Linear Algebra

$\langle \mathbf{u},\mathbf{v}\rangle = \mathbf{u}^\top\mathbf{v} = \sum_{i=1}^d u_i v_i$, $\mathbf{u},\mathbf{v}\in\mathbb{R}^d$
$\langle a\mathbf{x},\mathbf{y}\rangle = a\langle\mathbf{x},\mathbf{y}\rangle$, $\langle\mathbf{x}+\mathbf{y},\mathbf{z}\rangle=\langle\mathbf{x},\mathbf{z}\rangle+\langle\mathbf{y},\mathbf{z}\rangle$
$\langle\mathbf{x},\mathbf{x}\rangle = \|\mathbf{x}\|_2^2$
$\mathbf{AB}=\mathbf{C}\Rightarrow\sum_{k=1}^m a_{ik}b_{kj}$

Orthogonality

$\mathbf{u},\mathbf{v}$ are orthognal iff $\langle\mathbf{u},\mathbf{v}\rangle = 0$
$\mathbf{A}$ orthog.$\Rightarrow\mathbf{AA}^\top=\mathbf{A}^\top\mathbf{A}=\mathbf{I}$, $\mathbf{A}^\top=\mathbf{A}^{-1}$
$\mathbf{A}$ has orthonormal rows, i.e. $\|\mathbf{a}_i\|_2=1$
$\|\mathbf{Ax}\|_2=\|\mathbf{x}\|_2$, dist. & energy preserved
basis change $\mathbf{x}=\sum_{i=1}^n\langle\mathbf{x},\mathbf{u_i}\rangle\mathbf{u_i}=\mathbf{U}^\top\mathbf{x}$
and $\|\langle\mathbf{x},\mathbf{u_i}\rangle\mathbf{u_i}\|_2^2=|\langle\mathbf{x},\mathbf{u_i}\rangle|^2$
$\det(\mathbf{A})\in\{-1,1\}$, $\det(\mathbf{A}^\top\mathbf{A})=1$

Symmetry

$\mathbf{A}^\top=\mathbf{A}$. All eigenvalues of $\mathbf{A}>0$.
$\mathbf{A}=\mathbf{U}\Lambda\mathbf{U}^\top$ iff $\mathbf{A}$ symmetric where $\mathbf{U}$ is orthogonal.

Positive Semi-Definit

$\mathbf{v}^\top\mathbf{Av}\geq 0\forall\mathbf{v}$, $\mathbf{A}=\mathbf{B}^\top\mathbf{B}\Rightarrow\mathbf{A}$ is p.s.t.
$\mathbf{A}$ pst iff all its eigenvalues ≥ 0

Rank

dimension of vector space generated by $\mathbf{A}\in\mathbb{R}^{m\times n}$.
$\mathrm{rank}(\mathbf{A})=\#\sigma>0$, $\mathrm{rank}(\mathbf{A})\leq\min(m,n)$
$\mathrm{rank}(\mathbf{AB})=\min\{\mathrm{rank}(\mathbf{A}),\mathrm{rank}(\mathbf{B})\}$

Norms

$\|\mathbf{x}\|_0=|\{i|x_i\neq 0\}|$, $\|\mathbf{x}\|_p=(\sum_{i=1}^N|x_i|^p)^{\frac{1}{p}}$
nuclear $\|\mathbf{X}\|_\star=\sum_{i=1}^{\min(m,n)}\sigma_i$
eucl. $\|\mathbf{x}\|_2=\sqrt{\sum_{i=1}^N\mathbf{x}_i^2}=\sqrt{\mathbf{x}^\top\mathbf{x}}=\sqrt{\langle\mathbf{x},\mathbf{x}\rangle}=\sigma_1$
frob. $\|\mathbf{A}\|_F=\sqrt{\sum_{i=1}^M\sum_{j=1}^N|\mathbf{A}_{i,j}|^2}=$
$\sqrt{\mathrm{trace}(\mathbf{A}^\top\mathbf{A})}=\sum_{i=1}^{\min\{m,n\}}\sigma_i^2$, $\mathbf{A}\in\mathbb{R}^{M\times N}$
w. $\mathrm{trace}(\mathbf{ABC})=\mathrm{trace}(\mathbf{CAB})$ (cyclic)

Determinant

$\det(\mathbf{A})=a_{11}a_{22}-a_{12}a_{21}=\prod_i\sigma_i$
$\det(\mathbf{A}^\top)=\det(\mathbf{A})$, $\det(\mathbf{A}^{-1})=(\det(\mathbf{A}))^{-1}$

Eigenvectors & Eigenvalues

$\sigma\in\mathbb{R}$ s.t. $\mathbf{Au}=\sigma\mathbf{u}$ Find eigenvals: solve $\det(\mathbf{A}-\lambda\mathbf{I})=0$
Find eigenvecs $\mathbf{u}_i$: solve $(\mathbf{A}-\sigma_i\mathbf{I})\mathbf{u}_i=\mathbf{0}$

SVD - Singular Value Decomposition

$\mathbf{A}=\mathbf{U}\cdot\mathbf{D}\cdot\mathbf{V}^\top\ \forall\mathbf{A}\in\mathbb{R}^{m\times n}$
$\mathbf{U}\in\mathbb{R}^{m\times m}$ orth. matrix ($\mathbf{U}^\top\mathbf{U}=\mathbf{I}_m$)
$\mathbf{D}\in\mathbb{R}^{m\times n}$ diagonal matrix
$\mathbf{V}\in\mathbb{R}^{n\times n}$ orth. matrix ($\mathbf{V}^\top\mathbf{V}=\mathbf{I}_n$)
$\mathbf{U}$ has eigenvecs for $\mathbf{AA}^\top$, $\mathbf{V}$ for $\mathbf{A}^\top\mathbf{A}$.
$\mathbf{D}$ has eigenvalues for $\mathbf{AA}^\top$ and $\mathbf{A}^\top\mathbf{A}$, are sqrt of entries of Λ: $\mathbf{AA}^\top=\mathbf{V}\Lambda\mathbf{V}^\top$

