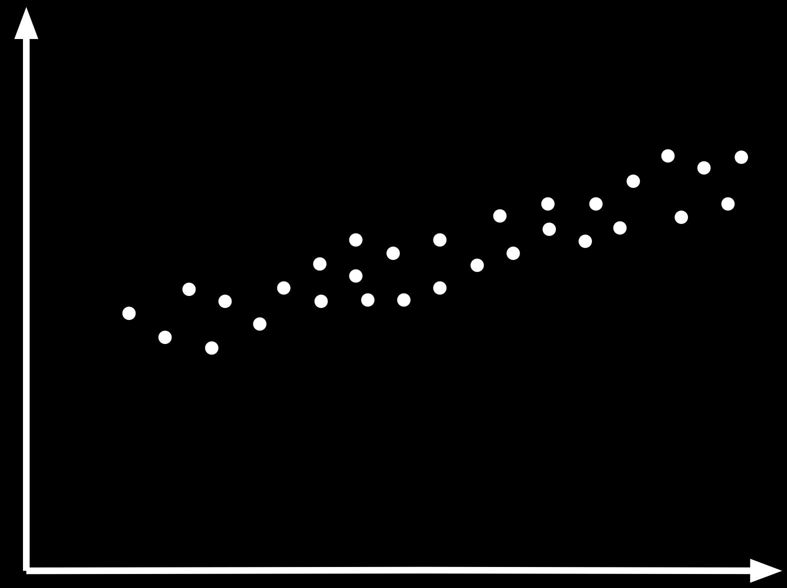
A Generalization of Principal Component Analysis to the Exponential Family

Michael Collins, Sanjoy Dasgupta, Robert E. Schapire

Presented by: Sameer Agarwal

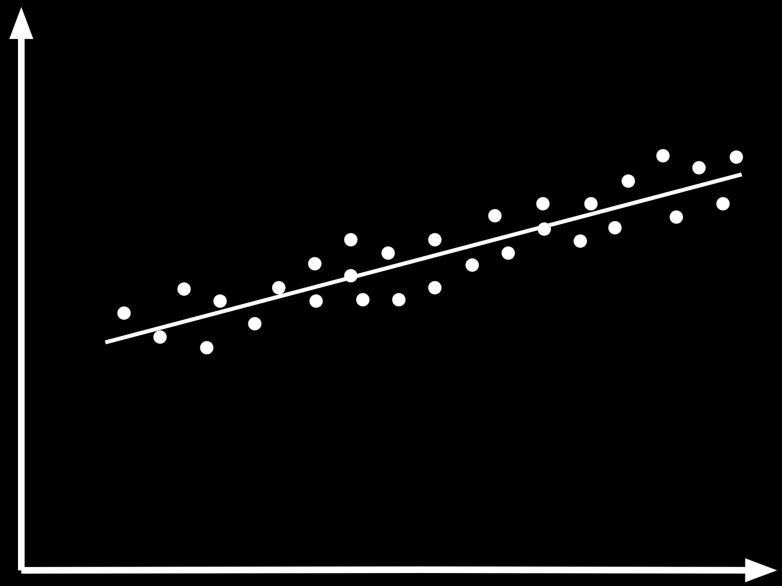
Artificial Intelligence Laboratory, UCSD

The shoe size problem



If you had only one variable to describe this data, what would it be ?

The shoe size problem



Fit a line to the data and take the projection of the data onto it.

The PCA problem statement

Given a set of data points $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, $\mathbf{x}_i \in \mathbb{R}^n$ and an integer $k < n$. Find a k-dimensional subspace $\mathbf{S}$ of $\mathbb{R}^n$ and the corresponding projections $\mathbf{\Theta} = \{\theta_1, \theta_2, \ldots, \theta_n\}$ of $\mathbf{X}$ onto $\mathbf{S}$, s.t.

