THE UNIVERSITY OF TEXAS AT AUSTIN
Department of Computer Science
College of Natural Sciences

Fixed-Length Poisson MRF: Adding Dependencies to the Multinomial

David I. Inouye, Pradeep Ravikumar, Inderjit S. Dhillon

Center for Big Data Analytics
http://bigdata.ices.utexas.edu

LPMRF Overview

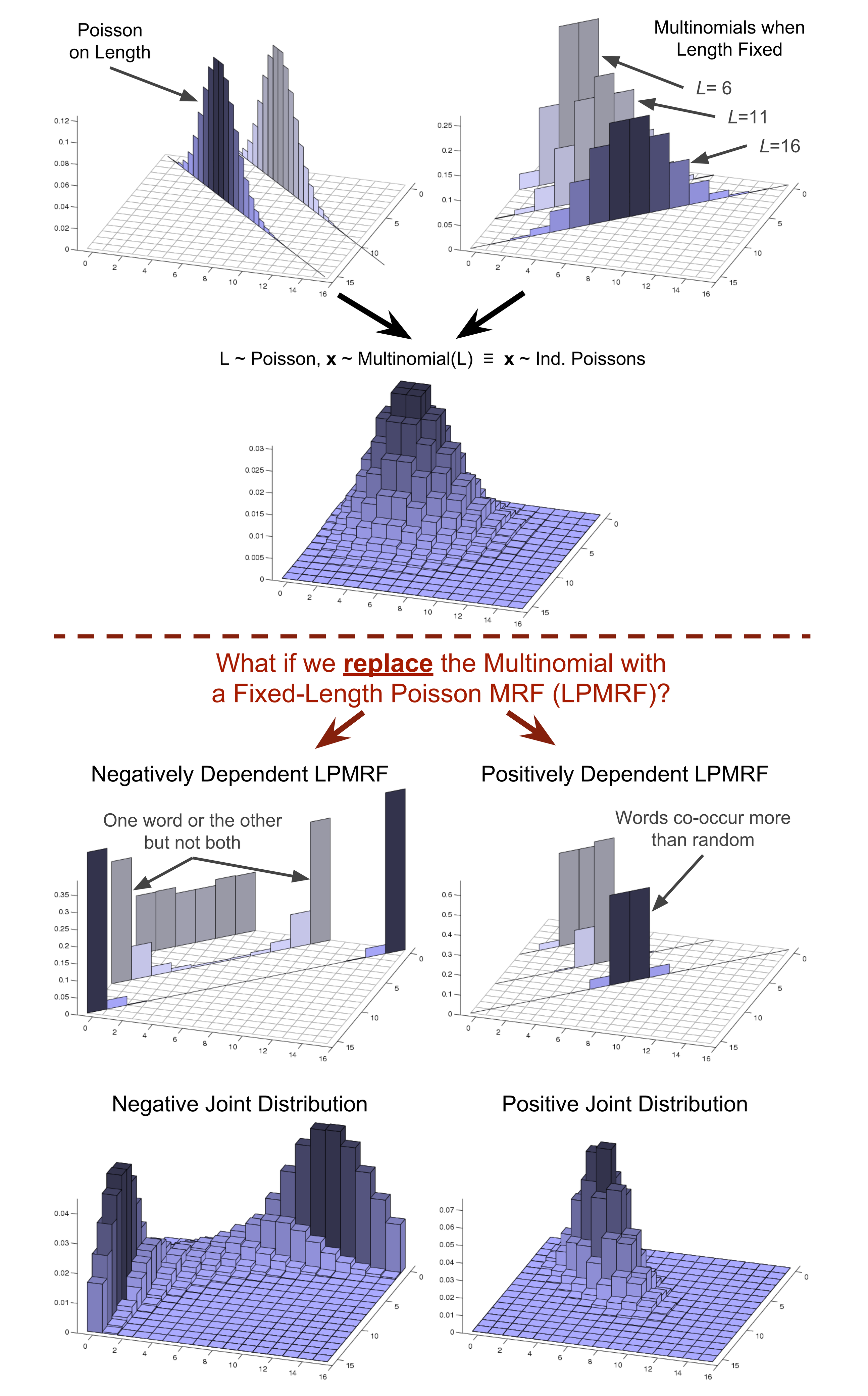


Figure : The Fixed-Length Poisson MRF (LPMRF) can replace the Multinomial by relaxing the independence assumption and allowing for positive and negative dependencies between words (or more generally between covariates).

