LATENT LINEAR MODELS

- LIMITATION OF MIXTURE MODELS → CATEGORICAL HIDDEN VARIABLES
- IN LLM HIDDEN VARS ARE REAL VALUED $z_i \in \mathbb{R}^L$

FACTOR ANALYSIS

- PRIOR $p(z_i) = \mathcal{N}(z_i | \mu_0, \Sigma_0)$ GAUSSIAN
- LIKELIHOOD ALSO GAUSSIAN $p(x_i|z_i) = \mathcal{N}(Wz_i + \mu, \Psi)$
  - W $D \times L$ FACTOR LOADING MATRIX
  - Ψ $D \times D$ COVARIANCE MATRIX, DIAGONAL BECAUSE WE WANT TO EXPLAIN CORRELATION AND NOT BAKE IT IN
  - IF $\Psi = \sigma^2 I$ → PPCA

- CAN BE SEEN AS WAY TO SPECIFY JOINT DENSITY WITH LOW NUMBER OF PARAMS; $\mathrm{cov}[x|\theta] = WW^T + \Psi \approx \mathrm{cov}[x]$
  $O(LD)$ PARAMS VS $O(D^2)$ FOR FULL GAUSSIAN AND $O(D)$ FOR DIAGONAL COVARIANCE

- HOPEFULLY LATENT FACTOR z REVEAL INTERESTING STUFF → INFERENCE. POSTERIOR OVER zs. $p(z_i|x_i,\theta) = \mathcal{N}(z_i|m_i, \Sigma)$
  LATENT FACTORS $m_i = \Sigma_i (W^T \Psi^{-1}(x_i - \mu) + \Sigma_0^{-1}\mu_0)$
  - LINEAR COMBINATIONS | PROJECTIONS OF ORIG. FEATURES IN LOWER DIMENSIONAL SPACE

- LATENT FACTORS ARE UNIDENTIFIABLE; CANNOT UNIQUELY IDENTIFY W TOO MANY DOF. $D + LD + L(L-1)/2 < D(D+1)/2$
  - BUT STILL ROTATIONAL AMBIGUITY.
  $L_{max} = LD + 0.5(1 - \sqrt{1+8D}$

  FIXES:
  - W FORCED ORTHONORMAL: COLUMNS ORDERED BY DECREASING VARIANCE → PCA. NOT NEC MORE INTERPRETABLE BUT UNIQUE
  - W FORCED LOWER TRIANGULAR: 1ST VISIBLE FEATURE → 1ST LATENT FACTOR, 2ND V.F → 1ST, 2ND LF, ...
    FIRST L VISIBLES AFFECT L.F. → FOUNDER VARIABLES. CAREFUL.
  - SPARSITY PROMOTING PRIORS ON W: ℓ_1 REGULARIZATION → SPARSE FACTOR ANALYSIS
  - INFORMATIVE ROTATION MATRIX: FIND R THAT MODIFY W TO BE SPARSE → VARIMAX
  - NON-GAUSSIAN PRIORS FOR LF: CAN ACHIEVE IDENTIFIABILITY → ICA

MIXTURES OF FA

APPROXIMATES DATA CURVED MANIFOLD W/ PIECEWISE LINEAR MANIFOLDS. LOW-RANK APPROX OF MIXTURE OF GAUSSIANS. $O(KLD)$ VS $O(KD^2)$ PARAMS
REDUCES OVERFITTING.

- LATENT INDICATOR TO PICK THE MIXTURE COMPONENT TO USE. W_k $p(x_i|z, q_i = k, \theta) = \mathcal{N}(x_i | \mu_k + W_k z_i, \Psi)$
  $p(z_i|\theta) = \mathcal{N}(z_i|0, I)$ $p(q_i|\theta) = \mathrm{Cat}(q_i|\pi)$

EM FOR FA

- KINDA STRAIGHTFORWARD E-STEP: $R_{ic} = p(q_i = c | x_i, \theta) = \pi \mathcal{N}(x_i | \mu_c, W_c W_c^T + \Psi)$ M STEP → $\hat{W}_c, \hat{\Psi}, \pi_c$
- MISSING DATA: USE MAP OR BAYESIAN INFERENCE

