The EM Algorithm and Principal Component Analysis

Anders Hansson and Martin Andersen

Linköping University and Technical Unversity of Denmark

April 18, 2024

Relative Entropy

Difference between pdf:s measured by *relative entropy* or equivalentely *Kullback-Leibler divergence*. Let

$$\mathcal{D}_n = \left\{ p \in \mathbf{R}^{\mathbf{R}^n} : \int_{\mathbf{R}^n} p(x)dx = 1,\ p(x) \geq 0,\ \forall x \in \mathbf{R}^n \right\}$$

Define realtive entropy $D : \mathcal{D}_n \times \mathcal{D}_n \to \mathbf{R}$ by

$$D(p, q) = -\int_{\mathbf{R}^n} p(x) \ln \frac{q(x)}{p(x)} dx$$

which is a convex function of p, since $q(x) \geq 0,\ \forall x$.

Gibb's Inequality

What p minimizes $D(p,q)$? Clearly $p = q$ makes the relative entropy zero.

For a convex function $\varphi : \mathbf{R} \to \mathbf{R}$, any function $f : \mathbf{R}^n \to \mathbf{R}$, and any pdf p it holds by Jensen's inequality that

$$\varphi\left(\int_{\mathbf{R}^n} f(x)p(x)dx\right) \leq \int_{\mathbf{R}^n} \varphi\left(f(x)\right)p(x)dx$$

We let $f(x) = q(x)/p(x)$, and $\varphi(f) = -\ln(f)$ and obtain

$$D(p,q) \geq -\ln \int_{\mathbf{R}^n} \frac{q(x)}{p(x)}p(x)dx = -\ln \int_{\mathbf{R}^n} q(x)dx = 0$$

with equality if and only if $p(x) = q(x)$ for allmost all x.

However, not a metric, since in general $D(p,q) \neq D(q,p)$.

Cross Entropy

Cross entropy defined as $C : \mathcal{D}_n \times \mathcal{D}_n \to \mathbf{R}$, where

$$C(p, q) = -\int_{\mathbf{R}^n} p(x) \ln q(x) dx$$

With $H(p)$ entropy of p

$$C(p, q) = D(p, q) + H(p)$$

Hence

$$C(p, q) \geq H(p)$$

with equality if and only if $p(x) = q(x)$ for almost all x.

Also

$$C(p, q) = -E_p(\ln q)$$

where $E_p : \mathcal{D}_n \to \mathbf{R}$ is defined by

$$E_p(f) = \int_{\mathbf{R}^n} f(x) p(x) dx$$

Cross Entropy and ML estimation

If we consider q to be parameterized with a parmeter $\theta \in \mathbf{R}^m$, i.e. $q : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}_+$, then the ML problem is equivalent to

$$\text{minimize} \ -\sum_{k=1}^{N} \ln q(x_k; \theta)$$

with variable θ, where x_k, $k \in \mathbf{N}_N$ are the observed data of x.

If we assume that the distribution for x is given by p, then the objective function above is proportional to a sample estimate of the cross entropy.

Latent Variables or Missing Data

Difficult ML problem with likelihood function $\ell : \mathbf{R}^N \times \mathbf{R}^p \to \mathbf{R}_+$ based on observations $y \in \mathbf{R}^N$ parameterized with $\theta \in \mathbf{R}^p$:

$$\text{maximize} \ln \ell(y; \theta)$$

with variable θ. With additional observations $z \in \mathbf{R}^M$ the ML problem with (y, z) as observations **easy**.

Infinite Dimensional Optimization Problem

Let $f_X : \mathbf{R}^N \times \mathbf{R}^M \times \mathbf{R}^p \to \mathbf{R}_+$ be the joint pdf for random variable $X = (Y, Z)$ with parameter θ.

Define conditional pdf $f_{Z|Y} : \mathbf{R}^N \times \mathbf{R}^M \times \mathbf{R}^p \to \mathbf{R}_+$:

$$f_{Z|Y}(z|y;\theta) = \frac{f_X(y,z;\theta)}{\ell(y;\theta)}$$

Consider arbitrary distribution $q \in \mathcal{D}_M$.