$$\sum_i \|\mathbf{x}_i - \theta_i\|^2 \tag{1}$$

is minimized. Subject to

$$\Theta\Theta^T \text{ is diagonal.}$$

The Solution

The subspace spanned by the k-largest eigenvectors of the matrix

$$C = (X - \overline{X})(X - \overline{X})^T$$

The PCA Algorithm

$\mathrm{PCA}\,(X, k)$

1. $[U, D, V] = \mathrm{svd}(X - \overline{X})$
2. $P = V[:, 1 : k]$

or,

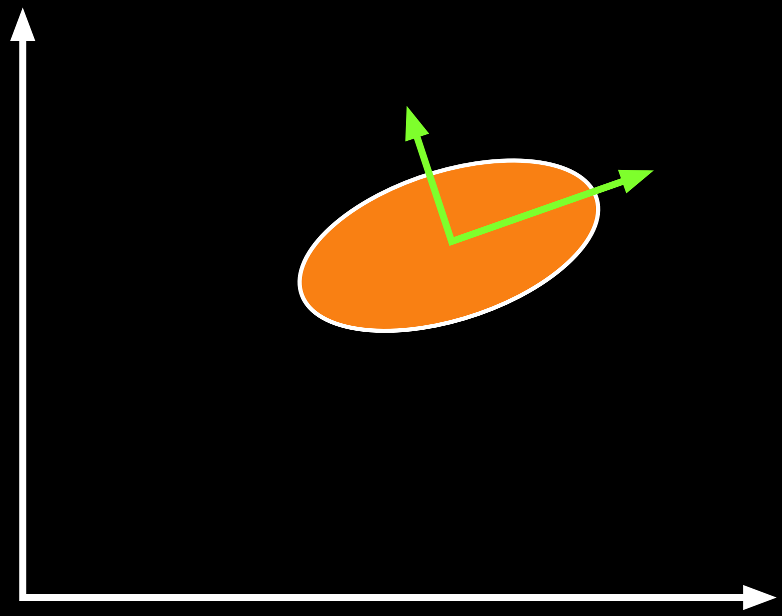
$\mathrm{PCA}\,(X, k)$

1. $C = (X - \overline{X})(X - \overline{X})^T$
2. $CV = \Lambda V$
3. $P = V[:, 1 : k]$

PCA in pictures

Assume that data is an ellipsoid.

PCA in pictures



Find the major and minor axes of the best fitting ellipsoid.

Why PCA?

1. Reduce the dimenionality of the data.
2. Optimal in terms of L_2 norm.
3. Output dimensions have zero correlation.
4. Denoises the data.

The age of the subspace

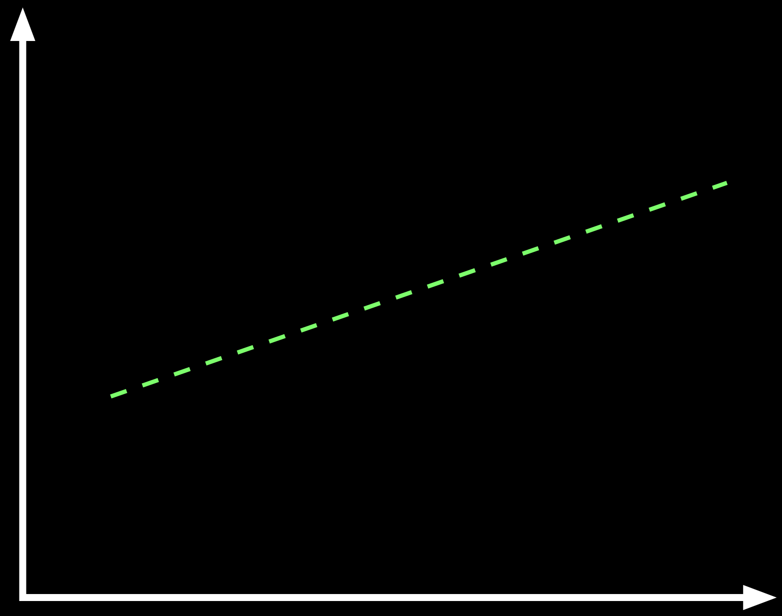
- Assumption : The underlying model generating the observations lives in a low dimensional subspace.

$$
\begin{aligned}
X &= AV \qquad (2)\\
\mathrm{size}(V,1) &< \mathrm{size}(X,2) \qquad (3)
\end{aligned}
$$

- PCA Rows of V are orthonormal.
- ICA Rows of V are independent.
- LDA Rows of V are such that there is maximum discrimination between classes.

