On the Model Shrinkage Effect of Gamma Process Edge Partition Models

Authored by:

Issei Sato
Iku Ohama
Takuya Kida
Hiroki Arimura

Abstract

The edge partition model (EPM) is a fundamental Bayesian nonparametric model for extracting an overlapping structure from binary matrix. The EPM adopts a gamma process ($Gamma$P) prior to automatically shrink the number of active atoms. However, we empirically found that the model shrinkage of the EPM does not typically work appropriately and leads to an overfitted solution. An analysis of the expectation of the EPM's intensity function suggested that the gamma priors for the EPM hyperparameters disturb the model shrinkage effect of the internal $Gamma$P. In order to ensure that the model shrinkage effect of the EPM works in an appropriate manner, we proposed two novel generative constructions of the EPM: CEPM incorporating constrained gamma priors, and DEPM incorporating Dirichlet priors instead of the gamma priors. Furthermore, all DEPM's model parameters including the infinite atoms of the $Gamma$P prior could be marginalized out, and thus it was possible to derive a truly infinite DEPM (IDEPM) that can be efficiently inferred using a collapsed Gibbs sampler. We experimentally confirmed that the model shrinkage of the proposed models works well and that the IDEPM indicated state-of-the-art performance in generalization ability, link prediction accuracy, mixing efficiency, and convergence speed.

1 Paper Body

Discovering low-dimensional structure from a binary matrix is an important problem in relational data analysis. Bayesian nonparametric priors, such as Dirichlet process (DP) [1] and hierarchical Dirichlet process (HDP) [2], have been widely applied to construct statistical models with an automatic model shrinkage effect [3, 4]. Recently, more advanced stochastic processes such as the Indian buffet process (IBP) [5] enabled the construction of statistical models for discovering overlapping structures [6, 7], wherein each individual in a data

matrix can belong to multiple latent classes. Among these models, the edge partition model (EPM) [8] is a fundamental Bayesian nonparametric model for extracting overlapping latent structure underlying a given binary matrix. The EPM considers latent positive random counts for only non-zero entries in a given binary matrix and factorizes the count matrix into two non-negative matrices and a non-negative diagonal matrix. A link probability of the EPM for an entry is defined by transforming the multiplication of the non-negative matrices into a probability, and thus the EPM can capture overlapping structures with a noisy-OR manner [6]. By incorporating a gamma process (?P) as a prior for the diagonal matrix, the number of active atoms of the EPM shrinks automatically according to the given data. Furthermore, by truncating the infinite atoms of the ?P with a finite number, all parameters and hyperparameters of the EPM can be inferred using closed-form Gibbs sampler. Although, the EPM is well designed to capture an overlapping structure and has an attractive affinity with a closed-form posterior 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

The proposed IDEPM successfully found expected 5 overlapped latent classes.

The EPM extracted many unexpected latent classes. (98 active classes)

(a) Synthetic data

(b) EPM solution

(c) Proposed IDEPM solution

Figure 1: (Best viewed in color.) A synthetic example: (a) synthetic 90 ? 90 data (white corresponds to one, and black to zero); (b) EPM solution; and (c) the proposed IDEPM solution. In (b) and (c), non-zero entries are colored to indicate their most probable assignment to the latent classes. inference, the EPM involves a critical drawback in its model shrinkage mechanism. As we experimentally show in Sec. 5, we found that the model shrinkage effect of the EPM does not typically work in an appropriate manner. Figure 1 shows a synthetic example. As shown in Fig. 1a, there are five overlapping latent classes (white blocks). However, as shown in Fig. 1b, the EPM overestimates the number of active atoms (classes) and overfits the data. In this paper, we analyze the undesired property of the EPM?s model shrinkage mechanism and propose novel generative constructions for the EPM to overcome the aforementioned disadvantage. As shown in Fig. 1c, the IDEPM proposed in this paper successfully shrinks unnecessary atoms. More specifically, we have three major contributions in this paper. (1) We analyse the generative construction of the EPM and find a property that disturbs its model shrinkage effect (Sec. 3). We derive the expectation of the EPM?s intensity function (Theorem 1), which is the total sum of the infinite atoms for an entry. From the derived expectation, we obtain a new finding that gamma priors for the EPM?s hyperparameters disturb the model shrinkage effect of the internal ?P (Theorem 2). That is, the derived expectation is expressed by a multiplication of the terms related to ?P and other gamma priors. Thus, there is no guarantee that the expected number of active atoms is finite. (2) Based on the analysis of the EPM?s intensity function, we propose two novel constructions of the EPM: the CEPM incorporating constrained gamma priors (Sec. 4.1) and the DEPM incorporating Dirichlet

priors instead of the gamma priors (Sec. 4.2). The model shrinkage effect of the CEPM and DEPM works appropriately because the expectation of their intensity functions depends only on the ?P prior (Sec. 4.1 and Theorem 3 in Sec. 4.2). (3) Furthermore, for the DEPM, all model parameters, including the infinite atoms of the ?P prior, can be marginalized out (Theorem 4). Therefore, we can derive a truly infinite DEPM (IDEPM), which has a closed-form marginal likelihood without truncating infinite atoms, and can be efficiently inferred using collapsed Gibbs sampler [9] (Sec. 4.3).