Define optimization problem

$$\text{maximize} \ln \ell(y;\theta) - D(q, f_{Z|Y})$$

with variables (θ, q), where D is the relative entropy. (This trivially has same solution as ML problem, and $q = f_{Z|Y}$ is optimal)

Block-Coordinate Accent

It holds

$$\ln \ell(y;\theta) - D(q, f_{Z|Y}) = \ln \ell(y;\theta) + H(q) - C(q, f_{Z|Y}) = H(q) - C(q, f_X)$$

where H is entropy and C is cross-entropy.

Iterate:

1. Fix θ and optimize with respect to q.
2. Fix q and optimize with respect to θ.

First step has trivial solution $q = f_{Z|Y}$. ($f_{Z|Y}$ depends on old value of θ denoted $\bar{\theta}$)

Second step equivalent to

$$\text{maximize} Q(\theta, \bar{\theta})$$

with variable θ, where

$$Q(\theta, \bar{\theta}) = -C(f_{Z|Y}, f_X) = E_Z(\ln f_X(Y, Z; \theta) \mid Y = y; \bar{\theta})$$

The Expectation Maximization (EM) Algorithm

Iterate:

1. Form $Q(\theta, \bar{\theta}) = E_Z(\ln f_X(Y, Z; \theta) \mid Y = y; \bar{\theta})$ (E-step)
2. Solve maximize$Q(\theta, \bar{\theta})$ (M-step)

Sometimes E-step is approximated by an empirical cross-entropy, i.e.

$$E_{f_{Z|Y}}(\ln f_X) \approx \frac{1}{M} \sum_{i=1}^{M} \ln f_i$$

where $f_i = f_X(y, z_i; \theta)$ are obtained by drawing samples z_i from the conditional pdf $f_{Z|Y}$. This is called *Monte Carlo EM*.

Gibbs Sampling

How to draw samples from a pdf $p : \mathcal{D}_1 \times \cdots \times \mathcal{D}_n \to \mathbf{R}_+$?

Define the conditional pdf $p_{i|\setminus i} : \mathcal{D}_i \to \mathbf{R}_+$, $i \in \mathbf{N}_n$ as the pdf for $x_i \in \mathcal{D}_i$ given $x_{\setminus i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$.

Let $x^{(0)} = \left(x_1^{(0)}, \ldots, x_n^{(0)}\right) \in \mathcal{D}_1 \times \cdots \times \mathcal{D}_n$ be given. Then for $k \in \mathbf{N}$ we compute $x_i^{(k+1)}$ by drawing a sample from $p_{i|\setminus i}(x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}, x_i, x_{i+1}^{(k)}, \ldots, x_n^{(k)})$.

Instead of cycling through the components make a random permutation of them for each k.

$x^{(k)}$ converges to a sample of p as k goes to infinity.

Boltzmann Machine

Ising distribution:

$$p(x) = \exp\left(-E(x) - A(\lambda, \Lambda)\right)$$

where $E(x) = -\lambda^T x - \mathbf{tr}\,\Lambda x x^T$ and $A(\lambda, \Lambda) = \ln \sum_x \exp(-E(x))$ for $x \in \mathcal{D}_x$.

Divide x into *visible* and *hidden* or latent variables, i.e. $x = (v, h)$.

Consider N observations v_i, $i \in \mathbf{N}_N$ of v.

Corresponding hidden variables are h_i, $i \in \mathbf{N}_N$.

Define $x_i = (v_i, h_i)$, and with abuse of notation $v = (v_1, \ldots, v_N)$, $h = (h_1, \ldots, h_N)$ and $x = (v, h)$.

Likelihood Function

Define $r : \mathcal{D}_x^N \times \mathbf{R}^m \times \mathbf{S}^m \to \mathbf{R}$ via

$$r(x, \lambda, \Lambda) = \prod_{i=1}^{N} p(x_i) = e^{-\bar{E}(x,\lambda,\Lambda) - NA(\lambda,\Lambda)}$$

where $\bar{E} : \mathcal{D}_x^N \times \mathbf{R}^m \times \mathbf{S}^m \to \mathbf{R}$ is given by $\bar{E}(x, \lambda, \Lambda) = \sum_{i=1}^{N} E(x_i)$.

