Sparse Linear Models

SELECTING SETS OF VARS WITH A MODEL-BASED APPROACH. IF LINEAR MODEL $p(y|x)=p(y|f(w^Tx))$ WE DO IT BY PROMOTING w TO BE SPARSE $\rightarrow$ LOTS OF ZEROS. **USEFUL FOR:** SMALL N, LARGE D; KERNEL MACHINES; SPARSE WAVELET REPRESENTATION FOR SIGNALS.

Bayesian Variable Selection

$$p(\gamma|D)=\frac{e^{-f(\gamma)}}{\sum_{\gamma'} e^{-f(\gamma')}}$$ POSTERIOR OVER MODELS. $f(\gamma)=-[\log p(D|\gamma)+\log p(\gamma)]$

MAP: $\hat{\gamma}=\arg\max p(\gamma|D)=\arg\min f(\gamma)$

MEDIAN MODEL: $\hat{\gamma}=\{j : \underbrace{p(\gamma_j=1|D)}_{\text{MARGINAL INCLUSION PROBS}}>0.5\}$

SPIKE AND SLAB MODEL $\log p(\gamma|D)\approx \log p(y|X,\hat{w}_\gamma,\hat{\sigma}^2)-\underbrace{\frac{\|\gamma\|_0}{2}\log N}_{\text{BIC APPROX PENALTY}}-\underbrace{\lambda\|\gamma\|_0}_{\text{PRIOR PENALTY}}+\text{const}$

- $\|A\|_0$ = ℓ_0 NORM, NUM OF NON-NULL ELEMENTS

BER/GAUSS MODEL

ℓ_0-REGULARIZATION: $f(w)=\|y-Xw\|_2^2+\lambda\|w\|_0$ $\rightarrow$ CONVERTS DISCRETE OPTIMIZATION PROBLEM OVER γ INTO A CONTINUOUS ONE OVER $w\in\mathbb{R}^D$

MAP STILL NONSMOOTH AND HARD 2 OPTIMIZE

Optimization Algos

SPACE OF MODELS IS 2^D $\rightarrow$ WE NEED HEURISTICS. FIT THE MODEL AT EACH POINT, COMPUTING $\underset{\text{ARGMAX}}{p(D|\gamma)}$ OR $\int p(D|w)p(w)dw$ $\rightarrow$ WRAPPER METHOD

FOR EFFICIENCY WE WANT TO BE ABLE TO COMPUTE γ' GIVEN γ $\rightarrow$ UPDATING SUFF. STATS. $\rightarrow$ OK IF γ' DIFFS BY 1 BIT AND $f(\gamma)$ DEPENDS ON DATA ONLY VIA X_γ

Greedy Search

- SINGLE BEST REPLACEMENT $\rightarrow$ GREEDY HILL CLIMBING BY MOVING IN NEIGHBORHOOD OF γ (1 BIT FLIPS). LATTICE SEARCH.
- ORTHOGONAL LEAST SQUARES $\rightarrow$ START WITH $\lambda=0$ AND EMPTY SET. ADD BEST FEATURE AT EACH STEP. $j^*=\arg\min \min\|y-(X_t\cup j)w\|^2$ | D-Dt LS PER STEP
- ORTHOGONAL MATCHING PURSUITS $\rightarrow$ OLS IS XPENSIVE. FREEZE WEIGHT AT CUR VAL THEN PICK NEXT $j^*=\arg\min\min\|y-Xw_t-\beta x_{:,j}\|^2$ ONLY 1 LS PER STEP BUT NOT AS ACCURATE
- MATCHING PURSUITS: $\rightarrow$ VERY GREEDY, ONLY ADDS COLUMN MOST CORRELATED W/ CURRENT RESIDUAL | LS BOOSTING
- BACKWARDS SELECTION: $\rightarrow$ STARTS WITH ALL VARS, REMOVES WORST ONE AT EACH STEP
- FoBa: SBR WITH OMP STEP FOR NEXT SELECTION
- BAYESIAN MATCHING PURSUITS: OMP BUT USES MARGINAL LIKELIHOOD CRITERION VS LEAST SQUARES. ALSO BEAM SEARCH.