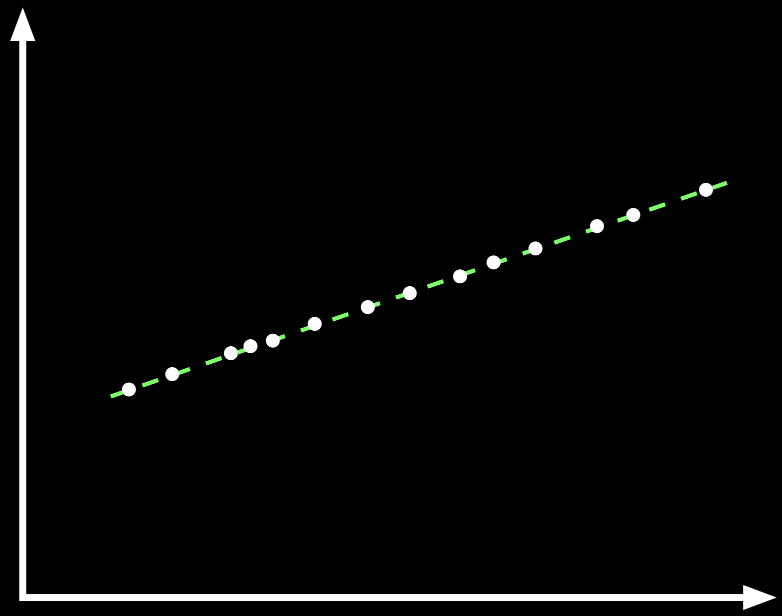
and variants thereof.

A gaussian view of PCA



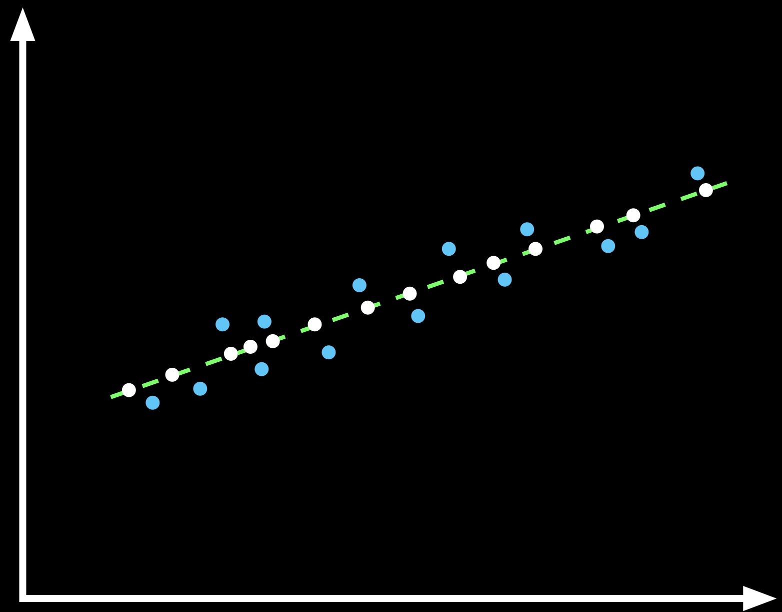
Take a Line.

A gaussian view of PCA



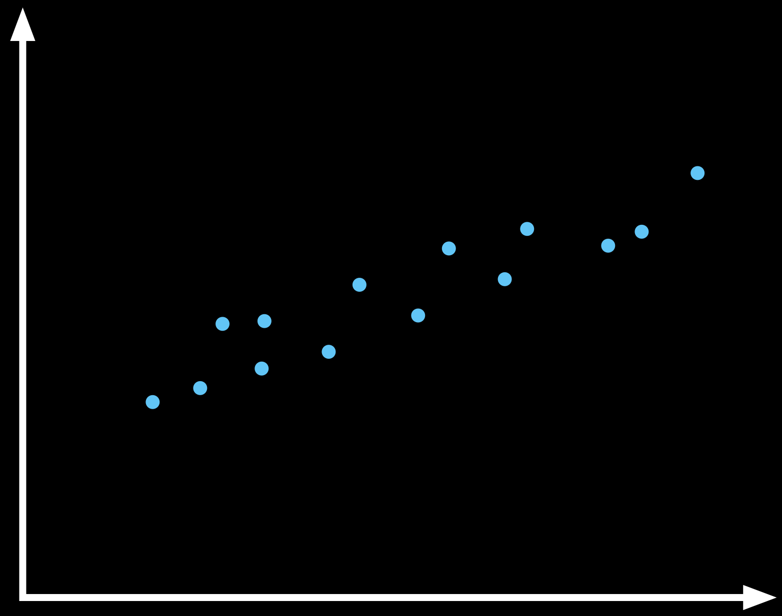
Sample points from it.

