MIXTURE MODELS

ASSUMES OBSERVED VARIABLES ARE CORRELATED B/C ARISE FROM HIDDEN COMMON CAUSE **LATENT VARIABLE MODELS.**

- HARDER TO FIT
- FEWER PARAMS
- CAN LEAD TO **BOTTLENECKS** → COMPRESSED REPRESENTATION OF DATA. UNSUPERVISED LEARNING, THINK AUTOENCODERS

MIXTURE MODELS

SIMPLEST FORM OF LVM. $z_i \in \{1 \dots K\}$ DISCRETE LATENT STATES. DISCRETE • PRIOR $p(z_i) = \text{Cat}(\pi)$ • LIKELIHOOD: $p(x_i|z_i = k) = p_k(x_i)$ KTH BASE DISTRIBUTION

- WE MIX TOGETHER B. DISTS $p(x_i|\theta) = \sum^{K} \pi_k p_k(x_i|\theta)$, CONVEX COMBINATION OF p_ks

$0 \le \pi_k \le 1$, $\sum^{K} \pi_k = 1$, π_k MIXING WEIGHTS

DIFFERENT TYPES OF MIXTURE MODELS DEPENDING ON TYPE OF $p(z_i)$ AND $p(x_i|z_i)$

MIXTURE OF GAUSSIANS

- PRIOR = DISCRETE
- LIKELIHOOD = MVN

$$p(x_k|\theta) = \sum^{K} \pi_k \cdot \mathcal{N}(x_i|\mu_k, \Sigma_k)$$

MIXTURE OF MULTINOULLIS

- PRIOR = DISCRETE
- LIKELIHOOD = DISCRETE

$$p(x_k|\theta) = \prod_j^{D} \text{Ber}(x_{ij}|\mu_{jk}) = \prod_1^{D} \mu_{jk}^{x_{ij}} (1-\mu_{jk})^{1-x_{ij}} \qquad z_i = k$$

μ_{jk} = PROB BIT j IS ON IN CLUSTER k

- $E[x] = \sum \pi_k \mu_k$ $\quad \text{COV}[x] = \sum \pi_k [\Sigma_k + \mu_k \mu_k^T] - E[x]E[x]^T$, $\Sigma_k = \text{DIAG}(\mu_{jk}(1-\mu_{jk}))$

USING MIXTURE MODELS FOR CLUSTERING

MM USED MAINLY FOR TWO THINGS: **BLACK BOX DENSITY MODELING** $p(x_i)$, DATA COMPRESSION, OUTLIER DETECTION, MODEL CLASS CONDITIONAL DENSITY $p(x|y=c)$

~~CLASSIFICATION~~

CLUSTERING: FIRST FIT MODEL, COMPUTE $p(z_i = k|x_i, \theta)$ → POSTERIOR OF POINT i BELONGING TO CLUSTER k

RESPONSIBILITY

$$r_{ik} = p(z_i = k|x_i, \theta) = \frac{p(z_i = k|\theta)\, p(x_i|z_i = k, \theta)}{\sum_{k'=1}^{K} p(z_i = k'|\theta)\, p(x_i|z = k', \theta)}$$

→ SOFT CLUSTERING. LIKE GEN CLASSIFIERS BUT HERE WE NEVER OBSERVE z_i

UNCERTAINTY $1 - \max_k r_{ik}$, IF SMALL → $z_i^* = \arg\max_k r_{ik} = \arg\max_k \log p(x_i|z_i = k, \theta) + \log p(z_i = k, \theta)$ HARD CLUSTERING

- WE CAN REPRESENT CLUSTERS USING PROTOTYPES / CENTROIDS → $\hat{\mu}_k$ FOR THE Ks (MAP)

MIXTURE OF EXPERTS

DISCRIMINATIVE MODELS FOR CLASSIFICATION / REGRESSION. APPLIED MIXTURE COMPONENT DEPEND ON INPUT VALUE, EACH IS AN EXPERT IN A PART OF THE INPUT SPACE

$p(y_i|x_i, z_i = k, \theta) = \mathcal{N}(y_i|w_k^T x_i, \sigma_k^2)$, $p(z_i|x_i, \theta) = \text{Cat}(z_i|S(V^T x_i))$, $p(z_i = k|x_i, \theta)$ GATING FUNCTION

