Expectation-Propagation for the Generative Aspect Model

Tom Minka and John Lafferty
Carnegie Mellon University

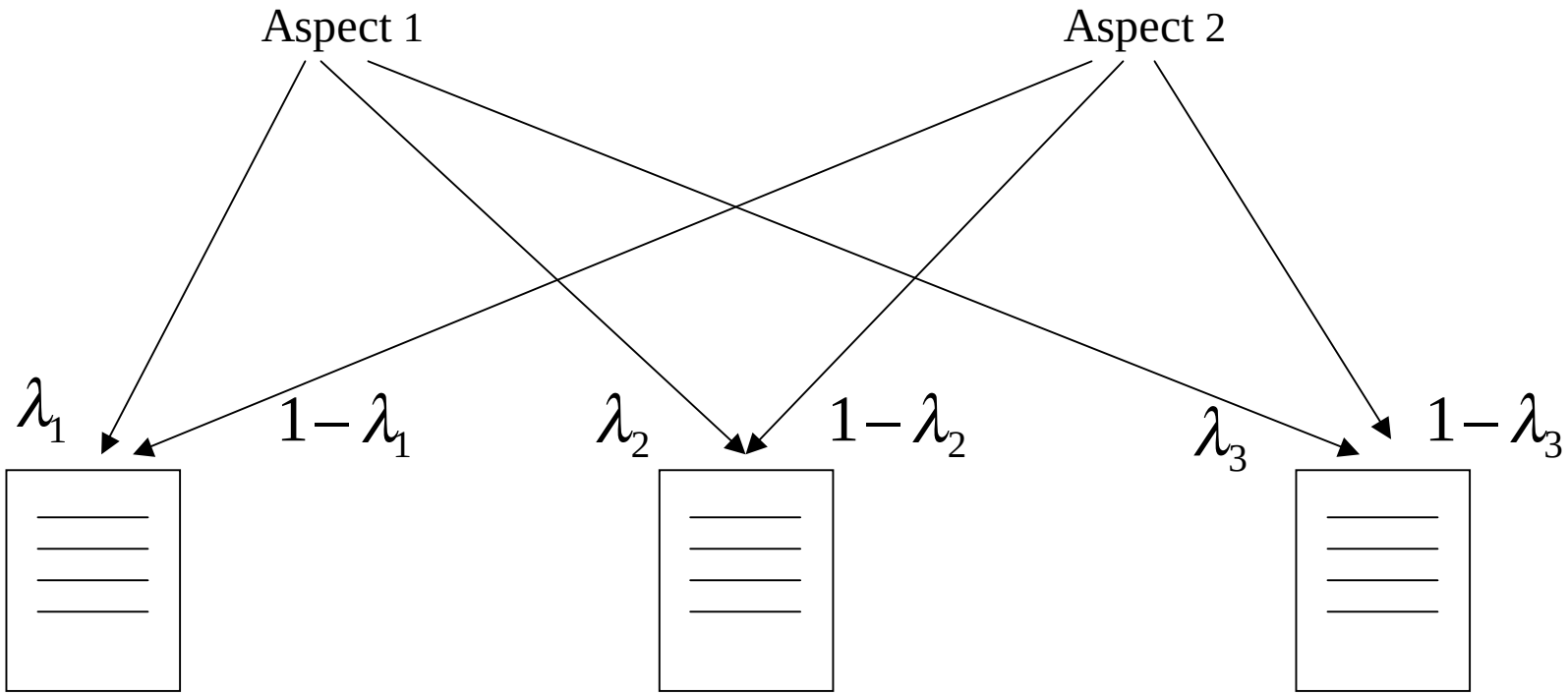
Main points

- Aspect model is important and difficult
  - PCA for discrete data
- How variational methods can fail
- Extensions of Expectation-Propagation
- Using EP for maximum likelihood

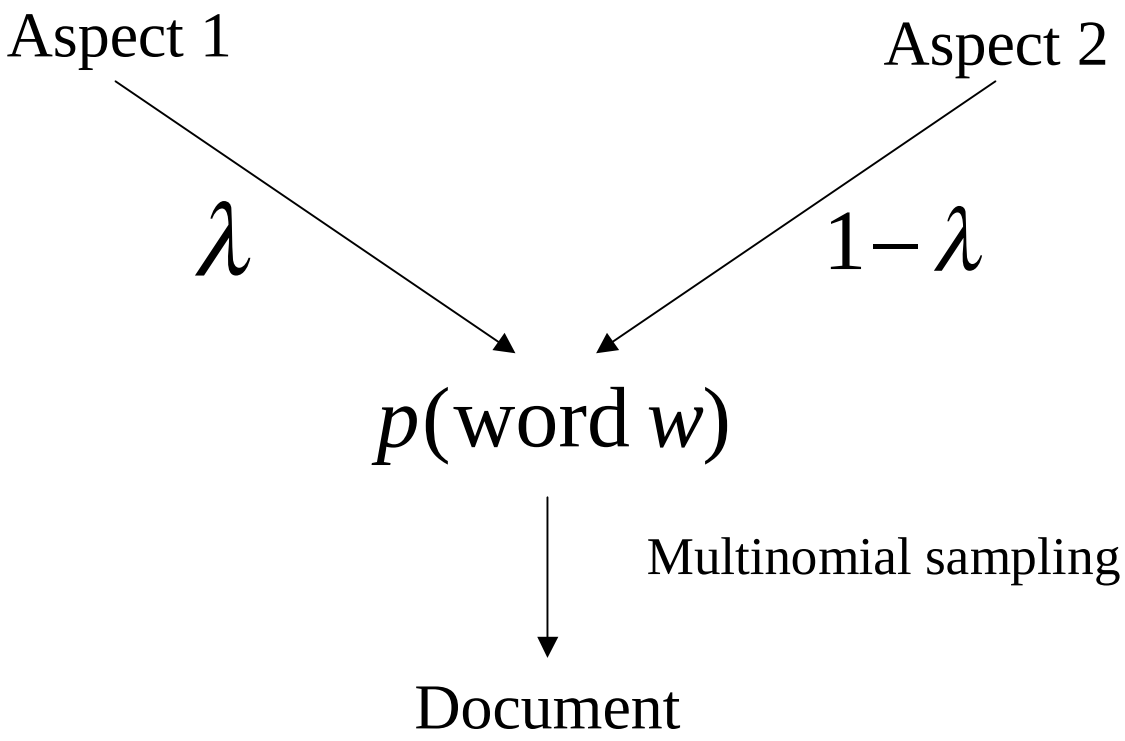
Generative aspect model

(Hofmann1999; Blei, Ng, & Jordan 2001)

