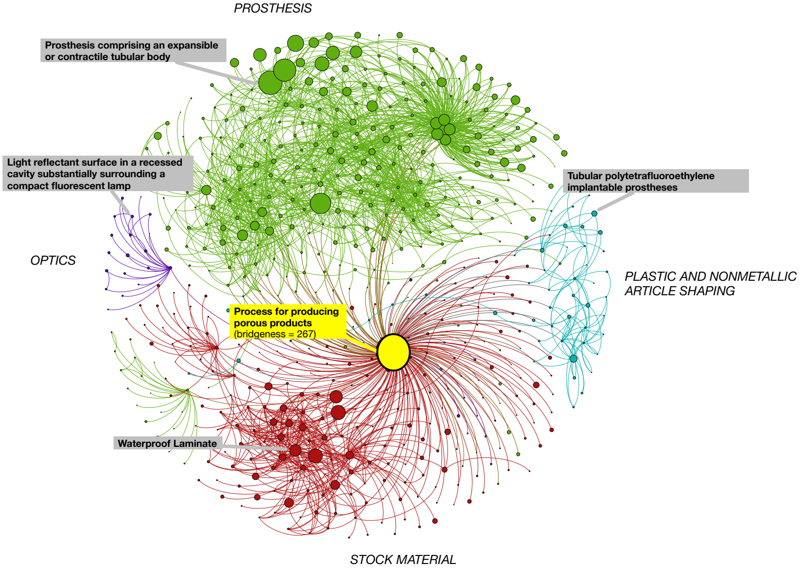
Stochastic Optimization and Variational Inference

David M. Blei

Princeton University

February 27, 2014



Communities discovered in a 3.7M node network of U.S. Patents

[Gopalan and Blei, PNAS 2013]

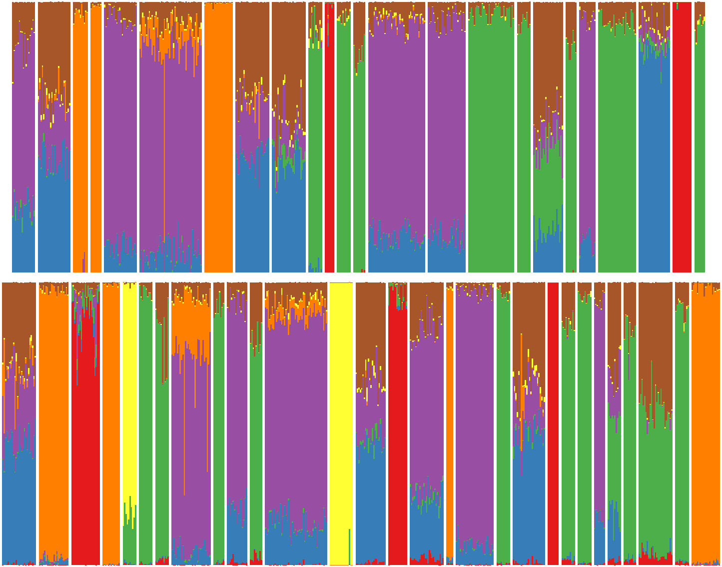
1	2	3	4	5
game	life	film	book	wine
season	know	movie	life	street
team	school	show	books	hotel
coach	street	life	novel	house
play	man	television	story	room
points	family	films	man	night
games	says	director	author	place
giants	house	man	house	restaurant
second	children	story	war	park
players	night	says	children	garden

6	7	8	9	10
bush	building	won	yankees	government
campaign	street	team	game	war
clinton	square	second	mets	military
republican	housing	race	season	officials
house	house	round	run	iraq
party	buildings	cup	league	forces
democratic	development	open	baseball	iraqi
political	space	game	team	army
democrats	percent	play	games	troops
senator	real	win	hit	soldiers

11	12	13	14	15
children	stock	church	art	police
school	percent	war	museum	yesterday
women	companies	women	show	man
family	fund	life	gallery	officer
parents	market	black	works	officers
child	bank	political	artists	case
life	investors	catholic	street	found
says	funds	government	artist	charged
help	financial	jewish	paintings	street
mother	business	pope	exhibition	shot

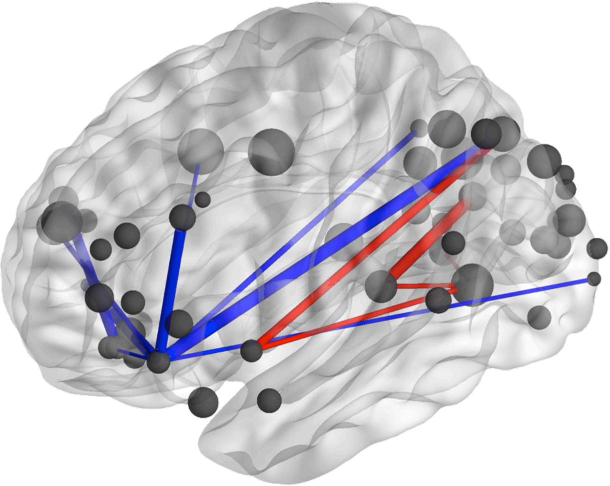
Topics found in 1.8M articles from the New York Times

[Hoffman, Blei, Wang, Paisley, JMLR 2013]