Conditional distribution for h given v is $s : \mathcal{D}_v^N \times \mathcal{D}_h^N \times \mathbf{R}^m \times \mathbf{S}^m \to \mathbf{R}$ given by

$$s(v, h, \lambda, \Lambda) = \frac{r(v, h, \lambda, \Lambda)}{\sum_h r(v, h, \lambda, \Lambda)} = \frac{e^{-\bar{E}(v,h,\lambda,\Lambda)}}{\sum_h e^{-\bar{E}(v,h,\lambda,\Lambda)}}$$

where $\mathcal{D}_v$ and $\mathcal{D}_h$ are defined such that $\mathcal{D}_v \times \mathcal{D}_v = \mathcal{D}_x$.

Q-function in EM Algorithm

Define $Q : \mathbf{R}^m \times \mathbf{S}^m \to \mathbf{R}$ via

$$Q(\lambda, \Lambda) = E_s \ln r = \sum_h s(v, h, \lambda^-, \Lambda^-) \left(-\bar{E}(v, h, \lambda, \Lambda) - NA(\lambda, \Lambda)\right)$$

$$= -\sum_h s(v, h, \lambda^-, \Lambda^-)\bar{E}(v, h, \lambda, \Lambda) - NA(\lambda, \Lambda)$$

where (λ^-, Λ^-) values of the parameters from the previous iteration in the EM lgorithm.

Define the conditional distribution for h_i given v_i as $s_i : \mathcal{D}_v \times \mathcal{D}_h \times \mathbf{R}^m \times \mathbf{S}^m \to \mathbf{R}$ via

$$s_i(v_i, h_i, \lambda, \Lambda) = \frac{p(x_i)}{\sum_{h_i} p(x_i)} = \frac{e^{-E(v_i, h_i, \lambda, \Lambda)}}{\sum_{h_i} e^{-E(v_i, h_i, \lambda, \Lambda)}}$$

Gradients

$$\frac{\partial Q}{\partial \lambda} = -\sum_{h} s(h, v, \lambda^{-}, \Lambda^{-}) \frac{\partial \bar{E}}{\partial \lambda} - N \frac{\partial A}{\partial \lambda}$$

$$= \sum_{h} s(h, v, \lambda^{-}, \Lambda^{-}) \sum_{i=1}^{N} x_i - N \sum_{\xi} \xi p(\xi)$$

$$= \sum_{i=1}^{N} \sum_{h_i} s_i(h_i, v_i, \lambda^{-}, \Lambda^{-}) x_i - N \sum_{\xi} \xi p(\xi)$$

$$\frac{\partial Q}{\partial \Lambda} = \sum_{i=1}^{N} \sum_{h_i} s_i(h_i, v_i, \lambda^{-}, \Lambda^{-}) x_i x_i^{T} - N \sum_{\xi} \xi \xi^{T} p(\xi)$$

Principal Component Analysis

Given data $x_i \in \mathbf{R}^n$, $i \in \mathbf{N}_N$ should be *approximated* with $W^T c_i$, where $W \in \mathbf{R}^{m \times n}$, $m < n$, and where $c_i \in \mathbf{R}^m$, $i \in \mathbf{N}_N$. Require $WW^T = I$.

Let $J : \mathbf{R}^{m \times n} \times \mathbf{R}^n \times \cdots \times \mathbf{R}^n \to \mathbf{R}_+$ be defined by

$$J(W, c_1, \ldots, c_N) = \frac{1}{2} \sum_{i=1}^{N} \|x_i - W^T c_i\|_2^2$$

Optimization problem:

$$\text{minimize } J(W, c_1, \ldots, c_N) \tag{1}$$

$$\text{subject to } WW^T = I \tag{2}$$

with variables $(W, c_1, \ldots, c_N)$. Convex problem for c_i when W fixed.