Stochastic Search

APPROXIMATES POSTERIOR, MCMC. BUT INEFFICIENT. USE $p(\gamma|D)=\frac{e^{-f(\gamma)}}{\sum_{\gamma's} e^{-f(\gamma)}}$

EM/Variational

EM ON SPIKE/SLAB, WITH NARROW GAUSSIAN INSTEAD OF DELTA. LOCAL MINIMA ISSUE.

EM ON BER/GAUSS $\rightarrow$ ALL VAR XPLAINED AWAY. USE MEAN FIELD APPROXIMATION

L_1 REGULARIZATION

PUTTING A 0 MEAN LAPLACE PRIOR ON THE PARAMS AND PERFORMING MAP. CAN BE COMBINED WITH ANY CONVEX/NONCONVEX NLL. $\begin{cases} \text{LIN REGR} \\ \text{LOGISTIC REGR} \\ \text{GLM} \\ \dots \end{cases}$

IT 'ENCOURAGES' $w_j = 0$

PRIOR: $p(w|\lambda) = \prod \mathrm{Lap}(w_j|0, 1/\lambda) \propto \prod_j e^{-\lambda|w_j|}$ NLL: $f(w) = -\lg p(D|w) - \lg p(w|\lambda) = NLL + \lambda\|w\|_1$ $\|w\|_1 = \sum|w_j|$ l_1 NORM

- CONVEX APPROX OF l_0 OBJECTIVE $\mathrm{ARGMIN}\ NLL(w) + \lambda\|w\|_0$
- IN LINEAR REGRESSION: $f(w) = \sum^N -\frac{1}{2\sigma^2}\left(y_i - (w_0 + w^T x)\right)^2 + \lambda\|w\|_1 = RSS(w) + \lambda'\|w\|_1,\ \lambda' = 2\sigma^2\lambda$
- YIELDS SPARSE SOLUTIONS

l_1 $\min RSS(w) + \lambda\|w\|_1 \longrightarrow \min RSS(w)\ \text{ST}\ \|w\|_1 < B$ LASSO EQUATION

NONSMOOTH — SMOOTH BUT CONSTRAINED

- MORE PROBABLE THAT CORNERS INTERSECT ELLIPSE $\longrightarrow$ CORNERS ARE SPARSE SOLUTIONS
- SPARSE SOLUTIONS COST LESS

l_2 $\min (RSS)(w)\ \text{ST}\ \|w\|_2^2 < B$

- NO CORNERS, NO PREFERENCE FOR SPARSITY
- COST, l_2 NORM, FOR SPARSE IS SAME AS FOR DENSE

$$\begin{cases} (c_j + \lambda)/a_j & c_j < -\lambda \\ 0 & -\lambda < c_j < \lambda \\ (c_j - \lambda)/a_j & c_j > \lambda \end{cases}$$

- SETS TO 0 OR SHRINKS
- SOFT THRESHOLDING

- $\lambda_{MAX} = \max_j |\nabla_j NLL(0)|$, IF $\lambda > \lambda_{MAX} \to \hat{w} = 0$

COMPARISON

- MLE $\hat{w}_k^{OLS} = x^T y$
- LASSO $w_k^{LASSO} = \mathrm{SIGN}(\hat{w}_k^{OLS})\left(|\hat{w}_k^{OLS}| - \frac{\lambda}{2}\right)$
- RIDGE $\hat{w}_k^{RIDGE} = \frac{\hat{w}_k^{OLS}}{1+\lambda}$
- SUBSET SELECTION $\begin{cases} w_k^{OLS} & \text{IF RANK} < K \\ 0 & \text{ELSE} \end{cases}$