Each document mixes aspects in different proportions

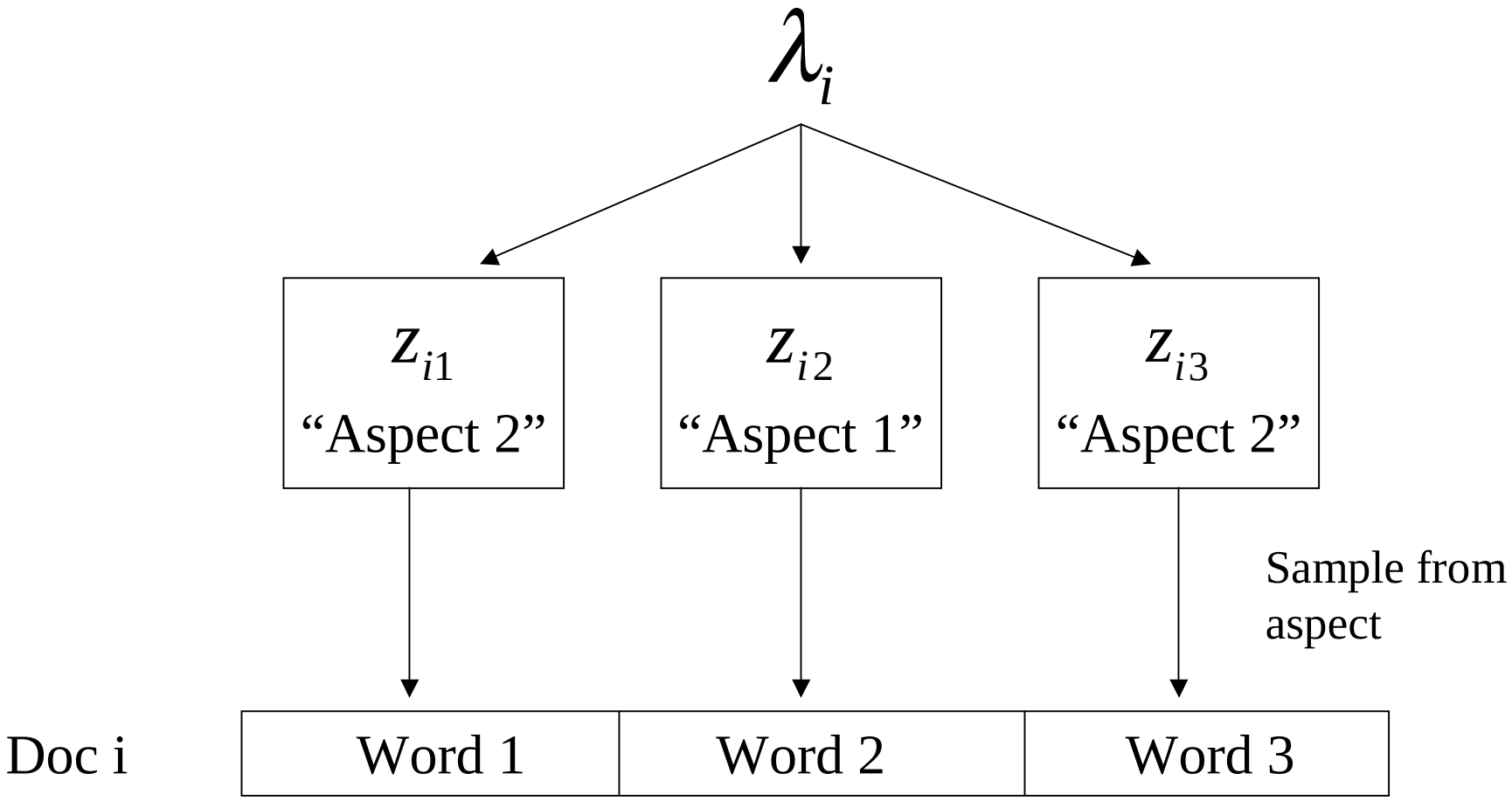


First interpretation



$p(\lambda) \sim Dirichlet(\alpha_1,...,\alpha_J)$

Second interpretation



Two tasks

Inference:

- Given aspects and document i, what is (posterior for) λ_i?

Learning:

- Given some documents, what are (maximum likelihood) aspects?

Variational method

- Blei, Ng, & Jordan, NIPS'01
- Uses (λ, z) interpretation
- Inference: factored variational distribution

  $$q(\lambda)q(z)$$

- Learning:
  - E-step (λ, z)
  - M-step (p, α) (aspects, aspect weights)

Expectation Propagation

- Uses (λ only) interpretation
- Inference: EP of Dirichlet posterior

  $q(\lambda)$
- Learning:
  - E-step (λ)
  - M-step (p, α) (aspects, aspect weights)
- Fewer latent variables $\rightarrow$ better

Geometric interpretation

- Likelihood is composed of terms of form

$$t_w(\lambda)^{n_w} = p(w)^{n_w} = (\sum_a \lambda_a p(w|a))^{n_w}$$

- Want Dirichlet approximation:

$$\tilde{t}_w(\lambda) = \prod_a \lambda_a^{\beta_{wa}}$$

Variational method

- Bound each term via Jensen's inequality:

$$\sum_a \lambda_a p(w|a) \geq \prod_a \lambda_a^{\beta_{wa}} \times (\text{const})$$

- Coupled equations:

$$\beta_{wa} \propto \exp(<\log \lambda_a>) p(w|a)$$

Moment matching

- Context function: all but one occurrence

$$q^{\setminus w}(\lambda) = t_w(\lambda)^{n_w - 1} \prod_{w' \neq w} t_{w'}(\lambda)^{n_{w'}}$$