PRINCIPAL COMPONENT ANALYSIS

[ORTHONORMALITY = $W^T W = 1$

- FA WHERE $\Psi = \sigma^2 I$ AND W ORTHONORMAL
  - $\sigma^2 = 0$ CLASSICAL PCA
  - $\sigma^2 > 0$ PROBABILISTIC/SENSIBLE PCA

PCA

DEFINITION: WE WANT ORTHOGONAL SET OF L LINEAR BASIS VECTORS $w_j \in \mathbb{R}^D$ AND SCORES $z_i \in \mathbb{R}^L$ SO TO MINIMIZE RECONSTRUCTION ERROR • $J(W,Z) = \frac{1}{N}\sum \|x_i - \hat{x}_i\|^2$ • WHERE $\hat{x}_i = W z_i$ IS CONSTRAINED ORTHONORMAL

$\longrightarrow J(W,Z) = \|X - WZ^T\|_F^2$; $Z = N \times L$ WITH z_i IN ROWS ; $\|A\|_F = \sqrt{\sum\sum a_{ij}^2} = \sqrt{TR(A^T A)}$ FROBENIUS NORM

SOLUTION: $\hat{W}_L = V_L$, V_L CONTAINS L EIGENVECTORS WITH LARGEST EIGENVALUES OF EMPIRICAL COVARIANCE MATRIX $\hat{\Sigma} = \frac{1}{N}\sum_i^N x_i x_i$

OPTIMAL LOW-DIMENSIONAL DATA ENCODING IS $\hat{z}_i = W^T x_i$, ORTHOGONAL PROJECTION OF DATA ONTO EIGENVECTORS' COLSPACE

- PRINCIPAL COMPONENTS: PRINCIPAL DIRECTIONS ALONG WHICH DATA SHOWS MAXIMAL VARIANCE $\longrightarrow$ STANDARDIZE DATA TO AVOID ISSUES DUE TO SCALE, ETC
  - PRINCIPAL COMPONENTS IS EIGENSTUFF (FACES/DIGITS), LIN. COMBS OF ORIGINAL DIMS

  $\longrightarrow$ USE CORRELATION MATRIX VS COVARIANCE MATRIX

  $\longrightarrow$ ALSO CENTER THE DATA

- PROOF • DERIVATIVE OF J WRT $W \longrightarrow$ OPTIMAL WEIGHTS ARE $\perp$ PROJ ON PRINCIPAL DIRECTIONS $\longrightarrow$ ARG MIN $J(w_1)$ = ARGMAX VAR $[z_1]$

  DERIVATIVE OF PROJECTION OF VARIANCE $= 0 \longrightarrow \hat{\Sigma} w_1 = \lambda_1 w_1 \longrightarrow$ EIGENVECTOR OF COV. MATRIX , VARIANCE $= w_1^T \hat{\Sigma} w_1 = \lambda_1$

  • ALT: USING SVD $\rightarrow$ OK FOR NONSQUARE MATRICES , SINGULAR VECTORS = EIGENVECTORS OF $\hat{\Sigma}$

[SVD $X = USV^T$, $X^T X = VS^T U^T U S V^T = VS^2 V^T$, $(X^T X) V = VD$ $\longrightarrow$ RIGHT SINGULAR VECTORS OF X ARE = TO λs OF $\hat{\Sigma}$

$(X^T X) U = UD$

PPCA

LL MAXIMA: $\hat{W} = V(\Lambda - \sigma^2 I)^{1/2} R$. R ORTHOGONAL , V COL EIGENVECTORS, Λ DIAGONAL EIGENVALUES $R \rightarrow I$ $\quad \hat{\sigma}^2 = \frac{1}{D-L}\sum_{L+1}^{D}\lambda_i$ AVG VARIANCE DISCARDED DIMENSIONS

- $\sigma^2 \rightarrow 0 \rightarrow$ PCA
- $\sigma^2 > 0 \rightarrow$ POSTERIOR MEAN $\hat{F} = \hat{W}^T\hat{W} + \sigma^2 I$ IS NOT ORTHOGONAL PROJECTION $\longrightarrow$ RECONSTRUCTION ERROR CLOSER TO DATA MEAN
- CAN BAYESIAN METHODS • OPTIMIZE PARAMS, INTEGRATE LATENTS • USEFUL WHEN $|D| \gg$