2

The Edge Partition Model (EPM)

In this section, we review the EPM [8] as a baseline model. Let x be an I ? J binary matrix, where an entry between i-th row and j-th column is represented by xi,j ? {0, 1}. In order to extract an overlapping structure underlying x, the EPM [8] considers a non-negative matrix factorization problem on latent Poisson counts as follows: ! K X (1) xi,j = I(mi,j,? ? 1), mi,j,? — U , V , ? ? Poisson Ui,k Vj,k ?k , k=1

where U and V are I ? K and J ? K non-negative matrices, respectively, and ? is a K ? K non-negative diagonal matrix. Note that I(?) is 1 if the predicate holds and is zero otherwise. The latent counts m take positive values only for edges (non-zero entries) within a given binary matrix and the generative model for each positive count is equivalently expressed as P a sum of K Poisson random variables as mi,j,? = k mi,j,k , mi,j,k ? Poisson(Ui,k Vj,k ?k). This is the reason why the above model is called edge partition model. Marginalizing m out from Eq. (1), the generative model of the EPM can be equivalently rewritten as 2

Q xi,j — U , V , ? ? Bernoulli(1 ? k e?Ui,k Vj,k ?k). As e?Ui,k Vj,k ?k ? [0, 1] denotes the probability that a Poisson random variable with mean Ui,k Vj,k ?k corresponds to zero, the EPM can capture an overlapping structure with a noisy-OR manner [6]. In order to complete the Bayesian hierarchical model of the EPM, gamma priors are adopted as Ui,k ? Gamma(a1 , b1) and Vj,k ? Gamma(a2 , b2), where a1 , a2 are shape parameters and b1 , b2 are rate parameters for the gamma distribution, respectively. Furthermore, a gamma process (?P) is incorporated as a Bayesian nonparametric prior for ? to make the EPM automatically shrink its number of atoms K. Let Gamma(?0 /T, c0) denote a truncated ?P with a concentration parameter ?0 and a rate parameter c0 , where T denotes a truncation level that should be set large enough to ensure a good approximation to the true ?P. Then, the diagonal elements of ? are drawn as ?k ? Gamma(?0 /T, c0) for k ? {1, . . . , T }. The posterior inference for all parameters and hyperparameters of the EPM can be performed using Gibbs sampler (detailed in Appendix conjugacy between P A). Thanks to the P gamma and Poisson distributions, given mi,?,k = j mi,j,k and m?,j,k = i mi,j,k , posterior sampling for Ui,k and Vj,k is straightforward. As the ?P prior is approximated by a gamma distribution, posterior sampling for ?k also can be performed straightforwardly. Given U , V , and ?, posterior sample for mi,j,? can be simulated using zero-truncated Poisson (ZTP) distribution [10]. Finally, we can obtain sufficient statistics mi,j,k by partitioning mi,j,? into T atoms using a multinomial distribution. Furthermore, all hyperparameters of

the EPM (i.e., ?0 , c0 , a1 , a2 , b1 , and b2) can also be sampled by assuming a gamma hyper prior Gamma(e0 , f0). Thanks to the conjugacy between gamma distributions, posterior sampling for c0 , b1 , and b2 is straightforward. For the remaining hyperparameters, we can construct closed-form Gibbs samplers using data augmentation techniques [11, 12, 2].

3

Analysis for Model Shrinkage Mechanism

The EPM is well designed to capture an overlapping structure with a simple Gibbs inference. However, the EPM involves a critical drawback in its model shrinkage mechanism. For the EPM, a ?P prior is incorporated as a prior for the non-negative diagonal matrix as ?k ? Gamma(?0 /T, c0). From the form of the truncated ?P, thanks to the additive property of independent gamma random P variables, the total sum of ?k over countably infinite atoms ? follows a gamma distribution as k=1P?k ? Gamma(?0 , c0), wherein the intensity function of ? the ?P has a finite expectation as E[k=1 ?k] = ?c00 . Therefore, the ?P has a regularization mechanism that automatically shrinks the number of atoms according to given observations.

However, as experimentally shown in Sec. 5, the model shrinkage mechanism of the EPM does not work appropriately. More specifically, the EPM often overestimates the number of active atoms and overfits the data. Thus, we analyse the intensity function of the EPM to reveal the reason why the model shrinkage mechanism does not work appropriately. P? Theorem 1. The expectation of the EPM?s intensity function k=1 Ui,k Vj,k ?k for an entry (i, j) is finite and can be expressed as follows: # "? X a2 ?0 a1 ? ? . (2) E Ui,k Vj,k ?k = b1 b2 c0 k=1

Proof. As U , V , and ? are independent of each other, the expected value operator is multiplicative for the EPM?s intensity function.PUsing the multiplicativity and the low of total P? ? expectation, P the proof is completed as E [k=1 Ui,k Vj,k ?k] = k=1 E[Ui,k]E[Vj,k]E[?k] = ? a1 a2 k=1 ?k]. b1 ? b2 ? E[As Eq. (2) in Theorem 1 shows, the expectation of the EPM?s intensity function is expressed by multiplying individual expectations of a ?P and two gamma distributions. This causes an undesirable property to the model shrinkage effect of the EPM. From Theorem 1, another important theorem about the EPM?s model shrinkage effect is obtained as follows: 3

Theorem 2. Given an arbitrary non-negative constant P? C, even if the expectation of the EPM?s intensity function in Eq. (2) is fixed as E [k=1 Ui,k Vj,k ?k] = C, there exist cases in which the model shrinkage effect of the ?P prior disappears.