REGULARIZATION PATH

PLOTS $\hat{w}_j(\lambda)$ VS λ FOR EACH FEATURE j. SPARSE SOLUTIONS FOR BOUND l_1 BETWEEN 0 AND $B_{MAX} = \|w_{OLS}\|_1$. INBETWEEN EACH LINEAR COEFF INCREASES OR DECREASES LINEARLY. PIECEWISE LINEAR. LARS ALGORITHM; SAME COST AS LEAST SQUARES FIT $O(\min(ND^2, DN^2))$

- LIMITATION IF $D > N$ LASSO ONLY GETS TO N BEFORE GOING TO COMPLETE SET FOR OPTIMAL OLS $\longrightarrow$ USE ELASTIC NET

MODEL SELECTION

RECOVER TRUE SPARSITY PATTERN. PICK λ. USE CV. (OR HEURISTICS) l_1 WEAK IF PERTURBED DATA. DO BOOTSTRAP AND COMPARE HOW EACH VAR IS PICKED IN RUNS AND DERIVE SPARSE ESTIMATOR $\longrightarrow$ BOLASSO PICK IF VAR RETURNED AT LEAST 90% OF TIMES FOR GIVEN λ

OPTIMAL λ FOR PREDICTION $\not\longrightarrow$ MODEL SELECTION CONSISTENCY

BAYESIAN INFERENCE: ONLY POSTERIOR MODE IS SPARSE. POSTERIOR MEAN BETTER IF WE MINIMIZE SQUARED ERROR. PMEAN + SPIKE AND SLAB > PMODE + LAP PRIOR

L_1 REGULARIZATION ALGORITHMS

- COORDINATE DESCENT: OPTIMIZE VARS 1-BY-1 $w_j = \mathrm{ARGMIN}\ f(w + z e_j) - f(w)$. PICK RANDOM OR ONE WHERE STEEPEST GRAD. OK IF ANALYTICAL SOLUTION FOR EACH 1D PROBLEM. $\longrightarrow$ SHOOTING ALGO
- LARS: ACTIVE SET $\to$ MORE THAN 1 VAR AT TIME. HOMOTOPY $\to$ STARTS FROM λ_{MAX} DOWN TO DESIRED λ_K.
  LEAST ANGLE REGRESSION & SHRINKAGE. STARTS AT λ WHERE ~~λ_{MAX}~~ THE SINGLE MOST CORRELATED VAR W/ y IS PICKED. λ DECREASES UNTIL SECOND VAR WITH SAME CORRELATION WITH RESIDUAL FROM FIRST IS FOUND. NEW λ FOUND ANALYTICALLY, GEOMETRICALLY. RINSE REPEAT.
  ANALYTICAL SOLUTIONS ONLY FOR LINREG, NOT OTHERS GLM.
- PROXIMAL & GRADIENT PROJECTION: GOOD FOR VERY LARGE PROBLEMS, CAN USE FOR OTHER TYPES BEYOND l_1 REG.
  CONVEX OBJECTIVE $f(\theta) = L(\theta) + R(\theta) \longrightarrow f(\theta) = R(\theta) + \frac{1}{2}\|\theta - y\|_2^2$. WE CALL THE MINIMIZER OF THIS PROXIMAL OPERATOR (DEPENDS ON $R(\theta)$)
  IN MOST CASES $\mathrm{PROX}_R(\theta)$ IS PROJECTION ON SET $C = \{\theta : \|\theta\|_1 \le B\} = \mathrm{PROJ}(\theta) \to$ COMPUTED IN $O(D)$ TIME. USED IN GRADIENT DESCENT ROUTINES
  QUADRATIC APPROX OF LOSS CENTERED ON θ_k

$$\begin{cases} \theta_{k+1} = \mathrm{PROX}_{t_k R}(u_k) & t_k \text{ STEPSIZE} \\ u_k = \theta_k - t_k g_k \\ g_k = \nabla L(\theta_k) \end{cases}$$