Population analysis of 2 billion genetic measurements

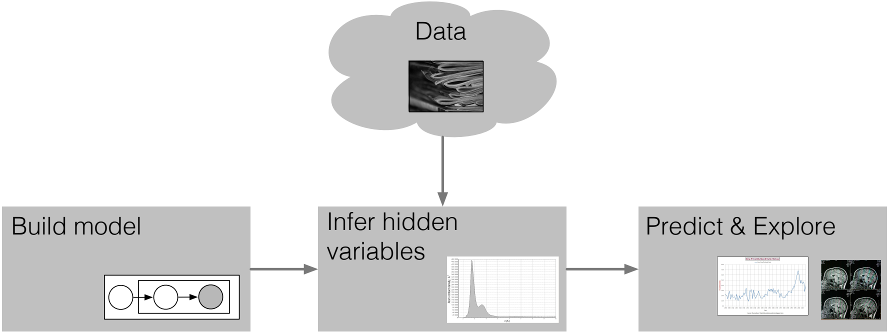
[Gopalan, Hao, Blei, Storey, in preparation]



Neuroscience analysis of 220 million fMRI measurements

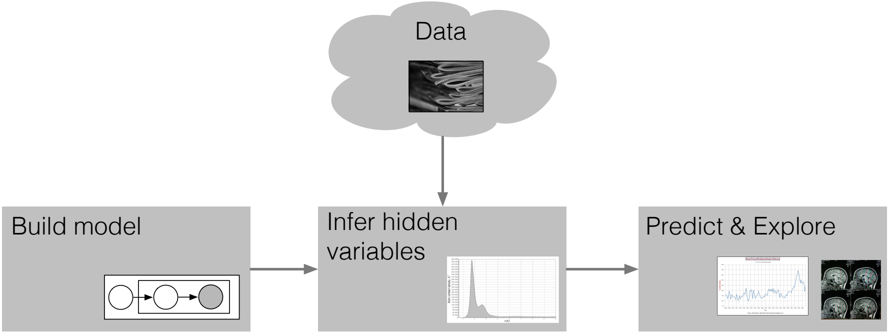
[Manning, Ranganath, Blei, Norman, submitted]

This talk



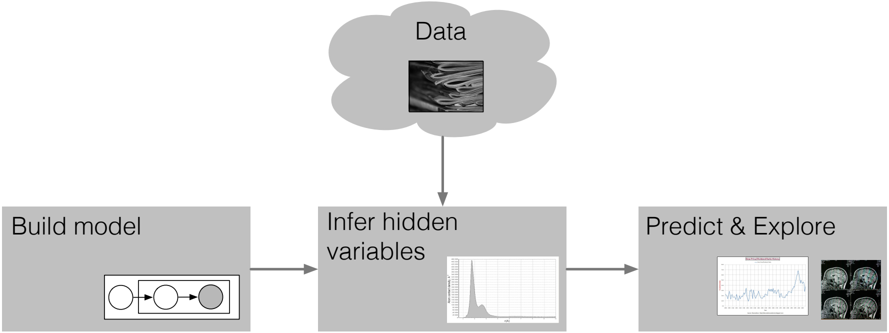
- Customized data analysis is important to many fields.
- Pipeline separates *assumptions*, *computation*, *application*
- Eases collaborative solutions to data science problems

This talk



- Graphical models are a language for expressing assumptions about data.
- Variational methods turn *inference* into *optimization*.
- Stochastic optimization *scales up* and *generalizes* variational methods.

This talk

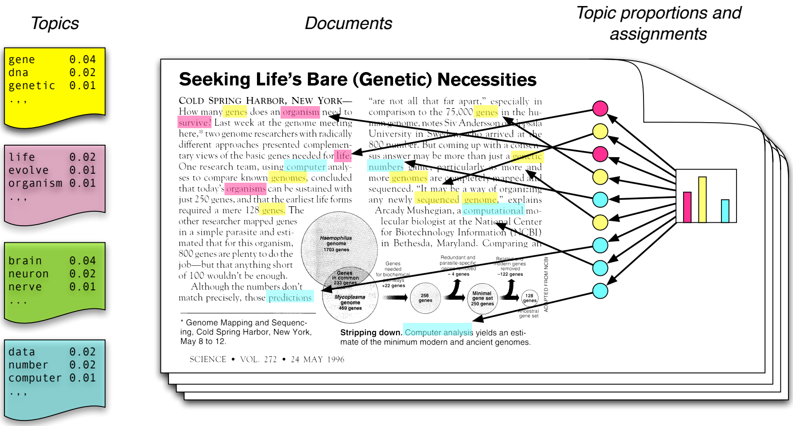


- Introduction to variational methods
- Scaling up with *stochastic variational inference* [Hoffman et al., 2013]
- Generalizing with *black box variational inference* [Ranganath et al., 2014]

Stochastic Variational Inference

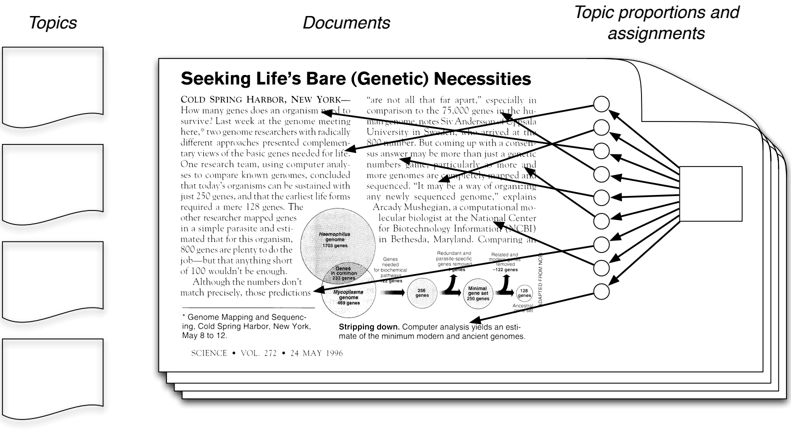
(with Matt Hoffman, Chong Wang, John Paisley)

