Probabilistic Topic Models

LINGZHE ZHAO
S1502670@SMS.ED.AC.UK

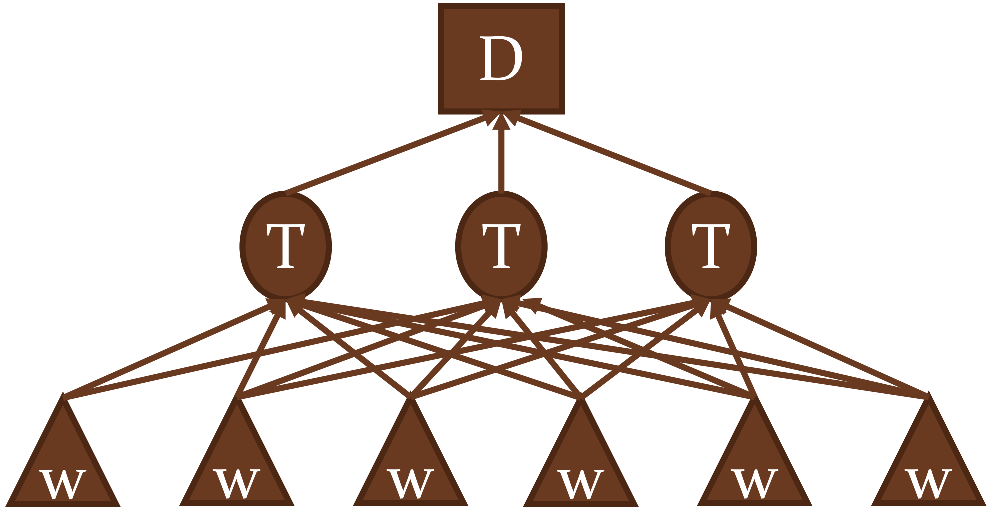
Topic modelling

Probabilistic Topic Models

Mark Steyvers
University of California, Irvine

Tom Griffiths
Brown University

Introduction



Topic 247

word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

Topic 5

word	prob.
RED	.202
BLUE	.099
GREEN	.096
YELLOW	.073
WHITE	.048
COLOR	.048
BRIGHT	.030
COLORS	.029
ORANGE	.027
BROWN	.027
PINK	.017
LOOK	.017
BLACK	.016
PURPLE	.015
CROSS	.011
COLORED	.009