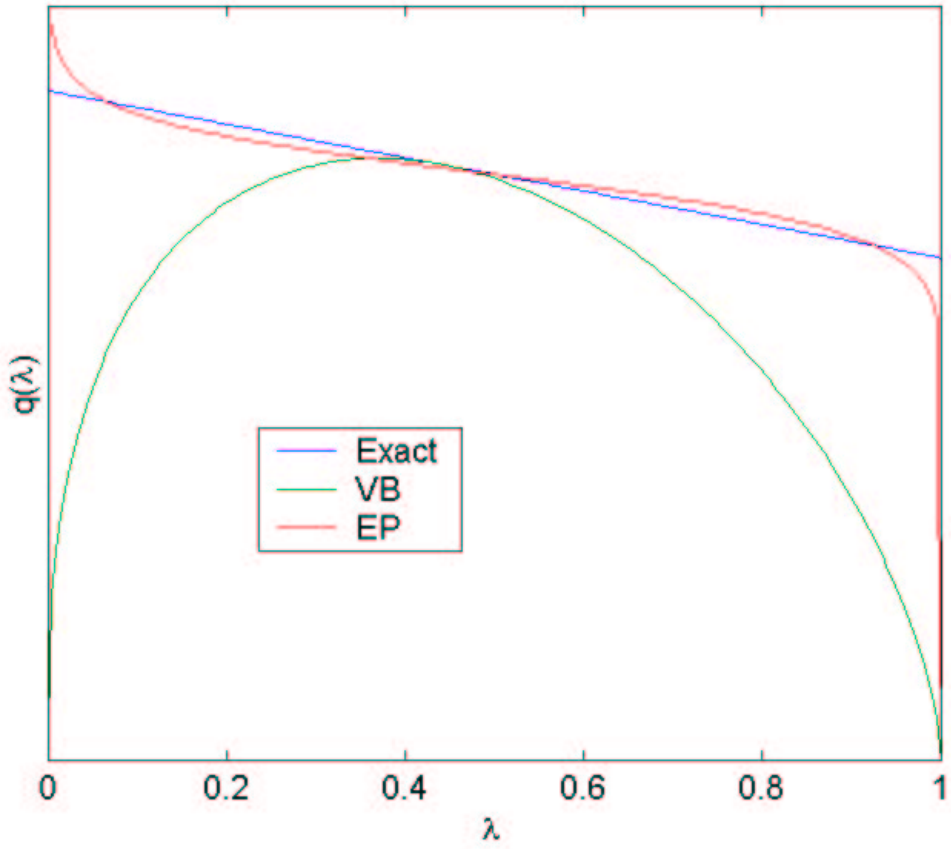
- Moment match:

$$t_w(\lambda) q^{\setminus w}(\lambda) \leftrightarrow \tilde{t}_w(\lambda) q^{\setminus w}(\lambda)$$