NESTEROV METHOD: $\begin{cases} \theta_{k+1} = \mathrm{PROX}_{t_k R}(\phi_k - t_k g_k) \\ g_k = \nabla L(\phi_k) \\ \phi_k = \theta_k + \frac{k+1}{k+2}(\theta_k - \theta_{k-1}) \end{cases}$

- APPROXIMATES IN POINT OTHER THAN θ_k
- NESTEROV + ITERATIVE SOFT THRESHOLDING + CONTINUATION = FISTA

FAST ITERATIVE SHRINKAGE THRESHOLDING

Proximal & Projected Gradient Methods

$f(\theta) = L(\theta) + R(\theta)$ — L loss, convex and differentiable — R regularizer not necessarily differentiable

- **Proximal operator** for convex fcn R: $\text{prox}_R(y) = \arg\min_z \left(R(z) + \frac{1}{2}\|z - y\|_2^2 \right)$
  - 'moves' fcn point towards minimum of fcn: $\text{prox}_R(y) \approx y - \lambda \nabla R(y)$
  - intuitively connected to gradient.
  - fixed point for prox are fcn minima

- **Example:** $L(\theta) = RSS$, $R(\theta) = I_C(\theta)$: Lasso
  - $R(\theta) = \lambda\|\theta\|_1 \longrightarrow$ prox = soft thresholding
  - $R(\theta) = \lambda\|\theta\|_0 \longrightarrow$ prox = hard thresholding
  - $R(\theta) = \lambda I_C(\theta) \longrightarrow$ prox = $\text{proj}_C(\theta) = \arg\min \|z - \theta\|_2^2$ (indicator fcn)

  proj is rather easy to compute

~~Projected~~ **Proximal gradient method**

- $\theta_{k+1} = \arg\min \left[t_k R(z) + \frac{1}{2}\|z - u_k\|_2^2 \right] = \text{prox}_{t_k R}(u_k)$, $u_k = \theta_k - t_k g_k$, $g_k = \nabla L(\theta_k)$, $t_k = 1/a_k$, $a_k I$ = Hesian approximation
- $R(\theta) = 0 \longrightarrow$ vanilla grad descent
- $R(\theta) = I_C(\theta) \longrightarrow$ **projected gradient descent**
- $R(\theta) = \lambda_1\|\theta\|_1 \longrightarrow$ iterative soft thresholding $\rightarrow$ ISTA $\rightarrow$ used for lasso

- can haz **Nesterov acceleration**

Conjugate Gradients

- maximal step in each dimension, optimal $\longrightarrow$ orthogonal. **In practice:** conjugate / A-orthogonal. **Conjugacy:** p_i, p_j (vectors) iff $p_i^T A p_j = 0$ for some matrix A

Zeros of: $\alpha_i = p_i^T \frac{(-\nabla f(x_{i-1}))}{p_i^T A p_i}$

How to find conjugate directions? Graham-Schmidt process: take candidate solution direction and remove parts in dir already done

$$p_k = v_k + \sum_{0}^{k-1} \beta_{ki} p_i$$

Two options for β: — $\beta_{i+1} = \frac{\nabla f(x_{i+1})^T \nabla f(x_{i+1})}{\nabla f(x_i)^T \nabla f(x_i)}$ **or** $(\nabla f(x_{i+1}) - \nabla f(x))$

Fletcher - Reeves | Polak - Ribière

- initial direction p_0 : i.e. steepest descent
- find α minimizing $f(x_i + \alpha p_i) \longrightarrow x_{i+1} = x_i + \alpha p_i$ $\xrightarrow{\text{Newton - Rhapson}}$ $\alpha = \frac{J(x)^T p}{p^T H(x) p}$
- $p_{i+1} = -\nabla f(x_{i+1}) + \beta_{i+1} p_i$
- run multiple times and pick best because finds different directions
- cool because requires no matrix inversions

