

Gaussianization

Ramesh A. Gopinath

IBM T. J. Watson Research, Yorktown Heights, NY 10598

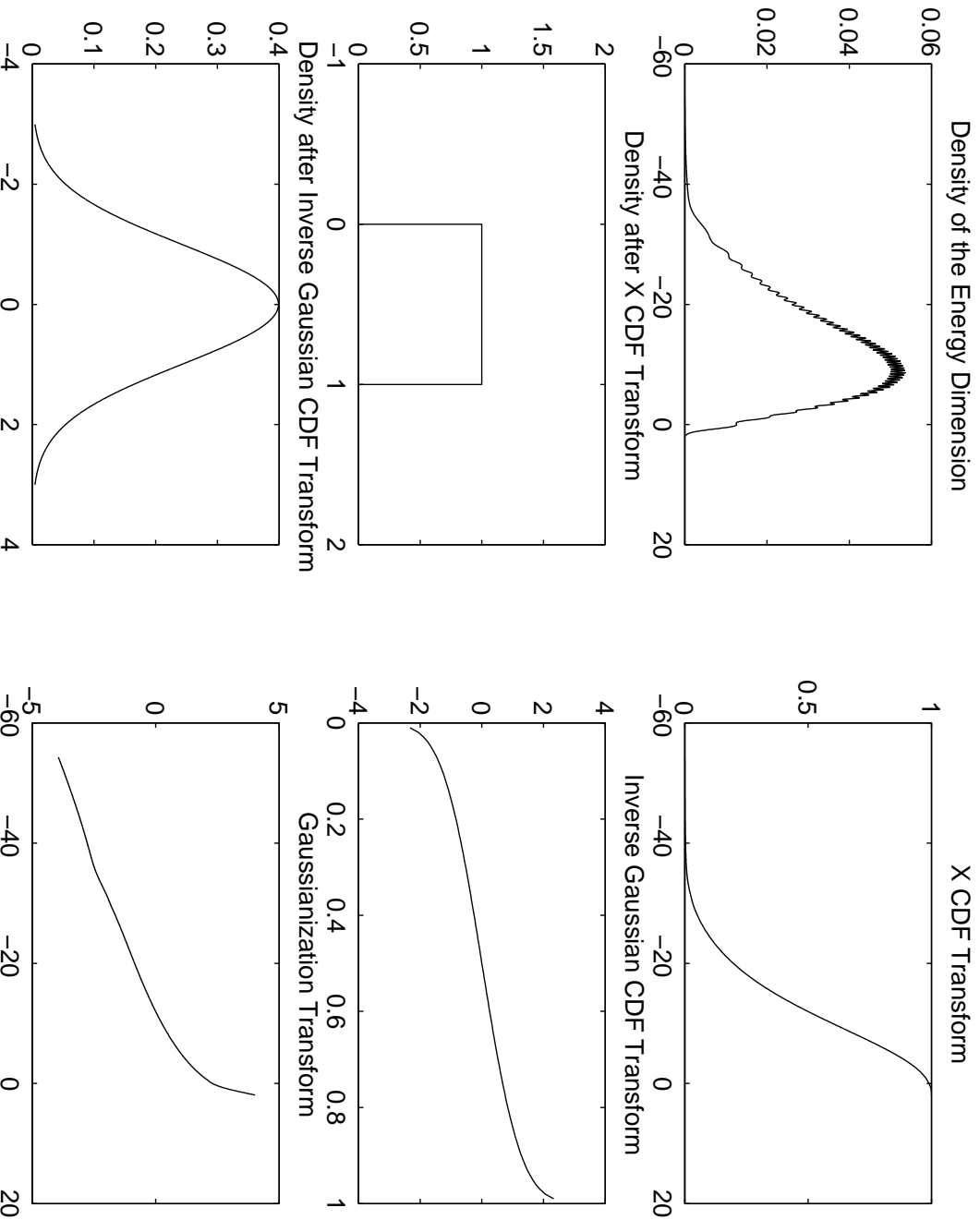
IMA Workshop Presentation, September 22, 2000

Joint work with Scott S. Chen, RenTech, NY

For a copy of the paper contact: rameshg@us.ibm.com

Univariate Gaussianization

Broadcast News Training Data (HUB4 1998): Energy Dimension



Simple Front-End Experiment

On 1997 DARPA HUB4 Broadcast News Transcription task:

- Motivation:
Logarithm \rightarrow *univariate Gaussianization*
- Experiment:
IBM speech recognizer with 135K Gaussians
- Results:

	Logarithm	1-d Gaussianization
Word Error Rate	18.5%	18.1%

Density Estimation

Univariate Density Estimation is pretty much solved!

- Kernel estimates
 - Variable kernel methods
- Radial basis function estimates
 - Gaussian mixture models
 - Wavelet density estimates
- ...

High Dimensional Density Estimation is hard!

- Curse of Dimensionality (Bellman 1961)
 - high dimensional data is sparse!