EM FOR PCA

HA IL SUO PERCHE' • E STEP: $\tilde{Z} = (W^T W)^{-1} W^T X$ ANALOGY: E STEPS MOVES POINTS TO ORTHOGONAL , M STEP 'ROTATES ROD/LINE'

• M STEP: $W = \tilde{X}\tilde{Z}^T(\tilde{Z}\tilde{Z}^T)^{-1} \longrightarrow$ LIKE LINEAR REGRESSION WITH EXPECTED LATENT VALUES INSTEAD OF OBSERVED INPUTS

- CONVERGES TO W IS SAME LINEAR SUBSPACE OF L EIGENVECTORS BUT HAS TO BE ORTHOGONALIZED AND ORDERED
- EM FASTER ESP $N, D \gg L$. DOMINATED BY E STEP , $O(TLND)$. EIGENVECTOR METHODS ARE $O(\min(ND^2, DN^2))$. LANCZOS METHOD COMPARABLE TO EM
- EM IS ONLINE - FRIENDLY
- EM CAN HANDLE MISSING DATA
- EM CAN BE USED FOR MIXTURE MODELS
- EM CAN BE UPGRADED TO VARIATIONAL METHODS FOR GREAT JUSTICE

FA/PPCA MODEL SELECTION

HOW TO PICK OPTIMAL NUMBER OF L

- SIMPLE: BIC, VARIATIONAL LOWER BOUNDS, CROSS VALIDATION → BUT EXPENSIVE.
- USUALLY EXHAUSTIVE SEARCH OVER VALUES OF L
- AUTOMATIC RELEVANCY DETERMINATION (ARD) + EM TO PRUNE OUT IRRELEVANT WEIGHTS
  - STILL SEARCH OVER K → BIRTH / DEATH MOVES, STOCHASTIC SAMPLING OF MODEL SPACES, GIBBS SAMPLING + NONPARAM PRIORS

PCA MODEL SELECTION

NON PROBABILISTIC → ABOVE METHODS NOT OK

- APPROX 4 RECONSTRUCTION ERROR: $E(D,L) = \frac{1}{|D|}\sum \|x_i - \hat{x}_i\|^2$, $\hat{x}_i = Wz_i + \mu$, $z_i = W^T(x_i - \mu)$
- RESIDUAL ERROR WITH L TERMS:

$$E(D_{train}, L) = \sum_{j=L+1}^{D} \lambda_j$$; SUM OF DISCARDED EIGENVALUES

→ CAN PLOT RETAINED EIGENVALUES / SCREE PLOT

FRACTION OF EXPLAINED VARIANCE: $\frac{\sum_{1}^{L} \lambda_j}{\sum_{1}^{L_{max}} \lambda_j'}$

- OBS: ERROR KEEPS GOING DOWN ON TEST SET TOO → PCA IS NOT GENERATIVE MODEL BUT COMPRESSION TECH!

  MORE DIMENSIONS → MORE ACCURATE APPROXIMATION NO MATTER WHAT

SOLUTION: PROFILE LIKELIHOOD $\ell(L) = \sum^{L} \log N(\lambda_k | \mu_1(L), \sigma^2(L)) + \sum^{K} \log N(\lambda_k | \mu_2(L), \sigma^2(L))$

~~PARTITION~~ MODEL OF SIZE K, ERROR λ_k, ERROR SO THAT $\lambda_1 \geq \dots \geq \lambda_{L_{MAX}}$. THRESHOLD L. PARTITION $k < L$, $k > L$

CHANGE POINT MODEL $\underset{\leq L}{\lambda_k} \sim N(\mu_1, \sigma^2)$, $\underset{> L}{\lambda_k} \sim N(\mu_2, \sigma^2)$; σ SAME. FIT FOR $L = 1 : L_{max}$, MLE, POOLED VARIANCE ESTIMATE

$L^* = \operatorname{ARGMAX} \ell(L)$

CATEGORICAL PCA

OBSERVED DATA IS CATEGORICAL. EACH y GENERATED FROM LATENT VAR $z_i \in \mathbb{R}^L$ W/ GAUSSIAN PRIOR, SOFTMAXED.