EM FOR LASSO

MODEL WITH MIXING DISTRIBUTION ON VARIANCES
↑
WE CAN REPRESENT LAPLACE DISTRIBUTION AS **GAUSSIAN SCALE MIXTURE**, DERIVE REPRESENTATION OF LASSO. $p(y, w, \tau, \sigma^2 | X) = \exp(-b/\sigma^2) \prod \exp\left(\frac{-\gamma^2}{2}\tau\right.$

E STEP INFERS τ^2 AND σ^2, **M STEP** ESTIMATES w

• SOMETIMES DOES NOT WORK FOR NUMERICAL REASONS

WHY EM?

- PROVIDES WAY TO DERIVE ℓ_1 REG. FOR OTHER MODELS
- ALLOWS FOR TRYING OUT OTHER PRIORS ON VARIANCES
- CAN COMPUTE FULL POSTERIOR AND NOT JUST MAP → **BAYESIAN LASSO**

L1 EXTENSIONS

- GROUP LASSO

WHEN VECTOR OF WEIGHTS AND NOT SINGLE, FOR EACH VAR. IE MULTINOMIAL LOGISTIC, LINEAR REGR WITH CATEGORICAL INPUTS, MULTITASK LEARNING

- PARTITIONS PARAM VECTOR INTO GROUPS
- $J(w) = NLL(w) + \sum^{G} \lambda_g ||w_g||_2$ → IN HERE TWO NORM RESULTS IN SPARSITY. ELSE USE ∞ NORM $||w_g||_\infty = \max[w_j]$. NLL IS LEAST SQUARES → LASSO
- CAN BE SEEN AS GSM → VARIANCE GROUP TERM COMES FROM A GAMMA PRIOR
- ALGOS: PROXIMAL GRADIENT DESCENT DECOMPOSED IN G, EM

- FUSED LASSO

IF WE WANT NEIGHBORING COEFFICIENTS TO BE SIMILAR TO EACHOTHER AND SPARSE. LOCATION-BASED PENALTY. GRAPHS, IMAGES, LATTICES ETC.
SIMILAR TO CHAIN-STRUCTURED GAUSSIAN MARKOV FIELDS, RANDOM VARIANCE. SOLVES WITH EM.

$$J(w) = J(w, \lambda_1, \lambda_2) = \sum^{N} (y_i - w_i)^2 + \lambda_1 \sum^{N} |w_i| + \lambda_2 \sum^{N-1} |w_{i+1} - w_i|$$

- ELASTIC NET

COMBINES LASSO AND RIDGE

- IF GROUP OF VARS CORRELATED → LASSO LIKELY PICKS ONE OF THEM
- $D > N$ → LASSO AT MOST N VARS
- $N > D$ + CORRELATED VARS → RIDGE DOES BETTER

• OBJECTIVE: $J(w, \lambda_1, \lambda_2) = ||y - Xw||^2 + \lambda_2 ||w||_2^2 + \lambda_1 ||w||_1$

STRICTLY CONVEX
GROUPING EFFECT = HIGHLY CORRELATED VARS HAVE EQUAL WEIGHTS

- ALGORITHM: LASSO ON MODIFIED DATA $\tilde{X} = c\begin{pmatrix} X \\ \sqrt{\lambda_2} I_D \end{pmatrix}$ $\tilde{y} = \begin{pmatrix} y \\ 0_{D \times 1} \end{pmatrix}$ $c = (1 + \lambda_2)^{-1/2}$
  - USE LARS → LARS-EN
  - IF WE STOP AT m VARS → $O(m^3 + Dm^2)$

$\tilde{w} = \arg\min ||\tilde{y} - \tilde{X}\tilde{w}||_2^2 + c\lambda_1 ||\tilde{w}||_1$

$w = c\tilde{w}$

• CV TO PICK λ_1, λ_2

• OBS! DOUBLE SHRINKAGE: RESCALE UP, UNDO ℓ_2, $\hat{w} = \sqrt{1 + \lambda_2}\, \tilde{w}$