P? Proof. Substituting E [k=1 Ui,k Vj,k ?k] = C for Eq. (2), we obtain C = ab11 ? ab22 ? ?c00 . Since a1 , a2 , b1 , and b2 are gamma random variables, even if the expectation of the EPM?s intensity function, C, is fixed, ?c00 can take an arbitrary value so that equation C = ab11 ? ab22 ? ?c00 holds. Hence, ?0 can take an arbitrary large value such that ?0 = T ? ? b0 . This implies that the ?P prior for the EPM degrades to a gamma distribution without model shrinkage effect as ?k ? Gamma(?0 /T, c0) = Gamma(b ?0 , c0). Theorem 2 indicates that the EPM might overestimate the number of active atoms, and

lead to overfitted solutions.

4

Proposed Generative Constructions

We describe our novel generative constructions for the EPM with an appropriate model shrinkage effect. According to the analysis described in Sec. 3, the model shrinkage mechanism of the EPM does not work because the expectation of the EPM?s intensity function has an undesirable redundancy. This finding motivates the proposal of new generative constructions, in which the expectation of the intensity function depends only on the ?P prior. First, we propose a naive extension of the original EPM using constrained gamma priors (termed as CEPM). Next, we propose an another generative construction for the EPM by incorporating Dirichlet priors instead of gamma priors (termed as DEPM). Furthermore, for the DEPM, we derive truly infinite DEPM (termed as IDEPM) by marginalizing out all model parameters including the infinite atoms of the ?P prior. 4.1

CEPM

In order to ensure that the EPM?s intensity function depends solely on the ?P prior, a naive way is to introduce constraints for the hyperparameters of the gamma prior. In the CEPM, the rate parameters of the gamma priors are constrained as b1 = C1 ? a1 and b2 = C2 ? a2 , respectively, where C1 ¿ 0 and C2 ¿ 0 are arbitrary constants. Based on the aforementioned constraints and Theorem 1,Pthe expectation of the intensity function for the CEPM depends ? only on the ?P prior as E[k=1 Ui,k Vj,k ?k] = C1?C02 c0 .

The posterior inference for the CEPM can be performed using Gibbs sampler in a manner similar to that for the EPM. However, we can not derive closed-form samplers only for a1 and a2 because of the constraints. Thus, in this paper, posterior sampling for a1 and a2 are performed using grid Gibbs sampling [13] (see Appendix B for details). 4.2

DEPM

We have another strategy to construct the EPM with efficient model shrinkage effect by re-parametrizing the factorization problem. Let us denote transpose of a matrix A by A? . According to the generative model of the EPM in Eq. (1), the original generative process for counts m can be viewed as a matrix factorization as m ? U ?V ? . It is clear that the optimal solution of the factorization problem is not unique. Let ?1 and ?2 be arbitrary K ? K non-negative diagonal matrices. If a solution m ? U ?V ? is globally optimal, then ?1 ? another solution m ? (U ?1)(??1 is also optimal. In order to ensure that 1 ??2)(V ?2) the EPM has only one optimal solution, we re-parametrize the original factorization problem to an equivalent constrained factorization problem as follows: m ? ??? ? ,

(3) P where ? denotes an I ? K non-negative matrix with l1 -constraints as Pi ?i,k = 1, ?k. Similarly, ? denotes an J ? K non-negative matrix with l1 -constraints as j ?j,k = 1, ?k. This parameterization ensures the uniqueness of the optimal solution for a given m because each column of ? and ? is constrained such that it is defined on a simplex. 4

According to the factorization in Eq. (3), by incorporating Dirichlet priors

instead of gamma priors, the generative construction for m of the DEPM is as follows: ! I T X z }— { mi,j,? — ?, ?, ? ? Poisson ?i,k ?j,k ?k , {?i,k }Ii=1 — ?1 ? Dirichlet(?1 , . . . , ?1), k=1

J

z }— { (4) — ?2 ? Dirichlet(?2 , . . . , ?2), ?k — ?0 , c0 ? Gamma(?0 /T, c0). P? Theorem 3. The expectation of DEPM?s intensity function ? ? ? depends sorely k=1 i,k j,k k P? ?0 . on the ?P prior and can be expressed as E[k=1 ?i,k ?j,k ?k] = IJc 0 {?j,k }Jj=1

Proof. The expectations of Dirichlet random variables ?i,k and ?j,k are I1 and J1 , respectively. Similar to the proof for Theorem 1, using the multiplicativity of independent random P? variables and the low of total expectation, the proof is completed as E [? ? ?k] = i,k j,k k=1 P? P? 1 1 E[?]E[?]E[?] = ? ? E[?]. i,k j,k k k=1 k=1 k I J

Note that, if we set constants C1 = I and C2 = J for the CEPM in Sec. 4.1, then the expectation of the intensity function for the CEPM is equivalent to that for the DEPM in Theorem 3. Thus, in order to ensure the fairness of comparisons, we set C1 = I and C2 = J for the CEPM in the experiments.