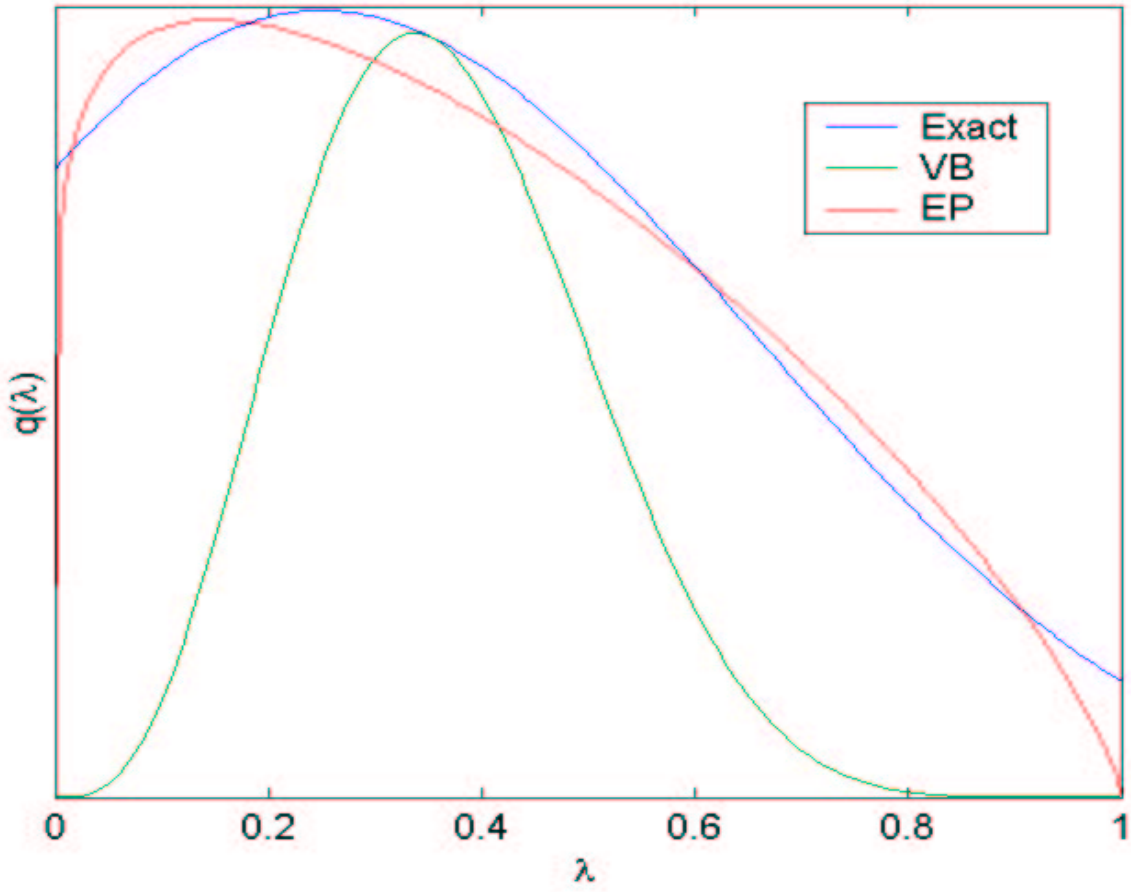
- Fixed point equations for β

One term

$$t_w(\lambda) = (\lambda)0.4 + (1-\lambda)0.3$$



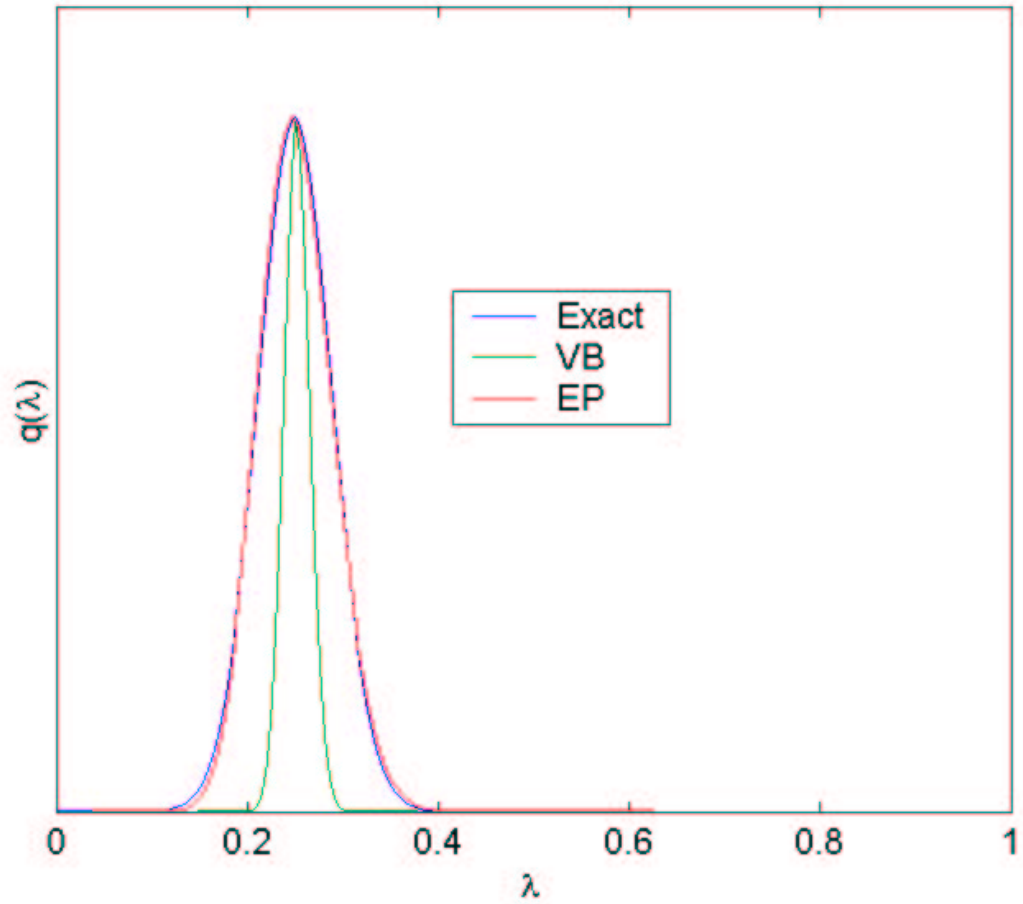
Ten word document



Moments for ten word doc

	Mean	Variance
Exact	0.393	0.0613
EP	0.393	0.0612
Variational	0.365	0.0178

1000 word document



Moments for 1000 word doc

	Mean	Variance
Exact	0.250	0.0015
EP	0.252	0.0015
Variational	0.252	0.0002

General behavior

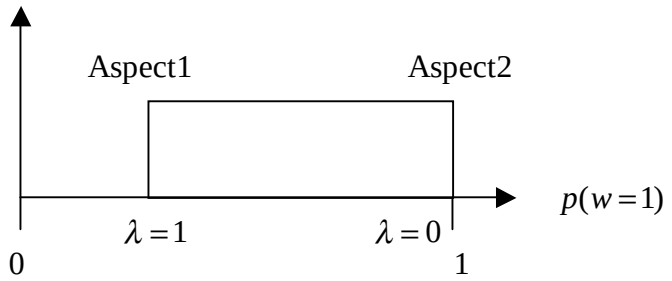
- For long documents, VB recovers correct mean, but not correct variance
- Optimizes global criterion, only accurate locally
- What about learning?

Learning

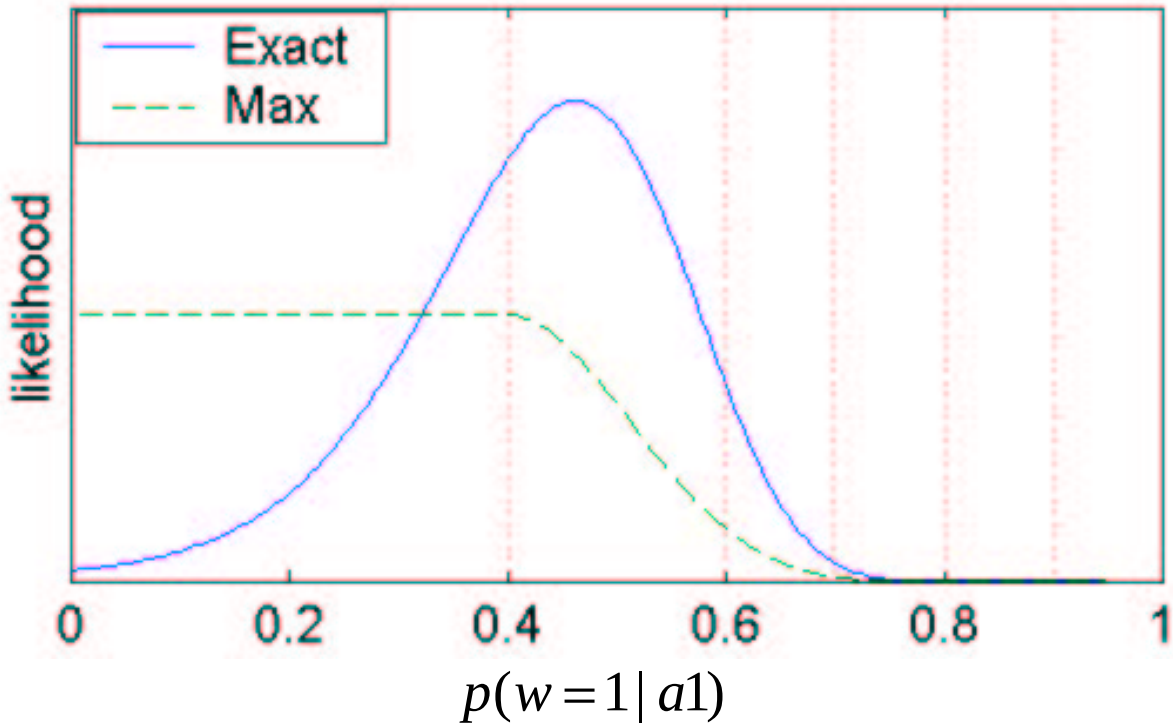
- Want MLE for aspects
- Requires marginalizing λ
- $p(doc)$ is probability of generating doc from a *random* λ
- Consider extreme approximation:
  - Probability of generating from *most likely* λ
  - Ignores uncertainty in λ ("Occam factor")
  - Used by Hofmann (1999)