The Curse of Dimensionality

Bias vs Variance:

- Small neighborhoods are almost empty : big variance!
- Big neighborhoods to achieve sufficient counts: big bias!

Kernel methods: $(\mathbf{x}_i \in \mathcal{R}^D : 1 \leq i \leq N)$

$$\hat{p}(\mathbf{x}) = \frac{1}{N|\det(H)|} \sum_{n=1}^N K(H^{-1}(\mathbf{x} - \mathbf{x}_n))$$

Error rate of adaptive kernel methods:

$$E \int |\hat{p}(\mathbf{x}) - p(\mathbf{x})|^2 d\mathbf{x} \sim N^{-4/(D+4)}$$

Equivalent sample size for achieving error ϵ :

$$N \sim \left(\frac{1}{\epsilon}\right)^{(D+4)/4}$$

High Dimensional Density Estimation

Parametric Estimation

- Dimensionality Reduction
 - LDA
 - Heteroscedastic LDA
 - ...
- Covariance modeling in Gaussian mixtures
 - Diagonal Covariances
 - Semi-Tied Covariances
 - ...

Nonparametric Estimation

- Projection Pursuit: Friedman 1987

Hilbert's 13th problem

High dimensional functions can be characterized by univariate functions!

Theorem 1 (Kolmogorov, 1957) Let $\mathbf{x} = (x_1, \dots, x_D) \in [0, 1]^D$. There exist D universal constants

$$\{ \lambda_d : 1 \leq d \leq D \}$$

and $(2D + 1)$ univariate functions

$$\{ \phi_j(\cdot) : 1 \leq j \leq 2D + 1 \}$$

such that for every continuous function $f(x_1, \dots, x_D)$ one can find an univariate continuous function $g_f(\cdot)$ such that

$$f(x_1, \dots, x_D) = \sum_{j=1}^{2D+1} g_f \left(\sum_{d=1}^D \lambda_d \phi_j(x_d) \right)$$

Projection Pursuit (Friedman 1987)

To overcome the curse of dimensionality by a series of 1-D projections.

whiten the data; let $X^{(0)} = X$; then iterate over (1) and (2):

(1) Find the most non-Gaussian 1D projection.

Let $\alpha^{(k)} X^{(k)}$ be the most non-Gaussian 1D projection. Let $U^{(k)}$ be the orthogonal completion

$$U^{(k)} = [\alpha^{(k)} \dots]$$

$$Y^{(k)} = U^{(k)} X^{(k)}$$

(2) Transform the projection to standard $N(0, 1)$.

$$X_1^{(k+1)} = \Psi(Y_1^{(k)})$$

$$X_d^{(k+1)} = Y_d^{(k)} \quad 2 \leq d \leq D$$

$$X^{(k)} \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{I})$$

Projection Index

Projection index

- characterization of $N(0, 1)$

$$X \sim N(0, 1) \implies \Phi(X) \sim U[0, 1] \implies 2\Phi(X) - 1 \sim U[-1, 1]$$

- departure from normality

$$I(\alpha^T \mathbf{X}) = \int_{-1}^1 [p_{[2\Phi(\alpha^T \mathbf{X})-1]}(r) - \frac{1}{2}]^2 dr$$

- estimate $p_{[2\Phi(\alpha^T \mathbf{X})-1]}(\cdot)$ by Legendre polynomial approximation.

Find the most non-Gaussian 1D projection by gradient descent

$$\max_{\alpha} I(\alpha^T \mathbf{X})$$

Gaussianizing 1D Projection is Trivial

Cumulative Distribution Function (CDF) of X :

$$F(x) = \int_{-\infty}^x p(y)dy.$$

Cumulative Distribution Function (CDF) of $N(0, 1)$:

$$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)dy.$$

Then

$$\Phi^{-1} \circ F(X) \sim N(0, 1)$$

In practice, estimate density of X by univariate Gaussian mixtures

$$\mathcal{T}_{\pi, \mu, \sigma}(X) = \Phi^{-1} \left[\sum_i \pi_i \Phi\left(\frac{X - \mu_i}{\sigma_i}\right) \right]$$

Projection Pursuit Density Estimation

Density Estimation:

$$X^{(K)} \sim N(\mathbf{0}, \mathbf{I})$$

$$p_X(\mathbf{x}) = \phi(\mathbf{x}^{(K)}) \Big| \frac{\partial \mathbf{x}^{(K)}}{\partial \mathbf{x}} \Big| = \phi(\mathbf{x}^{(K)}) \prod_{k=1}^K \Big| \frac{\partial \mathbf{x}^{(k)}}{\partial \mathbf{x}^{(k-1)}} \Big|$$

Projection Pursuit alleviates the curse of dimensionality:

- At each iteration, find the projection which is mostly non-Gaussian.
- At each iteration, perform univariate density estimation.