- COMPLETE PREDICTION MODEL $p(y_i|x_i, \theta) = \sum p(z_i = k|x_i, \theta)\, p(y_i|x_i, z_i = k, \theta)$
  - ANY MODEL CAN BE EXPERT
  - NN GATING, NN EXPERTS → MIXTURE DENSITY NETWORK
  - EACH EXPERT IS MoE ITSELF → HIERARCHICAL MOE
- MOE USEFUL IN INVERSE PROBLEMS → IE INVERSE KINEMATICS WITH A MANY-TO-ONE MAPPING

PARAMETER ESTIMATION

HOW TO LEARN THE θS • IF z_iS WERE OBSERVED → D-SEPARATION → POSTERIOR FACTORIZES → EASY BUT z_iS HIDDEN!
COND INDEP.

PROBLEM: PARAMETER UNIDENTIFIABILITY

- IF z_i WERE OBSERVED $p(\theta|D) = Dir(\pi|D) \prod^{K} NIW(\mu_k, \Sigma_k|D)$ → FIND GLOBAL MAP/MLE
- z_i HIDDEN → WE GET DIFFERENT UNIMODAL LIKELIHOOD FOR EACH WAY OF FILLING IN z_i. WHEN WE MARGINALIZE OVER z_i WE GET A MULTIMODAL POSTERIOR $p(\theta|D)$

  → DIFFERENT LABELING OF CLUSTERS. NO UNIQUE MLE/MAP. $K!$ POSSIBLE LABELINGS

  → FINDING OPTIMAL MLE FOR A GMM IS NP-HARD!
- ISSUES FOR BAYESIAN INFERENCE: IT'S POINTLESS TO AVG TOGETHER (FOR APPROX. POSTERIOR MEAN) SAMPLES FROM A MULTIMODAL POSTERIOR.
  – STILL OK TO AVG SAMPLES FROM POSTERIOR PREDICTIVE BECAUSE LIKELIHOOD IS INVARIANT TO MODE.
- HOW TO FIX

  OPTIMAL: MCMC

  – QUICK FIX: COMPUTE SINGLE LOCAL MODE/MAP APPROX. POSTERIOR UNCERTAINTY ABOUT PARAMS << POST. UNCERTAINTY ABOUT HIDDENS → $p(z_i|x_i, \hat{\theta})$

PROBLEM: NONCONVEX MAP ESTIMATE

$$\log p(D|\theta) = \sum_i \log\left[\sum p(x_i, z_i|\theta_i)\right]$$ CAN'T PUSH LOG INSIDE SUM, BUT SINCE STUFF IS IN EXPONENTIAL FAMILY WE CAN DO

$$\ell_c(\theta) = \sum_i \log p(x_i, z_i|\theta) = \theta^T\left(\sum \phi(x_i z_i)\right) - N Z(\theta)$$ COMPLETE DATA LOGLIKELIHOOD

$$\ell(\theta) = \sum_i \log\left[\sum e^{\theta^T \phi(z_i, x_i)}\right] - N \log(Z(\theta))$$ OBSERVED DATA LL (WHEN MISSING DATA)

- THE TWO TERMS ARE CONVEX BUT NO GUARANTEES ON THE DIFFERENCE
- SOLVE WITH HEURISTIC OPTIMIZATION ALGOS: SIMULATED ANNEALING, GENETIC ALGORITHMS, MULTIPLE RANDOM RESTARTS, EXPERIMENTAL CONVEX METHOD BASED ON ℓ_1 CONVEX PENALTIES

  UNSUPERVISED SPARSE KERNEL LOGISTIC REGRESSION

EXPECTATION – MAXIMIZATION ALGORITHM

WHEN MISSING DATA / LATENT VARIABLES ML/MAP IS HARD TO COMPUTE. GENERALLY WE CAN USE GRADIENT BASED OPTIMIZER AND FIND LOCAL MINIMUM OF NLL $NLL = -\frac{1}{N}\log p(D|\theta)$

WHEN THERE ARE CONSTRAINTS TO BE ENFORCED IT'S SIMPLE (NOT NEC-LY FASTER) TO USE EXPECTATION - MAXIMIZATION

- ITERATIVE, CLOSED-FORM UPDATES AT EACH STEP
- E STEP → INFERS MISSING DATA GIVEN PARAMS | MAX LL OVER LATENT PARAMS
- M STEP → OPTIMIZES THE PARAMS GIVEN THE 'FILLED IN DATA' | MAX OVER LL PARAMS