A gaussian view of PCA



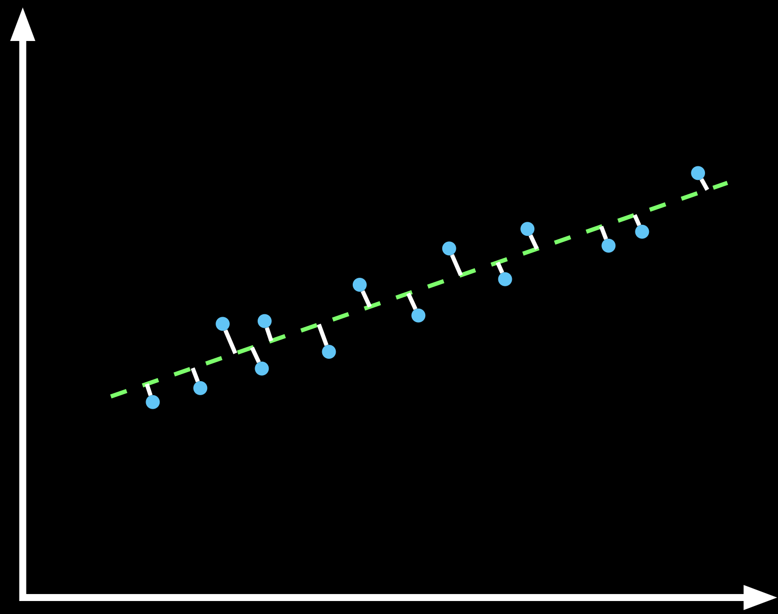
Add a sprinkling of unit variance gaussian noise.

A gaussian view of PCA



Remove all trace of the original model.

A gaussian view of PCA



Recover the best fitting subspace in the MLE sense, and the projection of the data onto this subspace.

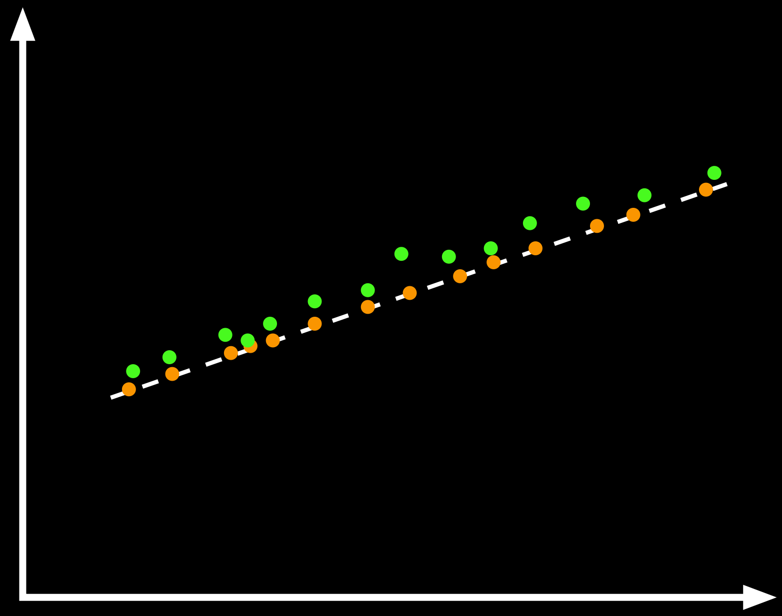
to summarize

Each data point $\mathbf{x}_i$ is a sample from a probability distribution $P(\mathbf{x}; \theta_i)$ of the form

$$P(\mathbf{x}; \theta_i) = \frac{1}{\sqrt{2\pi}} e^{\frac{(\mathbf{x}-\theta_i)}{2}^2}$$

Find $\boldsymbol{\Theta} = \{\theta_1, \theta_2, \ldots, \theta_n\}$, s.t. the likelihood of the data given the parameters Θ is maximized, subject to the constraint that Θ lies in a k-dimensional subspace.