Eckart-Young Theorem

$\arg\min_{\hat{\mathbf{X}}:\mathrm{rank}(\hat{\mathbf{X}})=k}\|\mathbf{X}-\hat{\mathbf{X}}\|_F^2=\mathbf{U}\Sigma_\mathbf{k}\mathbf{V}^\top$
$\Sigma=\mathrm{diag}(\sigma_1,\cdots,\sigma_s)$, $\sigma_1\geq\cdots\geq\sigma_s\geq 0$
$\min_{\mathrm{r}(\mathbf{B})}=\|\mathbf{A}-\mathbf{B}\|_F^2=\|\mathbf{A}-\mathbf{A}_k\|_F^2=\sum_{r=k+1}^{\mathrm{r}(\mathbf{A})}\sigma_r^2$
$\min_{\mathrm{r}(\mathbf{B})}=\|\mathbf{A}-\mathbf{B}\|_2=\|\mathbf{A}-\mathbf{A}_k\|_2=\sigma_{k+1}$
Those are not convex optimisation problems (combination of matrices of rank k is usually not rank k)

Probabilities

Expectation

$\mathbb{E}[X]=\int_\Omega xf(x)dx=\int_\omega xP[X{=}x]dx$
$\mathbb{E}_{Y|X}[Y]=\mathbb{E}_Y[Y|X]$
$\mathbb{E}_X[f(Y)]=f(Y)$ if $P(Y|X)=P(Y)$

Variance & Covariance

$Var(X)=\mathbb{E}[(X-\mathbb{E}[X])^2]=\mathbb{E}[X^2]-\mathbb{E}[X]^2$
$\mathrm{Var}[X+Y]=\mathrm{Var}[X]+\mathrm{Var}[Y]\quad XY\,iid$
$\mathrm{Var}[\alpha X]=\alpha^2\mathrm{Var}[X]$
$Cov(X,Y)=\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]$

Conditional Probabilities

$P[X|Y]=\frac{P[X,Y]}{P[Y]}$, $P[\overline{X}|Y]=1-P[X|Y]$
Bayes: $P(X|Y)=\frac{P(Y|X)P(X)}{P(Y)}$

Distributions

$\mathcal{N}(x|\mu,\sigma^2)=(\sqrt{2\pi\sigma^2})^{-1}\exp^{-(x-\mu)^2/(2\sigma^2)}$
$\mathcal{N}(x|\mu,\Sigma)=\frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}(\mathbf{x}-\mu)^\top\Sigma^{-1}(\mathbf{x}-\mu)}$
$\mathrm{Exp}(x|\lambda)=\lambda e^{-\lambda x}$, $\mathrm{Ber}(x|\theta)=\theta^x(1-\theta)^{(1-x)}$

Kullback-Leibler Divergence

divergence of distr P & Q $D_{KL}(P\|Q)=$
$-\sum_{x\in\mathcal{X}}P(x)\log\frac{Q(x)}{P(x)}=\sum_{x\in\mathcal{X}}P(x)\log\frac{P(x)}{Q(x)}$

Matrix Derivations

$\frac{\partial(\mathbf{b}^\top\mathbf{x})}{\partial\mathbf{x}}=\frac{\partial(\mathbf{x}^\top\mathbf{b})}{\partial\mathbf{x}}=\mathbf{b}$, $\frac{\partial(\mathbf{x}^\top\mathbf{x})}{\partial\mathbf{x}}=2\mathbf{x}$,
$\frac{\partial(\mathbf{b}^\top\mathbf{Ax})}{\partial\mathbf{x}}=\mathbf{A}^\top\mathbf{b}$, $\frac{\partial(\mathbf{x}^\top\mathbf{Ax})}{\partial\mathbf{x}}=(\mathbf{A}^\top+\mathbf{A})\mathbf{x}$,
$\frac{\partial(\mathbf{c}^\top\mathbf{Xb})}{\partial\mathbf{X}}=\mathbf{cb}^\top$, $\frac{\partial(\mathbf{c}^\top\mathbf{X}^\top\mathbf{b})}{\partial\mathbf{X}}=\mathbf{bc}^\top$, $\frac{\partial\log|\mathbf{A}|}{\partial\mathbf{A}}=\mathbf{A}^{-\top}$
$\frac{\partial(\|\mathbf{x}-\mathbf{b}\|_2)}{\partial\mathbf{x}}=\frac{\mathbf{x}-\mathbf{b}}{\|\mathbf{x}-\mathbf{b}\|_2}$, $\frac{\partial(\|\mathbf{x}\|_2^2)}{\partial\mathbf{x}}=\frac{\partial(\mathbf{x}^\top\mathbf{x})}{\partial\mathbf{x}}=2\mathbf{x}$,
$\frac{\partial(\|\mathbf{X}\|_F^2)}{\partial\mathbf{X}}=2\mathbf{X}$, $\frac{\partial(\|\mathbf{Ax}-\mathbf{b}\|_2^2)}{\partial\mathbf{x}}=2(\mathbf{A}^\top\mathbf{Ax}-\mathbf{A}^\top\mathbf{b})$
$\mathbf{X}^T\mathbf{X}$: invertible if eigvals $\neq 0$ but instable if big ratio last/first eigval.

Derivatives

chain rule $\frac{d}{dx}f(g(x)=f'(g(x))g'(x)$
$\frac{d}{dx}\cos(x)=-\sin(x)$ $\frac{d}{dx}\sin(x)=\cos(x)$
$\frac{d}{dx}\exp(x)=\exp$ $\frac{d}{dx}\log(x)=\frac{1}{x}$

Optimization

Gradient Descent

$\theta^{new}\leftarrow\theta^{old}-\eta\nabla_\theta\mathcal{L}(\theta)$
Convergence isn't guaranteed.
Less zigzag by adding momentum:
$\theta^{(l+1)}\leftarrow\theta^{(l)}-\eta\nabla_\theta\mathcal{L}+\mu(\theta^l-\theta^{(l-1)})$

Stochastic Gradient Descent

Assume $\mathcal{L}(\theta)=\sum_{n=1}^N\mathcal{L}_n(\theta)$
$\theta^{new}\leftarrow\theta^{old}-\eta\nabla_\theta\mathcal{L}_n(\theta)\qquad n{\sim}\text{uniform}$
this requires the function to be L-smooth to avoid oscillations

Jensen's Inequality

$\sigma(\frac{\sum a_i x_i}{\sum a_i})\geq\frac{\sum a_i\sigma(x_i)}{\sum a_i}$ if σ concave (log) for distribution a: $\sum_i a_i=1$

Convexity

f convex if $\forall x_1,x_2{\in}\mathbf{X},\forall t{\in}[0,1]$
$f(tx_1+(1-t)x_2)\leq tf(x_1)+(1-t)f(x_2)$
if f is twice diff., it is convex iff it's double Hessian is positive semi-definite A convex function has a unique minimum. sum of conv.is conv.