As the Gibbs sampler for ? and ? can be derived straightforwardly, the posterior inference for all parameters and hyperparameters of the DEPM also can be performed via closedform Gibbs sampler (detailed in Appendix C). Differ from the CEPM, l1 -constraints in the DEPM ensure the uniqueness of its optimal solution. Thus, the inference for the DEPM is considered as more efficient than that for the CEPM. 4.3

Truly Infinite DEPM (IDEPM)

One remarkable property of the DEPM is that we can derive a fully marginalized likelihood function. Similar to the beta-negative binomial topic model [13], we consider a joint distribumi,j,? ? {1, ? ? ? , T }mi,j,? tion for mi,j,? Poisson customers and their assignments zi,j =Q {zi,j,s }s=1 mi,j,? to T tables as P (mi,j,? , zi,j — ?, ?, ?) = P (mi,j,? — ?, ?, ?) s=1 P (zi,j,s — mi,j,? , ?, ?, ?). Thanks to the l1 -constraints we introduced in Eq. (3), the joint distribution P (m, z — ?, ?, ?) has a fully factorized form (see Lemma 1 in Appendix D). Therefore, marginalizing ?, ?, and ? out according to the prior construction in Eq. (4), we obtain an analytical marginal likelihood P (m, z) for the truncated DEPM (see Appendix D for a detailed derivation). Furthermore, by taking T ? ?, we can derive a closed-form marginal likelihood for the truly infinite version of the DEPM (termed as IDEPM). In a manner similar to that in [14], we consider the likelihood function for partition P [z] P instead of the assignments z. Assume we have K+ of T atoms for which m?,?,k = i j mi,j,k ¿ 0, and a partition of M (= P P m) customers into K subsets. Then, joint marginal likelihood of the IDEPM i,j,? + i j for [z] and m is given by the following theorem, with the proof provided in Appendix D: Theorem 4. The marginal likelihood function of the IDEPM is defined as P (m, [z])? = T! limT ?? P (m, [z]) = limT ?? (T ?K P (m, z), and can be derived as follows: +)! P (m, [z])? =

J I Y Y

i=1 j=1

1 mi,j,? !

?
K+ Y
k=1
I Y ?(I?1) ?(?1 + mi,?,k) ?(I?1 + m?,?,k) i=1 ?(?1)
?0 Y
K+ c0 ?(m?,?,k) ?(?2 + m?,j,k) ?(J?2) K+ ? ?0 , (5) ? ?(J?2 + m?,?,k) j=1 ?(?2) c0 + 1 (c0 + 1)m?,?,k k=1 k=1 P P P P where mi,?,k = j mi,j,k , m?,j,k = i mi,j,k , and m?,?,k = i j mi,j,k . Note that ?(?) denotes gamma function. K+
Y
J Y

From Eq. (5) in Theorem 4, we can derive collapsed Gibbs sampler [9] to perform posterior inference for the IDEPM. Since ?, ?, and ? have been marginalized out, the only latent variables we have to update are m and z. 5

Sampling z: Given m, similar to the Chinese restaurant process (CRP) [15], the posterior probability that zi,j,s is assigned to k ? is given as follows: ? (ijs) (ijs) ? m(ijs) ? ?1 +mi,?,k? ? ?2 +m?,j,k? if m(ijs) ¿ 0, (ijs) (ijs) ?,?,k? k? ? I?1 +m?,?,k? I?2 +m?,?,k? (6) P (zi,j,s = k — z(ijs) , m) ? ? (ijs) 1 1 ?0 ? I ? J if m?,?,k? = 0,

where the superscript (ijs) denotes that the corresponding statistics are computed excluding the s-th customer of entry (i, j).

Sampling m: Given z, posteriors for the ? and ? are simulated as {?i,k }Ii=1 — ? ? Dirichlet({?1 + mi,?,k }Ii=1) and {?j,k }Jj=1 — ? ? Dirichlet({?2 + m?,j,k }Jj=1) for k ? {1, . . . , K+ }. Furthermore, the posterior sampling of the ?k for K+ active atoms can be performed as ?k — ? ? Gamma(m?,?,k , c0 + 1). Therefore, similar to the sampler for the EPM [8], we can update m as follows:

?(0) if xi,j = 0, PK + mi,j,? — ?, ?, ? ? (7) ZTP(k=1 ?i,k ?j,k ?k) if xi,j = 1, ? ( )K+ ? ? ? ? K+ i,k j,k k ?, {mi,j,k }k=1 — mi,j,? , ?, ?, ? ? Multinomial ?mi,j,? ; PK+ (8) ? ?j,k ? ?k ? ? ? i,k k =1 k=1

where ?(0) denotes point mass at zero.