But what if ..



You knew that all your noise was positive ?

Beyond the Gaussian

1. Guassian error is not suitable for every problem.
2. Integer valued data $\sim$ Poisson.
3. Positive valued data $\sim$ Exponential.
4. Binary valued data $\sim$ Bernoulli.

The Exponential Family

$$
\begin{aligned}
P(\mathbf{x}|\theta) &= P_0(\mathbf{x})e^{\mathbf{x}\cdot\theta - G(\theta)} \\
\log P(\mathbf{x}|\theta) &= \log P_0(\mathbf{x}) + \mathbf{x}\cdot\theta - G(\theta)
\end{aligned}
$$

1. θ is the natural parameter.
2. $G(\theta) = \ln \int e^{\mathbf{x}\cdot\theta} P_0(\mathbf{x}) d\mathbf{x}$ is the Cumulant Function.
3. $G(\theta)$ is strictly convex.

Members of the Exponential family

Gaussian

$$
\begin{aligned}
P(x|\mu) &= \frac{1}{\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2}} = \frac{e^{-x^2}}{\sqrt{2\pi}} e^{x\theta - \frac{\theta^2}{2}} \\
\theta &= \mu
\end{aligned}
$$

Bernoulli

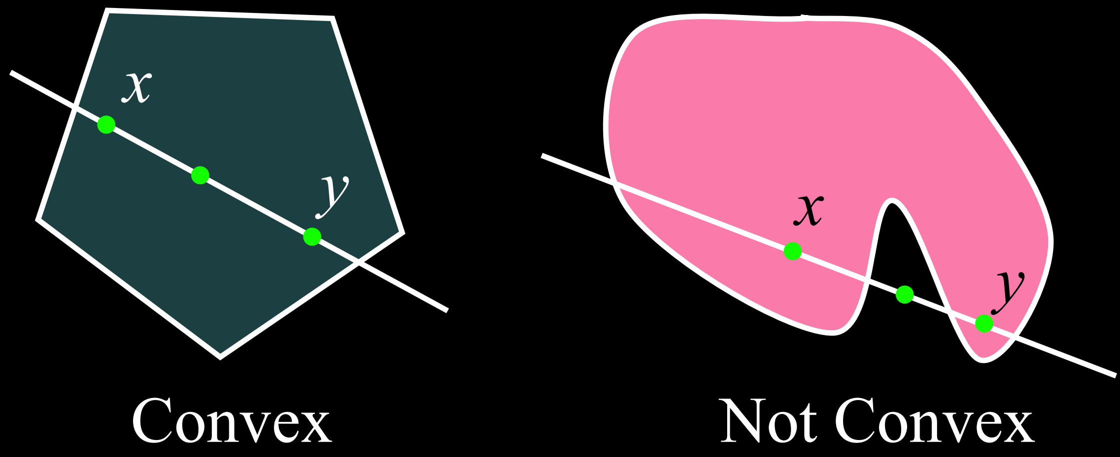
$$
\begin{aligned}
P(x|p) &= p^x (1-p)^{(1-x)} = 1 e^{x\theta - \log(1+e^\theta)} \\
\theta &= \log \frac{p}{1-p}
\end{aligned}
$$

Problem Statement

Given data points $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ and a member of the exponential family $P_G(\mathbf{x}|\theta)$, find a set $\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$, which maximizes the total likelihood of the data.

$$\begin{aligned}\Theta &= \operatorname*{argmax}_{\boldsymbol{\Theta}} \mathcal{L}(P_G, \mathbf{X}, \boldsymbol{\Theta}) \\ &= \operatorname*{argmax}_{\boldsymbol{\Theta}} \prod_i P(\mathbf{x}_i|\theta_i)\end{aligned}$$