EM IS SPECIAL CASE OF BOUND OPTIMIZATION / MINORIZE - MAXIMIZE ALGOS

GENERAL

OBSERVED LL

$\ell(\theta) = \sum \log\left[\sum p(x_i|z_i,\theta)\right]$ CAN'T PUSH LOG INSIDE SUM | $\ell_c(\theta) = \sum_i^N \log p(x_i, z_i|\theta)$ COMPLETE DATA L.L. NO CAN COMPUTE! | $Q(\theta,\theta^{t-1}) = E[\ell_c(\theta)|D,\theta^{t-1}]$

XPECTED COMPLETE DATA LL

→ AUXILIARY FCN

- E STEP COMPUTES TERMS INSIDE Q ON WHICH MLE DEPENDS ON → XPECTED SUFFICIENT STATISTICS. $Q(\theta|\theta^t) = E_{z|x}[\log L(\theta,X,Z)]$
- M STEP OPTIMIZES Q WRT TO θ $\quad \theta^t = \underset{\theta}{\operatorname{argmax}} Q(\theta,\theta^{t-1})$ $\quad \theta^t = \operatorname{argmax} Q(\theta,\theta^{t-1}) + \log p(\theta)$

MAP MLE MAP

- EM MONOTONICALLY INCREASES LL OF OBSERVED DATA (OR STAYS SAME). IF OBJ GOES DOWN → SOMETHING IS WRONG!

EM FOR GMM

- AUXILIARY FCN: $Q(\theta,\theta^{t-1}) = E\left[\sum_i \log p(x_i,z_i|\theta)\right] = \sum_i\sum_k R_{ik}\log\pi_k + \sum_i\sum_k R_{ik}\log p(x_i|\theta)$

R_{ik} = RESPONSIBILITY OF CLUSTER K FOR DATAPOINT I

- E STEP: $R_{ik} = \dfrac{\pi_k\, p(x_i|\theta_k^{t-1})}{\sum_{k'}\pi_{k'}\, p(x_i|\theta_{k'}^{t-1})}$
- M STEP: $\pi_k = \dfrac{R_k}{N}$ → WEIGHTED NO OF POINTS $\in K$. $\ell(\mu_k,\Sigma_k) = \sum_k\sum_i R_{ik}\log p(x_i|\theta)$ $\xrightarrow{\text{SOME MATH LATER}}$ $\mu_k = \dfrac{\sum_i R_{ik}x_i}{R_k}$

$\Sigma_k = \dfrac{\sum_i R_{ik} x_i x_i^T}{R_k} - \mu_k\mu_k^T$

- FINAL: $\theta^t = (\pi_k, \mu_k, \Sigma_k)$ FOR $k = 1:K$, RINSE & REPEAT

QUALITATIVELY MEAN OF CLUSTER K IS WEIGHTED AVG OF ALL POINTS $\in K$.
COV IS PROPORTIONAL TO EMPIRICAL SCATTER MATRIX.

K-MEANS

CAN BE SEEN AS VARIANT OF EM ON GMM WHERE $\Sigma_k = \sigma^2 I_D$ $\pi_k = 1/K$ ARE FIXED SO ONLY μ_k ARE TO BE ESTIMATED

- POSTERIOR APPROXIMATED TO $p(z_i = k|x_i,\theta) \approx \mathbb{1}(k=z)$ → HARD EM EQUAL SPHERICAL COVARIANCE MATRIX → $z_i = \underset{k}{\operatorname{argmin}} \|x_i - \mu_k\|_2^2$
- $\mu_k = \frac{1}{N_k}\sum x_i$ • $O(NKD)$ TIME • UNTIL CONVERGENCE
- INITIALIZATION: RANDOM OR FARTHEST POINT CLUSTERING / K-MEANS ++
  GUARANTEES DISTORTION $\leq O(\log K)$
  - CHOOSE CENTROIDS WITH PROBABILITY PROPORTIONAL TO DISTANCE OF ALREADY PICKED ONES

