**TEXT MINING FOR SOCIAL SCIENCES**

**PROBLEM SET 4**

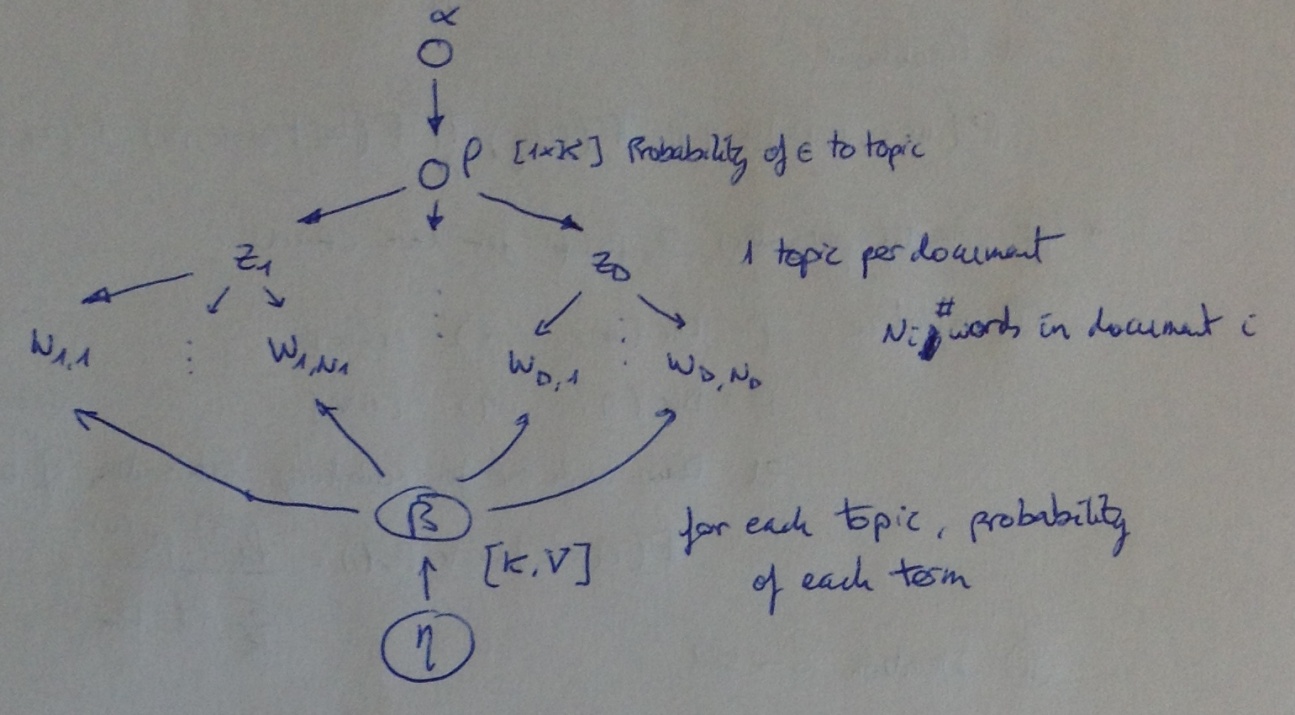
**Maria Fernández, Tim Kreienkamp, Joan Verdú**

*2. Consider the following Bayesian formulation of the multinomial mixture model for which you derived the EM algorithm in the previous assignment:*

* *Draw K topics from a Dirichlet, i.e. βk ~ Dirichlet(η). Note that βk 2 €ΔV where V is the number of unique vocabulary terms.*
* *Draw mixture probabilities ρ €ΔK from Dirichlet(α).*
* *For each document d draw a topic zd €{0;… ;K}, where Pr [ zd = k ] = ρ k.*
* *Draw each word in document d from βzd .*

