LATENT VARIABLE MODELS FOR DISCRETE DATA

- DISCRETE DATA ONLY, IE TEXT ANALYSIS BUT OTHER STUFF TOO. **TOKENS**: VARIABLE LENGTH SEQUENCES OF CATEGORICAL DATA $y \in \{1..V\}$ ONE-HOT ENCODING

BAG-OF-WORDS: IGNORES WORD ORDER, FIXED LENGTH VECTOR OF COUNTS. $N \times V$; N IS NO DOCUMENTS. RANGE BUT SAME. CAN HAVE MULTIPLE BOW x DOCUMENT → **CHANNELS | FEATURES TYPES**

GOAL: BUILDING JOINT MODELS FOR $P(y_i)$, $P(N_i)$ USING LATENTS

MIXTURE MODELS

ASSOCIATE SINGLE DISCRETE HIDDEN VAR PER DOCUMENT. PRIOR: $q_i \sim CAT(\pi)$ → **TOPIC**. LIKELIHOOD: $P(Y|_{1:L_i}, L_i \mid q_i = k) = \prod^{L} CAT(y_{il} \mid b_k)$ b_k IS TOPIC WORD DISTRIBUTION

- CAN HAVE DIFFERENT TOPIC MATRIX FOR EACH OUTPUT VARIABLE → 'UNSUPERVISED NAIVE BAYES'
  - **UNKNOWN LENGTH** → USE POISSON
  - **LENGTH KNOWN** → MULTINOMIAL

  } COUNT VECTORS

EXPONENTIAL FAMILY PCA

EXPONENTIAL FAMILY PCA: UNSUPERVISED ANALOG OF GLM

USES VECTOR OF CONTINUOUS, REAL-VALUED HIDDEN VARS.

- **LATENT SEMANTIC ANALYSIS | INDEXING**: GAUSSIAN PRIOR, GAUSSIAN LIKELIHOOD → APPLY PCA TO TERM-BY-DOCUMENT COUNT MATRIX
- **CATEGORICAL PCA**: USE MULTINOULLI | MULTINOMIAL + SOFTMAX RESPONSE
- **COUNTS**: IF WE HAVE THEM, MULTINOMIAL OR POISSON
- TRICKY TO FIT → DEGENERATE EM, VARIATIONAL EM, MCMC

- $P(y_{i,1:L}) = \int \left[\prod^{L} P(y_{il} \mid z_i, W) \right] N(z_i \mid \mu, \Sigma)\, dz_i$ ← **VISIBLE DISTRIBUTION**

LDA, mPCA

- WE WANT TO USE **DUAL PARAMETERS** OF EXPONENTIAL FAMILY. LOG ODDS → PROBABILITY VECTOR
- LATENT PRIOR $\pi_i \sim DIRICHLET(\alpha)$
- **COUNT VECTOR WITH KNOWN SUM** → mPCA, MULTINOMIAL. MARGINAL IS $P(n_i \mid L_i) = \int Mu(n_i \mid L_i, B\pi_i)\, DIR(\pi_i \mid \alpha)\, d\pi_i$
- **VARIABLE LENGTH SEQUENCE** (KNOWN LENGTH) → LATENT DIRICHLET ALLOCATION **LDA**. MARGINAL: $P(y_{i,1:L}) = \prod^{L_i} CAT(y_{il} \mid B\pi_i)$; PROBABILISTIC LSA

GaP MODEL

- DOES NOT CONSTRAIN COUNT VECTOR SUM TO BE OBSERVED. LATENT VARIABLES FORCED POSITIVE.
- **PRIOR**: $P(z_i^+) = \prod^{K} Ga(z_{ik}^+ \mid \alpha_k \beta_k)$ **LIKELIHOOD**: $P(n_i \mid z_i^+) = \prod^{V} Poi(n_{iv} \mid b_v^T z_i^+)$
- FIXED L → REDUCES TO mPCA
- $\alpha_k = \beta_k = 0$ → NON-NEGATIVE MATRIX FACTORIZATION, NOT PROPERLY PROBABILISTIC