Illustration with Real Documents

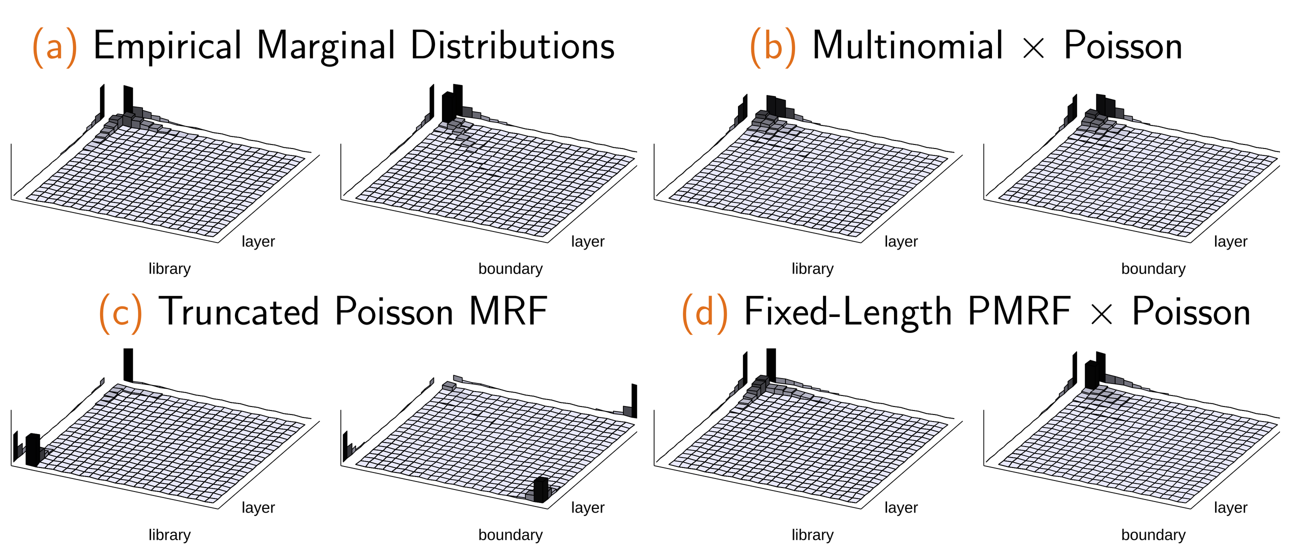


Figure : The simple empirical distribution clearly shows a strong dependency between "boundary" and "layer" but strong negative dependency of "boundary" with "library". Clearly, the word-independent Multinomial-Poisson distribution underfits the data. While the Truncated PMRF can modeled dependencies, it obviously has normalization problems because the normalization is dominated by the edge case. The LPMRF-Poisson distribution much more appropriately fits the empirical data.

Background: Poisson as Exp. Family

The standard Poisson(λ) can be transformed into an exponential family distribution with **natural** parameter η:

$$\Pr_{\text{Poiss}}(x \mid \lambda) = \frac{\lambda^x}{x!}\exp(-\lambda) = \exp(\log(\frac{\lambda^x}{x!}))\exp(-\lambda)$$
$$= \exp(\log(\lambda^x) - \log(x!) - \lambda) = \exp(\log(\lambda)x - \log(x!) - \lambda)$$
$$= \exp(\underbrace{\eta x}_{\log(\lambda)\equiv\eta} - \log(x!) - \underbrace{\exp(\eta)}_{\lambda\equiv\exp(\eta)}).$$