Example: Latent Dirichlet allocation



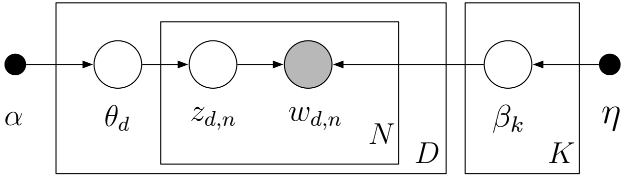
Generative process

Example: Latent Dirichlet allocation



Posterior inference

Classical variational inference

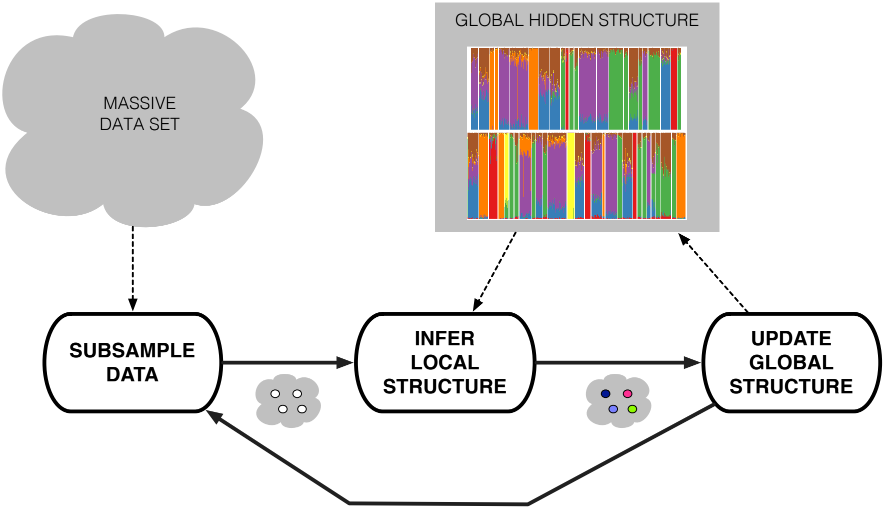


- Given data, estimate the conditional distribution of the hidden variables.
  - *Local* variables describe per-data point hidden structure.
  - *Global* variables describe structure shared by all the data.

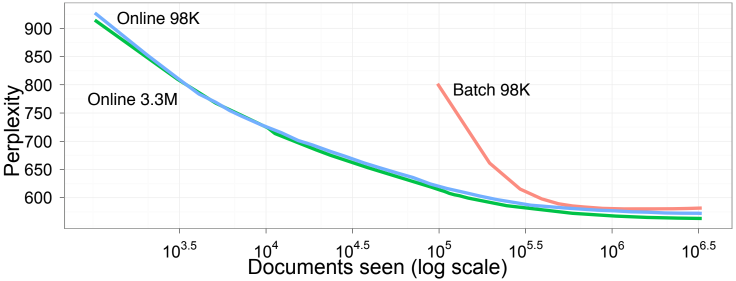
- Classical variational inference:
  - Do some local computation for each data point.
  - Aggregate these computations to re-estimate global structure.
  - Repeat.

- Inefficient, and cannot handle massive data sets.

Stochastic variational inference



Stochastic variational inference



Documents analyzed	**2048**	**4096**	**8192**	**12288**	**16384**	**32768**	**49152**	**65536**
Top eight words	systems	systems	service	service	service	business	business	business
	road	health	systems	systems	companies	service	service	industry
	made	communication	health	companies	systems	companies	companies	service
	service	service	companies	business	business	industry	industry	companies
	announced	billion	market	company	company	company	services	services
	national	language	communication	billion	industry	management	company	company
	west	care	company	health	market	systems	management	management
	language	road	billion	industry	billion	services	public	public

[Hoffman et al., 2010]

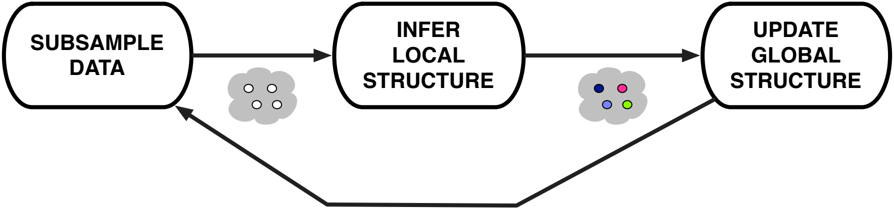
1	2	3	4	5
game	life	film	book	wine
season	know	movie	life	street
team	school	show	books	hotel
coach	street	life	novel	house
play	man	television	story	room
points	family	films	man	night
games	says	director	author	place
giants	house	man	house	restaurant
second	children	story	war	park
players	night	says	children	garden

6	7	8	9	10
bush	building	won	yankees	government
campaign	street	team	game	war
clinton	square	second	mets	military
republican	housing	race	season	officials
house	house	round	run	iraq
party	buildings	cup	league	forces
democratic	development	open	baseball	iraqi
political	space	game	team	army
democrats	percent	play	games	troops
senator	real	win	hit	soldiers

11	12	13	14	15
children	stock	church	art	police
school	percent	war	museum	yesterday
women	companies	women	show	man
family	fund	life	gallery	officer
parents	market	black	works	officers
child	bank	political	artists	case
life	investors	catholic	street	found
says	funds	government	artist	charged
help	financial	jewish	paintings	street
mother	business	pope	exhibition	shot