VECTOR QUANTIZATION

~~CAN BE~~ K-MEANS CAN BE SEEN AS GREEDY ALGO FOR OPTIMIZING LOSS RELATED TO DATA COMPRESSION

- ENCODE VECTORS x_i (REAL VALUED) WITH A DISCRETE SYMBOL $z_i = \{1..K\}$ INDEX INTO A CODEBOOK OF K PROTOTYPES μ_k. EACH VECTOR ENCODED W/ MOST SIMILAR PROTOTYPE

ENCODE(x) = $\underset{k}{\operatorname{argmin}} \|x_i - \mu_k\|^2$ RECONSTRUCTION ERROR: $J(\mu, z|K, X) = \frac{1}{N}\sum\|x_i - \text{DECODE}(\text{ENCODE}(x_i))\|^2$

- TAKES LESS SPACE $O(NDC) \to O(\log_2 K)$

MAP ESTIMATION

BECAUSE MLES MAY OVERFIT. COLLAPSING VARIANCE PROBLEM WHEN ONLY 1 POINT TO CLUSTER.

Q BECOMES XPECTED COMPLETE DATA LL + LOG PRIOR. **E** STEP STAYS THE SAME. **M** STEP CHANGES (PP. 356 - 357 MURPHY)

- MUCH MORE RESILIENT IN HIGH-DIMENSIONALITY SCENARIOS TO PROBLEMS DUE TO SINGULAR MATRICES
- HAS HYPERPARAMS

EM FOR MIXTURE OF EXPERTS

$Q(\theta,\theta^{old}) = \sum_i \sum_k R_{ik} \log[\pi_{ik} \cdot N(y_i | w_k^T x_i, \sigma_k^2)]$ $\pi_{ik} = S(V^T x_i)_k$ $R_{ik} = \pi_{ik}^{old} N(y_i | x_i^T w_k (\sigma_k^{old})^2)$

- **E** SAME AS GMM BUT $\pi_k \to \pi_{ik}$
- **M** MAXIMIZES $Q(\theta,\theta^{old})$ WRT W_k, σ_k^2, V
  - $W_k = (X^T R_k X)^{-1} X R_k y$ IS ANALOGOUS TO WEIGHTED LEAST SQUARES
  - $\sigma^2 = \dfrac{\sum^{N} R_{ik}(y_i - w^T x)^2}{\sum R_{ik}}$
  - $\ell(V) = \sum_i \sum_k R_{ik} \log \pi_{ik}$ IS LOGLIKELIHOOD FOR MULTINOMIAL LOGISTIC BUT WITH SOFT LABELS $\to$ V FITTING A LOG. REGR. MODEL

EM FOR DGM W/ HIDDEN VARS

FOR ARBITRARY GENERIC DGM. $E[\log p(D|\theta)] = \sum_t \sum_c \sum_k \bar{N}_{tck} \log \theta_{tck}$ $\bar{N}_{tck}$ = EMPIRICAL COUNTS $= \sum_i p(x_{it}=k, x_{i,pa(t)}=c | D)$

- FAMILY MARGINAL: $p(x_{it}, x_{i\,pa(t)} | D_i, \theta)$
- **E** PRODUCES $\bar{N}_{tjk}$ XPECTED SUFF STATS
- **M** $\hat{\theta}_{tck} = \dfrac{\bar{N}_{tck}}{\sum_{k'} \bar{N}_{tjk}}$

EM FOR STUDENT DISTRIBUTION

- GAUSSIAN MM ARE SENSITIVE TO OUTLIERS, STUDENT'S T IS MORE ROBUST BUT NO CLOSED FORM MLE! SO WE MUST OPTIMIZE ITERATIVELY $\to$ EM
- INTRODUCE 'ARTIFICIAL' HIDDEN VAR, WRITE STUDENT'S AS INFINITE MIXTURE OF GAUSSIANS W/ DIFF Σ
  - ↳ GAUSSIAN SCALE MIXTURE $T(x_i|\mu,\Sigma,v) = \int N(x_i|\mu, \Sigma/z_i) \cdot Ga(z_i | \frac{v}{2} \frac{v}{2}) dz_i$ TREAT z_i AS MISSING DATA
- $\ell_c(\theta) = L_N(\mu,\Sigma) + L_G(v)$ <PP. 359-360 MURPHY FOR MATH>
  - V KNOWN $\to$ IGNORE L_G, ONLY DO $E[z_i]$ WRT OLD PARAMS
  - V UNKNOWN $\to$ VERY COMPLEX EXPRESSION REQUIRING GRADIENT-BASED OPTIMIZATION (NO CLOSED-FORM UPDATE) GENERALIZED EM
- MIXTURE OF STUDENTS IS A THING, YO!