Background: Poisson MRF

By assuming that the conditional distribution of a variable x_s given all other variables $\mathbf{x}_{\setminus s}$ is a univariate Poisson, a joint Poisson distribution can defined [Yang et al. 2012]:

$$\Pr_{\text{PMRF}}(\mathbf{x} \mid \boldsymbol{\theta}, \Phi) \propto \exp\left\{\boldsymbol{\theta}^T\mathbf{x} + \mathbf{x}^T\Phi\mathbf{x} - \sum_{s=1}^{p}\log(x_s!)\right\},$$

where $\boldsymbol{\theta} \in \mathbb{R}^p$ and $\Phi \in \{\mathbb{R}^{p\times p} : \text{diag}(\Phi) = 0\}$.
Node conditionals (i.e. the distribution of one word given all other words) are 1-D Poissons:

$$\Pr(x_s \mid \mathbf{x}_{-s}, \theta_s, \Phi_s) \propto \exp\{\underbrace{(\theta_s + \mathbf{x}^T\Phi_s)}_{\eta_s} x_s - \log(x_s!)\}.$$

Fixed-Length PMRF (LPMRF)

$$\Pr_{\text{LPMRF}}(\mathbf{x}|\boldsymbol{\theta}, \Phi, L) = \exp(\boldsymbol{\theta}^T\mathbf{x} + \mathbf{x}^T\Phi\mathbf{x} - \textstyle\sum_s \log(x_s!) - \text{A}_L(\boldsymbol{\theta}, \Phi))$$
$$\text{A}_L(\boldsymbol{\theta}, \Phi) = \log\sum_{\mathbf{x}\in\mathcal{X}_L}\exp(\boldsymbol{\theta}^T\mathbf{x} + \mathbf{x}^T\Phi\mathbf{x} - \textstyle\sum_s\log(x_s!))$$
$$\mathcal{X}_L = \{\mathbf{x} : \mathbf{x} \in \mathbb{Z}_+^p, \|\mathbf{x}\|_1 = L\}$$

Note that the only—but critical—difference from the PMRF parametric form is the log partition function $\text{A}_L(\boldsymbol{\theta}, \Phi)$ which is conditioned on the set $\mathcal{X}_L$ (unlike the unbounded set for PMRF).

Extension to Topic Models

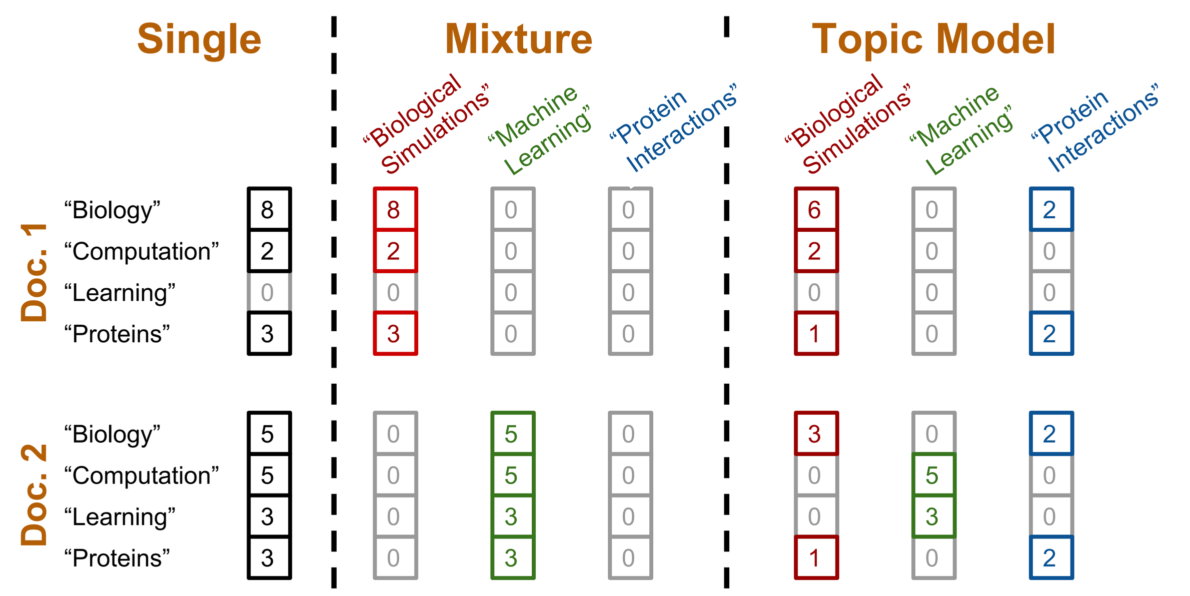


Figure : Topic models can be viewed as an extension of mixture models that allow words in the same document to come from different topics.