Toy problem

- Two aspects over two words
  - $p(w=1|a)$ describes each
- One aspect fixed to $p(w=1|a2)=1$
- λ is uniform, i.e. $\alpha=1$
- $p(w=1)=(\lambda)\,p(w=1|a1)+(1-\lambda)\,p(w=1|a2)$



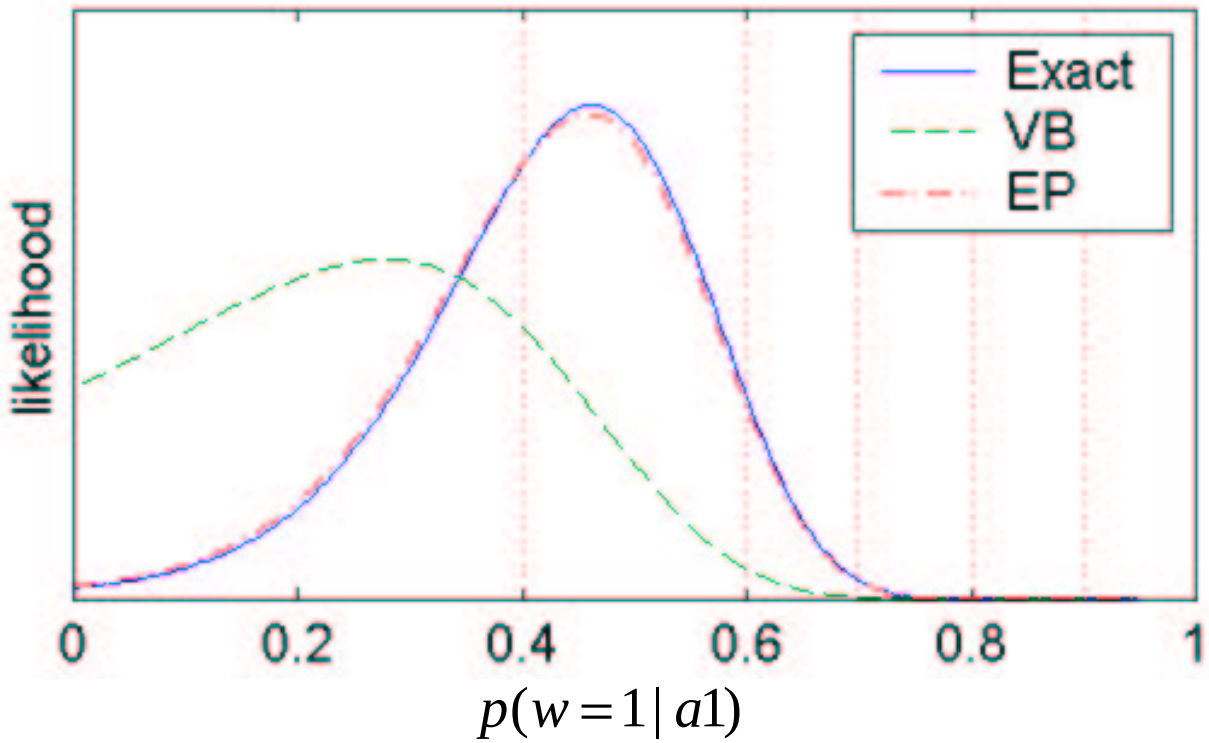
Exact likelihood vs. Max



Dotted are observed frequencies: n_{i1} / n_i

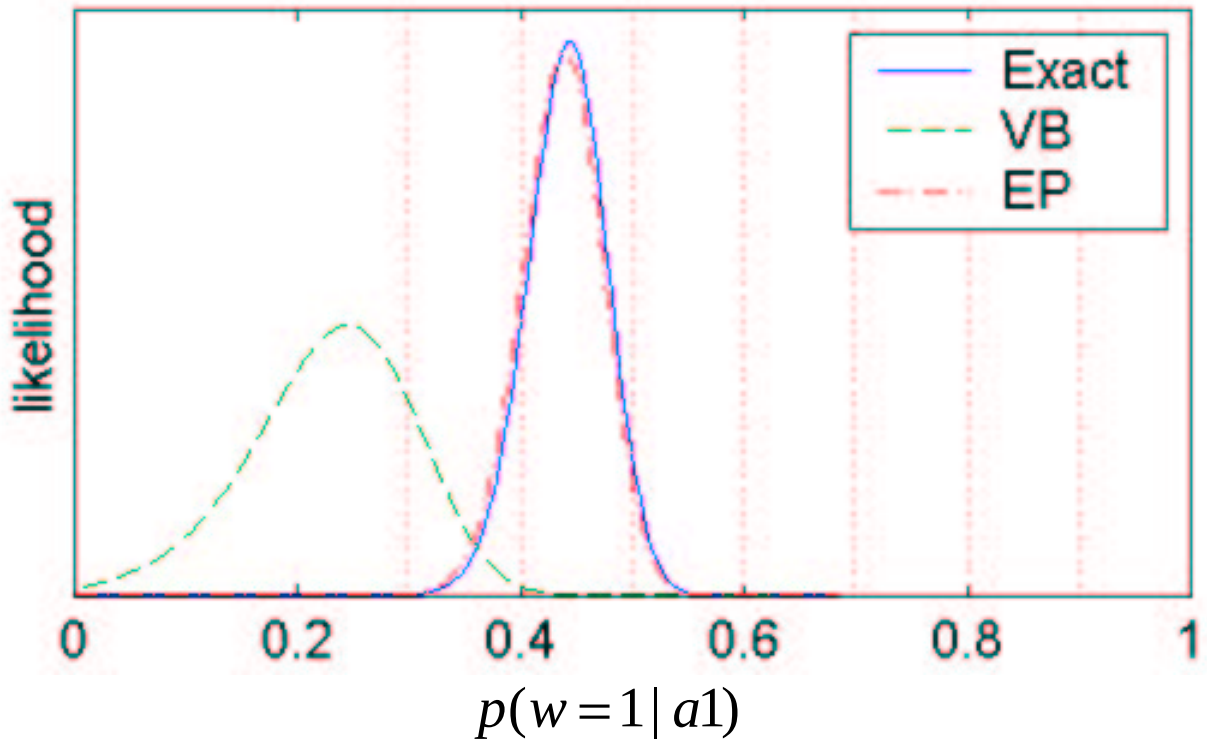
10 docs

EP likelihood vs. VB



