

Analytic note

Yidong Zhang

December 11, 2022

1 Model overview

treeLFA is the the abbreviation for “latent factor allocation with a tree structured prior for topics”. It is a topic model specifically designed for binary input data, which is based on the Bayesian mean-parameterized binary non-negative matrix factorization [1], and also incorporates an informative prior for topics constructed on a tree structure of individual words. It retains the basic “document-topic-word” configuration of topic models, but these components are assigned with new meanings to analyse the diagnostic data in biobanks. For treeLFA, each individual in the biobank is viewed as a document, and each disease a word. Topics of diseases capture constellations of diseases that frequently co-occur on the same individuals.

The input for treeLFA is a $D \times S$ binary matrix (W), where each row corresponds to an individual (d), and each column a disease code (s). W_{ds} in the input matrix records if disease code s is diagnosed for individual d . The output of treeLFA includes two matrices: the topic matrix ϕ ($T \times S$, ϕ_{ts} is the probability of disease code s in topic t), and the topic weight matrix θ ($D \times T$, θ_{dt} is the weight of individual d for topic t).

treeLFA is generalised from Bayesian mean-parameterized binary non-negative matrix factorization (BNMF), which is fundamentally different from the more commonly used topic model Latent Dirichlet Allocation (LDA) [2]. treeLFA models the presence and absence of all disease codes for all individuals with Bernoulli distributions, while LDA models the disease codes that are diagnosed (present) for individuals with Multinomial distributions. For treeLFA, the loading of a disease code in a topic is its Bernoulli probability, and the Bernoulli probability of this disease code for an individual is a mixture of the Bernoulli probability for this code in all topics, with the mixing coefficients specified by the individual’s topic weights. As a whole, the matrix of Bernoulli probability for all disease variables ($D \times S$ variables in total) is factorised into the topic weight matrix and the topic matrix that are mentioned above. To differentiate from LDA, we name our model latent factor allocation (LFA). In the next sections, we will introduce the generative process for treeLFA, starting from the generation of topics using a hierarchical prior.

2 The hierarchical prior for topics

In each topic, only a fraction of all disease codes are active, which means they have a relatively large Bernoulli probability in the topic. As a result, they will be likely to co-occur on the same individual. The remaining disease codes are inactive, meaning they have near-zero probability. Different Beta priors ($Beta(a_0^0, a_0^1)$ and $Beta(a_1^0, a_1^1)$) are put on ϕ_{ts} for inactive and active disease codes.

To introduce correlation in active diseases within each topic, we used an tree-structured prior. For disease code s in topic t a binary indicator variable I_{ts} is introduced to encode whether it is active in the topic (1 for active state, 0 for inactive state). The indicator variables for all disease codes in a topic are generated with a Markov process on the hierarchical/tree structure of disease codes specified by a medical ontology (such as the ICD-10 coding system). As is shown in Figure 1, each leaf node of the tree corresponds to the indicator variable of a disease code in a topic, while internal nodes on the tree correspond to different categories of diseases specified by the medical ontology (here we assume that they can also be diagnosed for individuals). I for all disease codes in a topic are sampled from the inactive root node (does not correspond to any disease code) to all leaf nodes, using a Markov process that has two states, active and inactive. ρ_{01} and ρ_{11} are the transition probability of the Markov process, corresponding to the probability of transitioning from inactive to active state and from active to active state, respectively.

The sparsity of a topic (the total number of active codes in a topic) can be controlled by tuning ρ_{01} and ρ_{11} of the Markov process, as what we can do with LDA by tuning the concentration parameter of the Dirichlet prior for topics. Furthermore, we can also control the distribution of active codes in a topic by tuning the transition probability. With a large ρ_{11} , most child codes of an active parent code will also be active. As a result, active codes will be concentrated on the same subtree. On the other hand, if ρ_{11} is small, active codes will be scattered across the whole tree.