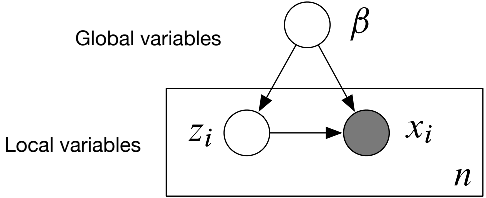
Topics found in 1.8M articles from the New York Times

Stochastic variational inference



1. A generic class of models
2. Classical mean-field variational inference
3. Stochastic variational inference

A generic class of models



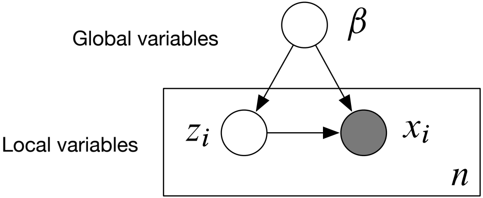
$$p(\beta, z_{1:n}, x_{1:n}) = p(\beta) \prod_{i=1}^{n} p(z_i \mid \beta) p(x_i \mid z_i, \beta)$$

A generic model with local and global variables:

- The observations are $x = x_{1:n}$.
- The **local** variables are $z = z_{1:n}$.
- The **global** variables are β.
- The ith data point x_i only depends on z_i and β.

Our goal is to compute $p(\beta, z \mid x)$.

A generic class of models



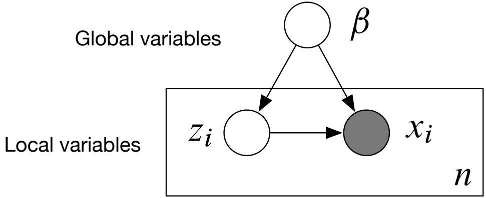
$$p(\beta, z_{1:n}, x_{1:n}) = p(\beta) \prod_{i=1}^{n} p(z_i \mid \beta) p(x_i \mid z_i, \beta)$$

- A **complete conditional** is the conditional of a latent variable given the observations and other latent variable.

- Assume each complete conditional is in the exponential family,

$$\begin{aligned} p(z_i \mid \beta, x_i) &= h(z_i) \exp\{\eta_\ell(\beta, x_i)^\top z_i - a(\eta_\ell(\beta, x_i))\} \\ p(\beta \mid z, x) &= h(\beta) \exp\{\eta_g(z, x)^\top \beta - a(\eta_g(z, x))\}. \end{aligned}$$

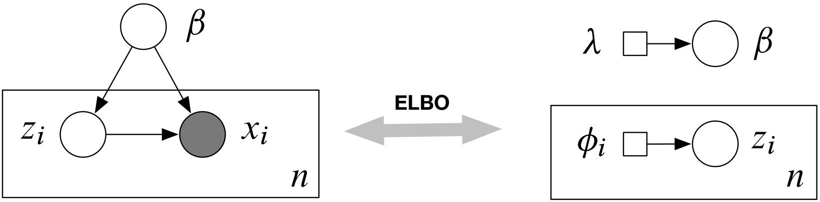
A generic class of models



$$p(\beta, z_{1:n}, x_{1:n}) = p(\beta) \prod_{i=1}^{n} p(z_i \mid \beta) p(x_i \mid z_i, \beta)$$

- Bayesian mixture models
- Time series models (variants of HMMs, Kalman filters)
- Factorial models
- Matrix factorization (e.g., factor analysis, PCA, CCA)
- Dirichlet process mixtures, HDPs
- Multilevel regression (linear, probit, Poisson)
- Stochastic blockmodels
- Mixed-membership models (LDA and some variants)

Mean-field variational inference

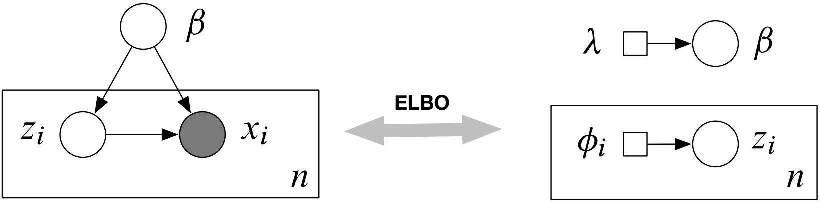


- Introduce a **variational distribution** over the latent variables $q(\beta, z)$.
- Optimize the **evidence lower bound** (ELBO) with respect to q,

$$\log p(x) \geq \mathrm{E}_q[\log p(\beta, Z, x)] - \mathrm{E}_q[\log q(\beta, Z)].$$

- Equivalent to minimizing the KL between q and the posterior
- *The ELBO links the observations/model to the variational distribution.*

Mean-field variational inference



- Set $q(\beta, z)$ to be a fully factored variational distribution,

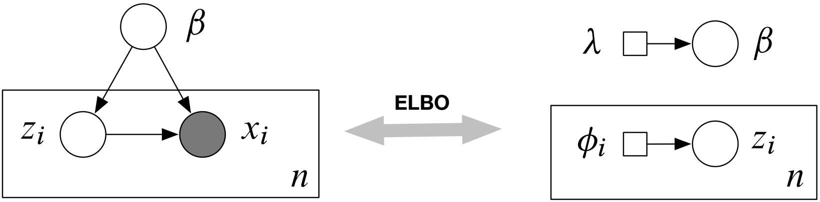
$$q(\beta, z) = q(\beta \mid \lambda) \prod_{i=1}^{n} q(z_i \mid \phi_i).$$

- Each component is in the same family as the model conditional,

$$\begin{aligned} p(\beta \mid z, x) &= h(\beta) \exp\{\eta_g(z, x)^\top \beta - a(\eta_g(z, x))\} \\ q(\beta \mid \lambda) &= h(\beta) \exp\{\lambda^\top \beta - a(\lambda)\}. \end{aligned}$$

- (Same for the local variational parameters)

Mean-field variational inference



- Optimize the ELBO with coordinate ascent. The ELBO is

$$\mathcal{L}(\lambda, \phi_{1:n}) = \mathrm{E}_q[\log p(\beta, Z, x)] - \mathrm{E}_q[\log q(\beta, Z)].$$

- With respect to the global parameters, the gradient is

$$\nabla_\lambda \mathcal{L} = a''(\lambda)(\mathrm{E}_\phi[\eta_g(Z, x)] - \lambda).$$

- This leads to a simple coordinate update [Ghahramani and Beal, 2001]

$$\lambda^* = \mathrm{E}_\phi\left[\eta_g(Z, x)\right].$$

Mean-field variational inference

Initialize λ randomly.

