Topic Models in ALVIS

Wray Buntine and Kimmo Valtonen
(Work with help from Antti Tuominen, Aleks Jakulin, Sami Perttu, ALVIS, CoSCo, ...)

Complex Systems Computation Group (CoSCo)
University of Helsinki &
Helsinki Inst. for Information Technology (HIIT)

July 8th, 2006

Overview

- **Background: PCA and ICA.**
- History, Religion, Interpretations, Algorithms.
- Wikipedia.
- Use in ALVIS for search.

† See `http://www.componentanalysis.org` and `http://wikipedia.hiit.fi`.

Bag of words as a Sparse Discrete representation for text

A page out of Dr. Zeuss's *The Cat in The Hat*:

> So, as fast as I could, I went after my net. And I said, "With my net I can bet them I bet, I bet, with my net, I can get those Things yet!"

In the *bag of words* representation as *word* (*count*):

> after(1) and(1) as(2) bet(3) can(2) could(1) fast(1) get(1) I(7) my(3) net(3) said(1) so(1) them(1) things(1) those(1) went(1) with(2) yet(1) .

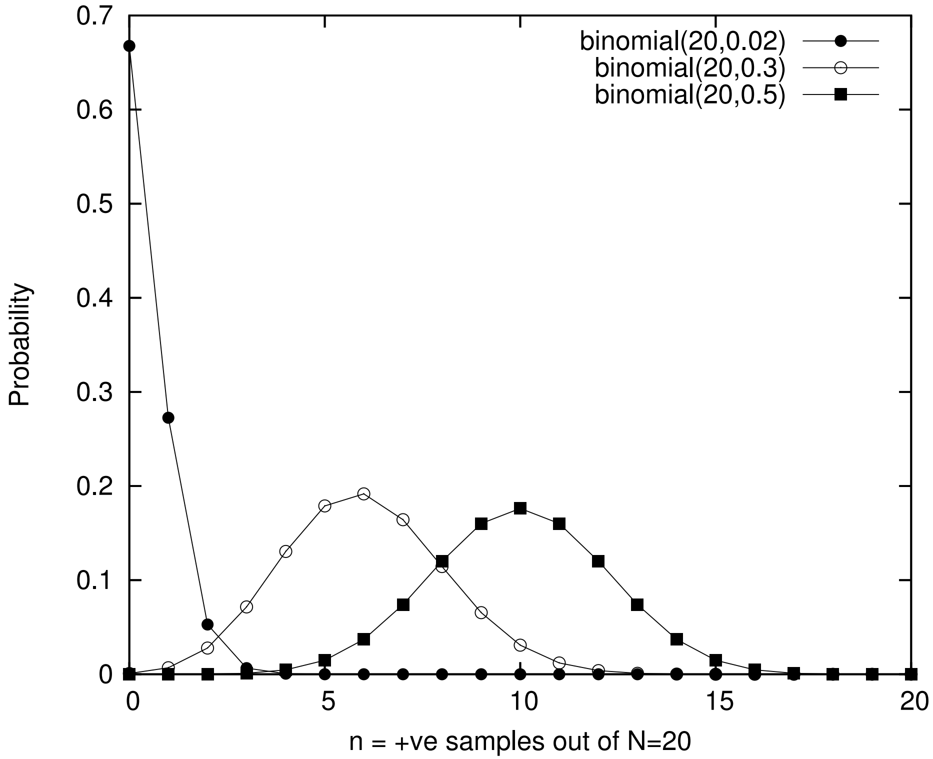
Notes:

- For the Reuters RCV1 collection from 2000: $I \approx 800k$ documents, $J \approx 400k$ different words (excluding those occurring few times), $S \approx 300Mb$ words total.
- Represent as sparse matrix/vector form with integer entries.
- Deleting words occurring less than < 50 times can shrink word dimension J by order of magnitude.

Principal Components Analysis (PCA)

- Invented by Karl Pearson, in 1901.
- Also known as Karhünen-Loève transform or Hotelling transform in image analysis, and latent semantic analysis (LSA) for text.
- Primarily used for *dimensionality reduction* prior to some other statistical processing.
- Has guarantees in terms of minimising least squares error in approximation.
- Also has a Gaussian interpretation using latent variables (Tipping and Bishop, 1999).
- Standard algorithm is to run an SVD and to throw away all but the top K eigenvectors and values, or to use sparse LAPACK style tools to extract just the top K without a full SVD.