Optimization of c_i

$$\frac{\partial J}{\partial c_i} = WW^T c_i - Wx_i = 0, \quad i \in \mathbf{N}_N$$

$$WW^T = I \Rightarrow c_i = Wx_i \Rightarrow$$

$$J(L, Wx_1, \ldots, Wx_N) = \frac{1}{2} \sum_{i=1}^{N} x_i^T \left(I - W^T W \right)^2 x_i$$

$$= \frac{1}{2} \sum_{i=1}^{N} x_i^T \left(I - W^T W \right) x_i$$

Equivalent optimization problem:

$$\text{maximize } \frac{1}{2} \mathbf{tr}\, WX^T XW^T \tag{3}$$

$$\text{subject to } WW^T = I \tag{4}$$

with variable W, where

$$X^T = \begin{bmatrix} x_1 & \cdots & x_N \end{bmatrix}$$

Equivalent Problem

Let $X = UDV^T$ be an SVD and define $Y = V^T W^T$ and $\tilde{J} : \mathbf{R}^{n \times m} \to \mathbf{R}_+$ where

$$\tilde{J}(Y) = \frac{1}{2} \mathbf{tr}\, Y^T D^2 Y$$

and equivalent problem

$$\text{maximize } \tilde{J}(Y) \tag{5}$$

$$\text{subject to } Y^T Y = I \tag{6}$$

with variable Y. Lagrangian $L : \mathbf{R}^{n \times m} \times \mathbf{S}^m \to \mathbf{R}$ where

$$L(Y, \Lambda) = \tilde{J}(L) + \mathbf{tr}\, \Lambda \left(I - Y^T Y \right)$$

Necessary condition for optimal Y: $\exists \Lambda$:

$$\frac{\partial L}{\partial Y} = D^2 Y - Y\Lambda = 0$$

Optimality Conditions

Let Z be such that $\begin{bmatrix} Y & Z \end{bmatrix}^T \begin{bmatrix} Y & Z \end{bmatrix} = I$.

Then equivalent optimality condition:

$$Y^T D^2 Y - \Lambda = 0 \tag{7}$$

$$Z^T D^2 Y = 0 \tag{8}$$

First equation always has a solution Λ for any Y.

Hence the necessary conditions are $\exists Z$:

$$\begin{bmatrix} Y & Z \end{bmatrix}^T \begin{bmatrix} Y & Z \end{bmatrix} = I \tag{9}$$

$$Z^T D^2 Y = 0 \tag{10}$$

One solution is $\begin{bmatrix} Y & Z \end{bmatrix} = I$. Also $\begin{bmatrix} Y & Z \end{bmatrix} = X_1 \oplus X_2$ for any orthogonal $X_1 \in \mathbf{R}^{m \times m}$ and $X_2 \in \mathbf{R}^{(n-m) \times m}$.

Proof of Optimality

Objective function:

$$\frac{1}{2}\sum_{i=1}^{N}\|Dy_i\|_2^2$$

where $Y = \begin{bmatrix} y_1 & \cdots & y_m \end{bmatrix}$, and

$$Dy_1 = \begin{bmatrix} d_1 y_{11} \\ d_2 y_{21} \\ \vdots \\ d_n y_{n1} \end{bmatrix}$$

Optimize with respect to y_1 fist. Since $d_i \geq d_j$ when $i < j$ and ${y_i}^T y_i = 1$ it is optimal to take $y_1 = e_1$.

Remaining y_i for $2 \leq i \leq m$ has to have its first components equal to zero in order to be orthogonal to y_1.

Because of this $y_2 = e_2$ and so on $\Rightarrow$ $Y = \begin{bmatrix} I \\ 0 \end{bmatrix}$ optimal.

Non-Uniqueness

With $Y = \begin{bmatrix} X_1 \\ 0 \end{bmatrix}$ where X_1 orthogonal it holds

$$\tilde{J}\left(\begin{bmatrix} X_1 \\ 0 \end{bmatrix}\right) = \frac{1}{2}\,\mathbf{tr}\left(\begin{bmatrix} X_1 \\ 0 \end{bmatrix}^T D^2 \begin{bmatrix} X_1 \\ 0 \end{bmatrix}\right) = \frac{1}{2}\,\mathbf{tr}\left(X_1^T D_1^2 X_1\right)$$
$$= \frac{1}{2}\,\mathbf{tr}\, X_1 X_1^T D_1^2 = \frac{1}{2}\,\mathbf{tr}\, D_1^2$$

where $D = \mathbf{bdiag}(D_1, D_2)$.