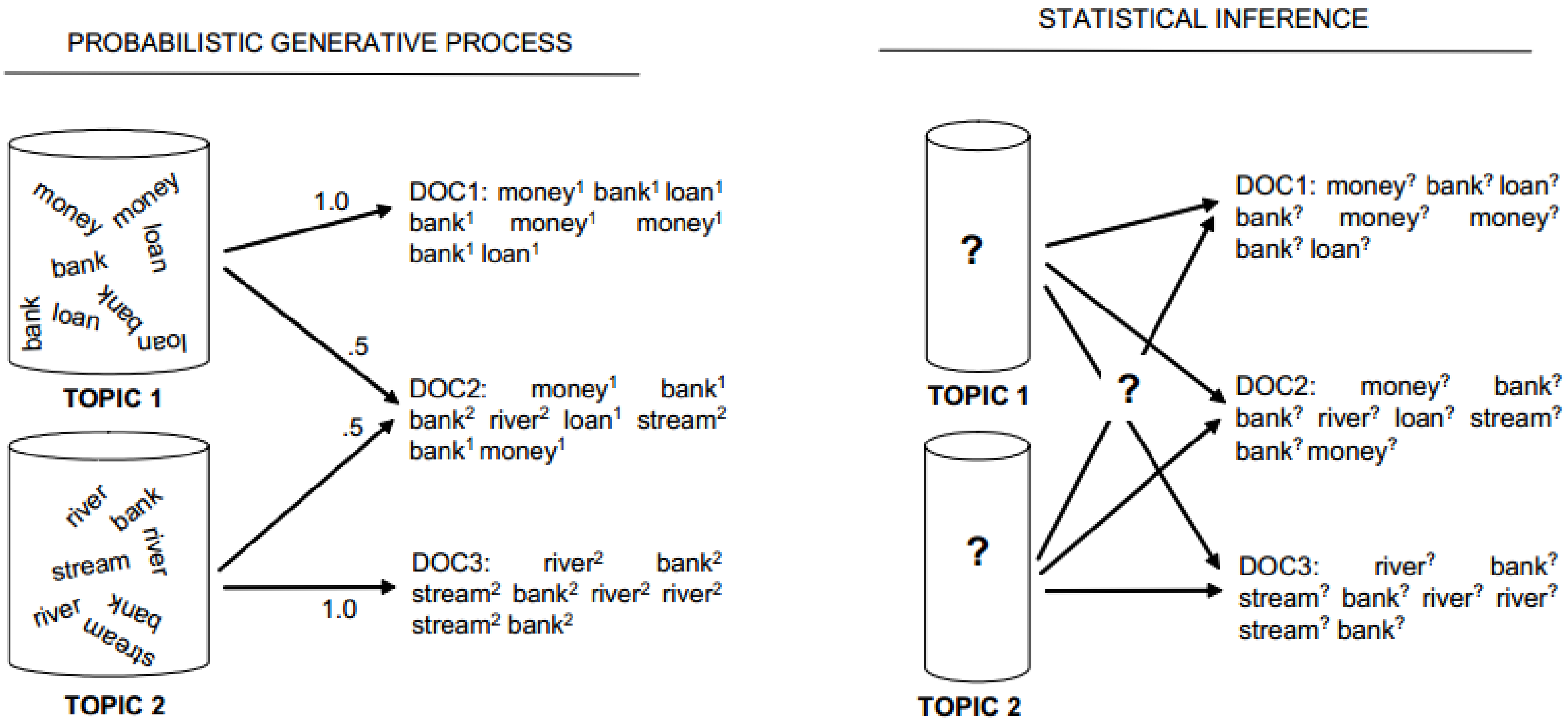
Introduction

Representation

Probabilistic topics | Axes of a spatial

- **Advantage** : Each topic is individually interpretable, providing a probability distribution over words that picks out a coherent cluster of correlated terms.

Generative Models



Probabilistic Topic Models

$$P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j) P(z_i = j)$$

Hofmann	Blei et. al	Griffiths and Steyvers
pLSI	LDA	Variant LDA
1999; 2001	2003	2002; 2003; 2004

Probabilistic Topic Models

LDA(Latent Dirichlet Allocation)

The probability density of a T dimensional Dirichlet distribution over the multinomial distribution $p = (p_1, \dots, p_T)$ is defined by:

$$\text{Dir}(\alpha_1, \dots, \alpha_T) = \frac{\Gamma\left(\sum_j \alpha_j\right)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{T} p_j^{\alpha_j - 1}$$

?

The Dirichlet is a convenient distribution on the simplex — it is in the exponential family, has finite dimensional sufficient statistics, and is conjugate to the multinomial distribution. In Section 5, these properties will facilitate the development of inference and parameter estimation algorithms for LDA.

Probabilistic Topic Models

Dirichlet distribution

$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

From PRML

posterior distribution for the parameters $\{\mu_k\}$ in the form

$$p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\alpha}) \propto p(\mathcal{D}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1}.$$

From PRML

We see that the posterior distribution again takes the form of a Dirichlet distribution, confirming that the Dirichlet is indeed a conjugate prior for the multinomial.

