

Machine learning for vision

Fall 2013

Roland Memisevic

Lecture 6, October 8, 2013



Gaussians and independence

- ▶ Spherical Gaussians have independent marginals:

$$\begin{aligned} p(x_1, \dots, x_n) &= \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\|\mathbf{x}\|^2\right) \\ &= \prod_i \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x_i^2\right) \end{aligned}$$

- ▶ So for Gaussian data, whitening will give us the independent components.
- ▶ For non-Gaussian data this is *not* the case.



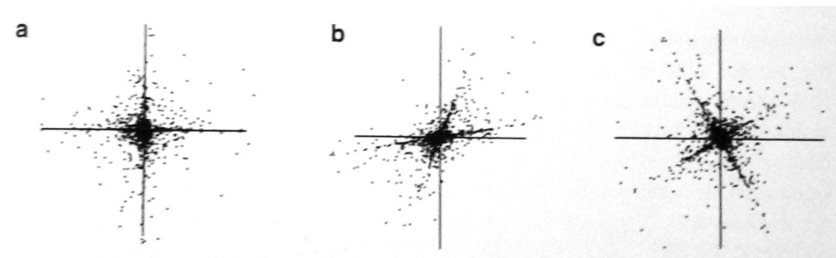
Correlation vs. dependence

Statistical independence implies uncorrelatedness.

Uncorrelatedness does *not* imply statistical independence.



A counter example

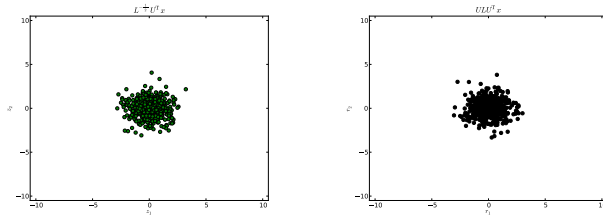


- ▶ a: Independent variables.
- ▶ b: A linear combination: not independent.
- ▶ c: After whitening the linear combination: still not independent.

from: Natural Image Statistics (Hyvarinen, Hurri, Hoyer; 2009)



Uncorrelatedness is not independence



- ▶ Any orthogonal transformation of white data is white.
- ▶ Therefore, PCA and ZCA are just two out of infinitely many whitening matrices.
- ▶ How can we find the matrix that maximizes *independence*?
- ▶ Since independence implies uncorrelatedness, it must still be a whitening matrix.

Independent components analysis

- ▶ Here, \mathbf{A} and \mathbf{W} are square matrices. We will later extend this to over- or undercomplete representations.
- ▶ Multiplying any component s_i by some scalar will have no effect if we divide the corresponding A_i by the same number.
- ▶ So we may assume the rows of \mathbf{A} to have some fixed length.
- ▶ It turns out that applying the model to already whitened components, \mathbf{z} , (e.g. from PCA or ZCA), will greatly simplify the learning.

Independent components analysis

- ▶ The ICA generative model: We have independent “source” variables s_i , which get mixed to yield the observed data

$$\mathbf{x} = \mathbf{A}\mathbf{s}$$

with

$$p(s_1, \dots, s_n) = \prod_i p_i(s_i)$$

- ▶ The corresponding analysis equation is thus

$$\mathbf{s} = \mathbf{W}^T \mathbf{x}$$

with $\mathbf{W}^T = \mathbf{A}^{-1}$

Maximum likelihood ICA

Densities under linear transformation

Given some random vector \mathbf{s} with density $p_s(\mathbf{s})$, the density of

$$\mathbf{x} = \mathbf{M}\mathbf{s}$$

is:

Maximum likelihood ICA

Densities under linear transformation

Given some random vector \mathbf{s} with density $p_s(\mathbf{s})$, the density of

$$\mathbf{x} = \mathbf{M}\mathbf{s}$$

is:

$$p_x(\mathbf{x}) = \frac{1}{|\det \mathbf{M}|} p_s(\mathbf{M}^{-1}\mathbf{x})$$



Maximum likelihood ICA

- ▶ For the ICA model, plug in $\mathbf{A} = \mathbf{W}^{-T}$ for \mathbf{M} .
- ▶ We have

$$\frac{1}{\det \mathbf{A}} = \frac{1}{\det \mathbf{W}^{-T}} = \frac{1}{\det \mathbf{W}^{-1}} = \det \mathbf{W}$$