Hence the principal component analysis picks out the components of x_i corresponding to the m largest singular values of X^T.

Low Rank Approximation

We approximate x_i with $W^T c_i$, where $c_i = W x_i$. Because of this X is approximated with $XW^T W$.

Since W is an orthogonal matrix, $W^T W$ is a projection matrix, and hence rank $XW^T W = m$.

We have $XW^T W = UDV^T V_1 V_1^T = UD \begin{bmatrix} I \\ 0 \end{bmatrix} V_1^T = U_1 D_1 V_1^T$,
where $U = \begin{bmatrix} U_1 & U_2 \end{bmatrix}$ and $V = \begin{bmatrix} V_1 & V_2 \end{bmatrix}$.

Mutual Information

For joint pdf $r : \mathbf{R}^m \times \mathbf{R}^n \to \mathbf{R}_+$ with marginal pdf:s $p : \mathbf{R}^m \to \mathbf{R}_+$ and $q : \mathbf{R}^n \to \mathbf{R}_+$ mutual information $I : \mathcal{D}_{m+n} \to \mathbf{R}_+$ is

$$I(r) = \int_{\mathbf{R}^m \times \mathbf{R}^n} r(x,y) \ln \frac{r(x,y)}{p(x)q(y)} dxdy$$

Consider two zero mean n-dimensional independent random variables X and E with normal distributions with covariances $\Sigma \in \mathbf{S}_+^n$ and I. Let $Z = WX$ and $Y = Z + E$, where $W \in \mathbf{R}^{m \times n}$.

(Y, Z) has zero mean normal pdf r with covariance

$$\begin{bmatrix} W\Sigma W^T + I & W\Sigma W^T \\ W\Sigma W^T & W\Sigma W^T \end{bmatrix}$$

We let p and q be the marginal pdf:s for Y and Z and define $J : \mathbf{R}^{m \times n} \to \mathbf{R}_+$ as

$$J(W) = I(r) = \frac{1}{2} \ln \det \left(I + W\Sigma W^T \right)$$

Optimization Problem

$$\text{maximize } J(W) \tag{11}$$
$$\text{subject to } WW^T = I \tag{12}$$

with variable W. Constraint will make objective bounded.

Let $V : V^T V = I$ and $\Sigma = VD^2V^T$ with D diagonal. Then

$$J(W) = \frac{1}{2} \ln \det \left(I + \bar{Y}^T D^2 \bar{Y} \right)$$

where $\bar{Y} = V^T W^T$. Assume $d_i \geq d_j$ for all $i < j$.

Equivalent Optimization Problem

$$\text{maximize } \tilde{J}(\bar{Y}) \tag{13}$$

$$\text{subject to } \bar{Y}^T \bar{Y} = I \tag{14}$$

with variable $\bar{Y}$, where $\tilde{J} : \mathbf{R}^{n\times m} \to \mathbf{R}_+$ with $\tilde{J}(\bar{Y}) = J(W)$.

Lagrangian $L : \mathbf{R}^{n\times m} \times \mathbf{S}^m \to \mathbf{R}$ where

$$L(\bar{Y}, \Lambda) = \frac{1}{2} \ln \det \left(I + \bar{Y}^T D^2 \bar{Y} \right) + \frac{1}{2} \mathbf{tr}\, \Lambda \left(I - \bar{Y}^T \bar{Y} \right)$$

The existence of Λ :

$$\frac{\partial L}{\partial \bar{Y}} = D^2 \bar{Y} \left(I + \bar{Y}^T D^2 \bar{Y} \right)^{-1} - \bar{Y} \Lambda = 0$$

is necessary condition for optimality of $\bar{Y}$.

Equivalent Optimality Conditions

Let $\bar{Z}$ be s.t. $\begin{bmatrix} \bar{Y} & \bar{Z} \end{bmatrix}$ is square and orthogonal.