Linear Autoencoder

Encoder $\mathbf{C}{\in}\mathbb{R}^{k\times m}$, Decoder $\mathbf{D}{\in}\mathbb{R}^{m\times k}$
lin. map: $\mathbf{F}:\mathbb{R}^m{\to}\mathbb{R}^m$ w. limited rank. performs low rank approximation
$\ell(\mathbf{x};\theta)=\frac{1}{2}\|\mathbf{x}-\hat{\mathbf{x}}(\theta)\|^2$
$\hat{\mathbf{x}}(\theta)=\mathbf{DCx}$, $\theta=(\mathbf{C},\mathbf{D})$

Reconstruction error

$J(\theta)=\frac{1}{n}\sum_{i=1}^n\ell(\mathbf{x}_i,\theta)=$
$\frac{1}{2n}\sum_{i=1}^n\|\mathbf{x}_i-\hat{\mathbf{x}}_i(\theta)\|=\frac{1}{2n}\|\mathbf{X}-\hat{\mathbf{X}}(\theta)\|_F^2$

Optimal Solution

$\mathbf{C}^*{=}\mathbf{U}_k^\top$, $\mathbf{D}^*{=}\mathbf{U}_k$, s.t. $\mathbf{X}=\mathbf{U}\Sigma\mathbf{V}^\top$
$\hat{\mathbf{X}}=\mathbf{D}^*\mathbf{C}^*\mathbf{X}=\mathbf{U_k}\mathbf{U}_\mathbf{k}^\top(\mathbf{U}\Sigma^\top)=\cdots=\mathbf{U}\Sigma_\mathbf{k}\mathbf{V}^\top$
opt. by EY. not the only optimal solution, i.e. limited interpretability

Weight Sharing

$\mathbf{D}=\mathbf{C}^\top$ reduces ambiguity but not modeling power. Mapping unique.

PCA - Principle Component Analysis

project data $\mathbf{X}=[\mathbf{x_1}\cdots\mathbf{x_N}]{\in}\mathbb{R}^{D\times N}$ to basis of orthogonal components.
Centralise data by substracting the mean $\overline{\mathbf{X}}=\mathbf{X}-[\overline{\mathbf{x}}\cdots\overline{\mathbf{x}}]$, $\overline{\mathbf{x}}=\frac{1}{N}\sum_{n=1}^N\mathbf{x}_n$.

Variance-Covariance Matrix

$\Sigma=\frac{1}{N}\sum_{n=1}^N(\mathbf{x}_n-\overline{\mathbf{x}})(\mathbf{x}_n-\overline{\mathbf{x}})^\top=\frac{1}{N}\overline{\mathbf{X}}\overline{\mathbf{X}}^\top$.
symmetric $\Sigma=\mathbf{U}\Lambda\mathbf{U}^\top$, $\mathbf{U}$ orth.
New orthogonal basis $\mathbf{U}_K$, $K<<D$
$\overline{\mathbf{Z}}_K=\mathbf{U}_K^\top\overline{\mathbf{X}}$ and $\tilde{\overline{\mathbf{X}}}=\mathbf{U}_k\overline{\mathbf{Z}}_K$ opt. rec.

Iterative View

Residual $\mathbf{r}_i$: $\mathbf{x}_i-\tilde{\mathbf{x}}_i=\mathbf{I}-\mathbf{uu}^\top\mathbf{x}_i$
Covariance $\frac{1}{N}\sum_{i=1}^N\mathbf{r}_i\mathbf{r}_i{}^\top=\Sigma-\lambda\mathbf{uu}^\top$
1^{st}eigvec of $\Sigma-\lambda\mathbf{uu}^\top=2^{nd}$eigvec of Σ
get d princ. eigvecs of Σ by iteration

Power Method

$\mathbf{v}_{t+1}=\frac{\mathbf{Av}_t}{\|\mathbf{Av}_t\|}$, $\lim_{t\to\infty}\mathbf{v}_t=\mathbf{u}_1$
assume $\langle\mathbf{u}_1,\mathbf{v}_0\rangle\neq 0$ and $|\lambda_1|>|\lambda_j|$

1-D PCA to a line

Line $\mu+\mathbb{R}\mathbf{u}\equiv\{\mathbf{v}{\in}\mathbb{R}^m{:}\exists z \text{ s.t. } \mathbf{v}=\mu+z\mathbf{u}\}$
$\hat{\mathbf{x}}=\mu+\langle\mathbf{x}-\mu,\mathbf{u}\rangle\mathbf{u}=\arg\min_{\hat{\mathbf{x}}\in\mu+\mathbb{R}\mathbf{u}}\|\mathbf{x}-\hat{\mathbf{x}}\|^2$
Center the data to find unique $\mathbf{u}$
$\mathbf{u}\leftarrow\arg\min[\frac{1}{n}\sum_{i=1}^n\|\langle\mathbf{u},\mathbf{x_i}\rangle\mathbf{u}-\mathbf{x_i}\|^2]$
yields $\mathbf{u}\leftarrow\arg\max[\frac{1}{n}\sum_{i=1}^n\langle\mathbf{u},\mathbf{x_i}{}^\top\rangle^2]=$
$\arg\max[\mathbf{u}^\top(\frac{1}{n}\sum_{i=1}^n\mathbf{x_i}\mathbf{x_i}{}^\top)\mathbf{u}^\top]=$
$\arg\max[\mathbf{u}\frac{1}{n}\mathbf{XX}^\top\mathbf{u}^\top]=\arg\max[\mathbf{u}\Sigma\mathbf{u}^\top]$

Lagrangian Optimization
$\mathcal{L}(\mathbf{u},\lambda)=\mathbf{u}^\top\Sigma\mathbf{u}+\lambda\langle\mathbf{u},\mathbf{u}^\top\rangle$
$\nabla_\mathbf{u}\mathcal{L}(\mathbf{u},\lambda)\overset{!}{=}0\Leftrightarrow\Sigma\mathbf{u}=\lambda\mathbf{u}$,ie $\mathbf{u}$ princ. vec