- $P(z_i) = N(0, I)$
- $P(y_i | z_i, \theta) = \prod^{R} \text{CAT}(y_{ir} | S(W_r^T z_i + w_{0r}))$
- FITTED WITH EM
- USED TO VISUALIZE HIGH-DIM CATEGORICAL DATA

PCA FOR PAIRED / MULTI-VIEW DATA

WHEN RELATED DATASETS WE WANT TO COMBINE IN LOW-D FOR IE. PREDICTION. **DATA FUSION**

EASY TO DO PAIRS → ~~DATA~~ SETS. FITTED USING EM OR GIBBS SAMPLING. DISCRETE/COUNT DATA → EXP. FAMILY RESPONSE INSTEAD OF GAUSS
→ APPROXIMATE INFERENCE (MCMC)

• SUPERVISED PCA / LATENT FACTOR REGRESSION

LIKE PCA BUT y_i IS TAKEN INTO ACCOUNT WHEN LEARNING LOW DIM. EMBEDDING

$$\begin{cases} p(z_i) = N(0, I_0) \\ p(y_i|z_i) = N(w_y^T z_i + \mu_y, \sigma_y^2) \\ p(x_i|z_i) = N(W_x z_i + \mu_x, \sigma_x^2 I_D) \end{cases}$$

• JOINTLY GAUSSIAN $y_i|x_i \sim N(x_i^T w, \sigma_y^2 + w_y^T C w_y)$, $w = \Psi^{-1} W_x C w_y$, $\Psi = \sigma_x^2 I_D$, $C^{-1} = I + W_x^T \Psi^{-1} W_x$

• DEPENDENCE OF PRIOR FOR W ON X ARISES FROM W IS DERIVED FROM JOINT MODEL OF X, Y

– DISCRETE VARS → GAUSS → EXP FAMILY. NO MORE CLOSED-FORM CONDITIONALS. BUT STILL **INFORMATION BOTTLENECK**:
FIND ENCODING DISTRIBUTION $p(z|x)$ S.T.O. MINIMIZE $I(x,z) - \beta I(x,y)$; $\beta > 0$

– **MULTILABEL**: y_i IS VECTOR OF RESPONSES, COLLAB. FILTERING

DISCRIMINATIVE SUPERVISED PCA

DIFFERENT WEIGHTS TO INPUTS x_i AND OUTPUTS y_i **WEIGHTED OBJECTIVE** $\ell(\theta) = \prod p(y_i|\eta_{iy})^{\alpha y} p(x_i|\eta_{ix})^{\alpha x}$ $\quad \eta_{im} = W_m z_i$

• IN CASE OF GAUSSIAN α_m CONTROLS NOISE VARIANCE • HARD TO ESTIMATE α_m BECAUSE IT CHANGES LIKELIHOOD NORMALIZATION

PARTIAL LEAST SQUARES

MORE 'DISCRIMINATIVE' SUPERVISED PCA. ALLOWS SOME INPUT COVARIANCE TO BE EXPLAINED BY THEIR OWN SUBSPACE z_i^x WHILE REST OF I/O COV IS z_i^s

$$\begin{cases} p(z_i) = N(z_i^s|0, I_{L_s}) N(z_i^x|0, I_{L_x}) \\ p(y_i|z_i) = N(W_y z_i^s + \mu_y, \sigma^2 I_{D_y}) \\ p(x_i|z_i) = N(W_x z_i^s + B_x z_i^x + \mu_x, \sigma^2 I_{D_x}) \end{cases}$$

• $p(v_i|\theta) = N(v_i|\mu, WW^T + \sigma^2 I)$ • L HAS TO BE PICKED 'LARGE ENOUGH'

• PROJECTS PREDICTED VARS y AND OBS x TO NEW SPACE

CANONICAL CORRELATION ANALYSIS (CCA)

LIKE SYMMETRIC UNSUPERVISED VERSION OF PARTIAL LEAST SQUARES. EACH OBS HAS ITS OWN PRIVATE SUBSPACE + A SHARED ONE. z_i^x, z_i^y, z_i^s

$$\begin{cases} p(z_i) = N(z_i^s|0, I_{L_s}) N(z_i^x|0, I_{L_x}) N(z_i^y|0, I_{L_y}) \\ p(x_i|z_i) = N(x_i|B_x z_i^x + W_x z_i^s + \mu_x, \sigma^2 I_{D_x}) \\ p(y_i|z_i) = N(y_i|B_y z_i^y + W_y z_i^s + \mu_y, \sigma^2 I_{D_y}) \end{cases}$$