We generalize topic models using fixed-length distributions:

Generic Topic Model	LPMRF Topic Model
$\mathbf{w}_i \sim \text{SimplexPrior}(\alpha)$	$\mathbf{w}_i \sim \text{Dirichlet}(\alpha)$
$L_i \sim \text{LengthDist}(\bar{L})$	$L_i \sim \text{Poisson}(\lambda = \bar{L})$
$\mathbf{m}_i \sim \text{PartitionDist}(\mathbf{w}_i, L_i)$	$\mathbf{m}_i \sim \text{Mult}(\mathbf{p} = \mathbf{w}_i\,;\, N = L_i)$
$\mathbf{z}_i^j \sim \text{FixedLength}(\phi^j; \|\mathbf{z}_i^j\| = m_i^j)$	$\mathbf{z}_i^j \sim \text{LPMRF}(\boldsymbol{\theta}^j, \Phi^j; N = m_i^j)$
$\mathbf{x}_i = \sum_{j=1}^k \mathbf{z}_i^j$	$\mathbf{x}_i = \sum_{j=1}^k \mathbf{z}_i^j.$

Log Partition Function (Likelihood)

LPMRF Gibbs Sampling

- Multinomial = Sum of individual variables drawn i.i.d.
- LPMRF = Sum of individual variables with conditional dependence
- Conditional distribution of one word given all other words:

$$\Pr(\mathbf{w}_\ell = e_s \mid \mathbf{w}_1, \dots, \mathbf{w}_{\ell-1}, \mathbf{w}_{\ell+1}, \dots, \mathbf{w}_L, \boldsymbol{\theta}, \Phi)$$
$$\propto \exp(\boldsymbol{\theta}_s + 2\Phi_s\mathbf{x}_{-\ell}).$$

LPMRF Annealed Importance Sampling

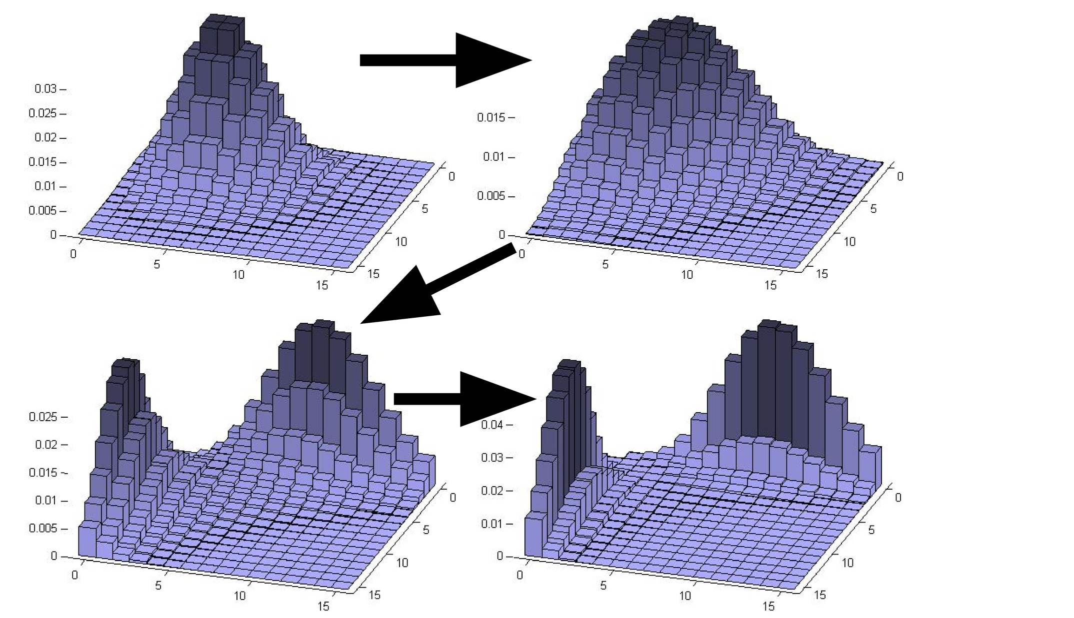


Figure : Annealed importance sampling starts from a well-known simple distribution (in our case a Multinomial distribution) and slowly moves toward the target distribution recording importance weights between each transition. This sampling can be used to approximate the log partition function.