Matrix Approximation & Reconst

Exact Rec. $\min_\mathbf{B}\|\mathbf{B}\|_*$ s.t.$\|\mathbf{A}-\mathbf{B}\|_\mathbf{G}=0$
Approx. Rec. $\min_\mathbf{B}\|\mathbf{A}-\mathbf{B}\|_\mathbf{G}^2$ s.t.$\|\mathbf{B}\|_*$,

SVD (e.g recommendation system)

$\mathbf{U}$ users to factors associations $\mathbf{V}$ items to factors associations Σ level of strength of each factor. **Limitations** matrix incomplete. can't run SVD. Lower rank might not be the best solution. Ratings should not be outside the expected range

Beyond SVD

optimize $\min_{\mathrm{rk}(\mathbf{B})=k}[\sum_{(i,j)\in\mathcal{I}}(a_{ij}-b_{ij})^2]$
$\mathcal{I}=\{(i,j)\}$ observed values
$\|\mathbf{X}\|_\mathbf{G}=\sqrt{\sum_{i,j}g_{ij}x_{ij}^2}$, $g_{ij}\in\{0,1\}$
$\min_{\mathrm{rk}(\mathbf{B})}\|\mathbf{A}-\mathbf{B}\|_\mathbf{G}^2$ NP hard

Alternating Least Square

parametrize $\mathbf{B}_{rk(k)}=\mathbf{UV}$, $\mathbf{U}{\in}^{m\times k}$, $\mathbf{V}{\in}^{k\times n}$
$f(\mathbf{U},\mathbf{V})=\frac{1}{|\mathcal{I}|}\sum_{(i,j)\in\mathcal{I}}(a_{ij}-\langle\mathbf{u}_i,\mathbf{v}_j\rangle)^2=$
$\sum_i[\sum_{j:(i,j)\in\mathcal{I}}(a_{ij}-\langle\mathbf{u}_i,\mathbf{v}_j\rangle)^2]=\sum_i f(\mathbf{U},\mathbf{v}_i)$.
f is convex for fixed $\mathbf{U}$ in $\mathbf{V}$ and vice-versa but not jointly
Regularise $\Omega(\mathbf{U},\mathbf{V})=\|\mathbf{U}\|_F^2+\|\mathbf{V}\|_F^2$
Minimise $f(\mathbf{U},\mathbf{V})+\mu\Omega(\mathbf{U},\mathbf{V})$ alternate between $\mathbf{U}$ and $\mathbf{V}$, ie:
$f(\mathbf{U},\mathbf{v}_i)=\sum_{(i,j)\in\mathcal{I}}(a_{i,j}-\langle\mathbf{u}_j,\mathbf{v}_i\rangle)^2$

Convex Relaxation

$\mathrm{rank}(\mathbf{B})\geq\|\mathbf{B}\|_*$ for $\|\mathbf{B}\|_2\leq 1$
$\min_{\mathbf{B}\in\mathcal{P}_k}\|\mathbf{A}-\mathbf{B}\|_\mathbf{G}^*$, $\mathcal{P}_k=\{\mathbf{B}:\|\mathbf{B}\|_*\leq k\}$
$\mathcal{P}_k\mathcal{Q}_k=\{\mathbf{B}:\mathrm{rank}\mathbf{B}=k\}$
$\mathbf{B}^*{=}\mathrm{shrink}_\tau(\mathbf{A}){=}$
$\arg\min_\mathbf{B}\{\frac{1}{2}\|\mathbf{A}-\mathbf{B}\|_F^2+\tau\|\mathbf{B}\|_*\}$ thus
$\mathbf{B}^*{=}\mathbf{UD}_\tau\mathbf{V}^\top$, $\mathbf{D}_\tau{=}\mathrm{diag}(\max(0,\sigma_i-\tau))$
Shrinkage Iterations for $\eta\geq 0$
$\mathbf{B}_{t+1}=\mathbf{B}_t+\eta_t\Pi(\mathbf{A}-\mathrm{shrink}_\tau(\mathbf{B}_t))$
$\Pi(\mathbf{X})=x_{ij}$ if $(i,j)\in\mathcal{I}$, 0 otw.
generalisation guarantees $\Pi(\mathbf{A})=\mathbf{A}^*$
ie $\mathbf{B}^*{=}\arg\min_\mathbf{B}\{\|\mathbf{B}\|_*\}$ s.t. $\Pi(\mathbf{A}-\mathbf{B})=0$

Topic Model

find low-dim representation of document from a corpus **Preprocessing** vocabulary extraction/tokenisation, filtering (of too frequent or rare words), normalisation (steming) e.g. argue(d) reduce to arg **bag of words** counts of cooc. ignores order. sparse!

pLSA

$p(w)=\sum_{z=1}^K p(w|z)p(z|d)$, for word w in doc d and given topics $z\in\{1\cdots K\}$
assumption $p(w|d,z){=}p(w|z)$
Log-Likel. $x_{ij}{:=}\{\#\text{ of }w_j\text{ in }d_i\}$, $\mathbf{X}{=}x_{ij}$
$\ell(\mathbf{U},\mathbf{V})=\sum_{ij}x_{ij}\log p(w_j|d_i)=$
$\sum_{(i,j)\in\mathcal{X}}x_{ij}\log\sum_{z=1}^K p(w_j|z)p(z|d_i)=$
$\sum_{(i,j)\in\mathcal{X}}x_{ij}\log\sum_{z=1}^K v_{jz}u_{zi}$, with $u_{iz}{\geq}0\sum_z u_{iz}{=}1$ and $v_{jz}{\geq}0\sum_j v_{jz}{=}1$
Lower Bound w. Jensen's inequ.
$q_{zij}{=}P(w_j\text{ in }d_i\text{ is from }z)$, $\sum_z q_{ijz}=1$
$\sum_{ij}x_{ij}\log\sum_{z=1}^K q_{ijz}\frac{u_{zi}v_{zj}}{q_{ijz}}\geq g(\mathbf{X}|\mathbf{U},\mathbf{V}){=}$
$\sum_{ij}x_{ij}\sum_{z=1}^K q_{ijz}[\log u_{zi}+\log v_{zj}-\log q_{ijz}]$
Lagrangian $\mathcal{L}_{\mathbf{U},\mathbf{V}}(\alpha,\beta)=-g(\mathbf{X}|\mathbf{U},\mathbf{V})+$
$\sum_j\alpha_j(\sum_z u_{zi}-1)+\sum_z\alpha_i(\sum_j v_{jz}-1)$
Optimal solution (EM)
E: $q_{zij}=\frac{u_{zi}v_{zj}}{\sum_{k=1}^K u_{zi}v_{zj}}=\frac{p(w_j|z)p(z|d_i)}{\sum_{k=1}^K p(w_j|k)p(k|d_i)}$
M: $u_{zi}=\frac{\sum_j x_{ij}q_{zij}}{\sum_j x_{ij}}$, $v_{zj}=\frac{\sum_j x_{ij}q_{zij}}{\sum_{i,l}x_{il}q_{zil}}$
conv. guaranteed but not global opt. fixed docs and words. Add a topic?

