Gaussianization

Ramesh A. Gopinath
IBM T. J. Watson Research, Yorktown Heights, NY

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:**

<table>
<thead>
<tr>
<th></th>
<th>Logarithm</th>
<th>1-d Gaussianization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word Error Rate</td>
<td>18.5%</td>
<td>18.1%</td>
</tr>
</tbody>
</table>

3
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!
Bias vs Variance:
- Small neighborhoods are almost empty: big variance!
- Big neighborhoods to achieve sufficient counts: big bias!

Kernel methods: \((x_i \in \mathcal{R}^D : 1 \leq i \leq N)\)

\[
\hat{p}(x) = \frac{1}{N|\text{det}(H)|} \sum_{n=1}^{N} K(H^{-1}(x - x_n))
\]

Error rate of adaptive kernel methods:

\[
E \int |\hat{p}(x) - p(x)|^2 dx \sim N^{-4/(D+4)}
\]

Equivalent sample size for achieving error \(\epsilon\):

\[
N \sim \left(\frac{1}{\epsilon}\right)^{(D+4)/4}
\]
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 0.1 (Kolmogorov, 1957)** Let \( x = (x_1, \cdots, x_D) \in [0, 1]^D \). There exist \( D \) universal constants

\[
\{ \lambda_d : 1 \leq d \leq D \}
\]

and \((2D + 1)\) universal univariate functions

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

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

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

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

\[
Y^{(k)} = U^{(k)} X^{(k)}
\]

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

\[
X^{(k+1)}_1 = \Psi(Y^{(k)}_1)
\]

\[
X^{(k+1)}_d = Y^{(k)}_d \quad 2 \leq d \leq D
\]

\[
X^{(k)} \xrightarrow{\mathcal{D}} N(0, I)
\]
Projection index

- characterization of $N(0, 1)$

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

- departure from normality

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

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

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

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

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

\[ X^{(K)} \sim N(0, I) \]

\[ p_X(x) = \phi(x^{(K)}) \left| \frac{\partial x^{(K)}}{\partial x} \right| = \phi(x^{(K)}) \prod_{k=1}^{K} \left| \frac{\partial x^{(k)}}{\partial x^{(k-1)}} \right| \]

Projection Pursuit alleviates the curse of dimensionality:

- At each iteration, find the projection which is mostly non-

- At each iteration, perform univariate density estimation.
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:

\[
Err = \frac{1}{M} \sum_{m=1}^{M} (\hat{p}(x_m) - p(x_m))^2
\]
2D Projection Pursuit

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

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

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

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

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

   - Rotate about the origin through angle $\gamma$:

     $$
     \begin{align*}
     Y_1' &= Y_1 \cos \gamma + Y_2 \sin \gamma \\
     Y_2' &= Y_2 \cos \gamma - Y_1 \sin \gamma
     \end{align*}
     $$

   - Gaussianize $Y_1'$ and $Y_2'$ individually.

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

   - Stop if the distributions stop becoming more Gaussian.
Gaussianization

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

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

Density estimation:

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

$$T_{\pi, \mu, \sigma}(X) = \Phi^{-1}\left[ \sum_{i} \pi_i \Phi\left(\frac{X - \mu_i}{\sigma_i}\right) \right]$$
Notation $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$.

∀ 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 $x_2|x_1$

\[ p(x_2|y_1 = 1) \xrightarrow{T(1, \cdot)} \phi(y_2) \]
\[ p(x_2|y_1 = 2) \xrightarrow{T(2, \cdot)} \phi(y_2) \]

\[ (y_1, y_2) = T(y_1, x_2) = (y_1, \Phi^{-1} \circ F_{x_2|y_1}(x_2)), \quad y_1 = \Phi^{-1} \]
\[ \Rightarrow p(y_1, y_2) = \phi(y_1)\phi(y_2) \]
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
  
  $$P(G \in A) = P(U(G) \in A)$$

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

- 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 $\rho$.
    * smoothly rotate the angle $\theta$ (to preserve the uniform distribution):
      
      $$\theta' = \theta + f(\rho)$$

  18
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:
   \[ Y = AX \]

2. Marginally Gaussianize the independent coordinates:
   \[ Z = \Psi(Y) \approx \Psi_{\pi, \mu, \sigma}(Y) \]

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:

\[ p_X(x) = |\text{det}(A)| \prod_{d=1}^{D} \sum_{i=1}^{I_d} \pi_{d,i} \phi(y_{d,i}, \mu_{d,i}, \sigma_{d,i}^2). \]

where \( y_d = a_d x \) and \( a_d \) 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
  \[ \rightarrow \text{stability and faster convergence} \]
Negentropy:

\[ J(X) = D(X \| N(0, I)) = D(UX \| UN(0, I)) = D(UX \| N(0, I)) \]

- Negentropy is invariant to an orthogonal linear transform \( U \).