Upper Bound on Log Partition

$$\text{A}_L(\boldsymbol{\theta}, \Phi) \le L^2\lambda_{\Phi,1} + L\log(\textstyle\sum_s e^{\theta_s}) - \log(L!).$$

Generalize to Multiple L

- What if we want estimates for different values of L?
- Introduce **weighting function**:

$$\omega(L) = 1 - \text{LogLogisticCDF}(L \mid \alpha_{\text{LL}}, \beta_{\text{LL}})$$

where $\alpha_{\text{LL}} \in [2, 3]$ and $\beta_{\text{LL}} = 2$ so that tail is $O(1/L^2)$.

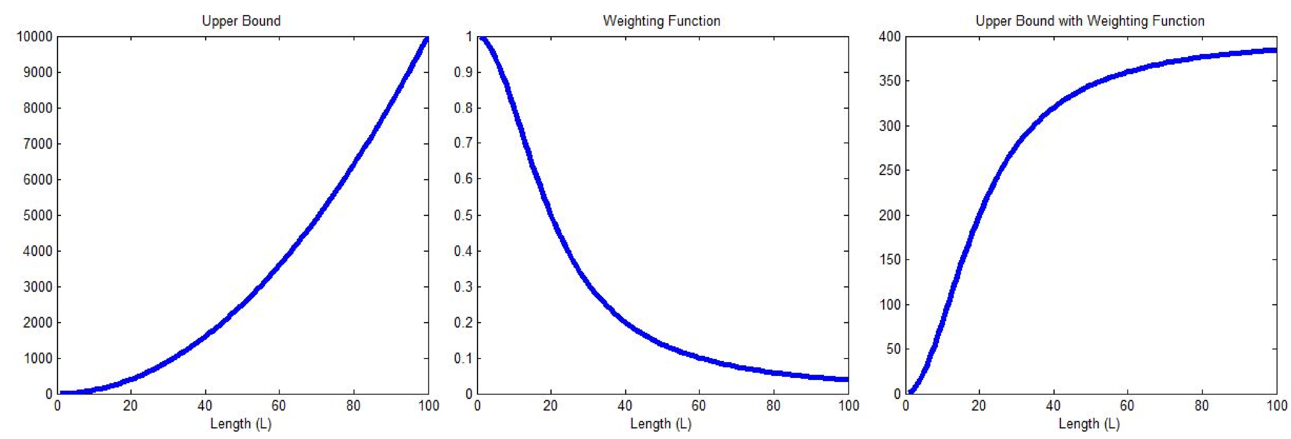


Figure : (Left) Quadratic upper bound. (Middle) Sigmoid weighting function based on Log Logistic CDF. (Right) Final functional form of log partition function for various L with weighting function.

Final Approximation

- 5,000 AIS samples - 100 AIS samples for 50 different test values of L linearly spaced between the $0.5\bar{L}$ and $3\bar{L}$.