Latent Dirichlet Allocation

$p(\mathbf{u}_i|\alpha)\propto\prod_{z=1}^K u_{zi}^{\alpha_z-1}$ generate topic weights. for doc. w. length $l=\sum_j x_j$
$p(\mathbf{x},|\mathbf{V},\mathbf{u})=\frac{l!}{\prod_j x_j!}\prod_j(\sum_z v_{zj}u_z)^{x_j}$
$p(\mathbf{x}|\mathbf{V},\alpha)=\int p(\mathbf{x}|\mathbf{V},\mathbf{u})p(\mathbf{u}|\alpha)d\mathbf{u}$

Non-Negative Matrix Factorization

factorize count matrix $\in\mathbb{Z}_{\geq 0}^{N\times M}$
$\mathbf{X}=\mathbf{U}^\top\mathbf{V}$. $\mathbf{U},\mathbf{V}$ non-neg. entries and L_1 column normalized.
Useful to model non-negative data like images (ink) and leads to part-based representation. pLSA is a kind of NMF.

Word Embeddings

latent vect model $w{\to}(\mathbf{x}_w,b_w){\in}\mathbb{R}^{d+1}$

Context Models

semantic from by co-occurences, e.g. skip-gram $p(w|w')$ w in context of w'

Log-likelihood
$\mathcal{L}(\theta|\mathbf{w})=\sum_{t=1}^{T}\sum_{\delta\in\mathcal{I}}\log p_\theta(w^{(t+\delta)}|w^{(t)})$
$\hat{\theta}=\arg\max_\theta \mathcal{L}(\theta|\mathbf{w})$ large cardinality
Log-bilinear model
$\log p(w|w')=\langle\mathbf{x}_w,\mathbf{x}_{w'}\rangle+b_w+c$
$b_w\uparrow\Rightarrow p_\theta(w|w')\uparrow$,
$\angle(\mathbf{x}_w\mathbf{x}_{w'})\downarrow\Rightarrow p_\theta(w|w')\uparrow$
Softmax $p_\theta(w|w')=\frac{\exp[\langle\mathbf{x}_w,\mathbf{x}_{w'}\rangle+b_w]}{Z_\theta(w')}$
where $Z_\theta(w')=\sum_{v\in\mathcal{V}}\exp[\langle\mathbf{x}_v,\mathbf{x}_{w'}\rangle+b_v]$
Context Vectors input & output context vectors to get rid of the bilinearity. $\log p_\theta(w|w')=\langle\mathbf{x}_w,\mathbf{y}_{w'}\rangle+b_w$
Negative Sampling
$\triangle^+$ observed pair, $\triangle^-\sim p_n$ where $(w_i,w_j)\sim p_n(i,j)$ rand. context words
$w_j\propto P(w_j)^\alpha,\alpha=\frac{3}{4}$oversample by $k\leq 20$
Maximize logistic Regression
$\mathcal{L}(\theta)=\sum_{(i,j)\in\triangle^+}\log\sigma(\langle\mathbf{x}_i,\mathbf{y}_j\rangle)+\sum_{(i,j)\in\triangle^-}\log\sigma(-\langle\mathbf{x}_i,\mathbf{y}_j\rangle),\sigma(z)=\frac{1}{1+\exp(z)}$
Bayes optimal discriminant for $\mathcal{L}$
$\langle\mathbf{x}_i,\mathbf{y}_j\rangle=\log\frac{p(w_i,w_j)}{p_n(w_i,w_j)}+\log\frac{\kappa}{1-\kappa},\kappa=\frac{1}{k+1}$
pointwise mutual information
for $k=1$ and $p_n(w_i,w_j)=p(w_i)p(w_j)$
$\langle\mathbf{x}_i,\mathbf{y}_j\rangle\approx\mathrm{PMI}(w_i,w_j)$

GloVe
co-occurence matrix $\mathbf{N}=\{n_{ij}\}\in\mathbb{R}^{|\mathcal{V}|\cdot|\mathcal{C}|}$
n_{ij}=# occ. of $w_i\in\mathcal{V}$ in ctxt of $w_j\in\mathcal{C}$
$\mathbf{N}$ is sparse & computed in one pass.
Objective least square $\mathcal{H}(\theta,\mathbf{N})=\sum_{i,j}f(n_{ij})(\log n_{ij}-\log\tilde{p_\theta}(w_i|w_j))^2$
unnormalised distribution (model): $\tilde{p_\theta}(w_i|w_j)=\exp[\langle\mathbf{x}_i,\mathbf{y}_j\rangle+b_i+c_j]$,
n_{ij} is the target and f weight func
$f(n)=\min\{1,(\frac{n}{n_{\max}})^\alpha\},\alpha\in[0,1]$,
limits the influence of large and small noisy counts
Solves $\min_{\mathbf{XY}}\|\mathbf{N}-\mathbf{X}^\top\mathbf{Y}\|_F^2$if $f=1$ But non-convex, hard to find optimum, full gradient descent too expensive to compute.$\Rightarrow$ use SGD