Comparative Study

Hwang (1992) performed extensive comparative study among:

- one dim projection pursuit density estimates
 - adaptive kernel density estimates
 - radial basis function density estimates
- projection pursuit density estimates outperform in most data sets!

Error criterion on test set:

$$Error = \frac{1}{M} \sum_{m=1}^M (\hat{p}(\mathbf{x}_m) - p(\mathbf{x}_m))^2$$

2D Projection Pursuit

whiten the data; let $\mathbf{X}^{(0)} = X$; then iterate over (1) and (2):

(1) find the most jointly non-Gaussian 2D projection.

$$I([\alpha\beta]^T \mathbf{X}) = \int_{-1}^1 [p_{[2\Phi([\alpha\beta]^T \mathbf{X}) - 1]}(y) - \frac{1}{4}]^2 dy$$

(2) jointly transform that 2D projected direction to standard $N(\mathbf{0}, \mathbf{I})$.

let $\mathbf{Y} = (Y_1, Y_2)^T$ be the 2D dimensional plane

– Rotate about the origin through angle γ :

$$Y'_1 = Y_1 \cos \gamma + Y_2 \sin \gamma$$

$$Y'_2 = Y_2 \cos \gamma - Y_1 \sin \gamma$$

– Gaussianize Y'_1 and Y'_2 individually.

– Repeat for several angles $\gamma = (0, \pi/8, \pi/4, 3\pi/8, \dots)$.

– Stop if the distributions stop becoming more Gaussian.

Gaussianization

For a random variable $\mathbf{X} \in \mathcal{R}^D$, we define its Gaussianization transform to be an invertible (and differentiable) transform $T(\mathbf{X})$ such that

$$T(\mathbf{X}) \sim N(\mathbf{0}, \mathbf{I})$$

Density estimation:

$$p_X(\mathbf{x}) = \left| \frac{\partial T(\mathbf{x})}{\partial \mathbf{x}} \right| \left(\frac{1}{\sqrt{2\pi}} \right)^D \exp\left(-\left\| \frac{1}{2} T(\mathbf{X}) \right\|^2\right).$$

Gaussianization In a Nutshell

Independence lifts the Curse of Dimensionality!

Key ideas leading to Gaussianization

- (1) Theory:
 - Linear Independent Component Analysis (Bell & Sejnowski 1995)
 - Projection Pursuit (Friedman 1987, Huber 1985)
- (2) Algorithm: Semi-Tied Covariances (Gales 1999)
- (3) By-products:
 - Density estimates sharper than kernel estimates
 - Efficient Solution for High dimensional projection pursuit
 - Friedman (1987) solved one dimensional projection pursuit.
 - Generalized Gaussian mixture models
 - Efficient Algorithm for Linear & Nonlinear ICA
 - Nonlinear ICA was an open problem

Univariate Gaussianization is Trivial

Cumulative Distribution Function (CDF) of X :

$$F(x) = \int_{-\infty}^x p(y)dy.$$

Cumulative Distribution Function (CDF) of $N(0, 1)$:

$$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)dy.$$

Then

$$\pm\Phi^{-1} \circ F(X) \sim N(0, 1)$$

In practice, estimate density of X by univariate Gaussian mixtures

$$T_{\pi, \mu, \sigma}(X) = \Phi^{-1} \left[\sum_i \pi_i \Phi\left(\frac{X - \mu_i}{\sigma_i}\right) \right]$$

High Dimensional Gaussianization Is Non-Trivial!

Notation $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$.

\forall two dimensional random variable $X = (x_1, x_2)$:

(1) marginally Gaussianize x_1

$$p(x_1, x_2) = p(x_1)p(x_2|x_1) = \phi(y_1)p(x_2|y_1)$$

(2) marginally Gaussianize the conditional $x_2|x_1$

$$\begin{array}{ccc} \dots & & \\ p(x_2|y_1 = 1) & \xrightarrow{T^{(1)}, \cdot} & \phi(y_2) \\ p(x_2|y_1 = 2) & \xrightarrow{T^{(2)}, \cdot} & \phi(y_2) \\ \dots & & \end{array}$$

$$\begin{aligned} (y_1, y_2) &= T(y_1, x_2) = (y_1, \Phi^{-1} \circ F_{x_2|y_1}(x_2)), \quad y_1 = \Phi^{-1} \circ F_{x_1}(x_1) \\ \Rightarrow p(y_1, y_2) &= \phi(y_1)\phi(y_2) \end{aligned}$$

Nonuniqueness of Gaussianization

Univariate Gaussianization is unique up to a flip of the sign.

High dimensional Gaussianization:

- A transform $U(\cdot)$ preserves the Gaussian measure if

$$P(\mathbf{G} \in A) = P(U(\mathbf{G}) \in A)$$

where $\mathbf{G} \sim N(\mathbf{0}, \mathbf{I})$.