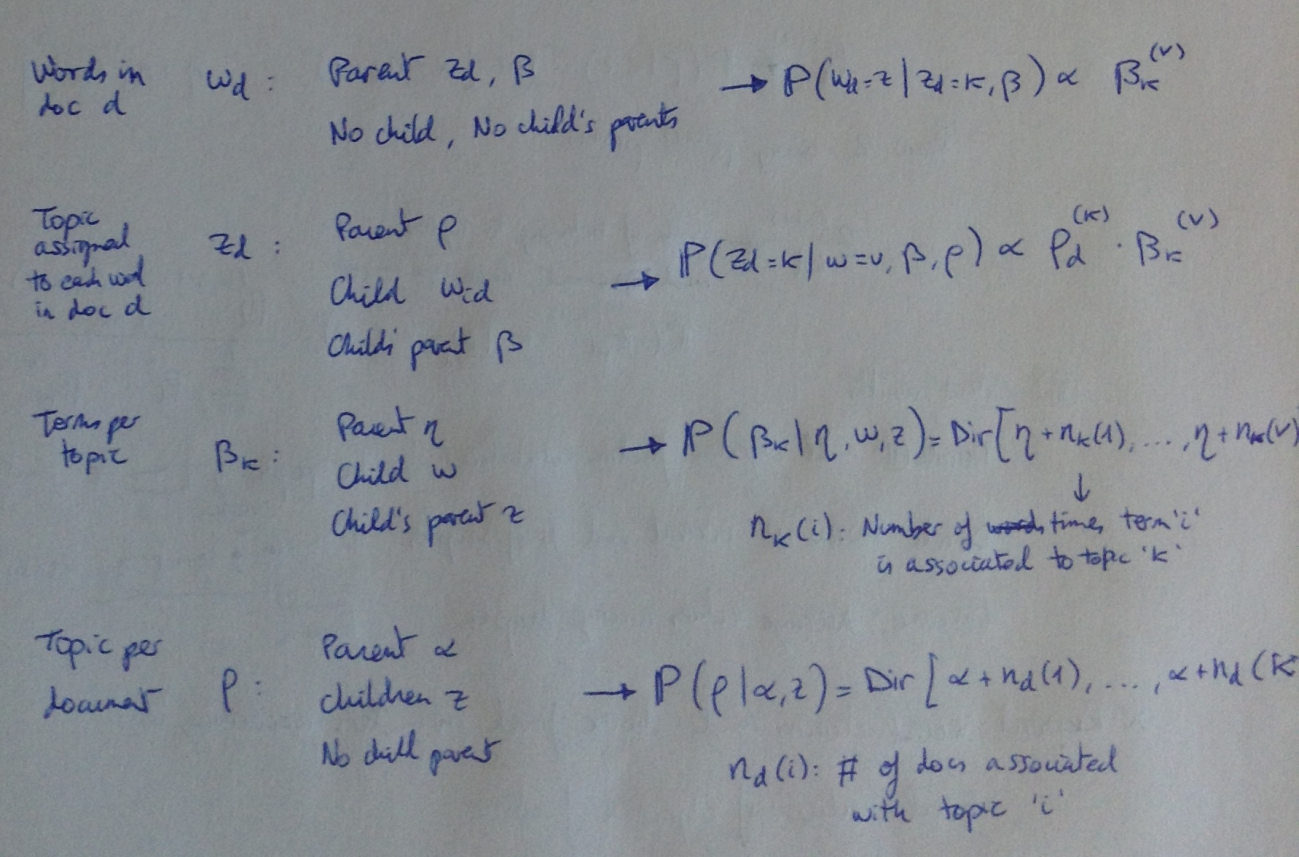
*Now answer the following questions:*

1. *Represent this data generating process as a directed graph.*



1. *What are the Markov blankets of the words in document d, the topic assignment zd, and the topics \_k?*

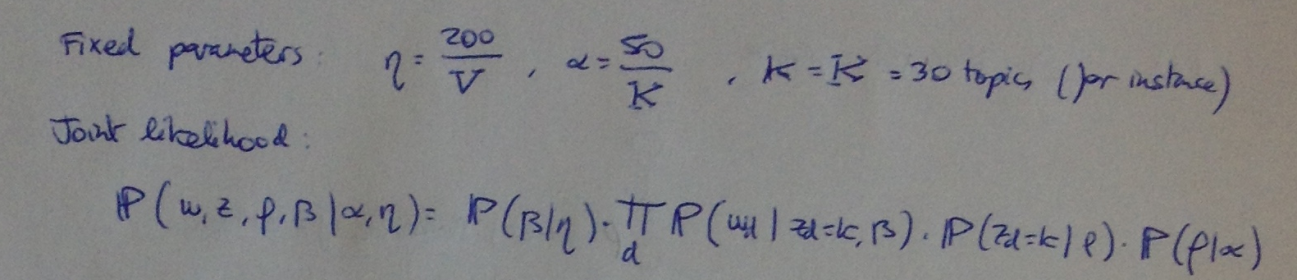
Here we show both Markov blankets and related posterior probability distributions. We also add *ρ*.