Marginal Negentropy:

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

Mutual Information:

\[ I(X) = \int p_X(x_1, \cdots, x_n) \log \frac{p_X(x_1, \cdots, x_n)}{p_{X_1}(x_1) \cdots p_{X_n}(x_n)} \]

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

Key Equality:

\[ J(X) = J_M(X) + I(X) \]
Minimizing the Negentropy

Gaussianization:

\[ J(T(\mathbf{X})) = 0 \]

Gaussianization with ICA assumption:

1. Linearly transform to the independent coordinates:

\[ J(AX) = J_M(AX) + I(AX) = J_M(AX) \]

2. Marginally Gaussianize each coordinate:

\[ J(\Psi(AX)) = J_M(\Psi(AX)) + I(\Psi(AX)) = J_M(AX) \]
Minimizing the KL Divergence:

\[
D(T_\theta(X) \parallel Y) = D(X \parallel T_\theta^{-1}(Y)) = \int p_X(x) \log \frac{p_X(x)}{p_{T_\theta^{-1}(Y)}(x)} \, dx = E_X \log p_X(X) - E_X \log p_{T_\theta^{-1}(Y)}(X) = H(X) - E_X(\log p_\theta(X))
\]

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

Maximum Likelihood:

\[
\max E_X(\log p_\theta(X)) \iff \max \sum_{n=1}^{N} \log p_\theta(x_n)
\]
Minimizing the negentropy:

\[
\min D(\Psi_{\pi,\mu,\sigma}(AX)||N(0, I))
\]

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

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

the same ML solution via EM!
Properties of Minimizing the Negentropy

If $\min J(\Psi(AX)) = 0$: finding the independent coordinates!

If $\min J(\Psi(AX)) > 0$:

- finding the least dependent coordinates:

  $$J(\Psi(AX)) = J_M(\Psi(AX)) + I(\Psi(AX)) = I(AX)$$

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

  $$\min I(AX) \iff \max J_M(AX).$$

  since

  $$J(X) = J(AX) = J_M(AX) + I(AX)$$
Linear transform \( Y = AX, \ B = A^{-1}. \)

- **PCA**: Model \( y_d \) as single Gaussian.
  \[
  p(y_d) = \phi(y_d, \mu_d, \sigma_d^2) \quad \text{i.e.} \quad X \sim N(B\mu, B \begin{pmatrix}
  \sigma_1^2 \\
  \vdots \\
  0
\end{pmatrix} B^T)
  \]

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

- **ICA**: Model \( y_d \) as non-Gaussian.
  Assuming univariate Gaussian mixture is adequate.

\[
\max_{A, \pi, \mu, \sigma} L \Rightarrow \min_A I(AX)
\]
Key idea: Independence + Marginal Univariate Gaussianization

Arbitrary random variable $\mathbf{X} \in \mathbb{R}^D$, 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:
\[
J(\mathbf{X}^{(0)}) > J(\mathbf{X}^{(1)}) > \cdots > J(\mathbf{X}^{(k)}) \to 0
\]

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:

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

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

\[ \mathbf{X}^{(k)} \rightarrow \mathcal{N}(\mathbf{0}, \mathbf{I}) \]
Iterative Gaussianization Density Estimation

*K*-step iterative Gaussianization:

\[ X^{(K)} \sim N(0, I) \]

Density estimation:

\[
\hat{p}_X^{(k)}(x) = \phi(x^{(K)}) \left| \frac{\partial x^{(K)}}{\partial x} \right| = \phi(x^{(K)}) \prod_{k=1}^{K} \left| \frac{\partial x^{(k)}}{\partial x^{(k-1)}} \right|
\]

\[
= \phi(x^{(K)}) \prod_{k=1}^{K} \left| \text{det}(A^{(k)}) \right| \prod_{d=1}^{D} \left( \sum_{i=1}^{I_d} \pi_{i,d} \phi \left( \frac{y_d^{(k)} - \mu_{i,d}}{\sigma_{i,d}} \right) \right)
\]

Convergence result:

\[
\lim_{k \to \infty} \hat{p}_X^{(k)}(x) = p_X(x)
\]
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 overfit.
Gaussian mixture models with diagonal covariances are inefficient to describe:

- Correlation among coordinates:
  more components along the principal axis of the underlying covariance matrix.

- Local non-Gaussianity of each coordinate: assume the truth is

\[
f(x_1, \ldots, x_D) = \prod_{d=1}^{D} \sum_{i=1}^{I_d} \pi_{(d,i)} \phi(x_d, \mu_{(d,i)}, \sigma^2_{(d,i)}).
\]

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

Modeled as a mixture of Gaussians with diagonal covariance:

\[
f(x_1, \ldots, x_D) = \sum_{i_1, \ldots, i_d} \prod_{d=1}^{D} \pi_{(d,i_d)} \phi(x_d, \mu_{(d,i_d)}, \sigma^2_{(d,i)}).
\]

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

35
Gales 1999 proposed to model the covariances of the Gaussian mixture model by

\[
\Sigma_k = B \begin{pmatrix} \sigma^2_{k,1} & 0 \\ \vdots & \ddots \\ 0 & \sigma^2_{k,D} \end{pmatrix} B^T.
\]

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

\[
p(x; \theta) = \sum_{k=1}^{K} \rho_k |\text{det}(A)| \prod_{d=1}^{D} \phi(a_d x, \mu_{k,d}, \sigma^2_{k,d})
\]

where \( A = B^{-1} \) and \( a_d \) is the \( d \)-th row of the matrix \( A \).

Efficient EM algorithm.

Significantly improve the word error rate.
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

\[ p(x; \theta) = \sum_{k=1}^{K} \rho_k |\text{det}(A)| \prod_{d=1}^{D} \sum_{i=1}^{I_{k,d}} \pi_{k,d,i} \phi(a_d x, \mu_{k,d,i}, \sigma_{k,d,i}^2) \]

Same efficient EM algorithm.
Two constraints: at each iteration

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

$$\min J(\Psi^l(AX^{(k)})) \Rightarrow \max J^l_M(AX^{(k)})$$

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