathbf{10^{-3}}$)	0.147 ($\pm$1.5 10^{-3})
	Grid	*0.3*	1	*0.200 ($\pm$4.1 10^{-3})*	*0.186 ($\pm$3.9 10^{-3})*	*0.173 ($\pm$3.7 10^{-3})*	*0.158 ($\pm$3.7 10^{-3})*
ZTP	VBEM	1.950	1	0.192 ($\pm$4.1 10^{-3})	0.178 ($\pm$3.7 10^{-3})	**0.167** ($\pm$**3.6** $\mathbf{10^{-3}}$)	**0.156** ($\pm$**3.8** $\mathbf{10^{-3}}$)
	Grid	*1*	1	*0.190 ($\pm$3.5 10^{-3})*	*0.178 ($\pm$3.0 10^{-3})*	*0.168 ($\pm$3.1 10^{-3})*	*0.158 ($\pm$3.3 10^{-3})*
Geo	VBEM	0.600	1	0.199 ($\pm$2.3 10^{-3})	**0.182** ($\pm$**1.8** $\mathbf{10^{-3}}$)	**0.167** ($\pm$**1.8** $\mathbf{10^{-3}}$)	**0.150** ($\pm$**1.2** $\mathbf{10^{-3}}$)
	Grid	*0.3*	1	*0.199 ($\pm$4.9 10^{-3})*	*0.185 ($\pm$4.6 10^{-3})*	*0.172 ($\pm$4.2 10^{-3})*	*0.159 ($\pm$4.0 10^{-3})*
Sh. NB	VBEM	0.873	$(1, 0.2)^T$	**0.201** ($\pm$**3.1** $\mathbf{10^{-3}}$)	**0.183** ($\pm$**2.5** $\mathbf{10^{-3}}$)	**0.166** ($\pm$**2.2** $\mathbf{10^{-3}}$)	0.147 ($\pm$1.5 10^{-3})
PFraw	.	.	.	0.156 ($\pm$3.0 10^{-3})	0.155 ($\pm$3.3 10^{-3})	0.150 ($\pm$3.5 10^{-3})	0.144 ($\pm$5.3 10^{-3})
PFbin	.	.	.	0.197 ($\pm$2.1 10^{-3})	0.177 ($\pm$1.5 10^{-3})	0.160 ($\pm$1.5 10^{-3})	0.140 ($\pm$1.3 10^{-3})

POSTERIOR PREDICTIVE CHECK

Posterior predictive check (PPC) of the distribution of the listening counts in the TP dataset.



- Percentages of non-zero values written in parentheses.
- Blue dots: histogram of the non-zero values in the train set.
- Colored curves: simulated histograms obtained from the different models, with estimated latent variables $\mathbf{W}$, $\mathbf{H}$ (and parameters κ and θ for dcPF).

TAKE-HOME MESSAGE

- Unified framework for dcPF. Four specific distributions to model self-excitation.
- Scalability + closed-form updates for the inference of the posterior.
- dcPF is a natural generalization of PF. It offers a continuum between PFraw and PFbin.

REFERENCES

P. Gopalan, J. M. Hofman, and D. M. Blei (2015). Scalable recommendation with hierarchical Poisson factorization. In *Proc. Conference on Uncertainty in Artificial Intelligence (UAI)*.

M. E. Basbug and B. E. Engelhardt (2016). Hierarchical compound Poisson factorization. In *International Conference on Machine Learning (ICML)*.

M. Zhou (2018). Nonparametric Bayesian negative binomial factor analysis. *Bayesian Analysis*.

Github:

ACKNOWLEDGEMENTS

Supported by the European Research Council (ERC FACTORY-CoG-6681839) and the ANR-3IA (ANITI).