Approximating Discrete Data



The plot shows different binomials in both its Gaussian regime and its Poisson regime.

Lesson: discrete data is only Gaussian-like in some contexts. When there are a lot of zeros, it is not Gaussian-like.

PCA: Issues

- PCA (with least squares or alternatively the Gaussian) is known to cause trouble in some contexts:
  - when values occur on a boundary (i.e., Gaussians don't admit boundaries).
  - with discrete and sparse values (i.e., outside the Gaussian regime).
- PCA has no realistic probabilistic interpretation in the discrete case.
  - Can be OK as a dimensionality reduction tool.
  - But no easy way to do probabilistic inference with the model.

NB. The ICA community can supply other issues!

Independent Components Analysis (ICA)

- Invented by Herault and Jutten in 1986.
- Intended for *blind source separation*, separation of independent signals in some data.
- Used for *dimensionality reduction* as well, based on the observation that PCA can perform poorly.
- Standard algorithm is the FastICA algorithm, developed by Hyvärinen and Oja.
- In effect, when the *dynamic range* is effectively 2/3/4-valued, we want to be carefully measuring the error, and the standard algorithm is poorly justified.
- Data often turned into real values before hand, for instance using `tf*idf` scores.

Discretizing ICA and PCA

- Can replace the Gaussians in the probabilistic intepretation of PCA with multinomials and Dirichlets.
- Alternative, Poissons and Gammas can be used, to get a robust version of ICA.
- Many variants.
- We call this Discrete Components Analysis (DCA).

See paper by Buntine and Jakulin for theory.

Viewing Components at the Word Level

From Blei, Ng, and Jordan, 2003.

Overview

- Background: PCA and ICA.

- **History, Religion, Interpretations, Algorithms.**

- Wikipedia.

- Use in ALVIS for search.

History

- Soft clustering, "grade of membership" (GOM), Woodbury and Manton, 1982.
- Admixture modelling in statistics, (198?).
- Hidden facets in image interpretation, Non-negative Matrix Factorization (NMF), Seung and Lee, 1999.
- Probabilistic Latent Semantic Analysis (PLSI), topics in text, Hofmann, 1999.
- Admixture modelling, fully Bayesian, population structure from genotype data, Pritchard, Stephens and Donnelly, 2000.
- Latent Dirichlet Allocation (LDA) Blei, Ng and Jordan, 2001. Variant of Pritchard *et al.* Introduced mean-field algorithm.
- Multi-aspect modelling: various 2001-2003.
- Gamma-Poisson model (GaP), Canny 2004 (extension of NMF).
- ...

Religious convictions

All manor of statistical beliefs and practices are permitted:

- Maximum likelihood or Kullback-Leibler divergence.
- Exponential family likelihoods or Bregman divergence.
- Regularised maximum likelihood.
- Bayesian and empirical Bayesian.

Catalogue of interpretations

- Approximating a discrete matrix as a product of lower dimension matrices.
- Multinomial version of the Gaussian interpretation of PCA (i.e., Roweis or Bishop style PCA).
- Multi-aspect modelling or soft clustering (documents have proportion/grade of membership).
- Admixture modelling (forming a mixture by mixing means, not distributions).
- Variation of ICA (independent component analysis) suitable for discrete data.
- Hidden topics for the individual words in a document, itself in a collection.

Correspondences

PLSI, Hofmann: regularised max.likelihood version, uses multinomial model for words, but no prior model for components, just regularisation. An instance of Gamma-Poisson with algorithmic customisation.

LDA, Blei *et al.*: Dirichlet-multinomial model, made sequential (i.e., sequence of words, not bag of words). An instance of Gamma-Poisson when hyper-parameters are fixed.

Algorithms

I	documents
J	lexemes
K	components
S	words in corpus

- Mean field and Gibbs on bag of words are O(JK+SK) in time per cycle and O(IK+JK) in space.
- Mean field DCA at dimension K a few times slower than incremental (i.e., fast) PCA at same dimension.
- Gibbs with Rao-Blackwellisation is O(SK) in time per cycle and O(S+JK) in space. Less then others.
- Gibbs with Rao-Blackwellisation much better time and space for very small "documents", e.g., analysis of sentences or noun-verb pairs, etc.
- Minka's Expectation-Propagation (EP) is an extra order of magnitude in space so is not practical.

