Linear Dependent Dimensionality Reduction

Nathan Srebro Tommi Jaakkola
Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology

Factor models are often natural in the analysis of multi-dimensional data. The underlying premise of such models is that **important aspects of the data can be captured via a low-dimensional representation.**

observations		small representation
y	⟵⟶	u
image	⟵···⟶	*varying aspects of scene*
gene expression levels	⟵···⟶	*description of biological process*
a user's preferences	⟵···⟶	*characterization of user*
document	⟵···⟶	*topics*

In many situations, including collaborative filtering and structure exploration, the "important" aspects are the dependencies between different attributes. Accordingly, we seek to identify a low-dimensional space that captures the ***dependent*** aspects of the data, and separate them from the ***independent*** variations. Our goal is to relax restrictions on the form of each these components, such as Gaussianity, additivity and linearity.

In this work we:

- Present a general framework for the problem: Dependent Dimensionality Reduction

Focusing on linear dependencies, we:

- Show that the standard approach (PCA) is consistent for additive i.i.d. noise, even if it is not Gaussian
- Show that a *variance-ignoring* estimator is appropriate for non-additive noise models
- Present a method for maximum likelihood estimation in the presence of Gaussian mixture additive noise
- Study the consistency of maximum likelihood estimation in this context, and show that the ML estimator is not always consistent (for example for Exponential-PCA).

Dependent Dimensionality Reduction

Starting point:
Identifying linear dependencies in the presence of i.i.d. Gaussian noise

observed		*parameters*		*random*
Y	=	X rank *k*	+	Z iid Gaussian $\sigma^2 I$
		dependent, low-dimensional component		independent, Gaussian component

Log-likelihood(X) $\square$ sum-squared (Frobenius) distance to Y
ML estimator is rank-k matrix minimizing $|X-Y|_{Fro}$
given by leading components of SVD of Y

This formulation valid, but displeasing:
- Entire matrix X are parameters, estimated with a single observation Y
- Number of parameters linear in data
- Even with more data, cannot estimate beyond a fixed precision

What we *can* estimate with more data rows is the rank-*k* subspace of X:

$$y = u \times \underbrace{V}_{\text{parameters being estimated}} + z$$

P_U — non-parametric, unconstrained; $\sigma^2 I$; nuisance

Standard parametric analysis: imposes some (parametric) distribution P_u.

We do not make any assumptions about the distribution of **u**:
- Model class is non-parametric
- Can estimate a parametric aspect of the model – the subspace V.

We do impose a strict form on the conditional distributions **y**[i]|**u**.
Goal: relax this, and concentrate on the structural constraint–
u captures all the dependencies in **y**

Relaxing Gaussianity leads to **Linear Additive** models:

$$y = u \times V + z \qquad P_Z=\Pi P_{Z[i]}$$

Relaxing additivity is appropriate, e.g., when the noise has a multiplicative component, or when the features in **y** are not real numbers, yielding **Linear Dependent Dimensionality Reduction**:

u → **x**[1], **x**[2], …, **x**[*d*]; **x**=**u**× V; **x**[i] → **y**[i]; $P_{y[i]|x[i]}$

Going beyond linear models, fitting a non-linear manifold by minimizing sum-squared distance can be seen as a ML estimator for:

$$y = g(u) + z, \quad z \sim \text{Gaussian } \sigma^2 I$$

g specifies smooth manifold

Combining these ideas leads us to discuss **y**[i]|**u** directly:

u → **y**[1], **y**[2], …, **y**[*d*]

Dependent Dimensionality Reduction:
Low-dimensional representation **u** such that coordinates of **y** are **independent given u**

Second Moment Methods

L2 approach to low-rank approximation: minimize sum-squared distance $|\mathbf{X}-\mathbf{Y}|_{Fro}$.
⇓
Subspace spanned by the leading *k* eigenvectors of empirical covariance of **y**.