• $p(v_i|\theta) = N(v_i|\mu, WW^T + \sigma^2 I_D)$
• CAN MLE VIA EM → EQUIV TO NON PROBABILISTIC
• CAN GENERALIZE TO $M > 2$ OBSERVED VARIABLES
• CAN CREATE MIXTURES OF CCA
• CAN SPARSIFY VIA ARD
• CAN GENERALIZE TO EXP FAMILY
• CAN BAYESIAN INFERENCE

• FINDS LIN. COMB OF x, y OF MAX CORRELATION

INDEPENDENT COMPONENT ANALYSIS - ICA

BLIND SIGNAL / SOURCE SEPARATION. RECONSTITUTE ORIGINAL SIGNAL(S) WHERE MANY LATENT SOURCES GET LINEARLY MIXED TOGETHER

- $x_t = Wz_t + \epsilon_t$; W IS DXL MIXING MATRIX, $\epsilon_t \sim N(0, \Psi)$. EACH TIMEPOINT IS INDEPENDENT OBSERVATION
- WE WANT TO INFER $p(z_t | x_t, \theta)$
- $L = D$ (SOURCES = SENSORS) → W IS SQUARE
- $|\Psi| = 0$ FOR SIMPLICITY

PRIOR: ANY NON-GAUSSIAN $p(z_t) = \prod_1^L p_j(z_{tj})$; CONSTRAINING VARIANCE TO 1. NO GAUSSIAN BECAUSE DOES NOT ALLOW UNIQUE RECOVERY OF SOURCES

- PCA LIKELIHOOD INVARIANT TO ORTHOGONAL TRANSFORMATION OF LIKELIHOOD.
- PCA RECOVERS BEST LINEAR SUBSPACE, NOT THE SIGNALS.
- IN SYMMETRIC GAUSSIAN POSTERIOR NO INFORMATION TO TELL US ANGLE WE NEED TO ROTATE
- PCA WHITENS DATA, SOLVES HALF THE PROBLEM
- ICA IDENTIFIES ROTATION
- ESTIMATE W AND p_j S
- SQUARE W → OK CLOSED FORM UNIQUE SIGNALS
- NONSQUARE W → NO UNIQUE SIGNALS BUT WE DO POSTERIOR $p(z_t | x_t, \hat{W})$

ESTIMATE: MLE

FOR NOISE-FREE SQUARE W_s, ALREADY WHITENED WITH PCA BECAUSE NOISEFREE, WHITENED AND CENTERED DATA.

- W MUST BE ORTHOGONAL BECAUSE $E[xx^T] = I$, AND $\text{cov}[x] = WW^T$ → $D(D-1)/2$ PARAMS AS OPPOSED TO D^2

- W GENERATIVE WEIGHTS
- V RECOGNITION WEIGHTS $= W^{-1}$

$$p_x = (Wz) = p_z(z_t) |\det(W^{-1})| = p_z(Vx_t) |\det V| \longrightarrow \text{LL: } \frac{1}{T} \log p(D|V) = \underbrace{\log|\det(V)|}_{\text{CONSTANT}} + \frac{1}{T} \sum^L \sum^T \log p_j(V_j^T x_t) \longrightarrow NLL(V) = \sum^L E[G_j(z_j)]$$

$z_j = V_j^T x$, $G_j(z) = -\log p_j(z)$

- MINIMIZE SO TO ROWS OF V ORTHOGONAL AND UNIT NORM → ORTHONORMAL
  - CAN GRADIENT DESCENT BUT SLOW
  - FASTICA
  - EM

~~FAST~~ ESTIMATE: FASTICA

APPROXIMATE NEWTON METHOD FOR FITTING ICA.

- ASSUME 1 LATENT FACTOR AND ALL SOURCE DISTRIBUTIONS KNOWN AND SAME → $G = -\log p(z)$
- $g(z) = \frac{d}{dz} G(z)$, GRADIENT, HESSIAN WRT W
- NEWTON UPDATE: $V^* = E[x g(V^T x)] - E[g'(V^T x)] V$
- IN PRACTICE WE REPLACE EXPECTATIONS WITH MC ESTIMATES FROM TRAINING SET
- MAXIMIZES NONGAUSSIANITY OF THE PROJECTION