• GSM: WITH PRODUCT OF GAUSSIAN AND LAPLACE PRIORS
MAP, MCMC, VARIATIONAL BAYES

NONCONVEX REGULARIZERS

LAPLACE PRIOR NOT SUPPRESSES NOISE. TOO MUCH SHRINKAGE ON LARGE VALUE COEFFS ⟶ INTRODUCE MORE FLEXIBLE PRIORS. NO GLOBAL OPTIMUM ANYMORE BUT OUTPERFORM IN PREDICTIVE ACCURACY AND VARIABLE SELECTION

– BRIDGE REGRESSION

GENERALIZES ℓ_1 $\quad \hat{w} = NLL(w) + \lambda \sum_j |w_j|^b$ • MAP WITH EXPONENTIAL POWER DISTRIBUTION  'BALL'

- $b=0$ $\ell_0 \to$ BSS
- $b=1$ = LAPLACE ⟶ LASSO
- $b=2$ = GAUSSIAN ⟶ RIDGE

• CAN GENERATE BIMODALS

– HIERARCHICAL ADAPTIVE LASSO

DIFFERENT PENALTY PARAMETER FOR EACH PARAMETER. PENALTIES MODELED AS RV COMING FROM CONJUGATE PRIORS $\gamma_j \sim IG(a,b)$, $\tau^2|\gamma_j \sim Ga(1, \gamma_j^2/2)$

$w_j|\tau_j^2 \sim N(0, \tau_j^2)$

- OFTEN WORKS MUCH BETTER THAN ℓ_1, FIT WITH EM
  - E-STEP: SAME FOR LASSO
  - M-STEP: WEIGHTED LASSO PROBLEM ⟶ USE LARS
- STARFISH ⟶ MORE AGGRESSIVE SPARSITY
- APPROXIMATES HARD-THRESHOLDING
- CAN USE WITH ALL SORTS OF COMBINATIONS OF $p(\tau_j^2), p(\gamma_j), p(w_j)$ FOR DIFFERENT RESULTS
- SCALED MIXTURE OF LAPLACIANS

AUTOMATIC RELEVANCE DETERMINATION / SPARSE BAYESIAN LEARNING
ARD / SBL

APPROACH BASED ON TYPE 2 ML, EMPIRICAL BAYES. • WE INTEGRATE OUT W AND MAXIMIZE MARGINAL LIKELIHOOD WRT τ. USING EM OR ℓ_1 SCHEME

- IS POINT ESTIMATION
- NON-FACTORIAL PRIOR
- VARIANCES ESTIMATED ⟶ PLUG-IN TO COMPUTE WEIGHT POSTERIOR MEANS $E[w|\hat{\tau}, D]$ ⟶ SPARSENESS
- 'PUNISHES' SOLUTIONS THAT WASTE PROBABILITY MASS

- WE INTEGRATE w OUT AND OPTIMIZE α, MLE STANDARD DOES OPPOSITE. PARAMS w_j BECOME CORRELATED IN POSTERIOR DUE TO EXPLAINING AWAY. ⟶ FOR α_j WE ~~ESTIM~~ BORROW INFO FROM ALL FEATURES
- PRIOR IS NON-FACTORIAL: $p(\hat{w}|\alpha)$ DEPENDS ON DATA D AND σ^2
- NON-FACTORIAL OBJECTIVE HAS ALWAYS FEWER LOCAL MINIMA THAN FACTORIAL ONES ⟶ BUT STILL GLOBAL MAXIMUM IS EQUAL TO ℓ_0
- ALGORITHMS: EM, FIXED-POINT ALGORITHM, REWEIGHTED ℓ_1 $w^{t+1} = \arg\min NLL(w) + \sum \lambda_j |w_j|$, VARIATIONAL APPROXIMATIONS
  - E μ, Σ
  - M α, β
- IN GAUSSIAN PROCESSES, NNS, ... • $k(x,x') = \theta_0 \exp\left[-\frac{1}{2}\sum_{1}^{2} \eta_i (x_i - x'_i)^2\right]$ $\eta \to 0$ FOR INSENSITIVE TO X VALUES
  USE ML TO ESTIMATE η ⟶ DETECT VARS WITH LITTLE POSTERIOR EFFECT, CAN DISCARD