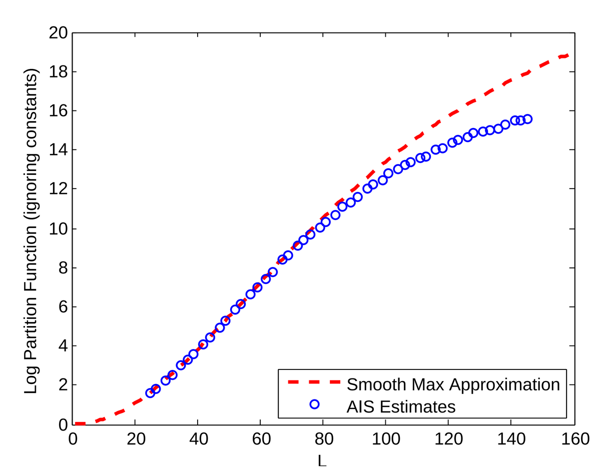
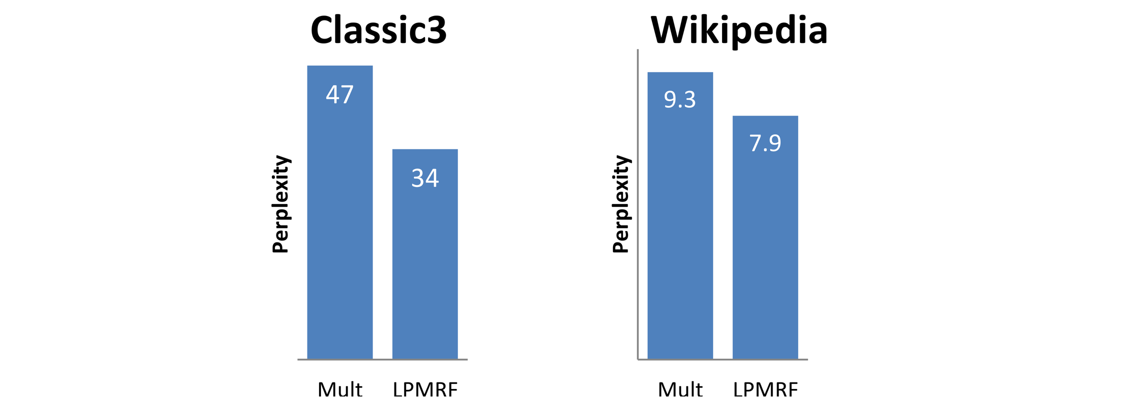


Figure : Example of log partition estimation for all values of L.

Perplexity Results

$$\text{Perplexity} = \exp(-\mathcal{L}(X^{\text{test}}|\theta^{1\dots k}, \Phi^{1\dots k})/N_{\text{test}})$$

(The inverse of the geometric mean of the per-word likelihood)



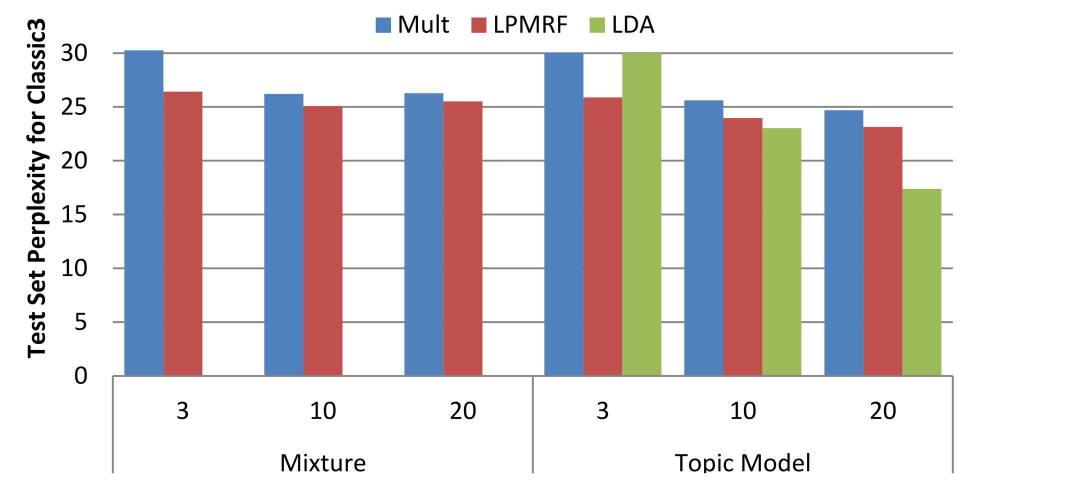
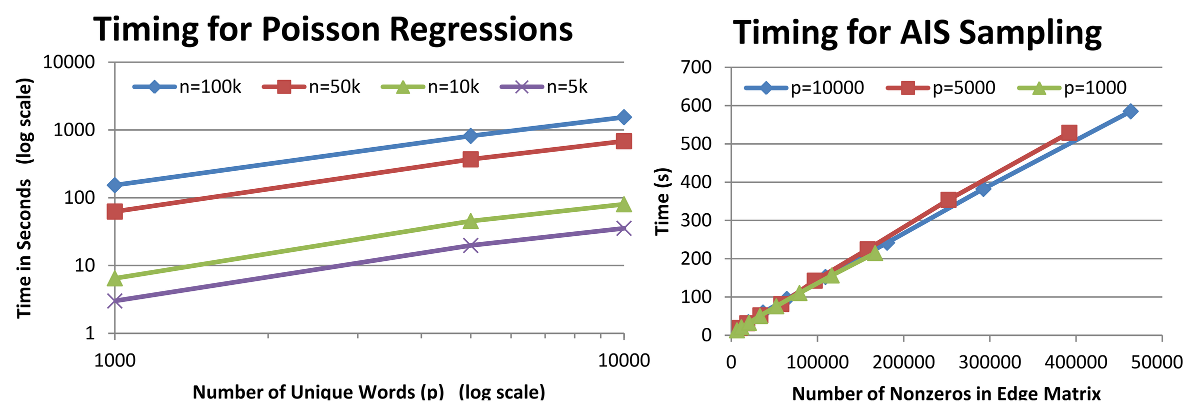