Repeat until the ELBO converges

1. For each data point, update the local variational parameters:

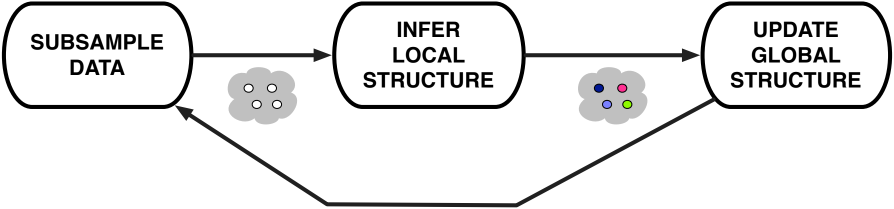
$$\phi_i^{(t)} = \mathrm{E}_{\lambda^{(t-1)}}[\eta_\ell(\beta, x_i)] \quad \text{for } i \in \{1, \ldots, n\}.$$

2. Update the global variational parameters:

$$\lambda^{(t)} = \mathrm{E}_{\phi^{(t)}}[\eta_g(Z_{1:n}, x_{1:n})].$$

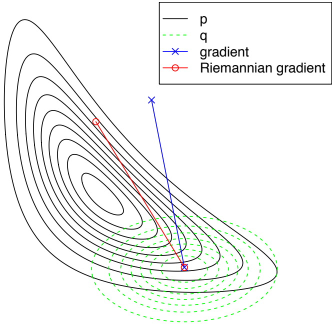
- Inefficient: We analyze the whole data set before completing one iteration.
- E.g.: In iteration #1 we analyze all documents with random topics.

Stochastic variational inference



- Stochastic variational inference stems from this classical algorithm
- Idea #1: **Natural gradients** [Amari, 1998]
- Idea #2: **Stochastic optimization** [Robbins and Monro, 1951]

Natural gradients



[Honkela et al., 2010]

- The **natural gradient** of the ELBO is

$$\hat{\nabla}_\lambda \mathcal{L} = \mathrm{E}_\phi[\eta_g(Z, x)] - \lambda.$$

- We can compute the natural gradient by computing the coordinate updates in parallel and subtracting the current variational parameters. [Sato, 2001]

Stochastic optimization

A STOCHASTIC APPROXIMATION METHOD[1]

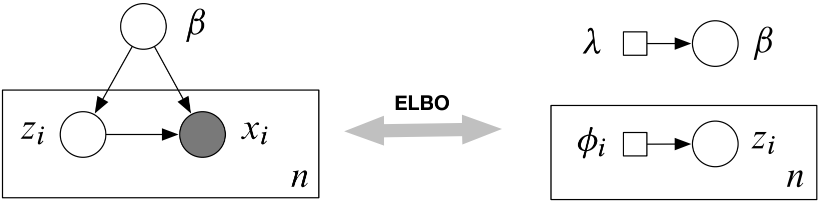
BY HERBERT ROBBINS AND SUTTON MONRO

University of North Carolina

1. Summary. Let $M(x)$ denote the expected value at level x of the response to a certain experiment. $M(x)$ is assumed to be a monotone function of x but is unknown to the experimenter, and it is desired to find the solution $x = \theta$ of the equation $M(x) = \alpha$, where α is a given constant. We give a method for making successive experiments at levels $x_1, x_2, \cdots$ in such a way that x_n will tend to θ in probability.

- Why waste time with the real gradient, when a cheaper noisy estimate of the gradient will do? [Robbins and Monro, 1951]
- **Stochastic optimization** follows noisy estimates of the gradient.
- Guaranteed to converge to a local optimum [Bottou, 1996]

Stochastic variational inference



- We will use stochastic optimization for global variables.

- Let $\nabla_\lambda \mathcal{L}_t$ be a realization of a random variable whose expectation is $\nabla_\lambda \mathcal{L}$.

- Iteratively set

$$\lambda^{(t)} = \lambda^{(t-1)} + \epsilon_t \nabla_\lambda \mathcal{L}_t$$

- This leads to a local optimum when

$$\sum_{t=1}^{\infty} \epsilon_t = \infty \qquad \sum_{t=1}^{\infty} \epsilon_t^2 < \infty$$

Stochastic variational inference

- With local and global variables, we decompose the ELBO

$$\mathcal{L} = \mathrm{E}[\log p(\beta)] - \mathrm{E}[\log q(\beta)] + \textstyle\sum_{i=1}^{n} \mathrm{E}[\log p(z_i, x_i \mid \beta)] - \mathrm{E}[\log q(z_i)]$$

- Sample a single data point t uniformly from the data and define

$$\mathcal{L}_t = \mathrm{E}[\log p(\beta)] - \mathrm{E}[\log q(\beta)] + n(\mathrm{E}[\log p(z_t, x_t \mid \beta)] - \mathrm{E}[\log q(z_t)]).$$

1. **The ELBO is the expectation of $\mathcal{L}_t$ with respect to the sample.**
2. **The gradient of the t-ELBO is a noisy gradient of the ELBO.**
3. **The t-ELBO is like an ELBO where we saw x_t repeatedly.**

Stochastic variational inference

- Let $\eta_t(Z_t, x_t)$ be the conditional distribution of the global variable for the model where the observations are n replicates of x_t.
- With this, the noisy natural gradient of the ELBO is

$$\hat{\nabla}_\lambda \mathcal{L}_t = \mathrm{E}_{\phi_t}[\eta_t(Z_t, x_t)] - \lambda\,.$$

- Notes:
  - It only requires the local variational parameters of one data point.
  - In contrast, the full natural gradient requires all local parameters.
  - Thanks to conjugacy it has a simple form.