3 Generative model specification

The graphical representation of the model is shown in Figure 2A. To generate a topic, I for all disease codes in this topic are firstly generated using a Markov process on the tree structure of disease codes (discussed in the previous section). Next, conditioned on these indicator variables, the probability ϕ for all disease codes in the topic are sampled from the corresponding Beta distributions for active and inactive codes. We then generate the topic weight vector θ_d for individual d using a Dirichlet distribution, and for him the topic assignment variables Z_d for all disease codes using the categorical distribution parameterized by θ_d . These two steps are the same as the corresponding steps in the generative process for LDA (Figure 2B). In the last step, disease variable W_{ds} is sampled from the Bernoulli distribution with probability $\phi_{Z_{ds},s}$ (this is the probability of

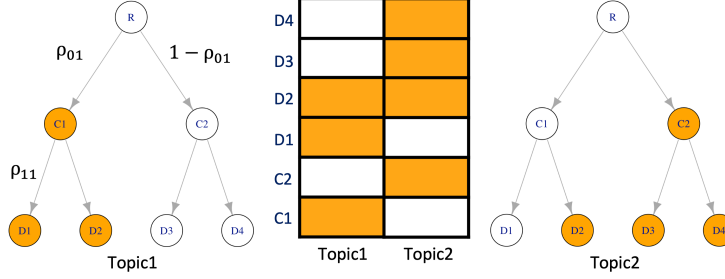


Figure 1: Schematic for the Markov process on the tree structure of disease codes. Values of the indicator variables for six disease codes in two topics are shown in the middle of the figure. Each cell corresponds to the indicator variable of a disease code in a topic, with white color representing inactive code, and orange representing active code. All indicator variables for a topic are generated by running a Markov process on the tree structure of disease codes. Each node (except for the root node) on the tree corresponds to a disease code in a topic. The top node on the tree labelled with R is the single root node, while C represents categories of diseases (internal nodes on the second layer of the tree), and D represents individual diseases (terminal nodes on the third layer of the tree). ρ_{01} and ρ_{11} are the two transition probabilities of this Markov process.

78 disease code s in the topic specified by topic assignment variable Z_{ds}). Notations
 79 for treeLFA are listed in Table 1.

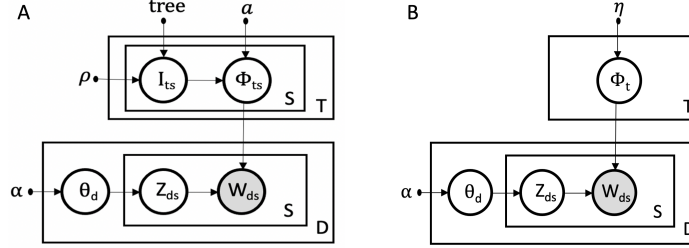


Figure 2: The probabilistic graphical model for treeLFA and LDA. A, The probabilistic graphical model for treeLFA. Each topic is a probability vector ($\phi_{t1} \dots \phi_{tS}$) for the Bernoulli distributions for S disease codes, and each probability variable ϕ_{ts} has a corresponding indicator variable I_{ts} , denoting if disease code s is active in topic t . Different Beta priors are put on ϕ of active and inactive disease codes. I for all disease codes in a topic are generated using a Markov process on the tree structure with transition probability vector ρ . The generative steps for the remaining hidden variables are the same as those for LDA. B, The graphical model for LDA.

80 The full treeLFA model is specified as below:

$$\begin{aligned} & P(W, \theta, Z, \phi, I \mid \rho, \alpha, a, tree) \\ = & P(I \mid tree, \rho) \cdot P(\phi \mid I, a) \cdot P(\theta \mid \alpha) \cdot P(Z \mid \theta) \cdot P(W \mid \phi, Z) \end{aligned} \quad (1)$$

$$P(I \mid tree, \rho) = \prod_{t=1}^T \prod_{s=1}^S P(I_{ts} \mid I_{ts}^{pa}, \rho)$$

$$P(\phi \mid I, a) = \prod_{t=1}^T \prod_{s=1}^S P(\phi_{ts} \mid I_{ts}, a)$$

$$P(\theta \mid \alpha) = \prod_{d=1}^D P(\theta_d \mid \alpha)$$

$$P(Z \mid \theta) = \prod_{d=1}^D \left\{ \prod_{s=1}^S P(Z_{ds} \mid \theta_d) \right\}$$

$$P(W \mid \phi, Z) = \prod_{d=1}^D \prod_{s=1}^S P(W_{ds} \mid \phi_{Z_{ds}, s})$$