Sampling hyperparameters: We can construct closed-form Gibbs sampler for all hyperparameters of the IDEPM assuming a gamma prior (Gamma(e0 , f0)). Using the additive property of the ?P, posterior sample for the sum of ?k over unused atoms is obtained as ??0 = P? k? =K+ +1 ?k? — ? ? Gamma(?0 , c0 + 1). Consequently, we obtain a closed-form posterior PK+ ?k). sampler for the rate parameter c0 of the ?P as c0 — ? ? Gamma(e0 + ?0 , f0 + ??0 + k=1 For all remaining hyperparameters (i.e., ?1 , ?2 , and ?0), we can derive posterior samplers from Eq. (5) using data augmentation techniques [12, 8, 2, 11] (detailed in Appendix E).

5

Experimental Results

In previous sections, we theoretically analysed the reason why the model shrinkage of the EPM does not work appropriately (Sec. 3) and proposed several novel constructions (i.e., CEPM, DEPM, and IDEPM) of the EPM with an efficient model shrinkage effect (Sec. 4). The purpose of the experiments

involves ascertaining the following hypotheses: (H1) The original EPM overestimates the number of active atoms and overfits the data. In contrast, the model shrinkage mechanisms of the CEPM and DEPM work appropriately. Consequently, the CEPM and DEPM outperform the EPM in generalization ability and link prediction accuracy. (H2) Compared with the CEPM, the DEPM indicates better generalization ability and link prediction accuracy because of the uniqueness of the DEPM?s optimal solution. (H3) The IDEPM with collapsed Gibbs sampler is superior to the DEPM in generalization ability, link prediction accuracy, mixing efficiency, and convergence speed. Datasets: The first dataset was the Enron [16] dataset, which comprises e-mails sent between 149 Enron employees. We extracted e-mail transactions from September 2001 and constructed Enron09 dataset. For this dataset, xi,j = 1(0) was used to indicate whether an e-mail was, or was not, sent by the i-th employee to the j-th employee. For larger dataset, we used the MovieLens [17] dataset, which comprises five-point scale ratings of movies submitted by users. For this dataset, we set xi,j = 1 when the rating was higher than three and xi,j = 0 otherwise. We prepared two different sized MovieLens dataset: MovieLens100K (943 users and 1,682 movies) and MovieLens1M (6,040 users and 3,706 movies). The densities of the Enron09, MovieLens100K and MovieLens1M datasets were 0.016, 0.035, and 0.026, respectively. 6