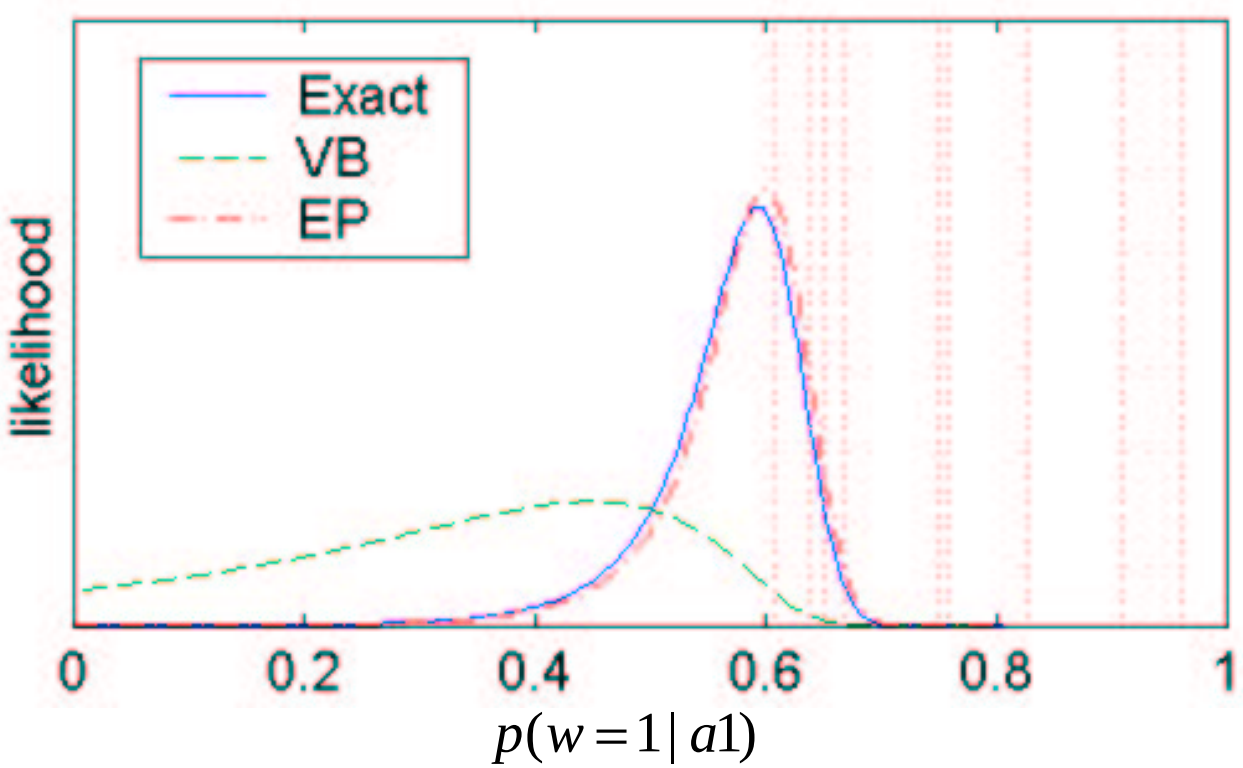
VB not much different from Max

100 documents



EP better, VB worse

Longer documents



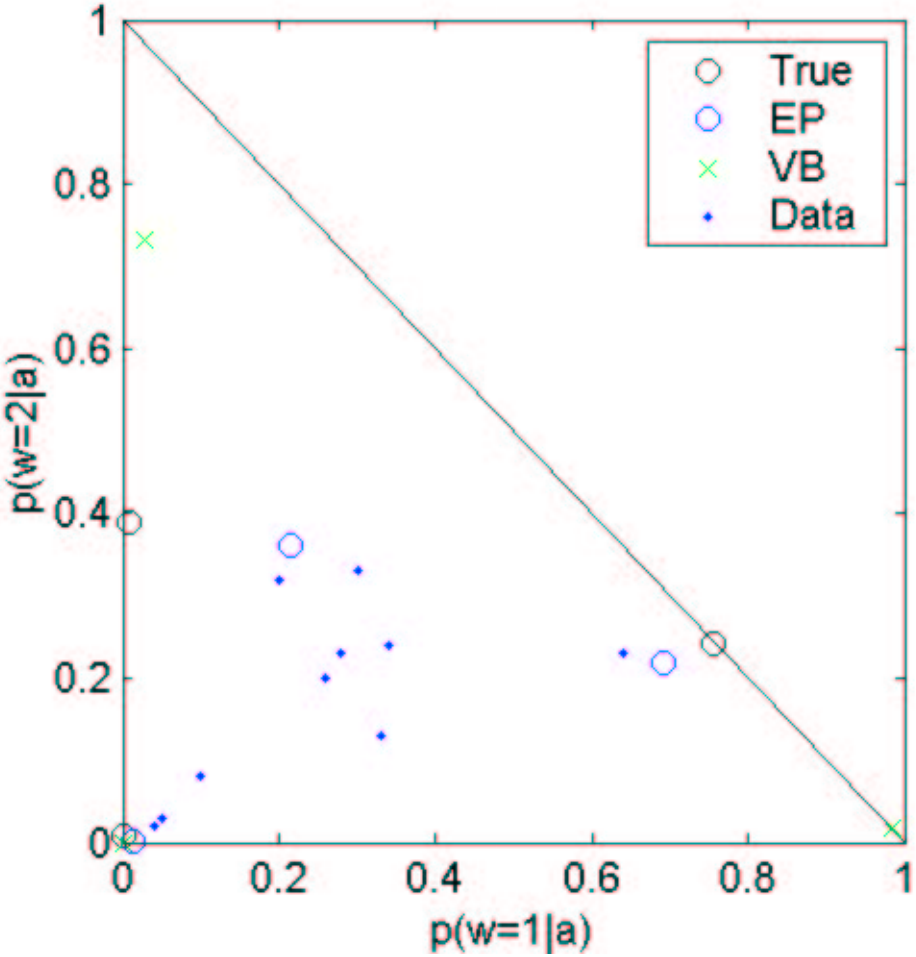
VB not helped

Why VB fails

- Doesn't capture posterior variance of λ
  - No Occam factor
- Gets *worse* with more documents
- Doesn't matter if posterior is sharp!
  - Longer documents
  - More distinct aspects