$$P(\theta_d \mid \alpha) \sim \text{Dirichlet}(\alpha)$$

$$P(Z_{ds} \mid \theta_d) \sim \text{Categorical}(\theta_d)$$

$$P(I_{ts} \mid I_{ts}^{pa} = 0, \rho) \sim \text{Bernoulli}(\rho_{01})$$

$$P(I_{ts} \mid I_{ts}^{pa} = 1, \rho) \sim \text{Bernoulli}(\rho_{11})$$

$$P(\phi_{ts} \mid I_{ts} = 0, a) \sim \text{Beta}(a_0^0, a_0^1)$$

$$P(\phi_{ts} \mid I_{ts} = 1, a) \sim \text{Beta}(a_1^0, a_1^1)$$

$$P(W_{ds} \mid \phi_{Z_{ds}, s}) \sim \text{Bernoulli}(\phi_{Z_{ds}, s})$$

Notation	Explanation
D	total number of individuals in the training dataset
S	total number of disease codes
T	total number of topics
d	individual d
s	disease code s
t	topic t
Z_{ds}	topic assignment variable for disease code s for individual d
ϕ_{ts}	probability variable for disease code s in topic t
I_{ts}	indicator variable for disease code s in topic t
I_{ts}^{pa}	indicator variable for the parent disease code of disease code s in topic t
I_{ts}^{ch}	indicator variables for all children disease codes of disease code s in topic t
θ_d	topic weight vector for individual d
W_{ds}	disease variable for disease code s for individual d
$tree$	the fixed tree structure of disease codes specified by a disease classification system
ρ_{01}	the probability of transitioning from an inactive code to an active code for the Markov process on the tree
ρ_{11}	the probability of transitioning from an active code to an active code for the Markov process on the tree
α	the parameter vector for the Dirichlet prior for θ
a_0^0, a_0^1	parameters of the Beta prior for ϕ of inactive codes
a_1^0, a_1^1	parameters of the Beta prior for ϕ of active codes

Table 1: Notations for “treeLFA”

81

82 3.1 Inference of treeLFA with Gibbs sampling

83 We use a partially collapsed Gibbs sampler to estimate the posterior distri-
84 butions of the latent variables of treeLFA. The collapsed Gibbs sampler performs
85 inference on the marginal distribution of disease incidences, with individuals’
86 topic weight vectors θ integrated out. The updating equations for each hidden
87 variable are derived as follows:

88 **3.1.1 Topic assignment variable for disease code s for individual d :**
89 Z_{ds}

$$\begin{aligned}
P(Z_{ds} = t* \mid Z_d^{-s}, \phi, I, W) &\propto \int P(Z_{ds} = t*, Z_d^{-s}, \phi, I, W, \theta) d\theta \\
&\propto P(W_{ds} \mid Z_{ds} = t*, \phi) \cdot \int P(\theta_d \mid \alpha) \cdot P(Z_d \mid \theta_d) d\theta_d \\
&= P(W_{ds} \mid Z_{ds} = t*, \phi) \cdot \int \left[\frac{\Gamma(\sum_{t=1}^T \alpha_t)}{\prod_{t=1}^T \Gamma(\alpha_t)} \prod_{t=1}^T \theta_{dt}^{\alpha_t-1} \right] \\
&\quad \prod_{s=1}^S \theta_{d, Z_{ds}} d\theta_d \\
&= P(W_{ds} \mid Z_{ds} = t*, \phi) \cdot \int \left[\frac{\Gamma(\sum_{t=1}^T \alpha_t)}{\prod_{t=1}^T \Gamma(\alpha_t)} \prod_{t=1}^T \theta_{dt}^{\alpha_t-1} \right] \\
&\quad \prod_{t=1}^T \theta_{dt}^{c_{dt}} d\theta_d \\
&= P(W_{ds} \mid Z_{ds} = t*, \phi) \cdot \frac{\Gamma(\sum_{t=1}^T \alpha_t)}{\prod_{t=1}^T \Gamma(\alpha_t)} \cdot \frac{\prod_{t=1}^T \Gamma(\alpha_t + c_{dt})}{\Gamma(\sum_{t=1}^T (\alpha_t + c_{dt}))} \\
&\propto P(W_{ds} \mid Z_{ds} = t*, \phi) \cdot \frac{\prod_{t=1}^T \Gamma(\alpha_t + c_{dt})}{\Gamma(\sum_{t=1}^T (\alpha_t + c_{dt}))} \\
&= P(W_{ds} \mid Z_{ds} = t*, \phi) \cdot \frac{\prod_{t \neq t*} \Gamma(\alpha_t + c_{dt}^{-s})}{\Gamma([\sum_{t=1}^T (\alpha_t + c_{dt}^{-s})] + 1)} \\
&\quad \cdot \Gamma(\alpha_{t*} + c_{dt*}^{-s} + 1) \\
&= P(W_{ds} \mid Z_{ds} = t*, \phi) \cdot \frac{\prod_{t=1}^T \Gamma(\alpha_t + c_{dt}^{-s})}{\Gamma([\sum_{t=1}^T (\alpha_t + c_{dt}^{-s})] + 1)} \\
&\quad \cdot (\alpha_{t*} + c_{dt*}^{-s}) \\
&\propto \phi_{t*s}^{W_{ds}} \cdot (1 - \phi_{t*s})^{1-W_{ds}} \cdot (\alpha_{t*} + c_{dt*}^{-s})
\end{aligned}$$