Figure : The LPMRF models always outperform the corresponding Multinomial models. However, the well-developed LDA topic model is competitive for larger number of topics.

Timing Results



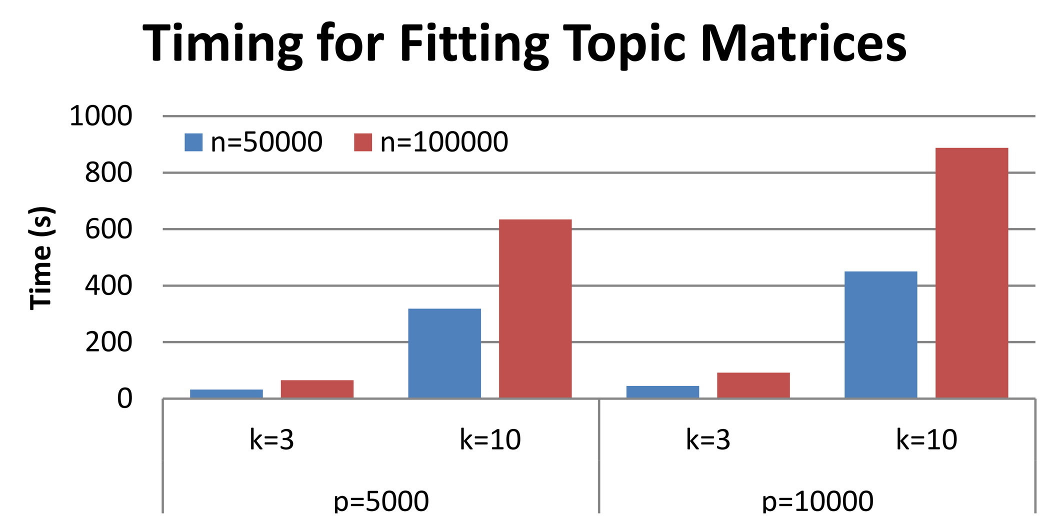


Figure : (Top Left) The timing for fitting p Poisson regressions shows an empirical scaling of $O(np)$. (Top Right) The timing for fitting topic matrices empirically shows scaling that is $O(npk^2)$. (Bottom) The timing for AIS sampling shows that the sampling is approximately linearly scaled with the number of non-zeros in Φ irrespective of p.

Qualitative Analysis of LPMRF

Topic 1			Topic 2		
Top words	Top Pos. Edges	Top Neg. Edges	Top words	Top Pos. Edges	Top Neg. Edges
patients	term+long	cells-patient	flow	supported+simply	flow-shells
cases	positive+negative	patients-animals	pressure	account+taken	number-numbers
normal	cooling+hypothermi	patients-rats	boundary	agreement+good	flow-shell
cells	system+central	hormone-protein	results	moment+pitching	wing-hypersonic
treatment	atmosphere+height	growth-parathyroid	theory	non+linear	solutions-turbulent
children	function+functions	patients-lens	method	lower+upper	mach-reynolds
found	methods+suitable	patients-mice	layer	tunnel+wind	flow-stresses
results	stress+reaction	patients-dogs	given	time+dependent	theoretical-drag
blood	low+rates	hormone-tumor	number	level+noise	general-buckling
disease	case+report	patients-child	presented	purpose+note	made-conducted