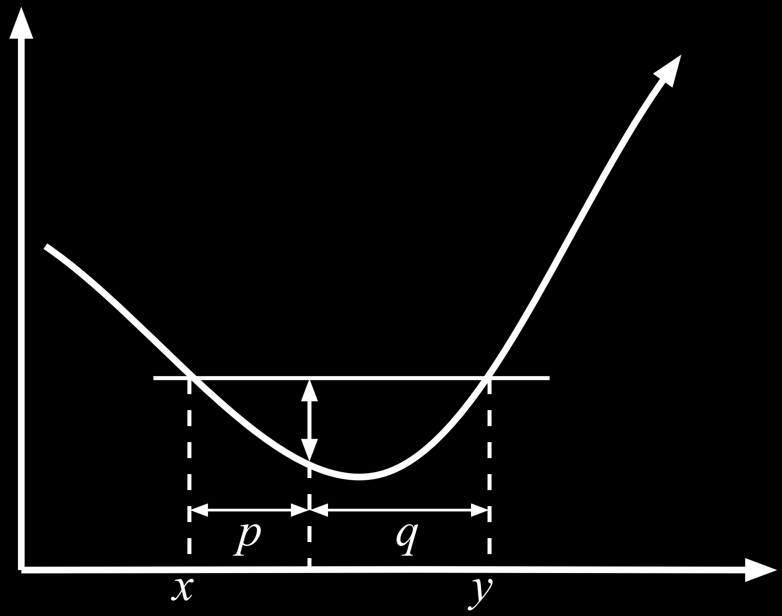
Convex Sets



$S \subseteq \mathbb{R}^n$ is a convex set if

$$x, y \in S, \quad p, q \geq 0, \quad p + q = 1 \Rightarrow pq + qy \in S$$

Convex Function



$$
\begin{aligned}
f\left(px+qy\right) &\leq pf(x)+qf(y) \\
p+q &= 1 \\
p,q &\geq 0
\end{aligned}
$$

Convex Optimization Problems

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0,\ i = 1, \ldots, m \\ & Ax = b,\ A \in \mathbb{R}^{p \times n} \end{array}$$

- $f_0, f_1, \ldots, f_m$ are convex.
- Affine equality constraints.
- Feasible set is convex.

and the big deal is ?

Theorem: For a convex optimization problem, any local solution is also a global solution.

$x \in C$, is locally optimal if it satisfies

$$y \in C, \ \|y - x\| \leq R \Rightarrow f_0(y) \geq f_0(x)$$

$x \in C$ is globally optimal if

$$y \in C \Rightarrow f_0(y) \geq f_0(x)$$

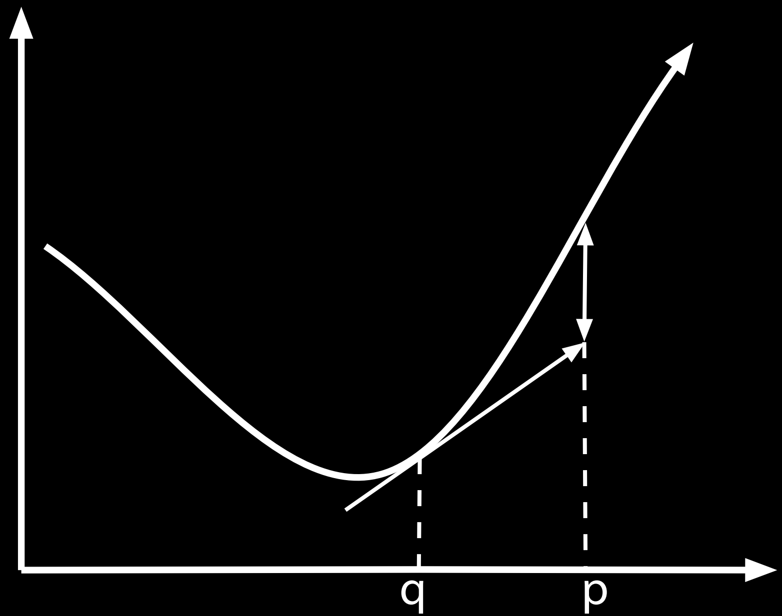
Proof

- Suppose x is locally optimal, but $y \in C,\ f_0(y) < f_0(x)$
- Let $z = \epsilon y + (1-\epsilon)x$ with a small $\epsilon > 0$.
- z is near x, with

$$\begin{aligned} f_0(z) &= f_0(\epsilon y + (1-\epsilon)x) \\ &\leq \epsilon f_0(y) + (1-\epsilon) f_0(x) \\ &< f_0(x) \end{aligned}$$

- A contradiction, since x is locally optimal.