EM FOR PROBIT

$p(y_i=1|z_i) = I(z_i > 0)$, $z_i \sim N(w^T x, 1)$ IS LATENT. $\ell(z, w | V_0) = \log p(y|z) + \log N(z|Xw, 1) + \log N(w|0, V_0)$

- **E** POSTERIOR $p(z_i | y_i, x_i, w) = \begin{cases} N(z_i | w^T x_i, 1) I(z_i > 0) & y_i = 1 \\ N(z_i | w^T x_i, 1)(z_i < 0) & y_i = 0 \end{cases}$ TRUNCATED GAUSSIAN. W ONLY LIN. DEP ON Z,

$E[z_i | w, x_i] = \begin{cases} \mu_i + \dfrac{\phi(\mu)}{\Phi(\mu)} & y=1 \\ \mu_i - \dfrac{\phi(\mu)}{\Phi(\mu)} & y=0 \end{cases}$

$\mu = w^T x_i$

- **M** ESTIMATES W WITH RIDGE REGRESSION $\mu = E[z]$ IS PREDICTION $\hat{w} = (V_0^{-1} + X^T X)^{-1} X^T \mu$
- SLOWER THAN DIRECT GRADIENT B/C POSTERIOR ENTROPY IS HIGH. USE STRONGER REGULARIZER TO CONSTRAIN Z VALUES $\to$ SPEEDUP CONVERGENCE

E-M THEORY 'XPECTED COMPLETE DATA LL IS LOWER BOUND'

- $Q(z_i)$ ARBITRARY DISTRIBUTION OVER HIDDEN VARIABLES
- OBSERVED DATA LL LOWER BOUND $Q(\theta, q) = \sum_i E_{q_i}[\log p(x_i, z_i | \theta)] + H(q_i)$ — $H(q_i)$: ENTROPY
  → PICK q W/ TIGHTER LOWER BOUND
- BUT ALSO $-KL(q_i(z_i) || p(z_i | x_i, \theta)) + \log p(x_i | \theta)$
- $Q(\theta, q^t) = \sum E_{q^t}[\log p(x_i, z_i | \theta)] + H(q_i^t)$
- M STEP: $\theta^{t+1} = \text{ARGMAX} \sum E_{q^t}[\log p(x_i, z_i | \theta)]$
- GUARANTEE IT MONOTONICALLY INCREASE OBS DATA LL UNTIL LOCAL MAX
- K-L $= 0$ ~~BECAUSE~~ FOR SOME REASON $\rightarrow L(\theta^t, q_i) = \log p(x_i | \theta)$
- $Q(\theta^t, \theta^t) = \sum_i \log p(x_i | \theta^t) = \ell(\theta^t)$ *
- LOWER BOUND TOUCHES FCN AT EVERY STEP. MAXIMIZING LB 'PUSHES UP' FCN → M STEP GUARANTEES TO INCREASE DATA LIKELIHOOD

– EM MONOTONICALLY INCREASES OBSERVED DATA LL

$$\underbrace{\ell(\theta^{t+1})}_{I} \geq \underbrace{Q(\theta^{t+1}, \theta^t)}_{II} \geq \underbrace{Q(\theta^t, \theta^t)}_{III} = \underbrace{\ell(\theta^t)}_{IV}$$

: II IS LB ON I; $II = \max_\theta Q(\theta, \theta^t) \geq III$; *

- IS SAME AS VARIATIONAL APPROACH, BUT Q HERE IS TRACTABLE

ONLINE EM

FOR STREAMING DATASETS

– BATCH EM: WE COMPUTE VECTOR OF SUFFICIENT STATISTICS (S), EXPECTATIONS (ϕ), AND THEIR SUMS (μ)

– INCREMENTAL EM: WE KEEP TRACK OF SUFF. STATS AND THEIR SUMS AFTER EACH DATAPOINT. S_i ALL COMPUTED INITIALLY AND μ INITIALIZED AS THE SUM THEN WE UPDATE S_{new}

– STEPWISE EM: WHENEVER S_i IS UPDATE, MOVE μ TOWARDS IT

- SPEED = STEPWISE > INCREMENTAL >> BATCH
- ACCURACY = STEPWISE => BATCH > INCREMENTAL

OTHER EM VARIANTS

– ANNEALED EM: DETERMINISTIC ANNEALING TO INCREASE CHANCE OF GLOBAL MAXIMUM. TEMPERATURE ACTS ON POSTERIOR.