$$P(Z_{ds} = t* \mid Z_d^{-s}, \phi, I, W) = \frac{\phi_{t*s}^{W_{ds}} \cdot (1 - \phi_{t*s})^{1-W_{ds}} \cdot (\alpha_{t*} + c_{dt*}^{-s})}{\sum_{t=1}^T [\phi_{t*s}^{W_{ds}} \cdot (1 - \phi_{t*s})^{1-W_{ds}} \cdot (\alpha_t + c_{dt}^{-s})]} \quad (2)$$

90 Z_d^{-s} : topic assignment variables for all disease codes excluding disease code s for individual d .
91 c_{dt} : the total number of disease codes assigned with topic t for individual d .
92 c_{dt}^{-s} : the total number of disease codes excluding disease s assigned with topic t for individual d .

93 We reduce the computation time by extracting common terms that are used repetitively. To
94 update Z_{ds} we need to calculate the numerator $\phi_{t*s}^{W_{ds}} \cdot (1 - \phi_{t*s})^{1-W_{ds}} \cdot (\alpha_{t*} + c_{dt*}^{-s})$ for all topics,
95 and then normalize them. Depending on whether W_{ds} is zero or one, only one of $\phi_{t*s}^{W_{ds}}$ and $(1 -$
96 $\phi_{t*s})^{1-W_{ds}}$ needs to be calculated. Take $\phi_{t*s}^{W_{ds}}$ as an example, when $W_{ds} = 1$ we need to calculate
97 $\phi_{t*s} \cdot (\alpha_{t*} + c_{dt*}^{-s})$, which equals $\phi_{t*s} \cdot \alpha_{t*} + \phi_{t*s} \cdot c_{dt*}^{-s}$. When sampling Z_{ds} , $\phi_{t*s} \cdot \alpha_{t*}$ have fixed
98 values. We can calculate and cache these values for all t and s before sampling Z . This simple trick
99 can significantly reduce the time we spend on sampling Z .

100 **3.1.2 Topic loading variable for disease code s in topic t : ϕ_{ts}**

$$\begin{aligned}
P(\phi_{ts} | \cdot) &\propto P(W | Z, \phi_{ts}) \cdot P(\phi_{ts} | a, I_{ts}) \\
&\propto \left[\prod_{d=1}^D I(Z_{ds} = t) \cdot P(W_{ds} | \phi_{ts}) \right] \cdot P(\phi_{ts} | a, I_{ts}) \\
&= \left[\prod_{d=1}^D I(Z_{ds} = t) \cdot (1 - \phi_{ts})^{1-W_{ds}} \cdot \phi_{ts}^{W_{ds}} \right] \cdot \text{Beta}(\phi_{ts} | a_{I_{ts}}^0, a_{I_{ts}}^1) \\
&\propto \phi_{ts}^{\sum_{d=1}^D \sum_{Z_{ds}=t} W_{ds}} \cdot (1 - \phi_{ts})^{\sum_{d=1}^D \sum_{Z_{ds}=t} (1-W_{ds})} \cdot (1 - \phi_{ts})^{a_{I_{ts}}^1 - 1} \cdot (\phi_{ts})^{a_{I_{ts}}^0 - 1} \\
P(\phi_{ts} | \cdot) &\sim \text{Beta}(a_{I_{ts}}^0 + N_{st}^1, a_{I_{ts}}^1 + N_{st}^0). \tag{3}
\end{aligned}$$