LATENT DIRICHLET ALLOCATION

- EVERY WORD IS ASSIGNED ITS OWN TOPIC. FROM DOCUMENT-SPECIFIC DISTRIBUTION **ADMIXTURE MIXTURE | MIXED MEMBERSHIP MODEL** DOCUMENT BELONGS TO DISTRIBUTION OVER TOPICS

TOPIC
$$\begin{cases} \pi_i \mid \alpha \sim DIR(\alpha 1_K) \\ q_{il} \mid \pi_i \sim CAT(\pi_i) \\ b_k \mid \gamma \sim DIR(\gamma 1_V) \\ y_{il} \mid q_{il} = k, B \sim CAT(b_k) \end{cases}$$

- **MARGINALIZE**: $P(y_{il} = v \mid \pi_i) = \sum_k \pi_{ik} b_{kv}$, π_i TOPIC DISTRIBUTION FOR DOCUMENT.
- **GEOMETRICALLY** IS DIMENSIONALITY REDUCTION. WE PROJECT A POINT IN V-SPACE (WORDS) TO K-SPACE (TOPICS). $V > K$

  TOPIC VECTORS 'LIVE' IN WORD SPACE. ALLOWS TO DISAMBIGUATE TOPIC BY LOOKING AT OTHER WORDS
- GIVEN WORD L IN DOC i WE CAN INFER LIKELY TOPIC $P(q_{il} = k \mid y_i, \theta)$

- USEFUL FOR TOPIC DISCOVERY. USES TAGS TO MATCH LDA TOPICS FOR IDENTIFIABILITY
- **UNIGRAMS ONLY** → NOT A VERY GOOD LANGUAGE MODEL
- **PERPLEXITY**: PERFORMANCE MEASURE FOR LANG MODELS; MEASURES **'BRANCHING FACTOR'** OF PREDICTIVE DISTRIBUTION
  - PERPLEXITY $(p, q) = 2^{H(p,q)}$, H IS CROSSENTROPY. FOR UNIGRAM MODELS $H = -\frac{1}{N}\sum_N \frac{1}{L_i}\sum_l \log q(y_{il})$. ALSO MEAN OF INVERSE PREDICTIVE PROBS.
  - **FOR LDA**: $P(v)$ GOTTEN BY PLUGGING IN B, POSTERIOR MEAN, AND INTEGRATE Q OUT WITH MEAN FIELD INFERENCE

FITTING LDA

- **GIBBS SAMPLING**, COLLAPSED VARIETY: WORD IN DOC IS ASSIGNED ON HOW OFTEN WORD IS IN TOPIC AND HOW OFTEN TOPIC IS IN DOCUMENT.
  OBTAIN FULL CONDITIONALS PER TOPICS $P(q_{il} = k \mid q_{-i,l}, y, \alpha, \gamma)$. RANDOMLY ASSIGN TOPIC TO WORDS. THEN FOR GIVEN WORD DECREMENT COUNTS, BASED ON CUR ASSIGNED TOPIC. ~~DECREMENT~~ DRAW NEW TOPIC, UPDATE COUNTS, REPEAT.
- **BATCH VARIATIONAL INFERENCE**: USE VARIATIONAL EM
  - **SEQUENCE VERSION**: USES VANILLA MEAN FIELD, FULLY FACTORED APPROX
    $Q(\pi_i, q_i) = DIR(\pi_i \mid \tilde{\pi}_i) \prod_l CAT(q_{il} \mid \tilde{q}_{il})$, $O((\sum_i L_i) VK)$
  - **COUNT VERSION**: ONLY $O(NVK)$. WORK ON mPCA PROBLEM BECAUSE LESS STORAGE
    $Q(\pi_i, c_i) = DIR(\pi_i \mid \tilde{\pi}_i) \prod_v Mu(c_{iv} \mid n_{iv}, \tilde{c}_{iv})$
  - **VB VERSION**: VB INSTEAD THAN EM, WE ALSO INFER PARAMETERS. ENCOURAGES SPARSITY.
    $Q(\pi_i, c_i, B) = DIR(\pi_i \mid \tilde{\pi}_i) \prod Mu(c_{iv} \mid n_{iv}, \tilde{c}_{iv}) \prod_k DIR(b_k \mid \tilde{b}_k)$ OF B
- **ONLINE VARIATIONAL INFERENCE**: *
  NORMALLY DO E STEP. COMPUTE PARAMS FOR B AS IF SINGLE DATA WERE FULL SET; PARTIAL UPDATES AGAIN FOR B WITH ρ_k WEIGHT ON NEW AND $(1-\rho_k)$ ON OLD. **SERIOUSLY, DO MINIBATCHES**
  → FASTEREST