Equivalent optimality conditions: $\exists \bar{Z}$ and Λ:

$$\bar{Y}^T D^2 \bar{Y} \left(I + \bar{Y}^T D^2 \bar{Y} \right)^{-1} - \Lambda = 0 \tag{15}$$

$$\bar{Z}^T D^2 \bar{Y} = 0 \tag{16}$$

with $\begin{bmatrix} \bar{Y} & \bar{Z} \end{bmatrix}$ orthogonal. From $A(I + A)^{-1} = I - (I + A)^{-1}$

$$I - \left(I + \bar{Y}^T D^2 \bar{Y} \right)^{-1} - \Lambda = 0 \tag{17}$$

$$\bar{Z}^T D^2 \bar{Y} = 0 \tag{18}$$

First equation has a solution in terms of Λ for any $\bar{Y}$.

Otimality conditions: $\exists \bar{Z} \in \mathbf{R}^{n \times (n-m)}$ such that

$$\bar{Z}^T D^2 \bar{Y} = 0 \tag{19}$$

$$\begin{bmatrix} \bar{Y} & \bar{Z} \end{bmatrix}^T \begin{bmatrix} \bar{Y} & \bar{Z} \end{bmatrix} = I \tag{20}$$

Relation to PCA

Same optimality conditions as for PCA if we identify Σ with $X^T X$.

However, the objective functions are not the same.

As in the previous section, $\begin{bmatrix} \bar{Y} & \bar{Z} \end{bmatrix} = \mathbf{bdiag}(X_1, X_2)$ with $X_1 \in \mathbf{R}^{m \times m}$ and $X_2 \in \mathbf{R}^{(n-m)\times(n-m)}$ orthogonal is solution to optimality conditions.

It holds

$$\tilde{J}(\bar{Y}) = \frac{1}{2} \sum_{k \in \mathbf{N}_m} \ln\left(1 + d_k^2\right)$$

It can be shown using Givens rotations that no other orthogonal $\begin{bmatrix} \bar{Y} & \bar{Z} \end{bmatrix}$ can be optimal.

Non-orthogonal y_i

$$\text{maximize } \tilde{J}(\bar{Y}) \tag{21}$$

$$\text{subject to } \|y_i\|_2^2 = 1, \qquad i \in \mathbf{N}_m \tag{22}$$

with variable $\bar{Y}$. Lagrangian $M : \mathbf{R}^{n\times m} \times \mathbf{R}^m \to \mathbf{R}$ where

$$M(\bar{Y}, \lambda) = \frac{1}{2} \ln\det\left(I + \bar{Y}^T D^2 \bar{Y}\right) + \frac{1}{2}\sum_{i=1}^{m} \lambda_i \left(1 - \|y_i\|_2^2\right)$$

We see that

$$\frac{\partial M}{\partial \bar{Y}} = D^2 \bar{Y}\left(I + \bar{Y}^T D^2 \bar{Y}\right)^{-1} - \bar{Y}\,\mathbf{diag}(\lambda)$$

and necessary conditions for optimality: $\exists \bar{Z}$ and $\lambda \in \mathbf{R}^m$:

$$\bar{Y}^T D^2 \bar{Y}\left(I + \bar{Y}^T D^2 \bar{Y}\right)^{-1} - \bar{Y}^T \bar{Y}\,\mathbf{diag}(\lambda) = 0 \tag{23}$$

$$\bar{Z}^T D^2 \bar{Y} = 0 \tag{24}$$

where $\bar{Z}$ is such that $\begin{bmatrix}\bar{Y} & \bar{Z}\end{bmatrix}$ not necessarily square has full column rank.

Equivalent optimality conditions

$$I - \left(I + \bar{Y}^T D^2 \bar{Y}\right)^{-1} - \bar{Y}^T \bar{Y}\,\mathbf{diag}(\lambda) = 0 \tag{25}$$

$$\bar{Z}^T D^2 \bar{Y} = 0 \tag{26}$$

$\begin{bmatrix}\bar{Y} & \bar{Z}\end{bmatrix} = I$ is a solution. However, any orthogonal $\begin{bmatrix}\bar{Y} & \bar{Z}\end{bmatrix}$ will not satisfy them, since for orthogonal $\bar{Y}$ it must hold that $I + \bar{Y}^T D \bar{Y}$ is diagonal for there to exist λ that satisfies the equation.

There are however non-orthogonal $\bar{Y}$ that satisfy them. Consider $\bar{Y} = \begin{bmatrix}\mathbf{1} & 0\end{bmatrix}^T$. Then $\lambda = d_1^2/(1 + m d_1^2)\mathbf{1}$ satisfies the necessary optimality conditions.