Probabilistic Topic Models

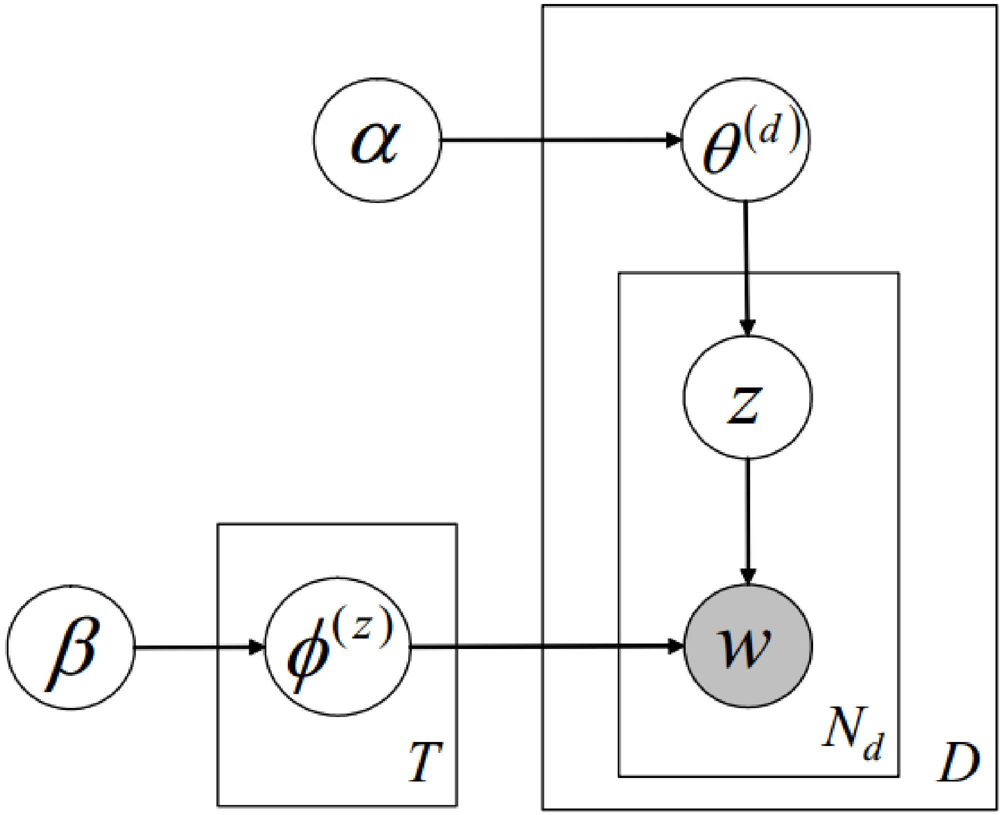
$\theta^{(d)} \sim \mathrm{P}(z_i = j)$
$\phi^{z} \sim \mathrm{P}(w_i | z_i = j)$
$\alpha, \beta \sim$hyperparameter

Variant LDA

Arrows indicate conditional dependencies between variables

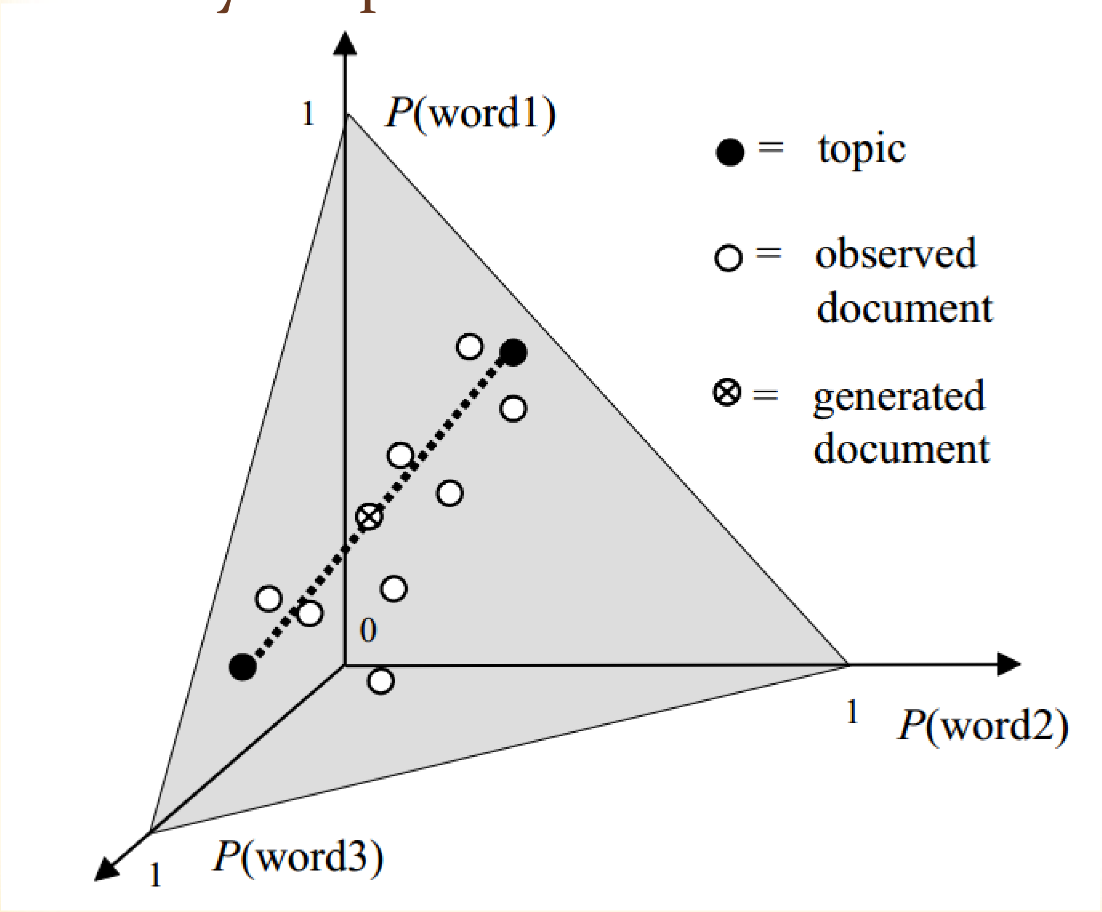
Plates (the boxes in the figure) refer to repetitions of sampling steps