Bregman divergences



$$B_F(p,q) = F(p) - F(q) - (p-q)\cdot\nabla F(q)$$

F is real, convex and differentiable.

Properties of Bregman divergences

1. $B_F(q,p)$ measures the convexity of F. It measures the increase in $F(q)$ over $F(p)$ above linear growth with slope $\nabla F(p)$.
2. $B_F(q,p) \geq 0$.
3. $B_F(q,p) = 0 \iff p = q$ and $F(p)$ is strictly convex.

Properties of the Exponential Family

Let ,

$$\lambda_G(\theta; \mathbf{x}) = \log P_G(\mathbf{x}; \theta) = \log P_0(x) + \mathbf{x} \cdot \theta - G(\theta)$$

and assume,

$$E_\theta[\nabla_\theta \lambda(\theta; \mathbf{x})] = 0$$

then,

1. $\nabla_\theta \lambda(\theta; \mathbf{x}) = \mathbf{x} - \nabla_\theta G(\theta)$
2. $E_\theta[\mathbf{x}] = \mu = \nabla_\theta G(\theta) = g(\theta)$
3. G stricly convex $\Rightarrow g(\theta)$ is invertible.

Properties of the Exponential Family

Define, $g^{-1}(\mu) = \theta$, then we can define the dual of $G(\theta)$

$$F(\mu) = \theta \cdot \mu - G(\theta)$$

$G(\theta)$ strictly convex $\Rightarrow F(\mu)$ is strictly convex too. now,

$$\begin{aligned} B_G(\theta, \hat{\theta}) &= G(\theta) + F(\hat{\mu}) - \theta \cdot \hat{\mu} = B_F(\hat{\mu}, \mu) \\ \lambda(\theta; \mathbf{x}) &= \log P_0(\mathbf{x}) + \theta \cdot \mathbf{x} - G(\theta) \\ \lambda(\theta; \mathbf{x}) &= \log P_0(\mathbf{x}) + F(\mathbf{x}) - B_F(\mathbf{x}, g(\theta)) \end{aligned}$$

ePCA

since for members of the exponential family,

$$-\log P(x|\theta) = -\log P_0(x) - F(x) + B_F(x, g(\theta))$$

we have

$$\begin{aligned} L(\Theta) &= -\sum_{ij} \log P_G(x_{ij}|\theta_{ij}) \\ &= C(X) + \sum_{ij} B_F(x_{ij}, g(\theta_{ij})) \\ &= C(X) + B_F(X, g(\Theta)) \\ \Theta_{pca} &= \operatorname*{argmin}_{\Theta} B_F(X, g(\Theta)) \end{aligned}$$

Subspace rank constraint

We want Θ to be constrained in a subspace of dimension k. Hence,

$$\Theta = VA$$

where

$$\begin{aligned} \text{size}(V) &= [d, k] \\ \text{size}(A) &= [k, n] \end{aligned}$$

V is a set of basis vectors for the subspace containing Θ.
A is the set of bregman projections of X onto V.
The optimization problem now is

$$\underset{\Theta}{\text{argmin}}\ L(\Theta) = \underset{\{V,A\}}{\text{argmin}}\ B_F(X, g(VA))$$

Minimizing $L(V, A)$: idea

1. Start Randomly
2. Hold V constant and minimize w.r.t A
3. Hold A constant and minimize w.r.t V
4. Repeat 2-3 until convergence.