(a) Enron09

Estimated # of K

128 64

(b) MovieLens100K

128

IDEPM DEPM-T CEPM-T EPM-T

64

64

32

32

16

16

16

8

8

8

4

4

4

2

2

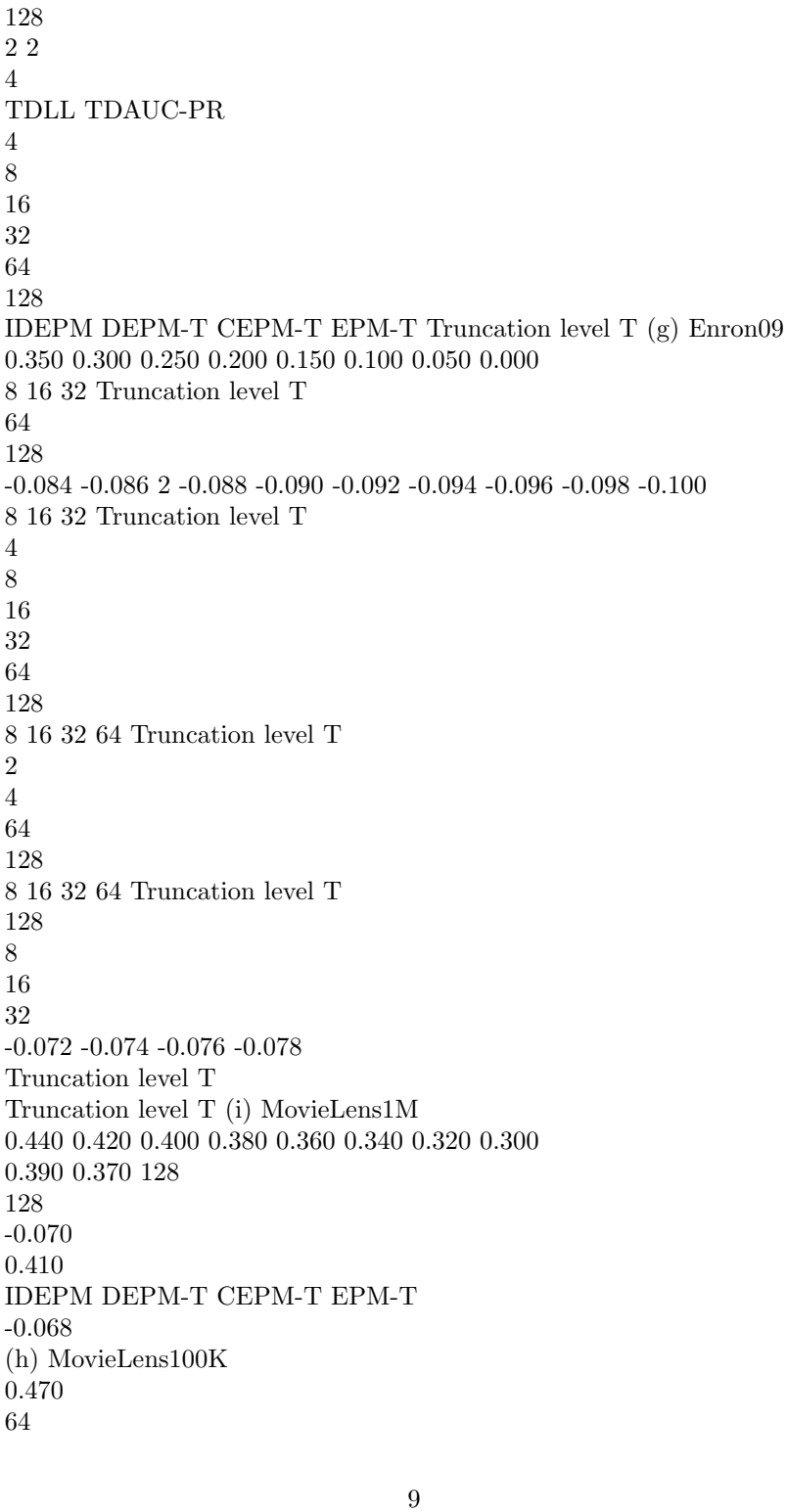
32

2

4

8 16 32 Truncation level T

64


128
2 2
4
TDLL TDAUC-PR
4
8
16
32
64
128
IDEPM DEPM-T CEPM-T EPM-T Truncation level T (g) Enron09
0.350 0.300 0.250 0.200 0.150 0.100 0.050 0.000
8 16 32 Truncation level T
64
128
-0.084 -0.086 2 -0.088 -0.090 -0.092 -0.094 -0.096 -0.098 -0.100
8 16 32 Truncation level T
4
8
16
32
64
128
8 16 32 64 Truncation level T
2
4
64
128
8 16 32 64 Truncation level T
128
8
16
32
-0.072 -0.074 -0.076 -0.078
Truncation level T
Truncation level T (i) MovieLens1M
0.440 0.420 0.400 0.380 0.360 0.340 0.320 0.300
0.390 0.370 128
128
-0.070
0.410
IDEPM DEPM-T CEPM-T EPM-T
-0.068
(h) MovieLens100K
0.470
64

(f) MovieLens1M
0.430
4
4
-0.066
0.450
2
2
(e) MovieLens100K
(d) Enron09 -0.040 -0.050 2 -0.060 -0.070 -0.080 -0.090 -0.100 -0.110 -0.120
(c) MovieLens1M
128
2
4
8 16 32 64 Truncation level T
128
2
4

Figure 2: Calculated measurements as functions of the truncation level T for each dataset. The horizontal line in each figure denotes the result obtained using the IDEPM. Evaluating Measures: We adopted three measurements to evaluate the performance of the models. The first is the estimated number of active atoms K for evaluating the model shrinkage effect of each model. The second is the averaged Test Data Log Likelihood (TDLL) for evaluating the generalization ability of each model. We calculated the averaged likelihood that a test entry takes the actual value. For the third measurement, as many real-world binary matrices are often sparse, we adopted the Test Data Area Under the Curve of the Precision-Recall curve (TDAUC-PR) [18] to evaluate the link prediction ability. In order to calculate the TDLL and TDAUC-PR, we set all the selected test entries as zero during the inference period, because binary observations for unobserved entries are not observed as missing values but are observed as zeros in many real-world situations. Experimental Settings: Posterior inference for the truncated models (i.e., EPM, CEPM, and DEPM) were performed using standard (non-collapsed) Gibbs sampler. Posterior inference for the IDEPM was performed using the collapsed Gibbs sampler derived in Sec. 4.3. For all models, we also sampled all hyperparameters assuming the same gamma prior (Gamma(e0 , f0)). For the purpose of fair comparison, we set hyper-hyperparameters as e0 = f0 = 0.01 throughout the experiments. We ran 600 Gibbs iterations for each model on each dataset and used the final 100 iterations to calculate the measurements. Furthermore, all reported measurements were averaged values obtained by 10-fold cross validation. Results: Hereafter, the truncated models are denoted as EPM-T , CEPM-T , and DEPM-T to specify the truncation level T . Figure 2 shows the calculated measurements. (H1) As shown in Figs. 2a?c, the EPM overestimated the number of active atoms K for all datasets especially for a large truncation level T . In contrast, the number of active atoms K for the CEPM-T and DEPM-T monotonically converges to a

specific value. This result supports the analysis with respect to the relationship between the model shrinkage effect and the expectation of the EPM?s intensity function, as discussed in Sec. 3. Consequently, 7