Variable in the lower right corner referring to the number of samples.



Interpretation

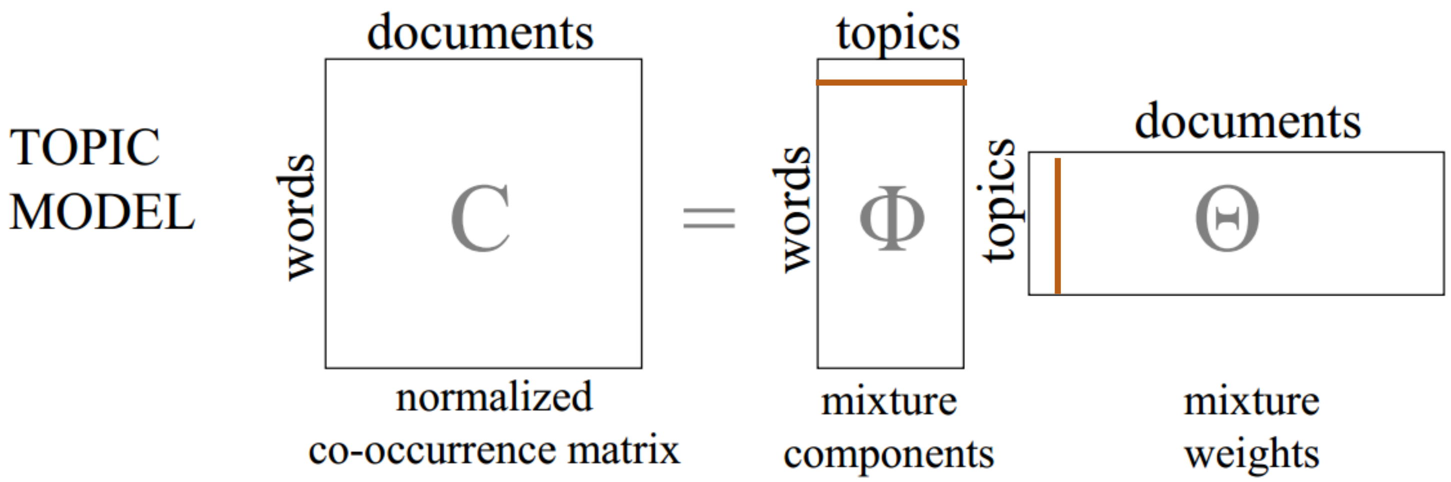
By Graph



- Each axis represents the probability of observing a particular word type.(with W dimentions)
- The W-1 dimensional simplex represents all probability distributions over words.
- Each **document** and **topic** in the text collection can be represented as a point on the simplex
- Each document that is generated by the model is a convex combination of the T topics

Interpretation

By Matrix Factorization



- Feature values are non-negative and sum up to one.
- Topic-word distributions are independent but not orthogonal

Algorithm for Extracting Topics

$$\theta^{(d)} \sim \mathrm{P}(z_i = j)$$
$$\phi^{z} \sim \mathrm{P}(w_i | z_i = j)$$

Hofmann
EM
local maxima of the likelihood function

Algorithm for Extracting Topics

θ, ϕ

Directly estimate the posterior distribution

Gibbs Sampling

Many text collections contain millions of word token.

The estimation of the posterior over z requires efficient estimation procedures.

Gibbs sampling is one of the best choices.