Stochastic variational inference

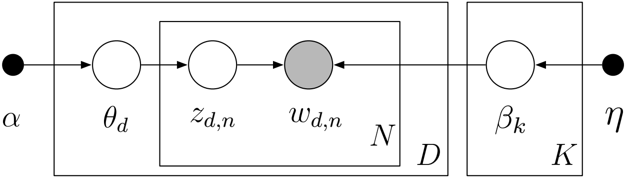
Initialize global parameters λ randomly.

Set the step-size schedule ϵ_t appropriately.

Repeat forever

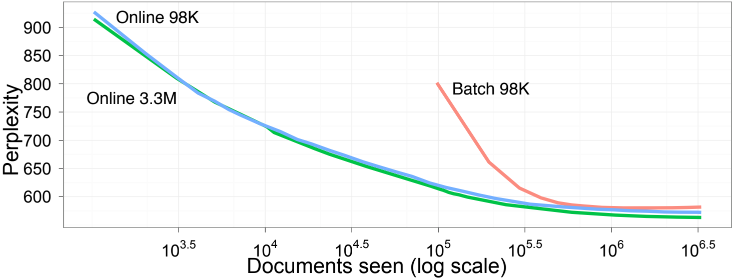
1. Sample a data point uniformly,
$$x_t \sim \mathrm{Uniform}(x_1, \ldots, x_n).$$
2. Compute its local variational parameter,
$$\phi = \mathrm{E}_{\lambda^{(t-1)}}[\eta_\ell(\beta, x_t)].$$
3. Pretend its the only data point in the data set,
$$\hat{\lambda} = \mathrm{E}_\phi[\eta_t(Z_t, x_t)].$$
4. Update the current global variational parameter,
$$\lambda^{(t)} = (1 - \epsilon_t)\lambda^{(t-1)} + \epsilon_t \hat{\lambda}.$$

Stochastic variational inference in LDA



1. Sample a document
2. Estimate the local variational parameters using the current topics
3. Form intermediate topics from those local parameters
4. Update topics as a weighted average of intermediate and current topics

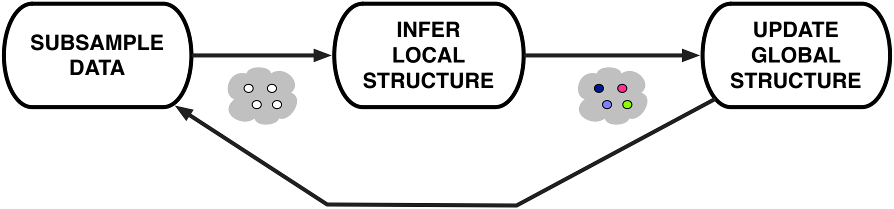
Stochastic variational inference in LDA



Documents analyzed	**2048**	**4096**	**8192**	**12288**	**16384**	**32768**	**49152**	**65536**
Top eight words	systems	systems	service	service	service	business	business	business
	road	health	systems	systems	companies	service	service	industry
	made	communication	health	companies	systems	companies	companies	service
	service	service	companies	business	business	industry	industry	companies
	announced	billion	market	company	company	company	services	services
	national	language	communication	billion	industry	management	company	company
	west	care	company	health	market	systems	management	management
	language	road	billion	industry	billion	services	public	public

[Hoffman et al., 2010]

Stochastic variational inference



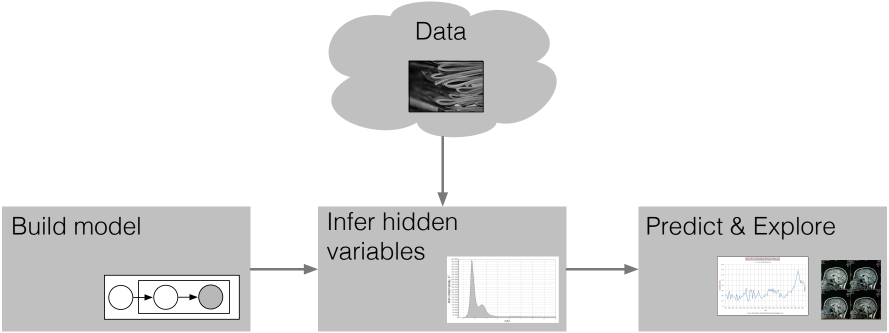
We defined a generic algorithm for scalable variational inference.