Mixtures
K-Means
Objective $J(\mathbf{U},\mathbf{Z})=\sum_i^N\sum_j^K z_{ij}\|\mathbf{x}_i-\mathbf{u}_j\|^2$
$=\|\mathbf{X}-\mathbf{U}^T\mathbf{Z}\|^2$, $\sum_{k=1}^K z_{ki}=1$
$\mathbf{X}=[\mathbf{x}_1\cdots\mathbf{x}_N]\in\mathbb{R}^{D\times N}$ data matrix
$\mathbf{U}=[\mathbf{u}_1\cdots\mathbf{u}_K]\in\mathbb{R}^{D\times K}$ centroids
Recursive optimal assignments
E: $z_{ij}^*=\{1$ if $j=\arg\min_k\|\mathbf{x}_i-\mathbf{u}_k\|^2$
M: $\nabla_{\mathbf{u}_j}J(\mathbf{U},\mathbf{Z})\stackrel{!}{=}0\Rightarrow\mathbf{u}_j^*=\frac{\sum_{i=1}^N z_{ij}\mathbf{x}_i}{\sum_{i=1}^N z_{ij}}$
guaranteed convergence but non-convex objective
K-means solves $\arg\min_Z\|\mathbf{X}-\mathbf{UZ}\|_F^2$
K-Means++
Init. with incremental D^2 sampling
sample first centroid rand. $U_1=\{\mathbf{x}_I\}$
$D_i=\min_{\mathbf{u}\in U_k\|\mathbf{x}_i-\mathbf{u}\|}$ $U_{k+1}=U_k\cup\{\mathbf{x}_I\}$
for $I\sim\mathrm{Cat}(\mathrm{p})$, $p_i=D_i^2/\sum_{j=1}^N D_j^2$
More expensive but better results.
K-Means Core Sets
sample core set of m centroids
$I\sim\mathrm{Cat}(\mathrm{p})$, $p_i=1/2N+D_i^2/2\sum_{j=1}^N D_j^2$
where $D_i^2=\|\mathbf{x}_i-\frac{1}{N}\sum_{i=1}^N\mathbf{x}_i\|^2$
give each sample weight $1/mp_i$ and weighted K-Mean on this core-set.
Finite Mixture Model
Probabilistic assignment to clusters.
$p(\mathbf{x},\theta)=\sum_{j=1}^K\pi_j p(\mathbf{x};\theta)$
where $\theta=(\pi,\theta_1,\cdots,\theta_K)$
and $\pi\geq 0,\sum_{i=1}^K\pi_i=1$
e.g Gaussian Mixture Model
$p(\mathbf{x};\theta_j)=p(\mathbf{x};\mu_\mathbf{j},\Sigma_\mathbf{j})$ (normal dist.)
Complete data distribution:
$p(\mathbf{x},\mathbf{z};\theta)=\prod_{j=1}^K(\pi_j p(\mathbf{x};\theta_j))^{z_j}$ where z_j latent and $P(z_j=1)=\pi_j$
Posterior Probabilities: $p(\mathbf{z}_k=1|\mathbf{x})=\frac{p(\mathbf{z}_k=1)p(\mathbf{x}|\mathbf{z}_k=1)}{\sum_{l=1}^K p(\mathbf{z}_l=1)p(\mathbf{x}|\mathbf{z}_l=1)}=\frac{\pi_k\mathcal{N}(\mathbf{x}|\mu_k,\Sigma_k)}{\sum_{l=1}^K\pi_l\mathcal{N}(\mathbf{x}|\mu_l,\Sigma_l)}$
Maximum Likelihood for MM
$\arg\max_\theta\sum_{i=1}^N log[\sum_{j=1}^K\pi_j p(\mathbf{x_i};\theta_j)]$
has no closed form solution
$log[\sum_{j=1}^K\pi_j p(\mathbf{x_i};\theta_j)]=$
$log[\sum_{j=1}^K q_j\frac{\pi_j p(\mathbf{x_i};\theta_j)}{q_j}]\geq$
$\sum_{j=1}^K q_j[\log p(\mathbf{x};\theta_j)+\log\pi_j-\log q_j]$
Lagrangian
$\mathcal{L}=\max_q\{\sum_{j=1}^K q_j[\log p(\mathbf{x};\theta_j)+\log\pi_j-\log q_j]+\lambda(\sum_{j=1}^K q_j-1)\}$
E-Step compute assignments
$\nabla q_j\stackrel{!}{=}0\Rightarrow q_j^*=\frac{\pi_j p(\mathbf{x};\theta_j)}{\sum_{l=1}^K\pi_l p(\mathbf{x};\theta_l)}=p(z_j=1|\mathbf{x})$
M-Step optimize clusters
$\mu_j^*=\frac{\sum_{i=1}^N q_{ij}\mathbf{x}_i}{\sum_{i=1}^N q_{ij}}$, $\pi_j:=\frac{1}{N}\sum_{i=1}^N q_{ij}$, and
$\Sigma_j=\frac{\sum_{i=1}^N q_{ij}(\mathbf{x}_n-\mu_j)(\mathbf{x}_j-\mu_j)^T}{\sum_{i=1}^N q_{ij}}$ EM requires more steps and computation to reach convergence than K-Mean

Singularities of GMM happen when a cluster shrinks to fit exactly one data point. log likelihood $\rightarrow\infty$. Bad convergence