SPARSE CODING

IDEA: SPARSE PRIORS FOR UNSUPERVISED LEARNING, THINK ICA BUT WITH SPARSITY PROMOTING PRIORS. VECTORS x_i AS SPARSE COMBINATION OF BASIS VECTORS

SPARSE CODING: W NOT ORTHOGONAL, SPARSE PRIOR ON LATENT FACTORS

W: HERE CALLED A DICTIONARY COLUMN OF W: ATOM • IF L>D OVERCOMPLETE
- FIXED: WAVELET / DCT BASIS
- LEARNED: $\log p(D|w) = \sum_i^N \log \int N(x_i|Wz_i, \sigma^2 I)\, p(z_i)\, dz_i$

SPARSE PCA: SPARSE ON Ws, PRIOR IS GAUSS

SPARSE MATRIX FACTORIZATION: SPARSE Ws + SPARSE LF

HOW TO LEARN DICTIONARY!

p(z)
LAPLACE: $NLL(W,Z) = \sum \frac{1}{2}\|x_i - Wz_i\|_2^2 + \lambda\|z_i\|_1$ CONSTRAIN ℓ_2 OF COLUMNS TO BE ≤ 1

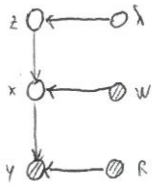
- FOR FIXED Z: LEAST SQUARE OPTIMIZATION OVER W
- FOR FIXED W LASSO PROBLEM OVER Z
- ANALYSIS/SYNTHESIS LOOP: ALTERNATE OPTIMIZE W AND Z

- OTHER MODELS ⟶ OTHER OPTIMIZATION PROBLEMS
- SMF: ELASTIC NET TYPE PENALTY ON WEIGHTS = $\min_{W,Z} \frac{1}{2}\sum^N \|x_i - Wz_i\|_2^2 + \lambda\|z_i\|_1$
- CAN USE OTHER SPARSITY INDUCING PRIORS THAN LAPLACE, IE BINOMIALS, BETA PROCESSES

WHAT TO DO WITH SPARSE CODING

- COOL BECAUSE LEARNED BASIS VECTOR LOOK LIKE BRAIN 'VISUAL FILTERS', ALSO ICA, BUT NOT PCA
- **COMPRESSED SENSING:** $y = Rx + \epsilon$, LOW-DIM PROJECTION OF DATA. R KNOWN SENSING MATRIX. IE MRI EACH BEAM DIRECTION IS ROW IN R

  → **INFER $P(x \mid y, R)$** VIA BAYESIAN INFERENCE. ASSUMING $x = W \cdot z$
  - L SPARSE PRIOR
  - DICTIONARY



WE OBSERVE LOW-DIM y PASSED THROUGH R.
x IS SPARSE WRT W AND z.

- **IMAGE INPAINTING / DENOISING**

  WE PARTITION IMG IN OVERLAPPING PATCHES y_i AND CONCATENATE TO FORM y.
  R DEFINED TO SELECT PATCH i WITH iTH ROW. V VISIBLE, H HIDDEN.

  WE COMPUTE $P(y_H \mid y_V; \theta)$ θ ARE PARAMS W AND SPARSITY LEVEL λ

  DICTIONARY LEARNED OR FIXED.

  ALTERNATIVE: FIELD OF EXPERTS → ENCODE CORRELATIONS BETWEEN NEIGHBORING PATCHES / NO LATENTS → COMPUTATIONALLY HEAVIER