→ PROJECT BACK ON CONSTRAINT SURFACE $V^{NEW} = \frac{V^*}{\|V^*\|}$

- CONVERGENCE: $|V^T V^{NEW}| \longrightarrow 1$

- NON-CONVEX OBJECTIVE ; XPLOITABLE TO LEARN MULTIPLE FEATURES, SEQUENTIALLY OR IN PARALLEL. (OK BECAUSE UNLIKE PCA 1ST FEATURE IS NOT MORE IMPORTANT THAN 2ND) $\perp$ TO V_j - PROJECT OUT

- IF $G(z)$ NOT KNOWN, WAT DO? SIGNAL PRIOR
- NOT GAUSSIAN - STH LEPTOKURTIC 'SUPER-GAUSSIAN' BECAUSE BRAINZ
- NOT CRITICAL XACT SHAPE: $G(z) = \sqrt{z}$, $\log \cosh(z)$
- LAPLACE $\log p(z) = -\sqrt{2}|z| - \log(\sqrt{2})$ MEAN 0 VAR 1 | NOT DIFF. LE
- LOGISTIC $\log p(z) = -2 \log \cosh\left(\frac{\pi}{2\sqrt{3}} z\right) - \log \frac{4\sqrt{3}}{\pi}$

ESTIMATE: EM

NO USE PARTICULAR FORM BUT FLEXY NONPARAMETRIC ESTIMATOR → GMM! (MIXTURE OF GAUSSIANS) CAN DO VIA $E[z_t, x_t, \theta]$ VIA SUMMING OVER K^L COMBINATIONS OF q^t PARAMS THEN ESTIMATE ALL SOURCES IN PARALLEL VIA FITTING A GMM TO $E[z_t]$ → MARGINALS $p_j(z_j)$ → ICA FOR W → RINSE REPEAT

ESTIMATE: OTHERS

COOL BUT EQUIVALENT TO MLE

- **MAX NONGAUSSIANITY!** NEGENTROPY $(z) = H(N(\mu, \sigma^2)) - H(z)$. AS GAUSSIAN IS MAXENTROPY, MAXIMIZING NEGENTROPY TAKES ME FAR.

$$J(V) = \sum \text{NEGENTROPY}(z_i) = \sum_i H(N(\mu_i, \sigma_i^2)) - H(z_i)$$ IF V ORTHOGONAL AND WHITENED DATA COV IS I

$$J(V) = \sum E[\log p(z_i)] + \text{const}$$ → EQUIV TO LL

- **MINIMIZING MUTUAL INFORMATION:** MULTI-INFORMATION $I(z) = KL(p(z) \| \prod p(z_i)) = \sum H(z_i) - H(z) \longrightarrow I(z) = \sum_i H(z_i)$
→ SAME AS MIN. NEGENTROPY

- **MAXIMIZING MUTUAL INFORMATION!** INFOMAX PRINCIPLE: MAX INFORMATION FLOW THROUGH A SYSTEM
MAXIMIZE MUTUAL INFO BETWEEN Y INTERNAL REP AND X INPUT. $I(x, y) = H(y) - H(y|x)$
→ SAME AS ML

LOCALLY LINEAR EMBEDDING

- IDEA: MINIMIZE RECONSTRUCTION ERROR BY MAKING SMALL PATCHES WHERE THERE IS NONLINEARITY IN DATA.
- DECIDE ON POINT NEIGHBORHOOD, COMPUTE WEIGHTS FOR POINTS AS LINCOMBS OF THEIR NEIGHBORS
- THEN FIND LOW-DIM EMBEDDING OF POINTS USING LINCOMBS. • UNROLLS STUFF
- EVECS OF QUADRATIC FORM MATRIX

ISOMAP

VARIANT OF MULTIDIMENSIONAL SCALING. DIMENSIONALITY REDUCTION TECHNIQUE ATTEMPTING TO PRESERVE DISTANCES IN LOWDIM SPACES. ASSUMES EUCLIDEAN | FLAT SPACE.

ISOMAP: CALCULATE DISTANCES. FIND NEIGHBORS (IE kNN, ...). CONSTRUCT NEIGHBOURHOOD GRAPH. ESTIMATE GEODESICS BY FINDING SHORTEST PATHS (DIJKSTRA, ...), APPLY MDS.
THIS INSERTS NONLINEAR MANIFOLD STRUCTURE INTO THE PROBLEM.

- T-SNE