Algorithm for Extracting Topics

MCMC

(a) (b)

From MLAPP

$$p(X_{1:T}) = p(X_1)p(X_2|X_1)p(X_3|X_2)\ldots = p(X_1)\prod_{t=2}^{T} p(X_t|X_{t-1})$$

From MLAPP

Algorithm for Extracting Topics

Gibbs Sampling

topic assignment of token i to topic j

topic assignments of all other word tokens

word indices

document indices

$$P(z_i = j \mid \mathbf{z}_{-i}, w_i, d_i, \cdot) \propto \frac{C^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha}$$

other known or observed information

word w is assigned to topic j

topic j is assigned to some word token in document d

Algorithm for Extracting Topics

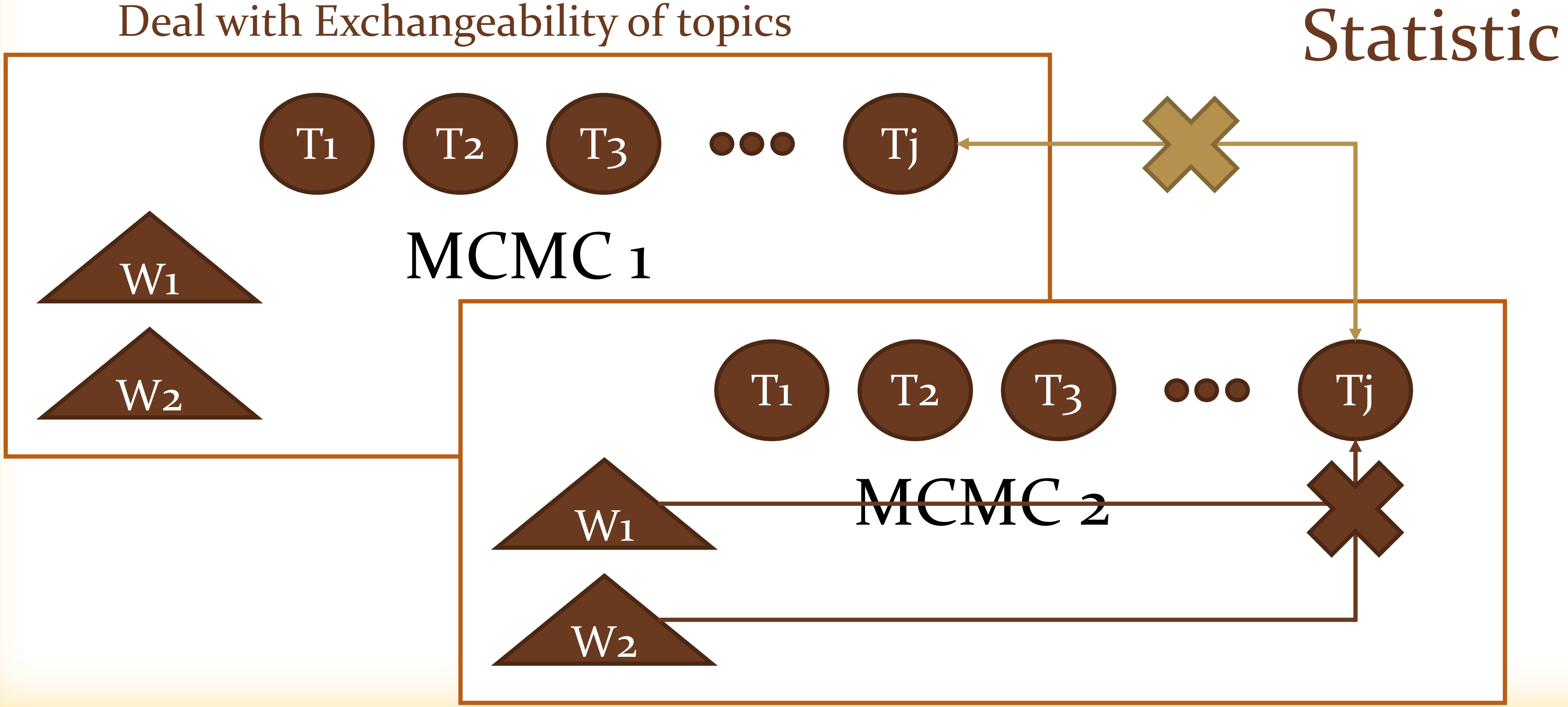
$$C_{WT}\begin{bmatrix} c_{11} & c_{12} & c_{13} & c_{14} & c1T \\ c_{21} & c_{22} & \cdots & c_{2t} & c_{2T} \\ c_{31} & \vdots & \ddots & \vdots & c_{3T} \\ c_{w1} & c_{w2} & \cdots & c_{wt} & c_{wT} \\ c_{W1} & c_{W2} & c_{W3} & c_{Wt} & c_{WT} \end{bmatrix}$$