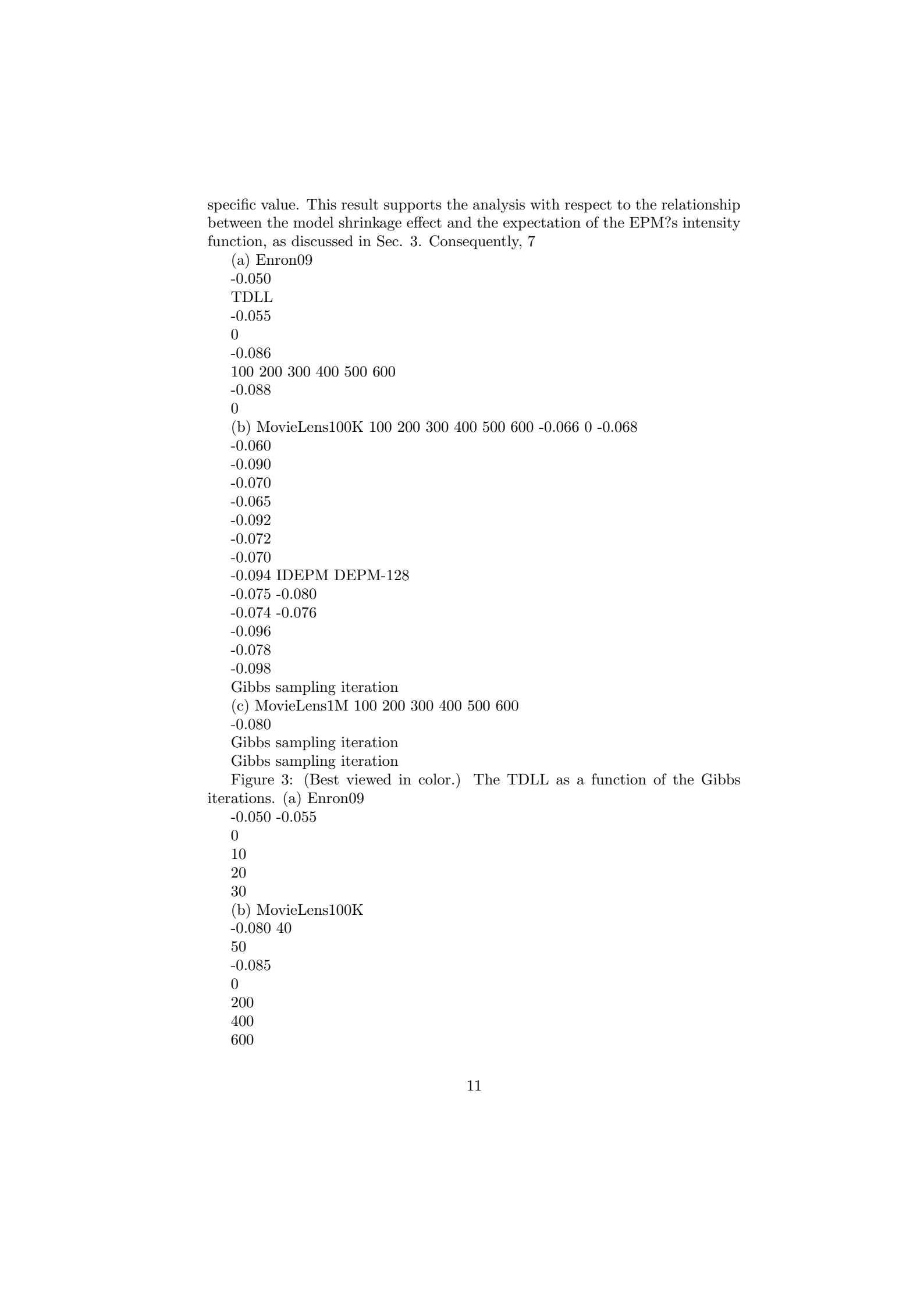


Figure 3: (Best viewed in color.) The TDLL as a function of the Gibbs iterations. (a) Enron09

800
(c) MovieLens1M
-0.065 1000
0
2000
4000
6000
8000 10000
TDLL
-0.070 -0.060
-0.090
-0.065
-0.095
-0.070 -0.075 -0.080
IDEPM DEPM-128 DEPM-64 DEPM-32 DEPM-16 DEPM-8 DEPM-4 DEPM-2 Elapsed time (sec)
-0.075
-0.100 -0.080 -0.105 -0.110
Elapsed time (sec)
-0.085
Elapsed time (sec)

Figure 4: (Best viewed in color.) The TDLL as a function of the elapsed time (in seconds). as shown by the TDLL (Figs. 2d?f) and TDAUC-PR (Figs. 2g?i), the CEPM and DEPM outperformed the original EPM in both generalization ability and link prediction accuracy. (H2) As shown in Figs. 2a?c, the model shrinkage effect of the DEPM is stronger than that of the CEPM. As a result, the DEPM significantly outperformed the CEPM in both generalization ability and link prediction accuracy (Figs. 2d?i). Although the CEPM slightly outperformed the EPM, the CEPM with a larger T tends to overfit the data. In contrast, the DEPM indicated its best performance with the largest truncation level (T = 128). Therefore, we confirmed that the uniqueness of the optimal solution in the DEPM was considerably important in achieving good generalization ability and link prediction accuracy. (H3) As shown by the horizontal lines in Figs. 2d?i, the IDEPM indicated the state-of-theart scores for all datasets. Finally, the computational efficiency of the IDEPM was compared with that of the truncated DEPM. Figure 3 shows the TDLL as a function of the number of Gibbs iterations. In keeping with expectations, the IDEPM indicated significantly better mixing property when compared with that of the DEPM for all datasets. Furthermore, Fig. 4 shows a comparison of the convergence speed of the IDEPM and DEPM with several truncation levels (T = {2, 4, 8, 16, 32, 64, 128}). As clearly shown in the figure, the convergence of the IDEPM was significantly faster than that of the DEPM with all truncation levels. Therefore, we confirmed that the IDEPM indicated a state-of-the-art performance in generalization ability, link prediction accuracy, mixing efficiency, and convergence speed.