- ▶ So we can write

$$p(\mathbf{x}) = |\det \mathbf{W}| p(\mathbf{W}\mathbf{x}) = |\det \mathbf{W}| \prod_{i=1}^n p_i(\mathbf{w}_i^T \mathbf{x})$$



Maximum likelihood ICA

- ▶ For IID observations, we get the following likelihood:

$$L(\mathbf{W}) = \prod_{t=1}^T p(\mathbf{x}_t) = \prod_{t=1}^T \left[|\det \mathbf{W}| \prod_{i=1}^n p_i(\mathbf{w}_i^T \mathbf{x}_t) \right]$$

- ▶ The log-likelihood is

$$\log L(\mathbf{W}) = T \log |\det \mathbf{W}| + \sum_{i=1}^n \sum_{t=1}^T \log p_i(\mathbf{w}_i^T \mathbf{x}_t)$$

- ▶ For whitened data, \mathbf{W} must be orthonormal, so

$$|\det \mathbf{W}| = 1$$

- ▶ In that case it is sufficient to maximize

$$\sum_{i=1}^n \sum_{t=1}^T \log p_i(\mathbf{w}_i^T \mathbf{z}_t)$$



Maximum likelihood ICA

- ▶ This leads to the constrained optimization problem

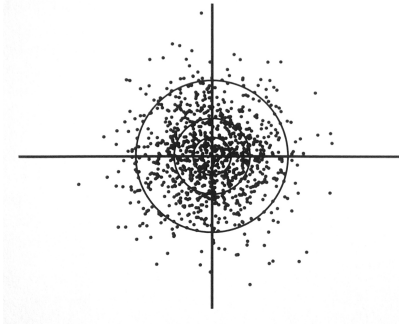
$$\begin{aligned} &\text{minimize} && \sum_t \sum_i \phi(\mathbf{w}_i^T \mathbf{z}_t) \\ &\text{s.t.} && \mathbf{W}^T \mathbf{W} = \mathbf{I} \end{aligned}$$

where $-\phi(s_i) = \log p(s_i)$ is the log-pdf of the independent sources.

- ▶ How to choose $\phi(s_i)$?



Maximum likelihood for Gaussian sources



- ▶ The independent Gaussian is spherically symmetric.
- ▶ So rotation (orthogonal \mathbf{W}) will have no effect on the objective.
- ▶ Any *non*-Gaussian source distribution will.

Gaussian scale mixtures

- ▶ A possible explanation for super-Gaussianity in natural images is that any one feature may occur at different (brightness-)scales.
- ▶ We can model an image patch using a Gaussian g_i whose value is modulated by some independent scale-variable d_i :

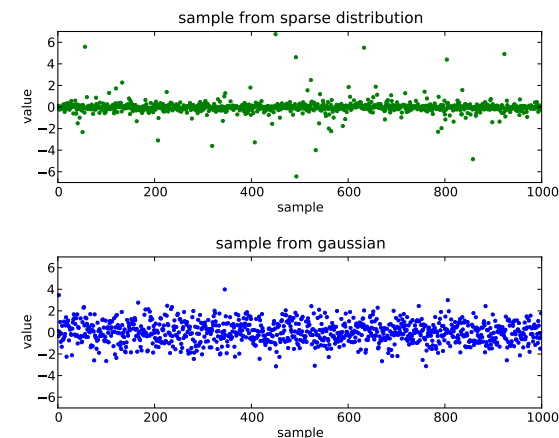
$$s_i = g_i d_i$$

- ▶ This yields a super-Gaussian distribution, because $p(s_i)$ will be a superposition of Gaussians each with different variance.
- ▶ In fact, contrast normalization seems to reduce sparsity (a bit).