It actually holds that we may take y_i equal to any basis vector for $\mathbf{R}^n$, and they may be linearly dependent.

What Stationary Points are Optimal?

Example where $m = 2$, $d_1 > d_2$.

For $\bar{Y} = \begin{bmatrix} \mathbf{1} & 0 \end{bmatrix}^T$ we have
$\ln\det(I + \bar{Y}^T D^2 \bar{Y}) = \ln(1 + 2d_1^2) \approx 2d_1^2$ for small values of d_1^2.

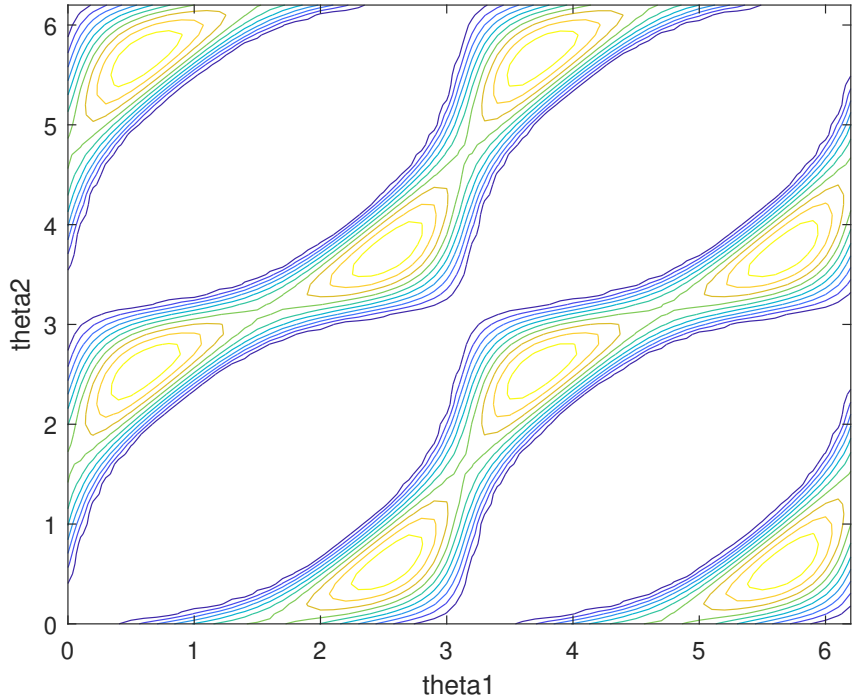
For $\bar{Y} = \begin{bmatrix} I \\ 0 \end{bmatrix}$ we have
$\ln\det(I + \bar{Y}^T D^2 \bar{Y}) = \ln(1 + d_1^2 + d_2^2 + d_1^2 d_2^2) \approx d_1^2 + d_2^2 + d_1^2 d_2^2$
for small values of d_1^2 and d_2^2.

Exist $d_1 > d_2$ for which second approximation is smaller than the first and for small SNR it is better to only consider the signal with the largest variance.

In general there are cases in-between, where one should pick out more than one of the largest, but not all m.

Non-Aligned and Non-Orthogonal Case

Example with $m = n = 2$ and let $y_1 = \begin{bmatrix} \cos\theta_1 \\ \sin\theta_1 \end{bmatrix}$ and $y_2 = \begin{bmatrix} \cos\theta_2 \\ \sin\theta_2 \end{bmatrix}$.



Several optima: e.g.

$$\bar{Y} = \begin{bmatrix} 0.83 & 0.83 \\ -0.56 & 0.56 \end{bmatrix}$$

which corresponds to $\theta_1 = 5.7$ and $\theta_2 = 0.6$. Other optima have same angle between y_1 and y_2.

Saddle Points and Large SNR

$$\bar{Y} = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}; \quad \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \quad \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}; \quad \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}$$

correspond to saddle points.