The 1-d Algorithm

1. $V = rand(d, 1)$
2. $A = rand(1, n)$
3. until convergence
4. $\qquad a_i^t = \underset{a}{\operatorname{argmin}} \sum_j B_F(x_{ij}, g(av_j^{(t-1)}))$
5. $\qquad v_j^t = \underset{v}{\operatorname{argmin}} \sum_i B_F(x_{ij}, g(a_i^t v))$
6. $\qquad t = t + 1$

An example : Gaussian Noise

$$
\begin{aligned}
a_i^t &= \underset{a}{\operatorname{argmin}}\, B_F(x_{ij}, g(av_j^{(t-1)})) \\
g(x) &= x \\
B_F(q,p) &= \frac{(p-q)^2}{2} \\
\rightarrow a_i^t &= \underset{a}{\operatorname{argmin}} \left[\sum_j \frac{(x_{ij} - av_j^{(t-1)})^2}{2} \right]
\end{aligned}
$$

Differentiating and equating to zero we get

$$
\sum_j (x_{ij} - av_j^{(t-1)}) v_j^{(t-1)} = 0
$$

An example (contd.)

Differentiating and equating to zero we get

$$
\begin{aligned}
a \sum_j {v_j^{(t-1)}}^2 &= \sum_j x_{ij} v_j^{(t-1)} \\
a \|V^{(t-1)}\|^2 &= X_i \cdot V^{(t-1)} \\
\rightarrow a_i^t &= \frac{X_i \cdot V^{(t-1)}}{\|V^{(t-1)}\|^2} \\
\rightarrow A^t &= \frac{X \cdot V^{(t-1)}}{\|V^{(t-1)}\|^2}
\end{aligned}
$$

An example (contd.)

Similarly

$$V^t = \frac{(A^t)^T X}{\|A^t\|^2}$$

combining the two we get

$$V^t = cV^{(t-1)} X^T X$$

equivalent to the power method of calculating the largest eigenvector.

Comments

1. Each step of the algorithm solves a convex optimization problem.
2. The over all optimization problem is not convex.
3. Convergence to global optima is difficult to prove in general.
4. Gaussian is a special case, where the power method converges to the global optima.

Avoiding Infinity

Problem : A local optima may exist at infnity.
Solution : Penalize large movements away from a fixed point in the range of $g(\theta)$.

$$L'(\Theta) = \sum_{ij} [B_F(x_{ij}, g(\theta_{ij})) + \epsilon B_F(\mu_0, g(\theta_{ij}))]$$

ϵ is a small positive constant. μ_0 is an arbitrary point in the range of g.

The alternating minimization procedure is guaranteed to find a bounded solution to $L'(\Theta)$.

The General Algorithm: code

1. $A = 0, V = 0$
2. For $n = 1, \ldots, N,\ c = 1, \ldots, l$
3. Initialize v_c^0 randomly.
4. $s_{ij} = \sum_{k \neq c} a_{ik} v_{kj}$
5. until convergence
6. $i = 1, \ldots, n$
7. $a_{ic}^t = \underset{a}{\operatorname{argmin}} \sum_j B_F(x_{ij}, g(av_{cj}^{(t-1)} + s_{ij}))$
8. $j = 1, \ldots, d$
9. $v_{cj}^t = \underset{v}{\operatorname{argmin}} \sum_j B_F(x_{ij}, g(a_{ic}^t v + s_{ij}) + s_{ij})$

Hindsight and GLMs

- Generalized Linear Models are extensions of standard regression

$$\mathbf{y} = A\mathbf{x} + \epsilon$$

where ϵ is a zero mean, constant variance normal. They generalize the model to

$$\mathbf{y} = g(A\mathbf{x}) + \eta$$

  - here, g is a real differentiable function known as the *link function*.
  - η is distributed according to a fixed member of the exponential family.

Related Work

- Hoffman et. al Probabilistc Latent Semantic Indexing
- Seung et. al Non-negative Matrix Factorization

Rely on explicit constraints on A and V.

Summary

- Interpret PCA using a generative model.
- Extend the model to the exponential family.
- Convert the Log Likelihood into a Bregman divergence.
- Minimize the divergence using an alternating minimization procedure.

References

- Collins and Schapire A Generalization of Principal Component Analsysis to the Exponential Family
- Tipping, M. E. & Bishop C. M. Probabilistic Principal Component Analysis.
- Azoury K. S. & Warmuth M. K. Relative loss bounds for on-line density estimation with the exponential family of distributions.

Acknowledgements

- Sanjoy Dasgupta for answering last minute questions.
- The music of Nusrat Fateh Ali Khan for keeping me company.