- Bayesian mixture models
- Time series models (variants of HMMs, Kalman filters)
- Factorial models
- Matrix factorization (e.g., factor analysis, PCA, CCA)
- Dirichlet process mixtures, HDPs
- Multilevel regression (linear, probit, Poisson)
- Stochastic blockmodels
- Mixed-membership models (LDA and some variants)

Black Box Variational Inference

(with Rajesh Ranganath and Sean Gerrish)

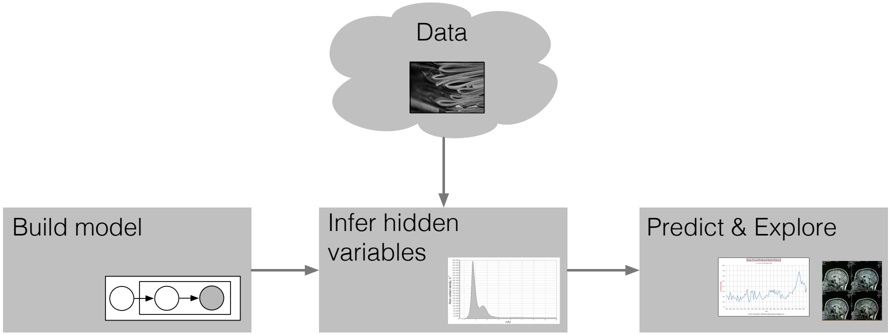
Black box variational inference



Our vision:

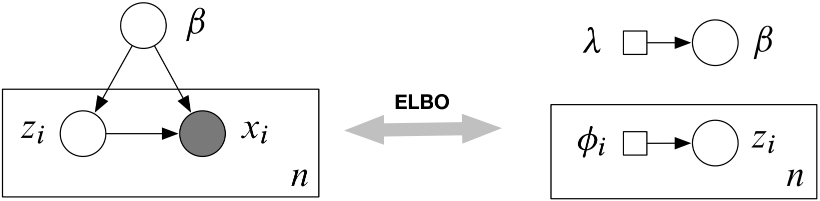
- Easily use variational inference with any model
- No requirements on the complete conditionals
- No mathematical work beyond specifying the model

Black box variational inference



- The original, but slow, black box: [Metropolis, 1953; Hastings, 1970]
- Tailored to models: [Jordan and Jaakkola, 1996; Braun and McAuliffe, 2008; others]
- Requires model-specific analysis: [Wang and Blei, 2013; Knowles and Minka, 2011]
- Similar goals: [Salimans and Knowles, 2012; Salimans and Knowles, 2014]

Black box variational inference



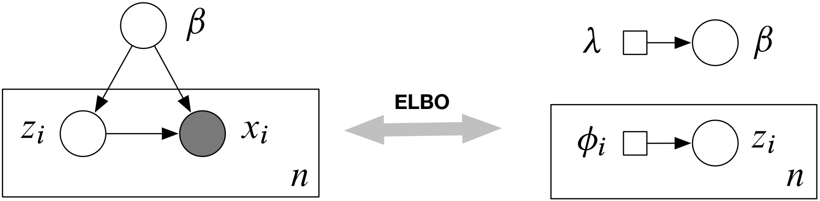
The ELBO:

$$\mathcal{L}(\nu) = \mathrm{E}_q[\log p(\beta, Z, x)] - \log q(\beta, Z \mid \nu)]$$

Its gradient:

$$\nabla_\nu \mathcal{L}(\nu) = \mathrm{E}_q[\nabla_\nu \log q(\beta, Z \mid \nu)(\log p(\beta, Z, x) - \log q(\beta, Z \mid \nu))]$$

Black box variational inference



A noisy gradient at ν:

$$\nabla_\nu \mathcal{L}(\nu) \approx \frac{1}{B} \sum_{b=1}^{B} (\nabla_\nu \log q(\beta_b, z_b \mid \nu)(\log p(\beta_b, z_b, x) - \log q(\beta_b, z_b \mid \nu))$$

where

$$(\beta_b, z_b) \sim q(\beta, z \mid \nu)$$

The noisy gradient

$$\boxed{\nabla_\nu \mathcal{L}(\nu) \approx \frac{1}{B} \sum_{b=1}^{B} (\nabla_\nu \log q(\beta_b, z_b \mid \nu)(\log p(\beta_b, z_b, x) - \log q(\beta_b, z_b \mid \nu))}$$

- We use these gradients in a stochastic optimization algorithm.
- Requirements:
  - Sampling from $q(\beta, z)$
  - Evaluating $\nabla_\nu \log q(\beta, z \mid \nu)$
  - Evaluating $\log p(\beta, z, x)$

The noisy gradient

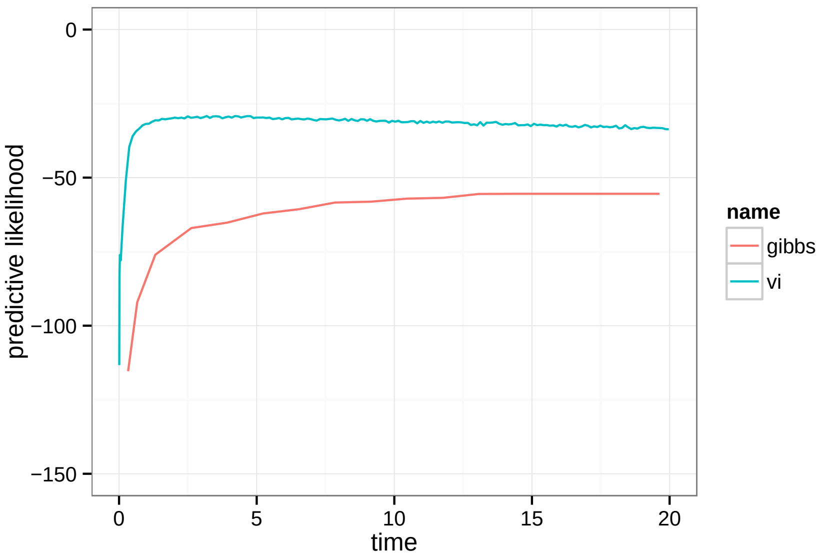
$$\boxed{\nabla_\nu \mathcal{L}(\nu) \approx \frac{1}{B}\sum_{b=1}^{B}(\nabla_\nu \log q(\beta_b, z_b \,|\, \nu)(\log p(\beta_b, z_b, x) - \log q(\beta_b, z_b \,|\, \nu))}$$