101 N_{st}^0 : among individuals who have $W_{ds}=0$, the total number of individuals whose disease variables
102 s are assigned with topic t .
103 N_{st}^1 : among individuals who have $W_{ds}=1$, the total number of individuals whose disease variables
104 s are assigned with topic t .

$$\begin{aligned}
N_{st}^0 &= \sum_{d=1}^D I(Z_{ds} = t, W_{ds} = 0). \\
N_{st}^1 &= \sum_{d=1}^D I(Z_{ds} = t, W_{ds} = 1).
\end{aligned}$$

105 **3.1.3 Indicator variable for disease code s in topic t : I_{ts}**

$$\begin{aligned}
P(I_{ts} = 0 | \cdot) &\propto P(\phi_{ts} | I_{ts} = 0) \cdot P(I_{ts} = 0 | I_{ts}^{pa}) \cdot P(I_{ts}^{ch} | I_{ts} = 0) \\
&= P(\phi_{ts} | I_{ts} = 0) \cdot ((1 - \rho_{11})^{I_{ts}^{pa}} \cdot (1 - \rho_{01})^{1-I_{ts}^{pa}}) \cdot (\rho_{01}^{N_{ts,ch}^1} \cdot (1 - \rho_{01})^{N_{ts,ch}^0}) \\
&= \text{Beta}(\phi_{ts} | a_0^0, a_0^1) \cdot ((1 - \rho_{11})^{I_{ts}^{pa}} \cdot (1 - \rho_{01})^{1-I_{ts}^{pa}}) \cdot (\rho_{01}^{N_{ts,ch}^1} \cdot (1 - \rho_{01})^{N_{ts,ch}^0}). \\
P(I_{ts} = 1 | \cdot) &\propto \text{Beta}(\phi_{ts} | a_1^0, a_1^1) \cdot (\rho_{11}^{I_{ts}^{pa}} \cdot \rho_{01}^{1-I_{ts}^{pa}}) \cdot (\rho_{11}^{N_{ts,ch}^1} \cdot (1 - \rho_{11})^{N_{ts,ch}^0}). \tag{4}
\end{aligned}$$

106 We have omitted the uninformative Bernoulli prior on I_{ts} which is cancelled out when normalising
107 $P(I_{ts} = 1 | \cdot)$ and $P(I_{ts} = 0 | \cdot)$ such that they sum to 1.
108 I_{ts}^{pa} : the indicator variable for the parent code of code s in topic t .
109 I_{ts}^{ch} : the indicator variables for all the children codes of code s in topic t .
110 $N_{ts,ch}^1$: the total number of indicator variables that equal 1 among all the chil-
111 dren codes of code s in topic t .
112 $N_{ts,ch}^0$: the total number of indicator variables that equal 0 among all the chil-
113 dren codes of code s in topic t .

114 4 Learn hyperparameters of treeLFA

115 There are three hyperparameters for treeLFA, including vector α that pa-
116 rameterise the Dirichlet prior for topic weight vector θ , the transition probability
117 vector ρ for the Markov process on the tree structure, and vector a that param-
118 eterise the Beta priors for the probability variables ϕ .

119 4.1 Beta prior for transition probability of the Markov 120 process on the tree structure

121 The inference of the transition probability vector ρ of the Markov process
122 is integrated into the Gibbs sampling framework for treeLFA by putting Beta
123 priors on them as follows:

$$\begin{aligned} P(\rho_{01}) &\sim \text{Beta}(b_{00}, b_{01}) \\ P(\rho_{11}) &\sim \text{Beta}(b_{10}, b_{11}) \end{aligned}$$

124 The updating equations for these transition probability variables are:

$$\begin{aligned} P(\rho_{01} | I) &\sim \text{Beta}(N_1^0 + b_{00}, N_0^0 + b_{01}) \\ P(\rho_{11} | I) &\sim \text{Beta}(N_1^1 + b_{10}, N_0^1 + b_{11}) \end{aligned} \tag{5}$$

125 N_1^0 : the total number of active indicator variables in all topics with an inactive parent
126 indicator variable.

127 N_1^1 : the total number of active indicator variables in all topics with an active parent
128 indicator variable.

129 N_0^0 : the total number of inactive indicator variables in all topics with an inactive
130 parent indicator variable.