If the signal to noise ratio is large, then $\bar{Y} = \begin{bmatrix} X_1 \\ 0 \end{bmatrix}$ with X_1 orthogonal is optimal also without orthogonality constraints, since first equation of optimality conditions in the limit when D is much larger than I is given by

$$I - \bar{Y}^T \bar{Y} \, \mathbf{diag}(\lambda) = 0$$

Relationship to PCA

Assume observations x_i, $i \in \mathbf{N}_N$ of x that we want to approximate with $W^T z_i$, where $z_i = W x_i$. We estimate the covariance matrix of X with $\Sigma = \frac{1}{N} \bar{X}^T \bar{X}$, where

$$\bar{X} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix}$$

We then let $\bar{X} = UDV^T$ be a singular value decomposition of $\bar{X}$. Then

$$J(W) = \frac{1}{2} \ln \det \left(I + \bar{Y}^T \frac{D^2}{N} \bar{Y} \right)$$

where $\bar{Y} = V^T W^T$.

The approximation of $\bar{X}$ will be $\bar{X} W^T W$, where $W^T W$ is not necessarily a projection matrix, but rank of $\bar{X} W^T W$ will still be m.

Why Mutual Information is Superior to PCA

Notice that we cannot relax the condition on orthogonality in the principal component analysis without obtaining $\bar{Y} = \begin{bmatrix} \mathbf{1}^T \\ 0 \end{bmatrix}$ as the optimal solution.

The signal to noise ratio will not help in making the correct choice from an information point of view. This is why optimizing mutual information seems to be more appropriate.

Cluster Analysis

Partition observations $x_i \in \mathbf{R}^n$, $i \in \mathbf{N}_N$ into groups called *clusters* so that the observations in each cluster are close to one another.

We may measure distance $d : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}_+$ as e.g. $d(x_i, x_j) = \|x_i - x_j\|^2$.

Assume $K < N$ clusters and define *encoder* $C : \mathbf{N}_N \to \mathbf{N}_K$ that assigns each observation to each cluster.

The sum of all distances within each cluster should be minimized:

$$\text{minimize } f(C)$$

where $f : \mathbf{N}_K \to \mathbf{R}_+$ is defined by

$$f(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} \sum_{j \in \mathcal{C}_k} d(x_i, x_j)$$

where $\mathcal{C}_k = \{i \in \mathbf{N}_N : C(i) = k\}$.

Equivalent Formulation

Define the mean vectors

$$m_k = \frac{1}{N_k} \sum_{i \in \mathcal{C}_k} x_i, \quad k \in \mathbf{N}_K$$

where $N_k = |\mathcal{C}_k|$. Then if d is the squared Euclidian norm:

$$f(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} \sum_{j \in \mathcal{C}_k} d(x_i, x_j) = \sum_{k=1}^{K} N_k \sum_{i \in \mathcal{C}_k} d(x_i, m_k)$$

We have for any $y_k \in \mathbf{R}^n$ that $\sum_{i \in \mathcal{C}_k} d(x_i, m_k) \leq \sum_{i \in \mathcal{C}_k} d(x_i, y_k)$ with equality for $y_k = m_k$ and hence an equivalent optimization:

$$\text{minimize } F(C, y_1, \ldots, y_K)$$

where $F : \mathbf{N}_N \times \mathbf{N}_K \times \mathbf{R}^n \times \cdots \times \mathbf{R}^n \to \mathbf{R}_+$ is defined as

$$F(C, y_1, \ldots, y_K) = \sum_{k=1}^{K} N_k \sum_{i \in \mathcal{C}_k} d(x_i, y_k)$$

Sets $\mathcal{C}_k$ are functions of encoder C.

K-Means Algorithm

Block coordinate descent:

1. For fixed C solve minimize$_{y_1,\ldots,y_K} F(C, y_1, \ldots, y_K)$
2. For fixed $(y_1, \ldots, y_K)$ solve minimize$_C F(C, y_1, \ldots, y_K)$

First problem is a LS problem.

Second problem has an explicit solution given by assigning observation x_i to cluster k if $d(x_i, y_k) \leq d(x_i, y_j)$ for all $j \neq k$.

The algorithm can be trapped in local minima.

Example in two dimensions with $N = 30$

Clustering using $K = 3$. We initialize y_k as the first three values of x_i that we are given. They actually come from the same cluster, showing that initialization is not extremely critical.