6

Conclusions

In this paper, we analysed the model shrinkage effect of the EPM, which is a Bayesian nonparametric model for extracting overlapping structure with an optimal dimension from binary matrices. We derived the expectation of the intensity function of the EPM, and showed that the redundancy of the EPM?s intensity function disturbs its model shrinkage effect. According to this finding, we proposed two novel generative construction for the EPM (i.e., CEPM and DEPM) to ensure that its model shrinkage effect works appropriately. Furthermore, we derived a truly infinite version of the DEPM (i.e, IDEPM), which can be inferred using collapsed Gibbs sampler without any approximation for the ?P. We experimentally showed that the model shrinkage mechanism of the CEPM and DEPM worked appropriately. Furthermore, we confirmed that the proposed IDEPM indicated a state-ofthe-art performance in generalization ability, link prediction accuracy, mixing efficiency, and convergence speed. It is of interest to further investigate whether the truly infinite construction of the IDEPM can be applied to more complex and modern machine learning models, including deep brief networks [19], and tensor factorization models [20]. 8

2 References

[1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14] [15] [16] [17] [18] [19] [20]

Thomas S. Ferguson. ?A Bayesian Analysis of Some Nonparametric Problems?. In: The Annals of Statistics 1.2 (1973), pp. 209?230. Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. ?Hierarchical Dirichlet Processes?. In: J. Am. Stat. Assoc. 101.476 (2006), pp. 1566?1581. Charles Kemp, Joshua B. Tenenbaum, Thomas L. Griffiths, Takeshi Yamada, and Naonori Ueda. ?Learning Systems of Concepts with an Infinite Relational Model?. In: Proc. AAAI. Vol. 1. 2006, pp. 381?388. Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. ?Mixed Membership Stochastic Blockmodels?. In: J. Mach. Learn. Res. 9 (2008), pp. 1981? 2014. Thomas L. Griffiths and Zoubin Ghahramani. ?Infinite Latent Feature Models and the Indian Buffet Process?. In: Proc. NIPS. 2005, pp. 475?482. Morten M?rup, Mikkel N. Schmidt, and Lars Kai Hansen. ?Infinite Multiple Membership Relational Modeling for Complex Networks?. In: Proc. MLSP. 2011, pp. 1? 6. Konstantina Palla, David A. Knowles, and Zoubin Ghahramani. ?An Infinite Latent Attribute Model for Network Data?. In: Proc. ICML. 2012, pp. 1607?1614. Mingyuan Zhou. ?Infinite Edge Partition Models for Overlapping Community Detection and Link Prediction?. In: Proc. AISTATS. Vol. 38. 2015, pp. 1135?1143. Jun S. Liu. ?The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem?. In: J. Am. Stat. Assoc. 89.427 (1994), pp. 958? 966. Charles J. Geyer. Lower-Truncated Poisson and Negative Binomial Distributions. Tech. rep. Working Paper Written for the Software R. University of Minnesota, MN (available: http://cran.r-project.org/web/packages/aster/vignettes/trunc.pdf), 2007. David Newman, Arthur U. Asuncion, Padhraic Smyth, and Max Welling.

?Distributed Algorithms for Topic Models?. In: J. Mach. Learn. Res. 10 (2009), pp. 1801?1828. Michael D. Escobar and Mike West. ?Bayesian Density Estimation and Inference Using Mixtures?. In: J. Am. Stat. Assoc. 90 (1994), pp. 577?588. Mingyuan Zhou. ?Beta-Negative Binomial Process and Exchangeable Random Partitions for Mixed-Membership Modeling?. In: Proc. NIPS. 2014, pp. 3455?3463. Thomas L. Griffiths and Zoubin Ghahramani. ?The Indian Buffet Process: An Introduction and Review?. In: J. Mach. Learn. Res. 12 (2011), pp. 1185?1224. David Blackwell and James B. MacQueen. ?Ferguson distributions via Polya urn schemes?. In: The Annals of Statistics 1 (1973), pp. 353?355. Bryan Klimat and Yiming Yang. ?The Enron Corpus: A New Dataset for Email Classification Research?. In: Proc. ECML. 2004, pp. 217?226. MovieLens dataset, http://www.grouplens.org/. as of 2003. url: http://www.grouplens.org/. Jesse Davis and Mark Goadrich. ?The Relationship Between Precision-Recall and ROC Curves?. In: Proc. ICML. 2006, pp. 233?240. Mingyuan Zhou, Yulai Cong, and Bo Chen. ?The Poisson Gamma Belief Network?. In: Proc. NIPS. 2015, pp. 3043?3051. Changwei Hu, Piyush Rai, and Lawrence Carin. ?Zero-Truncated Poisson Tensor Factorization for Massive Binary Tensors?. In: Proc. UAI. 2015, pp. 375?384.

9