- **PICKING K**

ANNEALED IMPORTANCE SAMPLING, CROSSVALIDATION, VAR LOWER BOUNDS, NONPARAMETRIC METHODS

EXTENSIONS TO LDA

- **CORRELATED TOPIC MODEL**

VANILLA LDA DOESN'T DO IT BECAUSE PRIOR FOR π_i IS DIRICHLET. REPLACE WITH LOGISTIC NORMAL, AS IN CATEGORICAL PCA.
FITTING IS TRICKY BECAUSE PRIOR NO LONGER CONJUGATED TO LIKELIHOOD. USE MEAN-FIELD OR VARIATIONAL MULTICLASS REGR.
VISUALIZATION BY OBTAINING $\hat{\Sigma}^{-1}$ AND PRUNE LOW-STRENGTH EDGES. GET SPARSE GRAPHICAL MODEL.

- **DYNAMIC TOPIC MODEL**

MODELS TOPIC DISTRIBUTIONS EVOLVING OVER TIME. USE DYNAMIC LOGISTIC NORMAL MODEL. ASSUME TOPICS EVOLVE WITH GAUSSIAN RANDOM WALK THEN MAP TO PROBABILITIES VIA SOFTMAX
word distribution of

- **LDA - HMM**

COMBINES LDA AND HMM, HMM MODELS/GENERATES STOPWORDS and/or SYNTAX. HAS SPECIAL STATE TO GENERATE WITH LDA. LDA DOES SEMANTIC WORDS

- **SUPERVISED LDA**
  - GENERATIVE: IE FOR SENTIMENT ANALYSIS. WE HAVE WORDS, AND ASSOCIATED LABEL. CLASS LABEL GENERATED FROM TOPICS
  $P(c_i|\bar{q}_i) = \mathrm{BER}(\mathrm{sigm}(w^T\bar{q}))$. $\bar{q}$ IS EMPIRICAL TOPIC. FIT WITH MONTECARLO EM
  - DISCRIMINATIVE: THE TOPIC PRIOR IS NOW INPUT DEPENDENT $P(q_{il}|\pi_i, c_i = c, \theta) = \mathrm{CAT}(A_c\pi)$. IMAGE TAGGING
  RETURN TAGS GIVEN INPUT. (CONVERTED) SIMPLEST: USE MIXTURE OF EXPERTS WITH MULTIPLE OUTPUTS.
  REPLACE DIRICHLET WITH $\pi_i = S(Wx_i)$ DETERMINISTIC. ← MULTINOMIAL REGRESSION LDA
  DETERMINISTIC π. OBSERVED → LITTLE INFLUENCE, TAG SCORE...
    - OTHERS: MAKE W RANDOM VARIABLE; RANDOM FX MIXTURE OF EXPERTS, CONDITIONAL TOPIC RANDOM FIELD, DIRICHLET MULTINOMIAL REGRESSION LDA, LABELED/PARTIALLY LABELED LDA
    MARKOV DYNAMICS
  - DISCRIMINATIVE CATEGORICAL: EXPAND CATEGORICAL PCA WITH INPUTS, LINEAR REGRESSION FOR IN/ MAPPING.
  CATPCA FOR HID/OUT MAPPING. LIKE AN ANN WITH 1 PROBABILISTIC HIDDEN LAYER BUT WITH BOTTLENECK (HIDDEN).

LVM FOR GRAPH-STRUCTURED DATA

- **STOCHASTIC BLOCK MODEL:** INFER 'BLOCK' STRUCTURE OF GRAPH FROM ADJACENCY MATRIX. EVEN IF 'PHENOTYPE' GRAPH HAS NO ARCS. PROPERTIES SUCH AS 'ALL N IN B CONNECT TO SAME NODE, ETC'. MODEL BLOCK AS LATENTS. • IF WE HAVE FEATURES FOR NODES, WE CAN MAKE DISCRIMINATIVE LIKE RELATIONAL EXTENSION OF MIXTURE OF EXPERTS

- **MIXED MEMBERSHIP SBM**

ALLOWS NODES TO BELONG TO MORE THAN ONE CLUSTER. AKIN TO SOFT/FUZZY CLUSTERING. SOCIAL NETWORK ANALYSIS.