Neural Networks
$F^\sigma(\mathbf{x};\mathbf{W})=\sigma(\mathbf{W}\mathbf{x})\Rightarrow F_j^\sigma(\mathbf{x};\mathbf{W})=\sigma(\mathbf{w}_j^\top\mathbf{x})$,
w. $\mathbf{W}=(\mathbf{w}_1\cdots\mathbf{w}_m)^T$ mapping $\mathbb{R}^n\rightarrow\mathbb{R}^m$ between 2 layers
$\mathbf{x}^{(l)}=\sigma(\mathbf{W}^{(l)}\mathbf{x}^{(l-1)}),1\leq l\leq L$
Activation Functions
Sig. $\sigma(x)=\frac{1}{1+\exp^{-x}},\nabla_x\sigma(x)=\sigma(x)(1-\sigma(x))$,
$\sigma^{-1}(x)=\log(1/(1-x))$, $1-\sigma(x)=\sigma(-x)$
ReLU $R(x)=\max\{0,x\}$ has simple derivative on $\mathbb{R}-\mathbf{0}$ and reduces vanishing gradient problem
Output Layer
Linear Regression $\mathbf{y}=\mathbf{W}^{(L)}\mathbf{x}^{(L-1)}$
Logistic binary classif. (one output)
$y_1=P(Y=1|\mathbf{x})=\frac{1}{1+\exp[-\mathbf{w}^T\mathbf{x}]}$
Soft-Max K-Multiclass
$y_k=P(Y=k|\mathbf{x})=\mathrm{e}[\mathbf{w}_k^T\mathbf{x}]/\sum_{j=1}^K\mathrm{e}[\mathbf{w}_j^T\mathbf{x}]$
Loss-Functions
Squared $l(y^*;y)=\frac{1}{2}(y^*-y)^2$
Cross-Entropy for classification
$l(y^*;y)=-y^*\log y-(1-y^*)\log(1-y)$
Empirical Risk
$\mathcal{L}(\theta;\mathcal{X})=\frac{1}{T}\sum_{l=1}^T l(y_t;y(\mathbf{x}_t;\theta))$ for weights $\theta=(\mathbf{W}^{(1)}\cdots\mathbf{W}^{(L)})$ and training data $\mathcal{X}=\{(\mathbf{x}_t,y_t),1\leq t\leq T\}$
Regularization
favors smaller weights
$\mathbf{L}_2$: $\mathcal{L}_\lambda(\theta;\mathcal{X})=\mathcal{L}(\theta;\mathcal{X})+\frac{\lambda}{2}\|\theta\|_2^2$
Dropout training with noise
Backpropagation
costs $\mathcal{O}(n)$ for NN with n nodes
$\frac{\partial x_i^{(l)}}{\partial x_k^{(l-n)}}=\sum_j\frac{\partial x_i^{(l)}}{\partial x_j^{(l-1)}}\frac{\partial x_j^{(l-1)}}{\partial x_k^{(l-n)}}=\sum_j\mathbf{J}_{ij}^{(l)}\frac{\partial x_j^{(l-1)}}{\partial x_k^{(l-n)}}$
$\frac{\partial\mathbf{x}^{(l)}}{\partial\mathbf{x}^{(l-n)}}=\sum_j\mathbf{J}^{(l)}\frac{\partial\mathbf{x}^{(l-1)}}{\partial\mathbf{x}^{(l-n)}}=\mathbf{J}^{(l)}\cdots\mathbf{J}^{(l-n+1)}$
Backprop: $\nabla_{\mathbf{x}^{(l)}}^T l=\nabla_\mathbf{y}^T l\cdot\mathbf{J}^{(l)}\cdots\mathbf{J}^{(l+1)}$
$\frac{\partial l}{\partial w_{ij}^{(l)}}=\frac{\partial l}{\partial x_i^{(l)}}\frac{\partial x_i^{(l)}}{\partial w_{ij}^{(l)}},\frac{\partial x_i^{(l)}}{\partial w_{ij}^{(l)}}=\sigma'(\mathbf{w}_i^{(l)T}\mathbf{x}^{(l-1)})x_j^{(l-1)}$
CNN
Convolutional Layers $F_{n,m}(\mathbf{x};\mathbf{w})$
$=\sigma(b+\sum_{k=-i}^i\sum_{l=-i}^i w_{kl}x_{n+k,m+l})$
Weight sharing and shift-invariant filtering thus less parameters and computational power required
Pooling Take avg or max over wind. Reduce size or extract features.

Generative Models
VAE - Variatonal Autoencoder
find $\mathbf{x}\in\mathbb{R}^n$ by sampling $\mathbf{z}\in\mathbb{R}^m$ from simple distribution and set $\mathbf{x}=F_\theta(\mathbf{z})$ where $F_\theta:\mathbb{R}^m\rightarrow\mathbb{R}^n$, deterministic DNN, since $\mathbb{E}_\mathbf{x}[f(\mathbf{x})]=\mathbb{E}_\mathbf{z}[f(F_\theta(\mathbf{z}))]$
Requires F_θ^{-1} often impossible to get.
ELBO - evidence lower bound
learn parameters of distribution p_θ instead of deterministic F_θ
$\log p_\theta(\mathbf{x})=\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(\mathbf{x})]=$
$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log\frac{p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})q_\theta(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}|\mathbf{x})q_\theta(\mathbf{z}|\mathbf{x})}]=$
$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]-$
$KL(q_\phi(\mathbf{z}|\mathbf{x})|p_\theta(\mathbf{z}))+KL(q_\phi(\mathbf{z}|\mathbf{x})|p_\theta(\mathbf{z}|\mathbf{x}))\geq$
$\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(\mathbf{x}|\mathbf{z})]-KL(q_\phi(\mathbf{z}|\mathbf{x})|p_\theta(\mathbf{z}))$
$=\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[-\log q_\phi(\mathbf{z}|\mathbf{x})+\log p_\theta(\mathbf{x},\mathbf{z})]=$
$=ELBO(\phi,\theta)=\mathcal{L}(\mathbf{x};\theta)$
maxim. wrt ϕ for inference model
maxim. wrt θ for generative model
stochastic approximation
Update for generative model:
$\nabla_\theta\mathbb{E}_{q_\phi}[\log p_\theta(\mathbf{x}|\mathbf{z})]=\mathbb{E}_{q_\phi}[\nabla_\theta\log p(\mathbf{x}|\mathbf{z})]$
$\approx\frac{1}{L}\sum_{r=1}^L\nabla_\theta\log p(\mathbf{x}|\mathbf{z}^{(r)})$, $z^{(r)}\sim q_\phi(\cdot|\mathbf{x})$
(by Monte Carlo approximation)
Update for inference model:
$\nabla_\phi\mathbb{E}_{q_\phi}[\mathcal{L}(\mathbf{x},\mathbf{z})]=\int\mathcal{L}(\mathbf{x},\mathbf{z})\nabla_\phi q_\phi(\mathbf{z}|\mathbf{x})dz$
$=\mathbb{E}_{q_\phi}[\mathcal{L}(\mathbf{x},\mathbf{z})\nabla_\phi\log q_\phi(\mathbf{z}|\mathbf{x})]$ hard!
variance in gradient usually high
Reparametrization Trick
$\mathbf{z}=g_\phi(\zeta|\mathbf{x}),\zeta\sim$ simple distribution
$\nabla_\phi\mathbb{E}_{q_\phi}[\mathcal{L}(\mathbf{x},\mathbf{z})]\approx\frac{1}{L}\sum_{r=1}^L\nabla_\phi\mathcal{L}(\mathbf{x},g_\phi(\zeta^{(r)}))$
easier to sample and differentiable by ϕ since the $\mathbf{z}$ are now deterministic. backprop. can be used
GAN-Generative Adversarial Net.
generator G and discriminator D learn to fool each others.
$\min_G\max_D\mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}[\log D(\mathbf{x})]$
$+\mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})}[\log(1-D(G(\mathbf{z}))]$
Hard to train, might not converge/learn to generate a few samples generates sharper images since it can predict high frequency details. But can only generate/sample