- If T is a Gaussianization transform, then, $U \circ T$ is also a Gaussianization transform.
 - U : Orthogonal transform
 - U : In the polar coordinate system,
 - * fix the radius ρ .
 - * smoothly rotate the angle θ (to preserve the uniform distribution):

$$\theta' = \theta + f(\rho)$$

Analogy to Universal Coding

Universal Coding:

Discrete High Dim Variable \rightarrow i.i.d Bernoulli($\frac{1}{2}$)

Gaussianization:

Continuous High Dim Variable \rightarrow i.i.d Gaussian(0, 1)

Maximum Entropy Property:

Bernoulli($\frac{1}{2}$) achieves **maximum entropy** among

$\{X : \text{supported on two points}\}$

Gaussian(0, 1) achieves **maximum entropy** among

$\{X : \text{supported on } \mathcal{R}, EX = 0, VarX = 1\}$

Gaussianization with Linear ICA Assumption

Key idea: Independence + Marginal Univariate Gaussianization

Linear ICA assumption: Independence after linear transform

Gaussianization can be achieved in two steps:

(1) Linearly transform to independent coordinates:

$$\mathbf{Y} = A\mathbf{X}$$

(2) Marginally Gaussianize the independent coordinates:

$$\mathbf{Z} = \Psi(\mathbf{Y}) \approx \Psi_{\pi, \mu, \sigma}(\mathbf{Y})$$

Equivalently, we model each independent coordinate by univariate Gaussian mixture:

$$p_{Y_d}(y_d) = \sum_{i=1}^{I_d} \pi_{d,i} \phi(y_d, \mu_{d,i}, \sigma_{d,i}^2)$$

Efficient EM Algorithm

Equivalent Maximum Likelihood Problem:

$$p_{\mathbf{X}}(\mathbf{x}) = |\det(A)| \prod_{d=1}^D \sum_{i=1}^{I_d} \pi_{d,i} \phi(y_d, \mu_{d,i}, \sigma_{d,i}^2) .$$

where $y_d = \mathbf{a}_d^T \mathbf{x}$ and \mathbf{a}_d is the d -th row of A .

Parameters $\theta = (A, \pi, \mu, \sigma)$

EM algorithm: M-step has to be solved by iterative methods

- Attias (1999): gradient descent with natural gradient
 - Row-by-row update of A - adapted from Gales (1999)
 - guarantee increasing the auxiliary function
 - no nuisance of choosing the stepsize
- **stability and faster convergence**

Negentropy

Negentropy:

$$J(\mathbf{X}) = D(X \| N(\mathbf{0}, \mathbf{I})) = D(UX \| UN(\mathbf{0}, \mathbf{I})) = D(UX \| N(\mathbf{0}, \mathbf{I})) = J(U\mathbf{X})$$

– Negentropy is invariant to an orthogonal linear transform U .

Marginal Negentropy:

$$J_M(\mathbf{X}) = \sum_{d=1}^D D(X_d \| N(0, 1))$$

Mutual Information:

$$I(\mathbf{X}) = \int p_{\mathbf{X}}(x_1, \dots, x_n) \log \frac{p_{\mathbf{X}}(x_1, \dots, x_n)}{p_{X_1}(x_1) \dots p_{X_n}(x_n)}$$

– coordinate-wise transform will not change the mutual information.

Key Equality:

$$J(\mathbf{X}) = J_M(\mathbf{X}) + I(\mathbf{X})$$

Minimizing the Negentropy

Gaussianization:

$$J(T(\mathbf{X})) = 0$$

Gaussianization with ICA assumption:

(1) Linearly transform to the independent coordinates:

$$J(A\mathbf{X}) = J_M(A\mathbf{X}) + I(A\mathbf{X}) = J_M(A\mathbf{X}) .$$

(2) Marginally Gaussianize each coordinate:

$$J(\Psi(A\mathbf{X})) = J_M(\Psi(A\mathbf{X})) + I(\Psi(A\mathbf{X})) = J_M(A\mathbf{X}) = 0 .$$

Minimizing the Negentropy as Maximum Likelihood

Minimizing the KL Divergence:

$$\begin{aligned} D(T_\theta(\mathbf{X})||\mathbf{Y}) &= D(\mathbf{X}||T_\theta^{-1}(\mathbf{Y})) = \int p_{\mathbf{X}}(\mathbf{x}) \log \frac{p_{\mathbf{X}}(\mathbf{x})}{p_{T_\theta^{-1}(\mathbf{Y})}(\mathbf{x})} d\mathbf{x} \\ &= E_{\mathbf{X}} \log p_{\mathbf{X}}(\mathbf{X}) - E_{\mathbf{X}} \log p_{T_\theta^{-1}(\mathbf{Y})}(\mathbf{X}) \\ &= H(\mathbf{X}) - E_{\mathbf{X}}(\log p_\theta(\mathbf{X})) \end{aligned}$$

where $p_\theta(\cdot)$ is the density of $T_\theta^{-1}(\mathbf{Y})$.