– VARIATIONAL EM: WHEN CANNOT EXACT INFERENCE FOR E STEP. APPROXIMATION STILL ENSURES LOWER BOUND TO LIKELIHOOD

– MONTECARLO EM: " " " " " " " SAMPLE FROM POSTERIOR $z_i^s \sim p(z_i | x_i, \theta^t)$ AND THEN COMPUTE SUFFICIENT STATS
MCMC FOR SAMPLING OR FOR FULL INFERENCE

– GENERALIZED EM: E STEP OK BUT M STEP NOT OK, PARTIAL M-STEP

– ECM(E): EXPECTATION - CONDITIONAL MAXIMIZATION. IF M PARAMS ARE DEPENDENT → OPTIMIZE THEM SEQUENTIALLY

– OVER - RELAXED EM: $\theta^{t+1} = \theta^t + \eta(M(\theta^t) - \theta^t)$. CAN LEAD TO FASTER CONVERGENCE, W/ LOTS OF MISSING DATA

MODEL SELECTION FOR EM

HOW TO PICK NUMBER OF LATENT VARIABLES / NUMBER OF CLUSTERS

— PROBABILISTIC MODELS

OPTIMAL $K^* = \arg\max_K p(D|K)$. LARGEST MARGINAL LIKELIHOOD

- TOUGH FOR LVM → BIC APPROXIMATION, X-VALIDATED LIKELIHOOD,
- LARGE MODEL SPACE → BAYES RAZOR, STOCHASTIC SAMPLING IE MCMC

— NON-PROBABILISTIC MODELS

NO LIKELIHOOD, IE NUMBER OF K IN K-MEANS → RELY ON RECONSTRUCTION ERROR $E(D,K) = \frac{1}{|D|}\sum \|x_i - \hat{x}_i\|^2$

- ERROR ON TEST SET DECREASES ALWAYS BECAUSE MORE MODELS → LARGER CHANCE OF CLOSE PROTOTYPE
- TRY TO IDENTIFY KNEE / KINK ON TRAINING DATA VS K PLOT, BECAUSE $< K^*$ R.ERROR GREATLY DECREASES, $> K^*$ NOT SO MUCH

FITTING MODELS W/ MISSING DATA

WE WANT JOINT DENSITY MODEL BY MLE BUT WE HAVE MISSING DATA $O_{ij} = 1$ IF THERE, 0 IF MISSING

$X_V = \{x_{ij} : O = 1\}$ $X_H = \{x_{ij}, O = 0\}$ $\xrightarrow{\text{WE WANT}}$ $\hat{\theta} = \arg\max p(X_V|\theta, O)$

MAR ASSUMPTIONS: $p(x_V|\theta,O) = \prod^N p(x_{iV}|\theta)$

$\log p(X_V|\theta) = \sum \log p(x_{iV}|\theta)$

$\log p(x_{iV}|\theta) = \sum p(x_{iV}, x_{iH}|\theta)$ → USUAL PROBLEM WITH NO CAN PUSH LOG INSIDE SUM → EM

(PP 373-374 MURPHY)

— FOR A MVN

- MLE FOR ~~ON~~ FULLY OBSERVED ROWS • E-STEP: $Q(\theta, \theta^{t-1})$ • M STEP: $\mu^t = \frac{1}{N}\sum [x_t]$, $\Sigma^t = \frac{1}{N}\sum_i E[x_i x_i^T] - \mu^t(\mu^t)^T$
- NOT EQUIVALENT TO JUST REPLACE VARS WITH EXPECTATIONS AND DO MLE. → IGNORES POSTERIOR VARIANCE
- WE COMPUTE EXPECTATIONS OF SUFF. STATS AND USE THOSE FOR MLE
- CAN ALSO USE FOR MAP WITH ESS INTO MAP EQUATIONS