Computational Issues

- While doing Gibbs sampling, we do *not* really want Gibbs sampling:
  - Components defined symmetrically, thus a true posterior mean would smear out all parameters to a bland uniform.
  - We should really be doing millions of major cycles, we just do 1000's when doing discovery (as opposed to hypothesis testing).

  i.e., we use Gibbs as an algorithm to generate an estimate, not as a method to do sound statistical inference.

- Text data as bags of words and the document intermediate variables (proportions $\vec{m}$, topic assignments $\vec{k}$, etc.) don't need to be kept in main memory and can be streamed over using `mmap()`.

- Memory constraints are thus two copies (current value and sufficient statistics) for $\mathbf{\Theta}$, and a typical desktop could handle $J = 200k$ and $K = 400$.

Overview

- Background: PCA and ICA.
- History, Religion, Interpretations, Algorithms.
- **Wikipedia.**
- Use in ALVIS for search.

Overview of the MPCA software

- Available from the website `http://www.componentanalysis.org`, and software releases announced on `Freshmeat.NET`, under GPL.

- Lots of options and diagnostics, train-test evaluations, displays, ...

- Implements mean field, Gibbs, and Gibbs with Rao-Blackwellisation.

- Support for parallel processing via MPI (mean field and Gibbs only), and for multiprocessors (but somewhat buggy)

- Manages gigabytes of text.

- Support for link matrices and subject-specific PageRank calculations.

Wikipedia

- We built a $K = 400$ component model of $I = 980k$ web pages from the English-language Wikipedia* dated December 2005.

- DCA is run on bags of lemmatised words organised by part of speech, and the external URLs at the page.

- Words or URLs occurring less than 10 times ignored, leaving feature dimension about $J = 1000k$.

- Used Gamma-Poisson model with *sparse* Gamma component priors (i.e., about 90% of component values are zero). Hyper-parameters for the component priors fitted (gradient ascent with trust regions).

- Thus each document is a sparse vector (perhaps 300 entries).

- Fitting used Gibbs with Rao-Blackwellisation, 1000 major cycles, on a dual CPU Opteron (64-bit) with 4Gb memory. About 6 days.

*`http://en.wikipedia.org`

Wikipedia, cont.

Example components on-line.

Overview

- Background: PCA and ICA.
- History, Religion, Interpretations, Algorithms.
- Wikipedia.
- **Use in ALVIS for search.**

Use in ALVIS

- Predefined topics (e.g., MESH, ODP, Dewey Decimal) best when they apply.

- Otherwise, use DCA to develop topics. (Hierarchical version TBD top-down).

- Use topics to allow mixed topic-browse and search.

- Allow users to enter block of text to indicate topical-preference, then combine this as the topical aspect of a fused language model. See demo. at *http://wikipedia.hiit.fi* of June 2004 Wikipedia.

Search engine topics.

Ads and banners; Advertising agency's appointments; Advertising and marketing; Affiliate program; America Online (AOL); Ask Jeeves and Google Answers; Blogs; Book search; Customer service; Dates; Desktop search; Domain name registration; DoubleClick, Inc.; Earnings of companies; E-mail spam; Film industry; Forums and discussions; Games; Google co-founders; Interactive media and advertising; Internet; Legal; Local search; Maps; Microsoft and Google; Mobile communication; Music search; News; Online multimedia; Organizations and standards; Paid inclusions and search; People; Privacy issue; "Acceptable" SEO; Regions; Research; Search Engine Marketing; Search engine optimization; Search engine optimization and marketing; Search marketing; Security issues; Shopping Search; Stock market; User interfaces; Users, people, communities; Web advertising; Wikipedia/Deja;

Overview of techniques

- Extracting relevant named-entities using an IR system: extension of a relevance language model:

$$p(name|query) = \sum_{d \in docs} p(d|query)p(name|d, query)$$

  $p(d|query)$ got from the IR score for a document. $p(name|d, query)$ is *ad hoc* based on location arguments.

- Extracting relevant topics using an IR system.

$$p(topic|query) = \sum_{d \in docs} p(d|query)p(topic|d)$$

  Efficient to compute (similar to IR retrieval) since topic probabilities ($p(topic|d)$) are sparse.

Wikipedia, cont.

Example search on-line.

Overview

- Background: PCA and ICA.
- History, Religion, Interpretations, Algorithms.
- Wikipedia.
- Use in ALVIS for search.

Thank you!