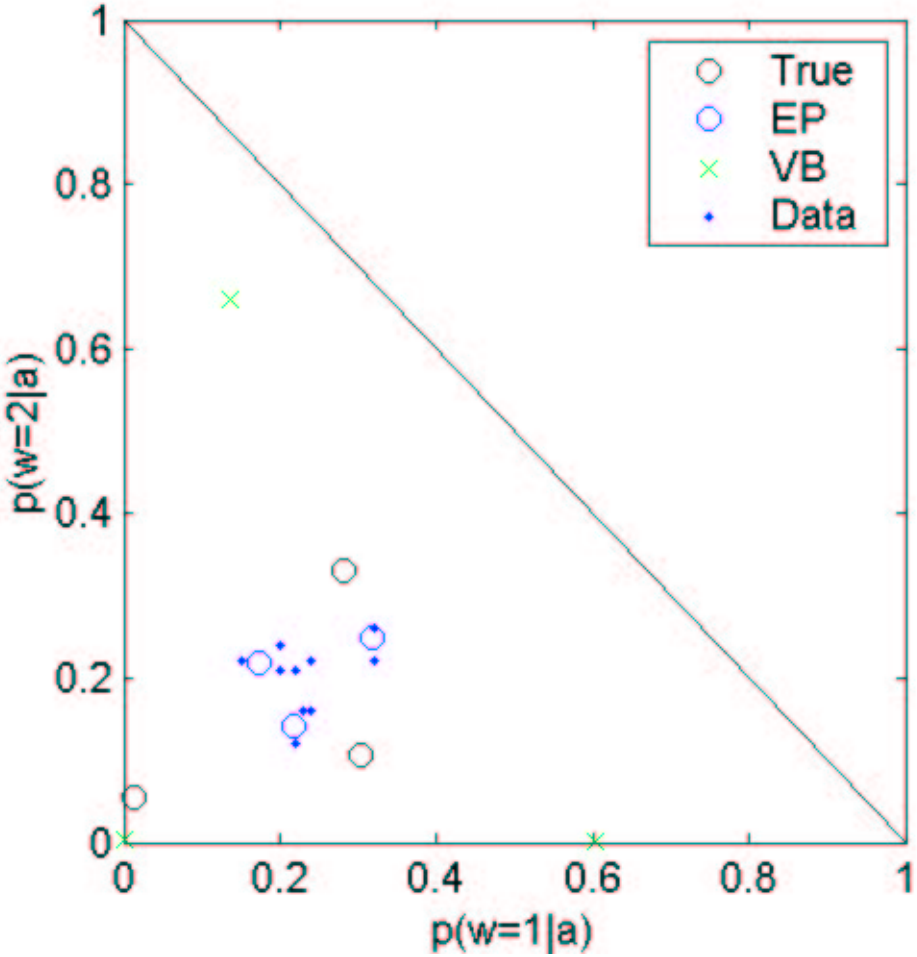
  Both sharpen posterior, but do not help VB