Maximum Likelihood:

$$\max E_{\mathbf{X}}(\log p_\theta(\mathbf{X})) \iff \max \sum_{n=1}^N \log p_\theta(\mathbf{x}_n)$$

Minimizing the Negentropy

Minimizing the negentropy:

$$\min D(\Psi_{\pi, \mu, \sigma}(A\mathbf{X}) || N(\mathbf{0}, \mathbf{I}))$$

Equivalent ML model: density of $A^{-1}\Psi_{\pi, \mu, \sigma}^{-1}(N(\mathbf{0}, \mathbf{I}))$

$$p_{\theta}(\mathbf{x}) = |\det(A)| \prod_{d=1}^D \sum_{i=1}^{I_d} \pi_{d,i} \phi(y_d, \mu_{d,i}, \sigma_{d,i}^2) .$$

the same ML solution via EM!

Properties of Minimizing the Negentropy

If $\min J(\Psi(A\mathbf{X})) = 0$: finding the independent coordinates!

If $\min J(\Psi(A\mathbf{X})) > 0$:

- finding the least dependent coordinates:

$$J(\Psi(A\mathbf{X})) = J_M(\Psi(A\mathbf{X})) + I(\Psi(A\mathbf{X})) = I(A\mathbf{X}).$$

- if A constrained orthogonal, finding the marginally most non-Gaussian coordinates:

$$\min I(A\mathbf{X}) \iff \max J_M(A\mathbf{X}).$$

since

$$J(\mathbf{X}) = J(A\mathbf{X}) = J_M(A\mathbf{X}) + I(A\mathbf{X})$$

PCA vs ICA

Linear transform $Y = AX$, $B = A^{-1}$.

- PCA: Model y_d as **single Gaussian**.

$$p(y_d) = \phi(y_d, \mu_d, \sigma_d^2) \quad \text{i.e.} \quad \mathbf{X} \sim N(B\mu, B \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_D^2 \end{pmatrix} B^T)$$

$$\max_{A, \mu, \sigma} L \quad \Rightarrow \quad \min_A \text{Cor}(AX)$$

- ICA: Model y_d as **non-Gaussian**.

Assuming univariate Gaussian mixture is adequate.

$$\max_{A, \pi, \mu, \sigma} L \quad \Rightarrow \quad \min_A I(AX)$$

Iterative Gaussianization

Key idea: Independence + Marginal Univariate Gaussianization

Arbitrary random variable $\mathbf{X} \in R^D$, let $\mathbf{X}^{(0)} = \mathbf{X}$. At each iteration,

(1) Linear transform :

$$\mathbf{Y}^{(k)} = A\mathbf{X}^{(k)}$$

(2) Nonlinear marginal Gaussianization:

$$\mathbf{X}^{(k+1)} = \Psi(\mathbf{Y}^{(k)}) \approx \Psi_{\pi, \mu, \sigma}(\mathbf{Y}^{(k)})$$

where the parameters $\theta = (A, \pi, \mu, \sigma)$ are chosen via

$$\min_{\theta} J(\mathbf{X}^{(k+1)})$$

Intuition:

$$J(\mathbf{X}^{(0)}) > J(\mathbf{X}^{(1)}) > \dots > J(\mathbf{X}^{(k)}) \rightarrow 0$$

Engine: *Minimizing the negentropy is solved by the same ML algorithm!*

Properties of Iterative Gaussianization

At each iteration minimizing $J(\Psi(A\mathbf{X}^{(k)}))$ is equivalent to

- finding the least dependent coordinates.
- if A constrained orthogonal, finding the marginally most non-Gaussian coordinates.

Convergence result:

$$X^{(k)} \xrightarrow{\mathcal{D}} N(0, \mathbf{I})$$

adapted from the convergence proof of projection pursuit (Huber 1985).

A Relaxed Algorithm

less dependent coordinates instead of least dependent coordinates:

At each iteration, linearly transform the data into coordinates which are **less** dependent:

$$I(\mathbf{X}^{(k)}) - I(A\mathbf{X}^{(k)}) \geq \epsilon \left[I(\mathbf{X}^{(k)}) - \inf_A I(A\mathbf{X}^{(k)}) \right]$$

where the constant $\epsilon > 0$.

Convergence result still holds!

$$\mathbf{X}^{(k)} \rightarrow N(\mathbf{0}, \mathbf{I})$$

Iterative Gaussianization Density Estimation

K -step iterative Gaussianization:

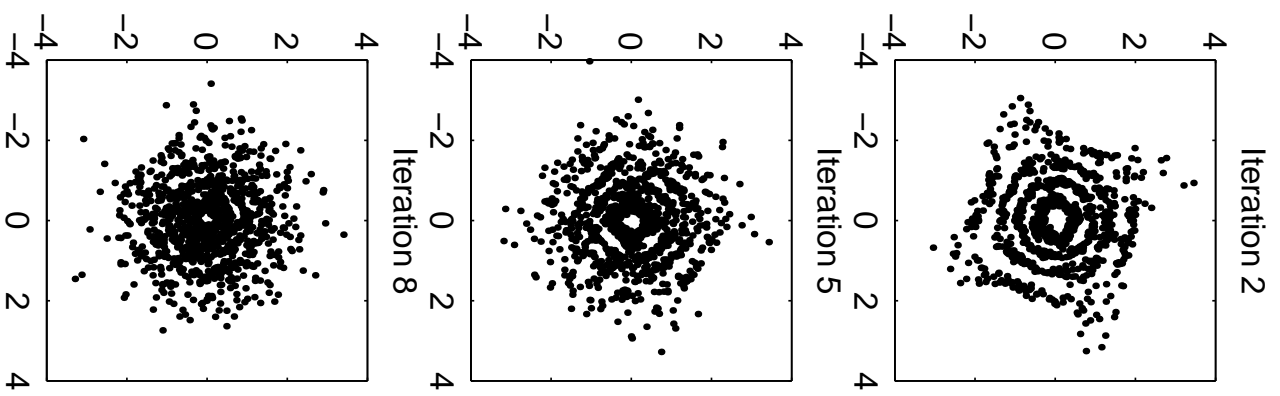
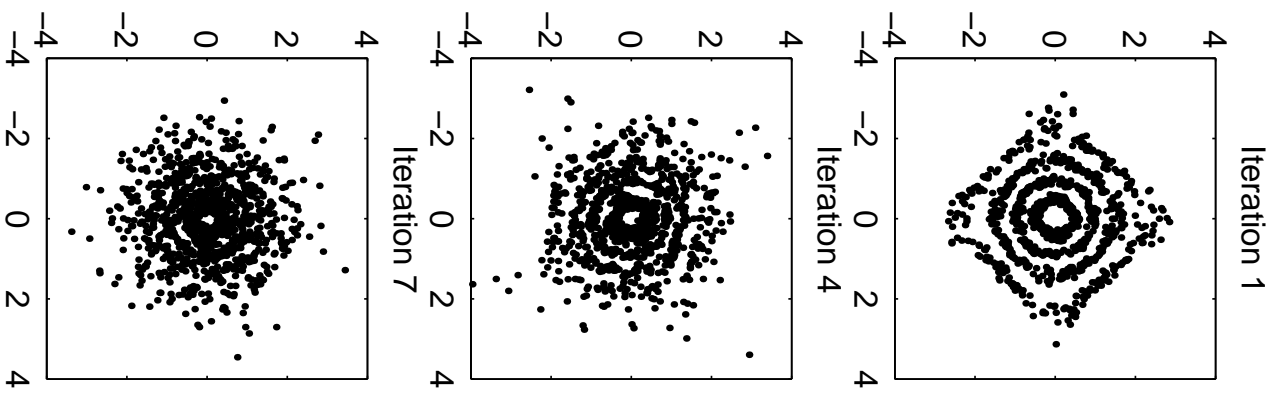
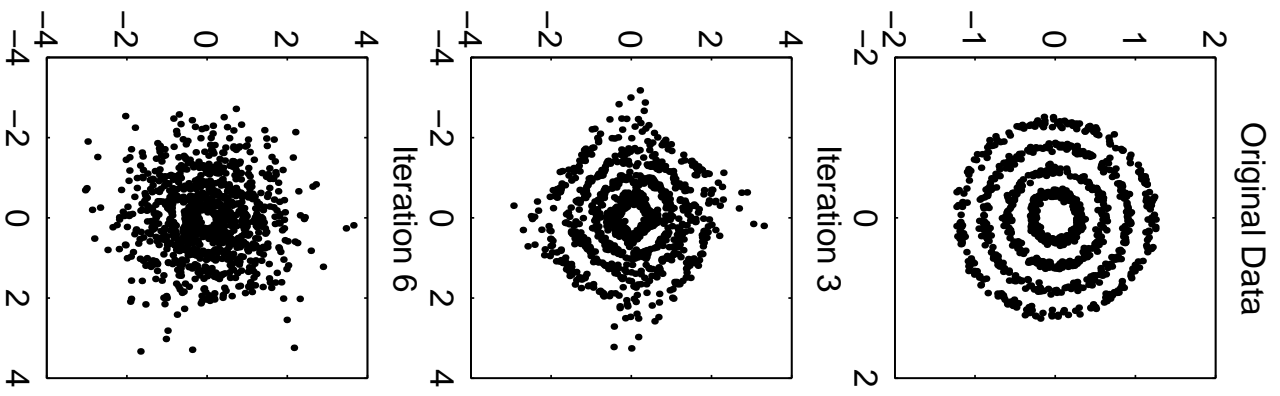
$$\mathbf{X}^{(K)} \sim N(\mathbf{0}, \mathbf{I})$$

Density estimation:

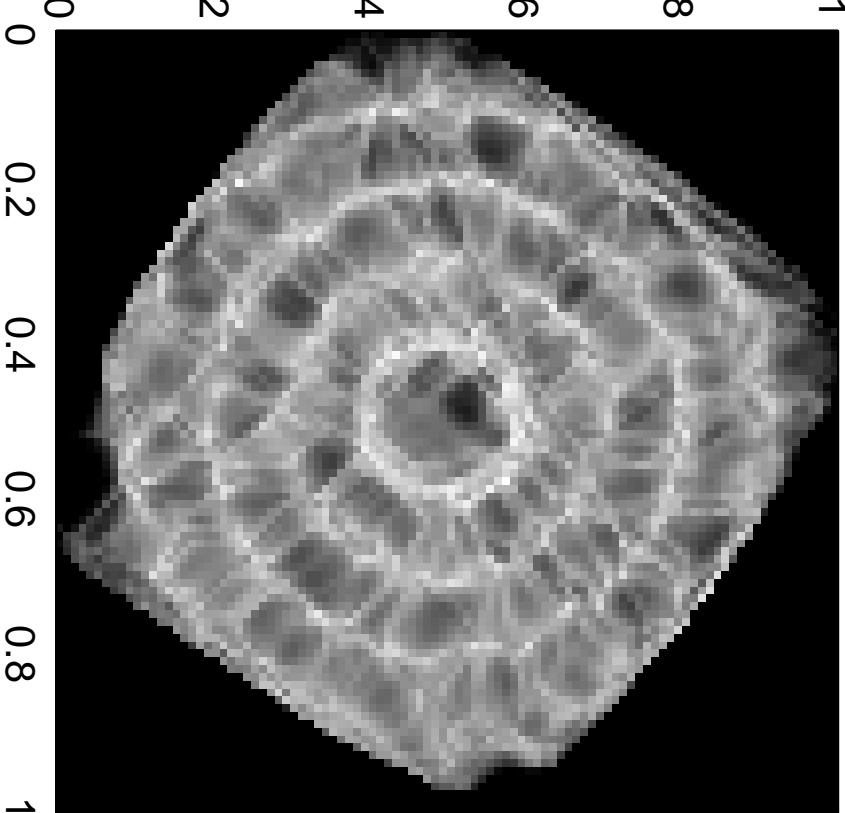
$$\begin{aligned} \hat{p}_{\mathbf{X}}^{(k)}(\mathbf{x}) &= \phi(\mathbf{x}^{(K)}) \left| \frac{\partial \mathbf{x}^{(K)}}{\partial \mathbf{x}} \right| = \phi(\mathbf{x}^{(K)}) \prod_{k=1}^K \left| \frac{\partial \mathbf{x}^{(k)}}{\partial \mathbf{x}^{(k-1)}} \right| \\ &= \phi(\mathbf{x}^{(K)}) \prod_{k=1}^K |\det(A^{(k)})| \prod_{d=1}^D \left(\sum_{i=1}^{I_d} \pi_{i,d} \phi\left(\frac{y_d^{(k)} - \mu_{i,d}}{\sigma_{i,d}}\right) \right) / \phi(\mathbf{x}_d^{(k+1)}) \end{aligned}$$

Convergence result:

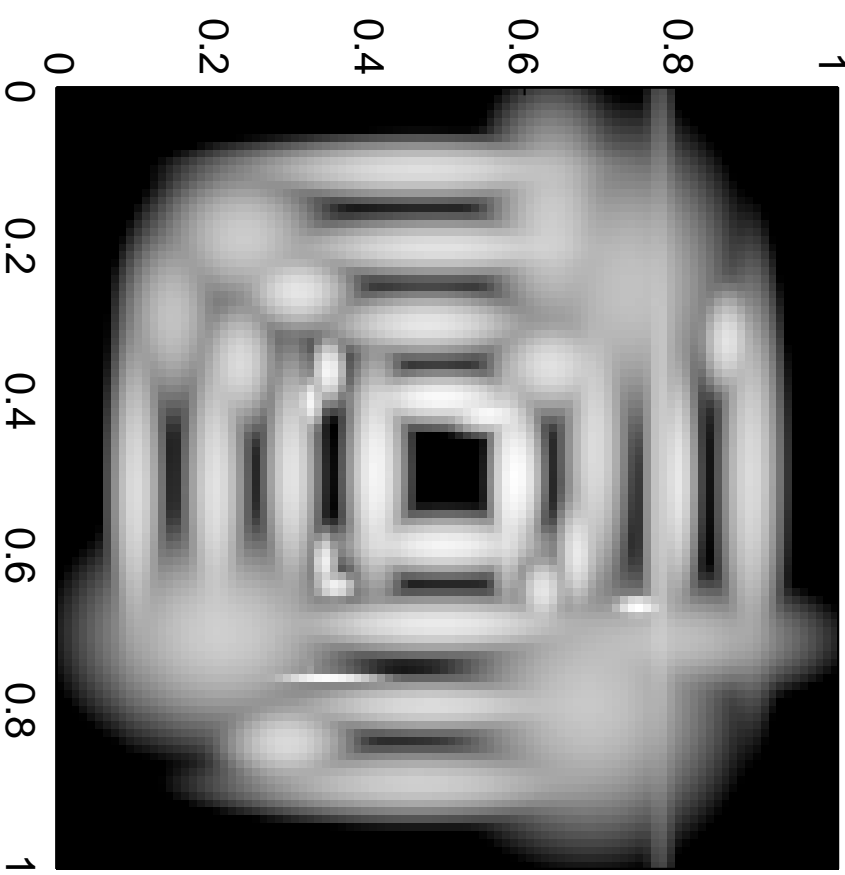
$$\lim_{k \rightarrow \infty} \hat{p}_{\mathbf{X}}^{(k)}(\mathbf{x}) = p_{\mathbf{X}}(\mathbf{x})$$