For any i.i.d. additive noise:

$$y = \underbrace{u \times V}_{x} + z \qquad P_Z=\Pi P_{Z[i]}$$

$$\hat{\Sigma}_{\mathbf{y}} \ \square\ \Sigma_{\mathbf{y}} = \underbrace{\Sigma_{\mathbf{x}} + \square^2_{\square}\square}_{s_1+\sigma^2,\, s_2+\sigma^2,\, \dots,\, s_k+\sigma^2,\, \sigma^2,\, \sigma^2,\, \dots,\, \sigma^2}$$

$\Sigma_{\mathbf{x}}$: $s_1, s_2, \dots, s_k, 0, 0, \dots, 0$

L_2 estimation of the low-rank subspace (PCA) is **consistent** in the presence of **any i.i.d. additive noise** with finite variance

Independent, non-identical additive noise:

$$\hat{\Sigma}_{\mathbf{y}} \ \square\ \Sigma_{\mathbf{y}} = \underbrace{\Sigma_{\mathbf{x}}}_{\text{rank } k} + \underbrace{\Sigma_{\mathbf{z}}}_{\text{diagonal}}$$

When the additive noise is independent, but not identically distributed, the L2 estimator is biased towards the high-variance coordinates. Instead, the **Variance Ignoring Estimator** seeks a rank-*k* matrix approximating (minimizing the sum-squared distance to) the non-diagonal entries of the empirical covariance. This is a **Weighted** Frobenius Low Rank Approximation (**WLRA**) problem.

Unbiased Non-additive Noise:

u → **x**[1], **x**[2], …, **x**[*d*] → **y**[1], **y**[2], …, **y**[*d*]; **x**=**u**× V

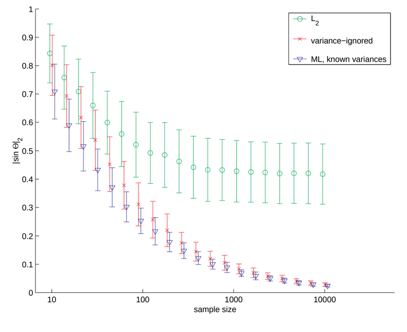
$$\mathbf{E}[\mathbf{y}_i|\mathbf{x}_i] = \mathbf{x}_i$$

$$\begin{aligned}\mathrm{Cov}[y_i y_j] &= \mathbf{E}[y_i y_j] - \mathbf{E}[y_i]\mathbf{E}[y_j]\\ &= \mathbf{E}[\mathbf{E}[y_i y_j|x]] - \mathbf{E}[\mathbf{E}[y_i|x]]\mathbf{E}[\mathbf{E}[y_j|x]]\\ &= \mathbf{E}[\mathbf{E}[y_i|x]\mathbf{E}[y_j|x]] - \mathbf{E}[x_i]\mathbf{E}[x_j]\\ &= \mathbf{E}[x_i x_j] - \mathbf{E}[x_i]\mathbf{E}[x_j]\\ &= \mathrm{Cov}[x_i x_j]\end{aligned}$$

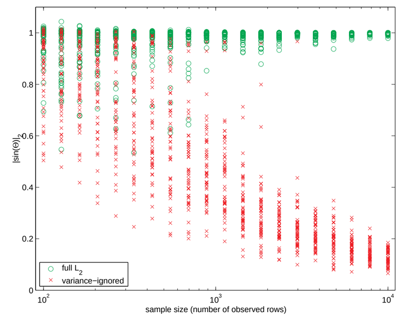
$$\begin{aligned}\mathrm{Var}[y_i] &= \mathbf{E}[y_i^2] - \mathbf{E}[y_i]^2\\ &= \mathbf{E}[\mathbf{E}[y_i^2|x_i] - \mathbf{E}[y_i|x_i]^2] + \mathbf{E}[\mathbf{E}[y_i|x_i]^2] - \mathbf{E}[y_i]^2\\ &= \mathbf{E}[\mathrm{Var}[y_i|x_i]] + \mathbf{E}[x_i^2] - \mathbf{E}[x_i]^2\\ &= \mathbf{E}[\mathrm{Var}[y_i|x_i]] + \mathrm{Var}[x_i]\end{aligned}$$

$$\hat{\Sigma}_{\mathbf{y}} \ \square\ \Sigma_{\mathbf{y}} = \underbrace{\Sigma_{\mathbf{x}}}_{\text{rank } k} + \underbrace{\mathbf{E}[\mathrm{Var}[\mathbf{y}_i|\mathbf{x}_i]]}_{\text{diagonal}}$$

Variance Ignoring Estimator is appropriate for any **unbiased** independent conditional model



Rank 2 subspace in $\square^{10}$. Additive Gaussian noise of different scale on different coordinates (between 0.17 and 1.7 signal variances)



Exponentially distributed observations with means in rank-2 subspace of $\square^{10}$