131 N_0^1 : the total number of inactive indicator variables in all topics with an active parent
132 indicator variables.

133

134 4.2 Beta prior for probability of disease codes in topics

135 We use pre-specified beta distributions as the prior for ϕ . In Figure 3 we
136 show the histograms of 10,000 samples drawn from $\text{Beta}(0.3, 80)$ and $\text{Beta}(2, 4)$,
137 which are the prior distributions we choose for ϕ of inactive and active codes in
138 topics used for simulation and inference on the top-100 UKB dataset.

139 4.3 Optimization of asymmetric Dirichlet prior for topic 140 weights

141 We choose to estimate hyperparameter α to allow for the flexibility that dif-
142 ferent topics are assigned to individuals' disease codes with varying probability.
143 For LDA it is common to use a symmetric (in most cases an uninformative)
144 Dirichlet prior [3] parameterized by α , which implicitly assumes that there are

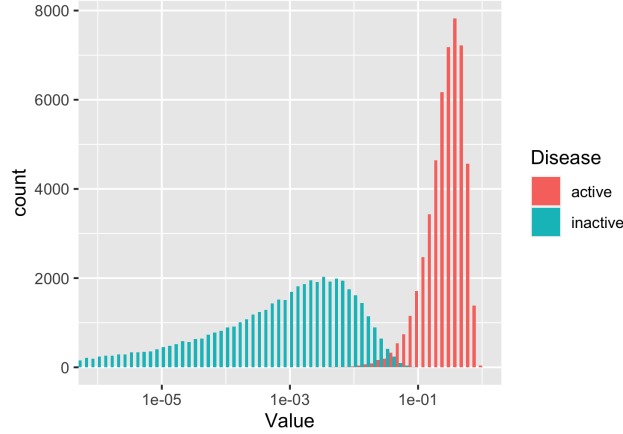


Figure 3: Histogram of samples from the Beta priors for probability of disease codes in topics. Different Beta priors are used for the probability variable ϕ for active and inactive codes in topics, such that active codes tend to have large probability, while inactive codes have negligible small probability. 10,000 samples are drawn from Beta(0.3,80) and Beta(2,4) respectively, and their histograms are plotted.

roughly same number of words assigned to each topic across documents. However, the prevalence of diseases in biobanks varies greatly, from one in thousands to more than 0.1 (for example, 28.2 % individuals in UKB are diagnosed with essential hypertension), calling for different topic frequencies. In addition, an empty topic will always be inferred by treeLFA on real-world diagnostic data, and assigned to most disease variables that have value zero. Consequently, the empty topic will have a much larger weight in α than other disease topics, as most individuals only have a few diagnosed diseases. This also requires an asymmetric Dirichlet prior distribution for θ to be used.

Wallach *et al.* studied the use of complex Dirichlet prior for topic models, and advocated using asymmetric Dirichlet prior for topic weight vector θ and symmetric Dirichlet prior for topic loading ϕ for LDA. He verified with experiments that this strategy of choosing hyperparameters gave the best inference result. Besides, he also suggested the use of an optimization based method to estimate α from the data, since it gave result as good as that can be achieved by using a full Bayesian approach, and was much simpler to implement and time-saving [4].

We use a stochastic EM algorithm, named “Gibbs-EM” to optimise α [5]. The objective function for this method is the marginal likelihood of the data: $P(W \mid \alpha)$, which is maximised with respect to α by alternating between two steps. In the E-step, we fix the value of α and maximise a lower bound of the marginal likelihood by approximating the posterior distribution of hidden variables with G samples given by the Gibbs sampler for treeLFA:

$$\begin{aligned}
\log P(W \mid \alpha) &\geq \sum_Z P(Z \mid W, \alpha) \cdot \log P(W, Z \mid \alpha) \\
&\approx \frac{1}{G} \sum_{g=1}^G \log P(W, Z^g \mid \alpha) \\
&= \frac{1}{G} \sum_{g=1}^G \left[\sum_{d=1}^D \log P(W_d, Z_d^g \mid \alpha) \right].
\end{aligned} \tag{6}$$

168 In the above equation, hidden variables ϕ , I and θ are omitted, and only Z is
169 retained, since it is the only hidden variable related to the optimization of α in
170 the M-step.