Sparse Coding
Basis Transformation
signals often allow sparse representation. Vanishing coeffs due to regularity. **Find orthonormal dictionary** $\mathcal{U}=\{\mathbf{u}_1,\cdots,\mathbf{u}_L\}$ and change of basis.
$\mathbf{x}=\mathbf{Uz}\rightarrow\mathbf{z}=\mathbf{U}^T\mathbf{x}\rightarrow\hat{\mathbf{x}}=\mathbf{U}\hat{\mathbf{z}}$
$\hat{\mathbf{x}}=\sum_{d\in\sigma}\hat{z}_d(\mathbf{x})\mathbf{u}_d$, $\hat{z}_d(\mathbf{x})=\langle\mathbf{x},\mathbf{u}_d\rangle$
Error $\|\mathbf{x}-\hat{\mathbf{x}}\|^2=\sum_{d\notin\sigma}\langle\mathbf{x},\mathbf{u_d}\rangle^2$ e.g. fourier transform denoises well but not for localised signals $\mathbf{O}(D\log D)$
Haar Wavelets
$\psi_{n,k}(t)=2^{n/2}\psi(2^n t-k),0\leq k\leq 2^n$
good for localized signal but poor for denoising smooth signal $\mathbf{O}(D\log D)$
Overcomplete Dictionaries
dictionaries generalisation
overcompleteness
$\mathbf{U}=[\mathbf{u_1}\cdots\mathbf{u_L}],\mathbf{U}\in\mathbb{R}^{D\times L},L>D$
$\mathbf{z}\in\mathbb{R}^L$ contains coeffs. of signal $\mathbf{x}\in\mathbb{R}^D$ in base $\mathbf{U}$. $\mathbf{z}$ is not unique.
Search $\mathbf{z}^*=\arg\min_{\mathbf{z}\in\mathbb{R}^l}\|\mathbf{z}\|_0$ s.t. $\mathbf{Uz}=\mathbf{x}$
Problem is NP-hard/Ill-posed.
Coherence
$L/D\uparrow\Rightarrow$ sparsity $\uparrow,\mathbf{u_i}$ dependency $\uparrow$
$m(\mathbf{U})=\max_{i,j:i\neq j}|\mathbf{u_i}^T\mathbf{u_j}|$ coherence
$m(\mathbf{B})=0$ for orhtog. $\mathbf{B}$
$m([\mathbf{Bu}]\geq\frac{1}{\sqrt{D}}$ if $\mathbf{u}$ added to $\mathbf{B}$
Matching Pursuit greedy algorithm at each step chose dimension with max projection onto residual.
Init $\mathbf{r}_0=\mathbf{x}$, $\hat{\mathbf{x}}_0=\mathbf{0}$. Then repeat:
find $j^*=\arg\max_j|\langle\mathbf{r}_i,\mathbf{u}_j\rangle|$
$\hat{\mathbf{x}}_{i+1}\leftarrow\hat{\mathbf{x}}_i+\langle\mathbf{r}_i,\mathbf{u}_{j^*}\rangle\mathbf{u}_{j^*}$
$\mathbf{r}_{i+1}\leftarrow\mathbf{r}_i-\langle\mathbf{r}_i,\mathbf{u}_{j^*}\rangle\mathbf{u}_{j^*}$
Convergence greedily reduces residual energy at each step.
$\|\mathbf{r}_i\|_2^2=\|\mathbf{r}_{i+1}\|_2^2+|\langle\mathbf{r}_i,\mathbf{u}_{j^*}\rangle|^2$ (by energy conservation and linearity)
Convex Optimization with l_1-norm
$\mathbf{z}^*=\arg\min_{\mathbf{z}\in\mathbb{R}^l}\|\mathbf{z}\|_1$ s.t. $\mathbf{Uz}=\mathbf{x}$
can approximate l_0 even same results

Dictionary Learning
learn one dictionary for $\mathbf{x}_1\cdots\mathbf{x}_N$ obj. $(\mathbf{U}^*,\mathbf{Z}^*)\in\arg\min_{\mathbf{UZ}}\|\mathbf{x}-\mathbf{UZ}\|_F^2$ not jointly convex in $(\mathbf{U},\mathbf{Z})$
Greedy Convex Minimization
1.Coding $\mathbf{Z}^{t+1}\in\arg\min_\mathbf{Z}\|\mathbf{X}-\mathbf{U}^t\mathbf{Z}\|_F^2$
where $\mathbf{Z}$ sparse and $\mathbf{U}$ fixed
column separable, $\forall n=1\cdots N$
$\mathbf{z}_n^{t+1}\in\mathrm{armin}_\mathbf{z}\|\mathbf{z}\|_0\mathrm{st}\|\mathbf{x}_n-\mathbf{U}^t\mathbf{z}\|^2\leq\sigma\|\mathbf{x}_n\|_2$
2.Update $\mathbf{U}^{t+1}\in\arg\min_\mathbf{U}\|\mathbf{X}-\mathbf{UZ}^{t+1}\|_F^2$
where $\|\mathbf{u}_l\|_2=1,\forall l$ and $\mathbf{Z}$ fixed.
not separable $\mathbf{R}_l^t$ residual of atom $\mathbf{u}_\ell$
$\|\mathbf{X}-[\mathbf{u}_1^t\cdots\mathbf{u}_\ell\cdots\mathbf{u}_L^t]\mathbf{Z}^{t+1}\|_F^2=$
$\|\mathbf{X}-(\sum_{e\neq\ell}\mathbf{u}_e^t(\mathbf{z}_e^{t+1})^T+\mathbf{u}_\ell\mathbf{z}_\ell^{t+1T})\|_F^2=$
$\|\mathbf{R}_\ell^t-\mathbf{u}\ell(\mathbf{z}_\ell^{t+1T})\|$ approx. by SVD.
$\mathbf{R}_\ell^t=\tilde{\mathbf{U}}\Sigma\tilde{\mathbf{V}}^T$, $\mathbf{u}_\ell^*=\tilde{\mathbf{u}}_1$ first. sing vec.