Gaussianization Density Estimation



Gaussian Mixture Density Estimation



Advantages of Gaussianization Density Estimation

Iterative Gaussianization alleviates the curse of dimensionality:

- At each iteration, linearly transform the current data to less dependent coordinates.
- At each iteration, perform univariate density estimation.

Gaussianization density estimates can outperform kernel methods!

Robust to overfitting as the number of iterations increases: the density estimate seems always improving!

- Since our algorithm is greedy, the parameters in each iteration are estimated independently.
- If the parameters are estimated jointly by ML, one would expect overfitting.

Diagonal Gaussian Mixture Models

Gaussian mixture models with diagonal covariances are inefficient to describe

- Correlation among coordinates:
more components along the principal axis of the underlying cov matrix.
- Local non-Gaussianity of each coordinate: assume the truth is

$$f(x_1, \dots, x_D) = \prod_{d=1}^D \sum_{i=1}^{I_d} \pi_{(d,i)} \phi(x_d, \mu_{(d,i)}, \sigma_{(d,i)}^2) .$$

$$\#terms = \sum_{d=1}^D I_d$$

Modeled as a mixture of Gaussians with diagonal covariances

$$f(x_1, \dots, x_D) = \sum_{i_1, \dots, i_D} \prod_{d=1}^D \pi_{(d,i_d)} \phi(x_d, \mu_{(d,i_d)}, \sigma_{(d,i_d)}^2) .$$

$$\#terms = \prod_{d=1}^D I_d$$

Semi-Tied Covariances

Gales 1999 proposed to model the covariances of the Gaussians in the mixture model by

$$\Sigma_k = B \begin{pmatrix} \sigma_{k,1}^2 & & & 0 \\ & \ddots & & \\ & & & \\ 0 & & & \sigma_{k,D}^2 \end{pmatrix} B^T .$$

Equivalently the semi-tied covariances can be viewed as feature space transform. The density can be rewritten as

$$p(\mathbf{x}; \theta) = \sum_{k=1}^K \rho_k |det(A)| \prod_{d=1}^D \phi(\mathbf{a}_d \mathbf{x}, \mu_{k,d}, \sigma_{k,d}^2) \quad (1)$$

where $A = B^{-1}$ and \mathbf{a}_d is the d -th row of the matrix A .

Efficient EM algorithm.

Significantly improve the word error rate.

Generalized Gaussian Mixture Models via Gaussianization

Assume that the Linear ICA assumption holds locally:

- Locally the data can be linearly transformed to independent coordinates
- Locally the transformed coordinates modeled as univariate Gaussian mixtures

Density

$$p(\mathbf{x}; \theta) = \sum_{k=1}^K \rho_k |det(A)| \prod_{d=1}^D \sum_{i=1}^{I_{k,d}} \pi_{k,d,i} \phi(\mathbf{a}_{d\mathbf{x}}, \mu_{k,d,i}, \sigma_{k,d,i}^2) \cdot$$

Same efficient EM algorithm.

Projection Pursuit via Constrained Gaussianization

Two constraints: at each iteration

- Orthogonal linear transform A
- Marginally Gaussianize the first l coordinates.

$$\min J(\Psi^l(A\mathbf{X}^{(k)})) \Rightarrow \max J_M^l(A\mathbf{X}^{(k)})$$

Advantages:

- Find the most **Marginally** non-Gaussian l -D projection
- **Marginally** Gaussianize that l -D variable
- much simplified EM algorithm

Conclusion

Independence lifts the Curse of Dimensionality!

Development of Gaussianization

- (1) Theory:
 - Linear Independent Component Analysis (Bell & Sejnowski 1995)
 - Projection Pursuit (Friedman 1987, Huber 1985)
- (2) Algorithm: Semi-Tied Covariances (Gales 1999)
- (3) By-products:
 - Iterative Gaussianization Density Estimation
 - Generalized Gaussian mixture models
 - Efficient Solution for Linear and Nonlinear ICA
 - Efficient Solution for High dimensional projection pursuit
 - Efficient Algorithm for Laplacianization, Uniformization, etc.