171 In the M-step, we use a fixed point method to optimise α such that the
172 expectation calculated in the E-step is maximised. Entries of α are optimised
173 one at a time:

$$\alpha_t^{new} = \alpha_t \cdot \frac{\sum_g \sum_d \{\Psi(N_{dt}^g + \alpha_t) - \Psi(\alpha_t)\}}{\sum_g \sum_d \{\Psi(S + \sum_k (\alpha_k)) - \Psi(\sum_k (\alpha_k))\}}. \tag{7}$$

174 N_{dt}^g : the total number of disease variables for individual d that are assigned with topic
175 t in posterior sample g ;
176 $\Psi(\cdot)$: the digamma function.
177 S : the total number of disease codes.

178 5 Model selection: predictive likelihood on the 179 testing dataset

180 We use predictive likelihood on the held out test data to evaluate models
181 with different topic numbers. We use posterior samples of ϕ and the optimal
182 α estimated within the training data to compute the predictive likelihood on
183 the test data. With test data W' and one posterior sample of ϕ , the predictive
184 likelihood on the test data could be expressed as [6]:

$$P(W' | W) = \prod_{d=1}^D P(W'_d | \phi, \alpha)$$

185 Ideally, calculation of the predictive likelihood requires integrating out all
186 latent variables. Since this is analytically intractable, we use a Monte-Carlo
187 approximation for this integral. For topic weight variable θ_d , we draw P samples
188 from its prior distribution and use them to approximate the integral:

$$\begin{aligned} \theta_d^p &\sim \text{Dirichlet}(\alpha) \\ P(W'_d | \phi, \alpha) &\approx \frac{1}{P} \sum_{p=1}^P P(W'_d | \theta_d^p, \phi, \alpha) \end{aligned}$$

189 Conditioned on one sample of θ_d for individual d , we calculate the likelihood of each
190 disease variable (W_{ds}) independently. We do this by summing out Z_{ds} for all topics:

$$\begin{aligned} P(W'_d | \theta_d^p, \phi) &= \prod_{s=1}^S \left[\sum_{Z_{ds}=1}^T P(W'_{ds}, Z_{ds} | \theta_d^p, \phi) \right] \\ &= \prod_{s=1}^S \left[\sum_{Z_{ds}=1}^T P(W'_{ds} | \phi_{Z_{ds},s}) \cdot P(Z_{ds} | \theta_d^p) \right] \end{aligned}$$

191 In summary, the predictive likelihood on the full test data could be expressed as:

$$P(W' | W) \approx \prod_{d=1}^D \left\{ \frac{1}{P} \sum_{p=1}^P \left\{ \prod_{s=1}^S \left[\sum_{Z_{ds}=1}^T [P(W'_{ds} | \phi_{Z_{ds},s}) \cdot P(Z_{ds} | \theta_d^p)] \right] \right\} \right\} \quad (8)$$

192 References

- 193 [1] Alberto Lumbreras, Louis Filstroff, and Cédric Févotte. Bayesian mean-
194 parameterized nonnegative binary matrix factorization. *Data Mining and*
195 *Knowledge Discovery*, 34(6):1898–1935, 2020.
- 196 [2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet alloca-
197 tion. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
- 198 [3] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings*
199 *of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.
- 200 [4] Hanna M Wallach, David M Mimno, and Andrew McCallum. Rethinking
201 lda: Why priors matter. In *Advances in neural information processing sys-*
202 *tems*, pages 1973–1981, 2009.
- 203 [5] Thomas Minka. Estimating a dirichlet distribution, 2000.
- 204 [6] Andrew Kachites McCallum. Mallet: A machine learning for language
205 toolkit. *http://mallet.cs.umass.edu*, 2002.