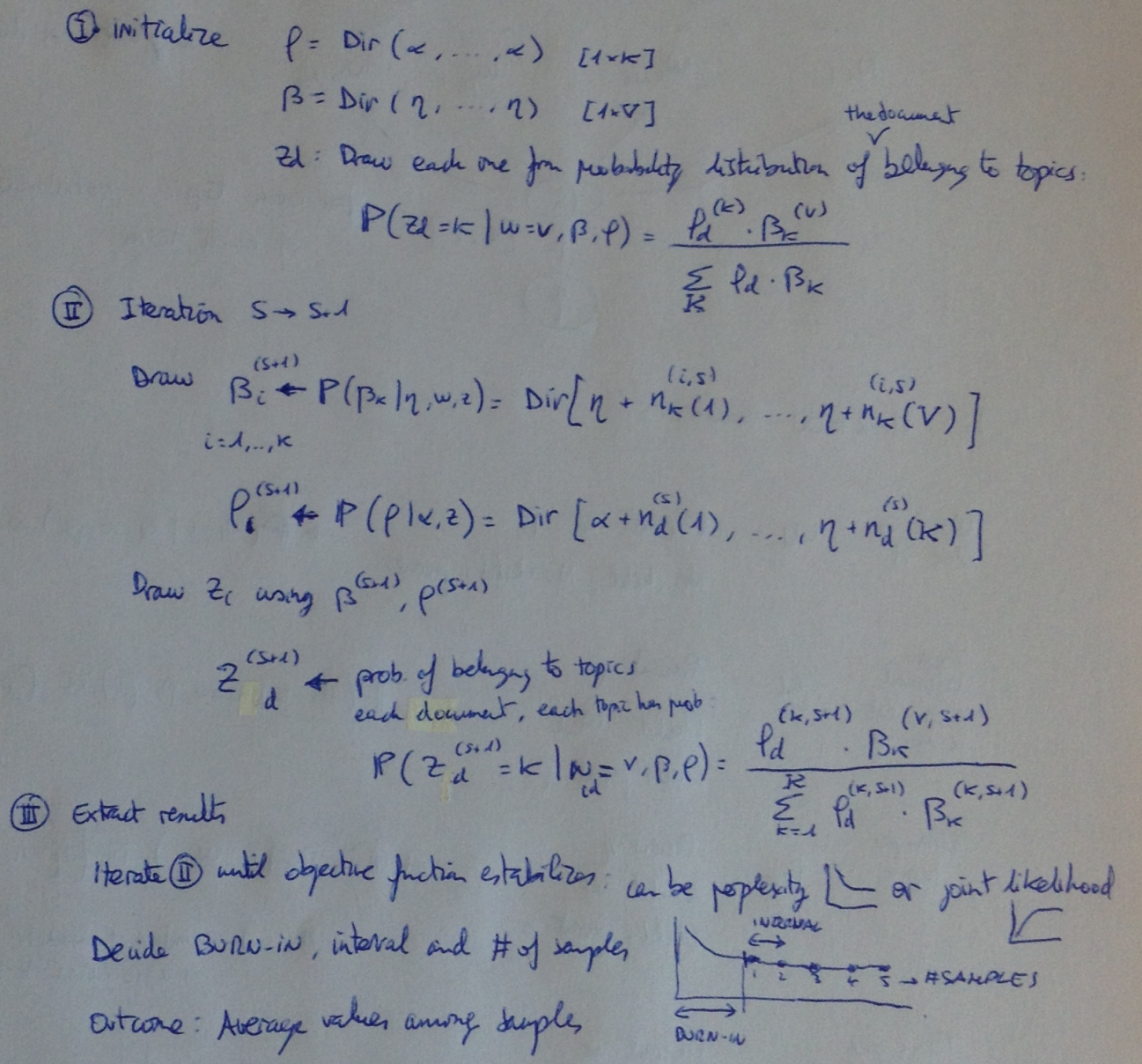
Zd is the topic assigned to document d for **all** words in this document d.

1. *Use your answer to the above to propose an (uncollapsed) Gibbs sampler for posterior inference for the Bayesian multinomial mixture model.*

Fixed parameters are η, α and K. The words of each doc are known (wd).



The parameters to estimate are β, ρ and z. Here is the proposed uncollapsed Gibbs sampler to estimate parameters (one chain).



The estimated parameters would be the average values across samples.

Then we should run several chains to optimize the result, and choose the final parameters according to several criteria, such as:

* Good value of objective function (low if perplexity, high if joint likelihood used)
* Good interpretability of topics
* Cross validation with out-of-sample documents