- **RELATIONAL TOPIC MODEL**

EXTENDS SUPERVISED LDA. WHEN NODES HAVE ATTRIBUTES. IE PAPERS, CITATIONS. WE WANT LINKS GIVEN TEXT OR VICEVERSA.
ATTRIBUTE: FULLTEXT.
$P(R_{ij}=1|\bar{q}_i, \bar{q}_j, \theta) = \mathrm{sigm}(w^T(\bar{q}_i \otimes \bar{q}_j) + w_0)$, FORCES ~~GRAPH~~ LEARNED LATENT SPACE TO BE PREDICTIVE OF GRAPH STRUCTURE AND WORDS, UNLIKE LDA

LVM FOR RELATIONAL DATA

MULTIPLE TYPES OF OBJECTS (VARS), MULTIPLE TYPE OF RELATIONS. RELATION IS TYPED. REPRESENTED WITH BINARY MATRICES. STATISTICAL RELATIONAL LEARNING

- **INFINITE RELATIONAL MODEL:** EXTENDS SBM, ASSOCIATES LATENT VAR $q^i \in \{1..K\}$ WITH EACH ENTITY i OF TYPE T. PROBABILITY OF RELATION BETWEEN TWO ENTITIES BY LOOKING UP PROB OF RELATION BETWEEN TYPES. MULTIDIM.
- CAN BE USED TO INFER/LEARN ONTOLOGIES FROM DATA. • CAN BE USED TO LEARN CLUSTERS BASED ON RELATIONS AND FEATURES AT SAME TIME
- **COLLABORATIVE FILTERING**

NETFLIX DATA EXAMPLE → PROBABILISTIC MATRIX FACTORIZATION, REPLACES DISCRETE LATENT VARIABLES WITH UNCONSTRAINED CONTINUOUS ONES. USERS AND MOVIES EMBEDDED IN SAME LOW-DIM CONTINUOUS SPACES. PMF IS CLOSE TO SVD BUT MISSING DATA → NONCONVEX OBJECTIVE
MINIMIZE NLL WITH GRADIENT DESCENT METHODS, REGULARIZE WITH GAUSSIAN PRIORS. • MOVIE/USER SPECIFIC EFFECTS MODELED WITH BIAS TERM, CAN MAKE IT ADAPT OVER TIME. EXPLOIT SIDE INFORMATION

LATENT DIRICHLET ALLOCATION – ADDENDUM

- FOR EACH DOCUMENT
  - DRAW TOPIC DISTRIBUTION $\theta_d \sim \mathrm{DIR}(\alpha)$ → 'DRAWS ME A MULTINOMIAL'
  - FOR EACH WORD IN DOCUMENT
    - DRAW SPECIFIC TOPIC $z_{dn} \sim \mathrm{MULTINOMIAL}(\theta_d)$
    - DRAW WORD $w_{dn} \sim \beta_{z_{dn}}$ FROM β DISTRIBUTION OF WORDS IN CHOSEN TOPIC

- DOCUMENT MARGINAL AS CONTINUOUS MIXTURE

$$P(w \mid \alpha, \beta) = \int P(\theta \mid \alpha) \left(\prod P(w_n \mid \theta, \beta) \right) d\theta$$

- LDA SPACE IS ALL POSSIBLE DISTRIBUTIONS OVER WORDS
  - AXES ARE WORDS

- KEY INFERENCE:

POSTERIOR OF HIDDEN GIVEN DOCUMENT

$$P(\theta, z \mid w, \alpha, \beta) = \frac{P(\theta, z, w \mid \alpha, \beta)}{P(w \mid \alpha, \beta)}$$

- HIERARCHICAL DIRICHLET PROCESS

GENERALIZATION OF LDA. USES CRP TO AUTOTUNE NO OF TOPICS FROM DATA! CLUSTER → TOPIC