Larger vocabulary



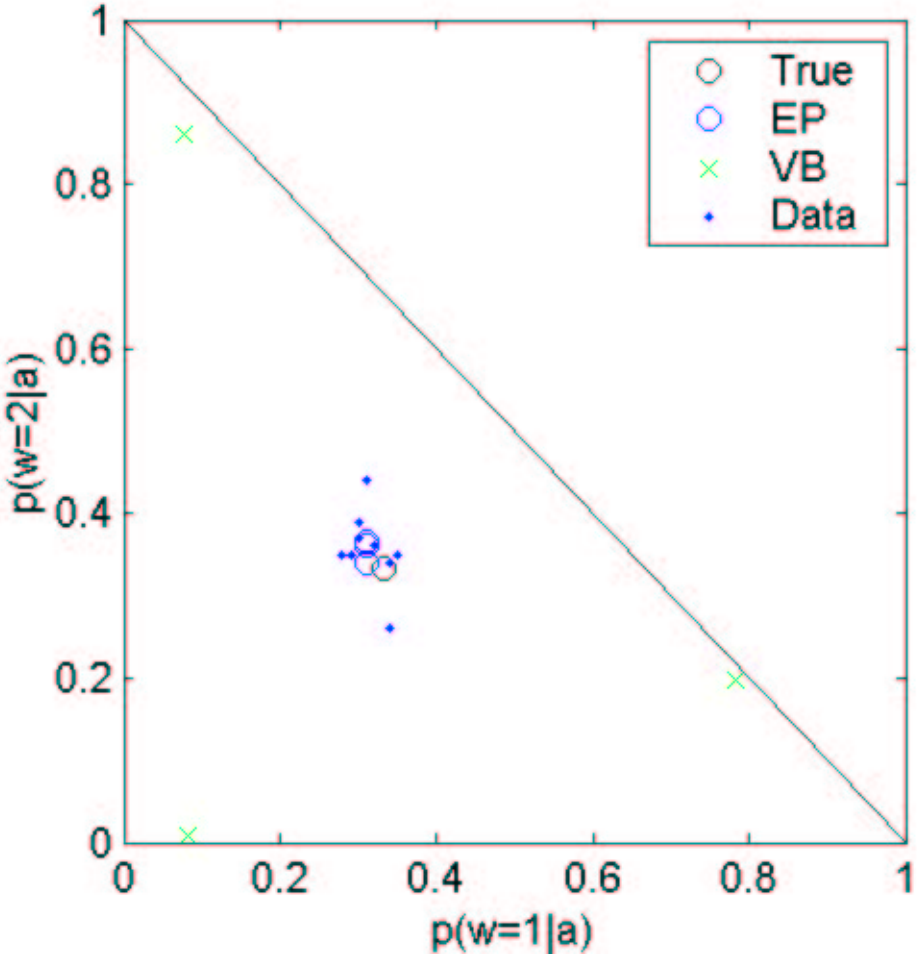
10 docs,
Length 10

Larger vocabulary



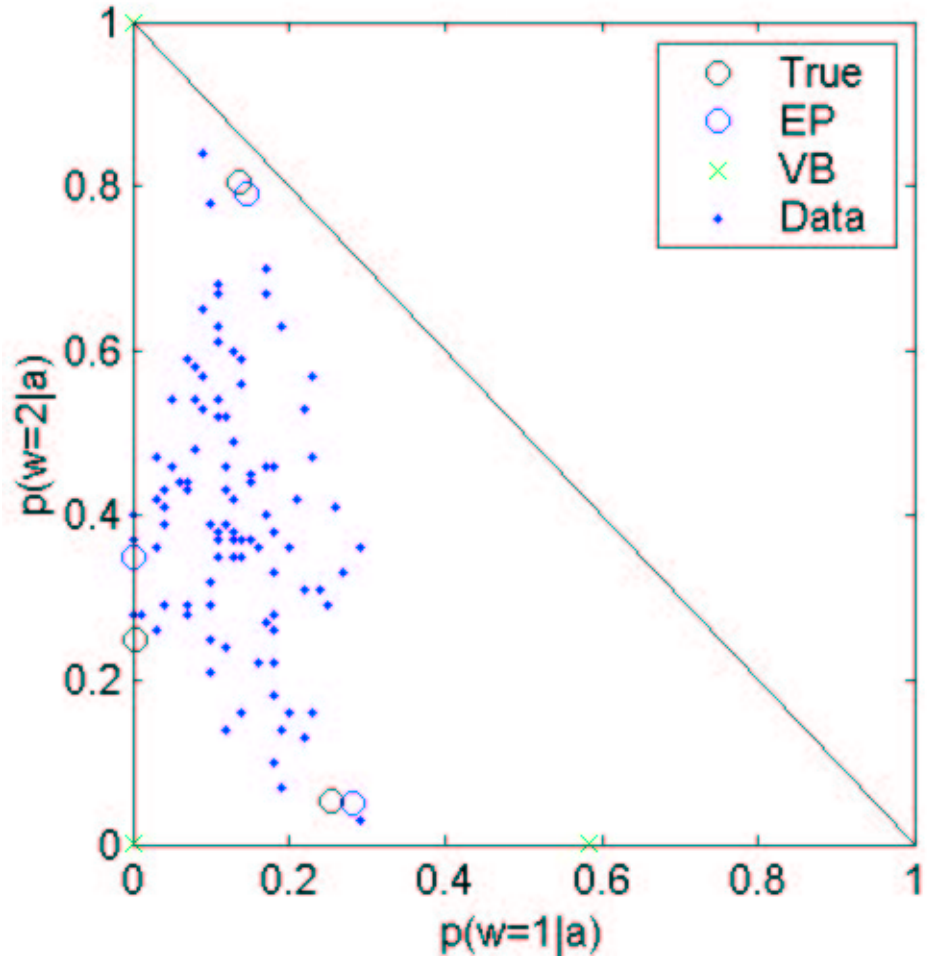
10 docs,
Length 10

Larger vocabulary



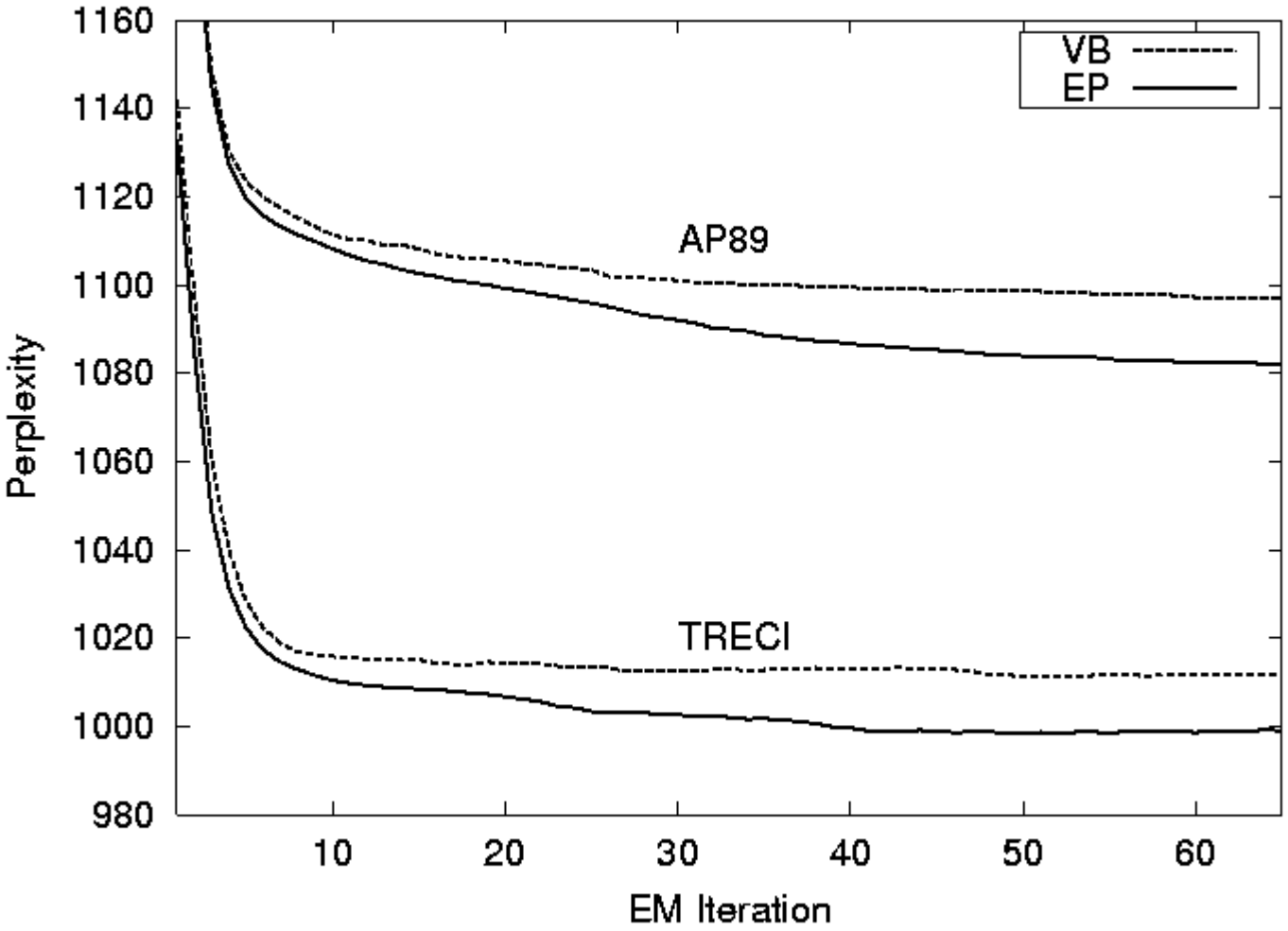
10 docs,
Length 10

More documents



100 docs,
Length 10

TREC documents



Perplexity

- EP and VB models have nearly same perplexity (!)
- Perplexity is document probability, normalized by length
  - Dilutes the penalty of extreme aspects
- Better measure: classification error

Summary & Future

- Approximate inference can lead to biased learning
  - Must preserve “Occam factor” terms
- Moment matching does well
- Simpler methods may also work
  - Area of aspect simplex
- Make visualizations!