$$C_{DT}\begin{bmatrix} c_{11} & c_{12} & c_{13} & c_{14} & c1T \\ c_{21} & c_{22} & \cdots & c_{2t} & c_{2T} \\ c_{31} & \vdots & \ddots & \vdots & c_{3T} \\ c_{d1} & c_{d2} & \cdots & c_{dt} & c_{dT} \\ c_{D1} & c_{D2} & c_{D3} & c_{Dt} & c_{DT} \end{bmatrix}$$

$$\begin{bmatrix} P_{11} & P_{12} & P_{13} & P_{14} & P1T \\ P_{21} & P_{22} & \cdots & P_{2t} & P_{2T} \\ P_{31} & \vdots & \ddots & \vdots & P_{3T} \\ P_{w1} & P_{w2} & \cdots & P_{wt} & P_{wT} \\ P_{W1} & P_{W2} & P_{W3} & P_{Wt} & P_{WT} \end{bmatrix}$$

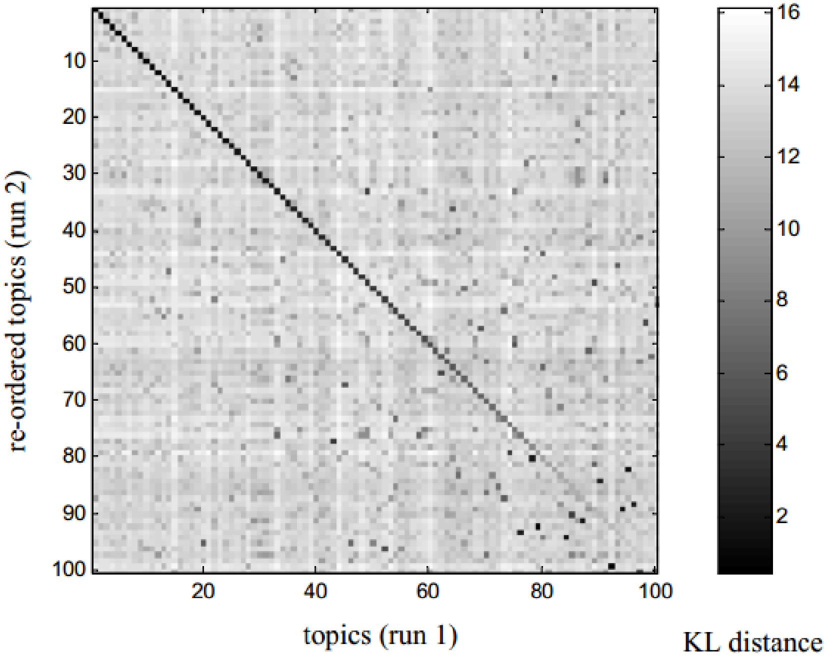
Algorithm for Extracting Topics

Deal with Exchangeability of topics



Algorithm for Extracting Topics

Stability of Topics



Run 1		Run 2	
MONEY	.094	MONEY	.086
GOLD	.044	PAY	.033
POOR	.034	BANK	.027
FOUND	.023	INCOME	.027
RICH	.021	INTEREST	.022
SILVER	.020	TAX	.021
HARD	.019	PAID	.016
DOLLARS	.018	TAXES	.016
GIVE	.016	BANKS	.015
WORTH	.016	INSURANCE	.015
BUY	.015	AMOUNT	.011
WORKED	.014	CREDIT	.010
LOST	.013	DOLLARS	.010
SOON	.013	COST	.008
PAY	.013	FUNDS	.008

$$KL(j_1, j_2) = \frac{1}{2}\sum_{k=1}^{W} \phi'^{(j_1)}_k \log_2 \phi'^{(j_1)}_k / \phi''^{(j_2)}_k + \frac{1}{2}\sum_{k=1}^{W} \phi''^{(j_2)}_k \log_2 \phi''^{(j_2)}_k / \phi'^{(j_1)}_k$$

Conclusion

- Generative Models
- Different representations
- Improvement