- A black box:
  - Requirements around $q(\cdot)$ can be reused across models.
  - Evaluating $\log p(\beta, z, x)$ is akin to defining the model.
- But the variance of the estimator is high

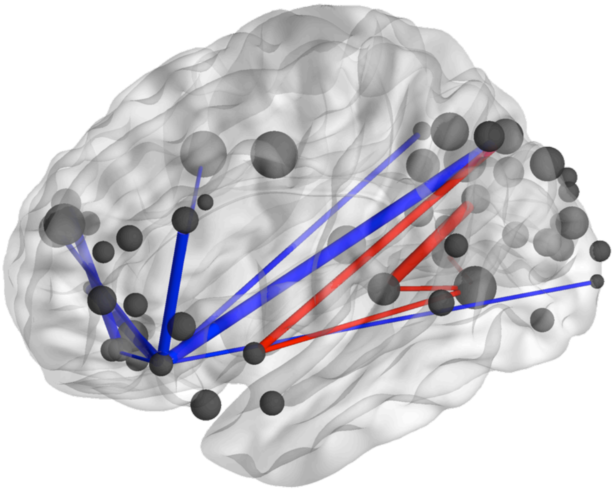
The noisy gradient

$$\boxed{\nabla_\nu \mathcal{L}(\nu) \approx \frac{1}{B}\sum_{b=1}^{B}(\nabla_\nu \log q(\beta_b, z_b \mid \nu)(\log p(\beta_b, z_b, x) - \log q(\beta_b, z_b \mid \nu))}$$

- Rao-Blackwellization for each component of the gradient
- Control variates, again using $\nabla_\nu \log q(\beta, z \mid \nu)$
- AdaGrad, for setting learning rates
- Stochastic variational inference, for handling massive data



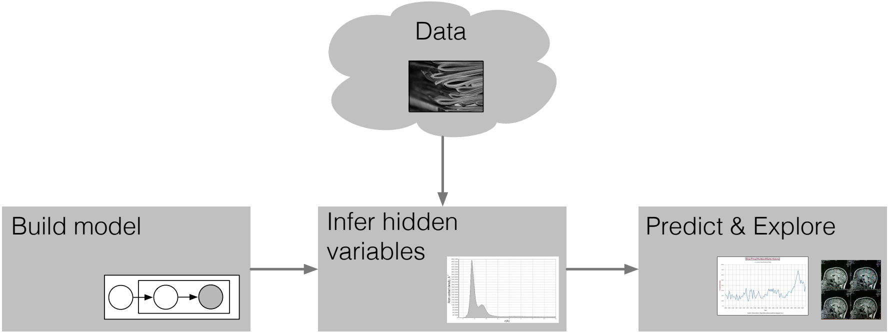
A nonconjugate Normal-Gamma time-series model



Neuroscience analysis of 220 million fMRI measurements

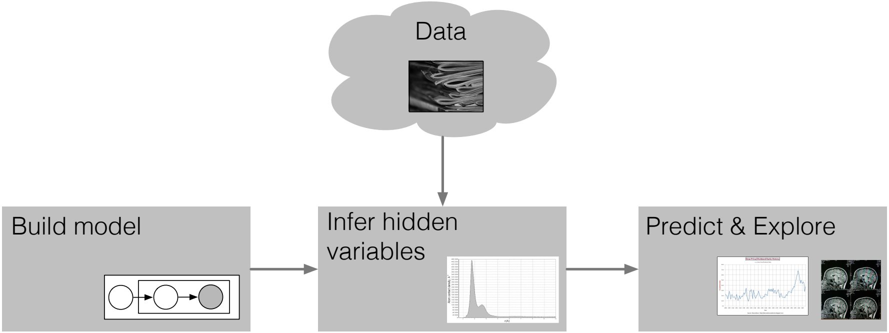
[Manning, Ranganath, Blei, Norman, submitted]

This talk



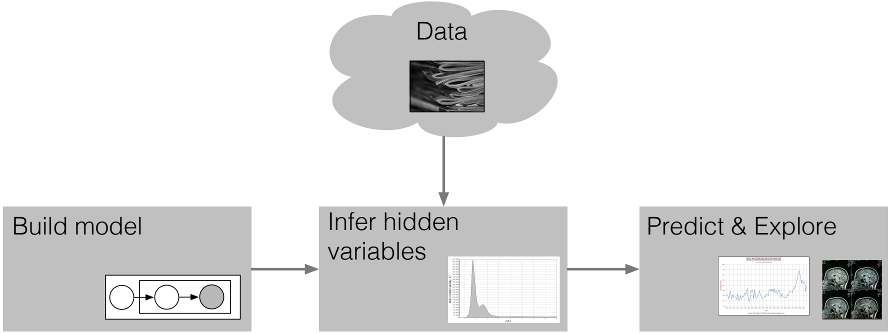
- Customized data analysis is important to many fields.
- Pipeline separates *assumptions*, *computation*, *application*
- Eases collaborative solutions to data science problems

This talk



- Graphical models are a language for expressing assumptions about data.
- Variational methods turn *inference* into *optimization*.
- Stochastic optimization *scales up* and *generalizes* variational methods.

This talk



- Scaling up with *stochastic variational inference* [Hoffman et al., 2013]
- Generalizing with *black box variational inference* [Ranganath et al., 2014]
- Please help us.