Maximum Likelihood Estimation with Gaussian Mixture Noise

Model additive noise as a Gaussian mixture:

$$Y = \underbrace{X}_{\text{rank } k} + Z \qquad C \text{ (hidden)}$$

Z: Mixture of Gaussians: $\{p_c, \mu_c, \sigma_c\}$

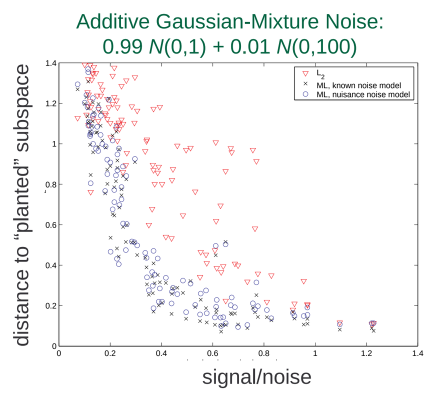
E step: calculate posteriors of *C*
M step:

$$\mathbf{E}_{\square|Y}[\log \Pr(\square = \square + \square|\square;\square)]$$
$$= -\sum_{i\square} \mathbf{E}_{\square_{i\square}|Y_{i\square}}\left[\frac{1}{2}\log 2\square\square^2_{\square_{i\square}} + \frac{\left((X_{i\square} - Y_{i\square}) - \square_{\square_{i\square}}\right)^2}{2\square^2_{\square_{i\square}}}\right]$$
$$= -\sum_{i\square}\sum_{\square} \frac{\Pr(\square_{i\square}=\square|Y_{i\square})}{2\square^2_{\square}}\left(\square_{i\square} - (\square_{i\square} + \square_{\square})\right)^2 + \text{Const}$$
$$= -\frac{1}{2}\sum_{i\square} \square_{i\square}\left(\square_{i\square} - \square_{i\square}\right)^2 + \text{Const}$$

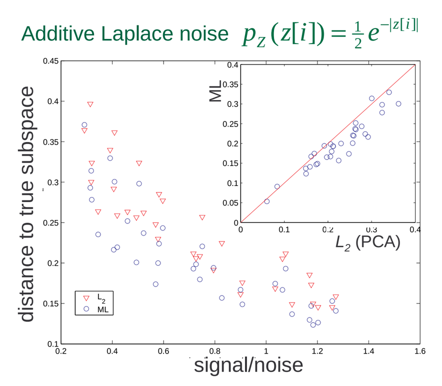
Weighted Low-Rank Approximation with:

$$\square_{i\square} = \sum_{\square} \frac{\Pr(\square_{i\square}=\square|\square_{i\square})}{\square^2_{\square}} \qquad \square_{i\square} = \square_{i\square} + \sum_{\square} \frac{\Pr(\square_{i\square}=\square|\square_{i\square})\square_{\square}}{\square^2_{\square}\square_{i\square}}$$

and update mixture parameters.



Additive Gaussian-Mixture Noise: 0.99 *N*(0,1) + 0.01 *N*(0,100)



Additive Laplace noise $p_z(z[i]) = \frac{1}{2}e^{-|z[i]|}$

Noise modeled a bounded **Gaussian Scale Mixture**
(mixture of zero-mean Gaussians with variance bounded away from zero)
Captures many distributions, including heavy tailed (non log-concave)

Instead of Gaussian mixture modeling:
Newton's method on log-Likelihood
[Gordon NIPS02] [SJ ICML 03]
Distribution must be log-concave – not applicable to heavy tailed distributions

Weighted Low Rank Approximation

[ICML03]

Given W, A, find rank-*k* X minimizing *weighted* sum-square distance:

$$\sum_{ij} W_{ij}\left(A_{ij} - X_{ij}\right)^2, \quad X \text{ rank } k$$

$$J(UV') = \sum_{ij} W_{ij}(A - UV')^2_{ij}$$

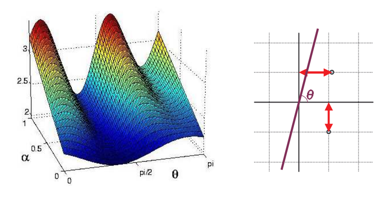
For fixed *V*, find optimal *U*
For fixed *U*, find optimal V

$$J^*(V) = \min_U J(UV')$$
$$\frac{\partial}{\partial V} J^*(V') = 2U^{*\prime}((U^*V' - A)\otimes W)$$

Conjugate gradient descent on *J**

$X \leftarrow \mathrm{LRA}(W \otimes A + (1-W) \otimes X)$

When weights introduced:
- Not incremental with *k*
- Eigenmethods do not apply
- Local minima:



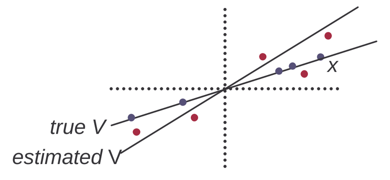
$$A = \begin{bmatrix}1 & 1.1\\ 1 & -1\end{bmatrix} \qquad W = \begin{bmatrix}1+\alpha & 1\\ 1 & 1+\alpha\end{bmatrix}$$

Consistency of Maximum Likelihood Estimation with a Known Noise model

General Conditions:

Assume a known additive noise model, and consider maximum likelihood estimation with respect to that model:

$$y = \underbrace{u \times V}_{x} + z \qquad P_Z=\Pi P_{Z[i]}$$



$$\Phi(V) = \max_u \log p_{\mathbf{z}}(\mathbf{y} - uV)$$

ML estimator is consistent ⇄ $\mathbf{E}_{\mathbf{y}=\mathbf{u}V+\mathbf{z}}[\Phi(V)]$ is maximized on true V

expected contribution of *x* to likelihood of *V*: $\Psi(V;x) = \mathbf{E}_{\mathbf{z}}\left[\max_u \log p_{\mathbf{z}}((x+\mathbf{z}) - uV)\right]$

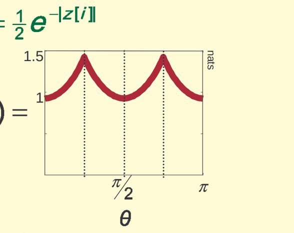
ML estimator is consistent for any P_u ⇄ for all *x*, V maximizes Ψ(V;*x*) iff V spans *x*
(⇐ for p_z positive, continuous, finite variance)

ML estimator is consistent for any P_u ⟹ Ψ(V;0) is constant for all V

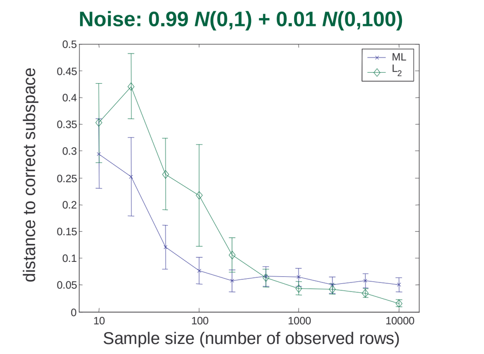
Gaussian Additive Noise:
Ψ(V;*x*) = E[L_2 distance of *x*+*z* from V]
ML estimator is consistent

Laplace Additive Noise:
$p_z(z[i]) = \frac{1}{2}e^{-|z[i]|}$

$-\Psi(V;0) =$



The ML estimator is **not** consistent!



Noise: 0.99 *N*(0,1) + 0.01 *N*(0,100)

Maximum Likelihood Low-Rank estimation with non-Gaussian noise is not, in general, consistent

Non-Additive Models:

u → **x**[1], **x**[2], …, **x**[*d*] → **y**[1], **y**[2], …, **y**[*d*]; **x**=**u**× V; $P_{y[i]|x[i]}$

These conditions can also be used to investigate the consistency of ML estimators with non-additive known conditionals $\mathbf{y}_i|\mathbf{x}_i$, where:

$$\Psi(V;x) = \mathbf{E}_{\mathbf{y}|x}\left[\max_u \log p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|uV)\,\middle|\, x\right]$$

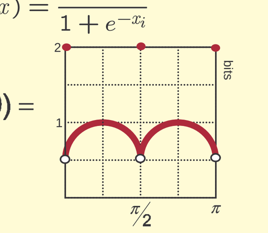
Of particular interest is "Exponential PCA", where the distribution $\mathbf{y}_i|\mathbf{x}_i$ forms an exponential family with $\mathbf{x}_i$ the natural parameters [Collins Dasgupta Schapire, NIPS01].

"Exponential-PCA" is not, in general, consistent

When the *mean* parameters form a low-rank subspace, the variance-ignoring estimator is applicable, but when the *natural* parameters form a low-rank subspace no generally consistent estimator is known.

Logistic Low-Rank Approximation:
$$P_{\mathbf{y}_i|\mathbf{x}_i}(+1|x) = \frac{1}{1+e^{-x_i}}$$

$-\Psi(V;0) =$



The ML estimator is **not** consistent!

Challenge: Find a consistent estimator for the low-rank subspace of natural parameters

http://www.ai.mit.edu/~nati/LowRank