Forms non-Gaussianity

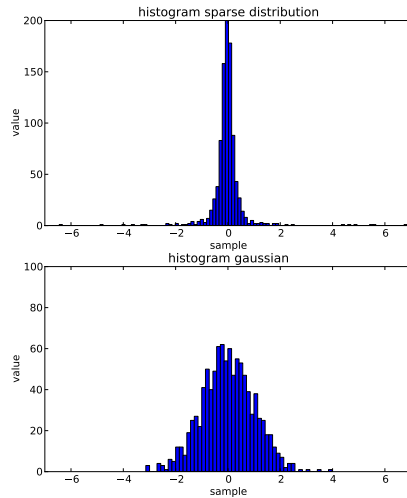
- ▶ Three possible forms of non-Gaussianity are
 1. Super-Gaussian (= *sparsity*): Distribution is peaked at zero (positive kurtosis)
 2. Sub-Gaussian: Distribution is “flat” at zero (negative kurtosis)
 3. Skew: Distribution is unsymmetric.
- ▶ Of these, super-Gaussianity is generally assumed to be the best match for (most) image data.
- ▶ Sparse representations are beneficial in many ways, so it is a natural choice of non-Gaussian density also for practical reasons.

Sparseness

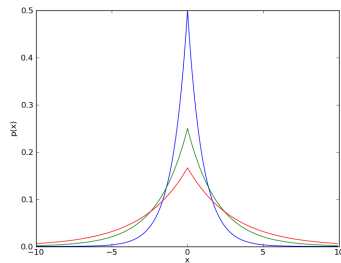


- ▶ Top: Samples from a student-T-distribution (sparse)
- ▶ Bottom: Samples from a normal distribution of the same variance (not sparse).

Histograms



Sparse source densities



- ▶ A popular choice of sparse source density is the zero-mean Laplacian:

$$p(s_i) = \frac{1}{2b} \exp\left(-\frac{|s_i|}{b}\right)$$

- ▶ Another, differentiable, one is the log cosh function.

Sparse does not mean “small values”

- ▶ Sparsity should not be confused with “small”.
- ▶ A normal distribution may be scaled to take on small values, too. This doesn’t make it sparse.
- ▶ Sparsity should always be thought of as *relative to a given variance*.

Sparse coding

- ▶ With a Laplacian source density, the optimization problem becomes

$$\begin{aligned} & \text{minimize} && \sum_t \sum_i |w_i^T z_t| \\ & \text{s.t.} && W^T W = I \end{aligned}$$

- ▶ This can be solved by alternating gradient steps with projections to enforce the constraint.

Orthogonality from reconstruction error

- ▶ An alternative to solving the constrained optimization problem is to enforce the orthogonality constraint

$$\mathbf{W}^{-1} = \mathbf{W}^T$$

implicitly.

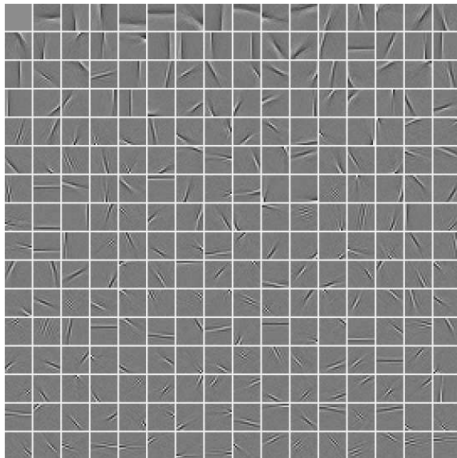
- ▶ By adding a reconstruction term we can encourage this as follows:

$$\text{minimize} \quad \sum_t \|\mathbf{W}\mathbf{W}^T \mathbf{z}_t - \mathbf{z}_t\|^2 + \lambda \sum_t \sum_i |\mathbf{w}_i^T \mathbf{z}_t|$$

- ▶ Note that we have $\mathbf{A} = \mathbf{W}$ now.



ICA filters



from: Natural Image Statistics (Hyvarinen, Hurri, Hoyer; 2009)



Search based inference

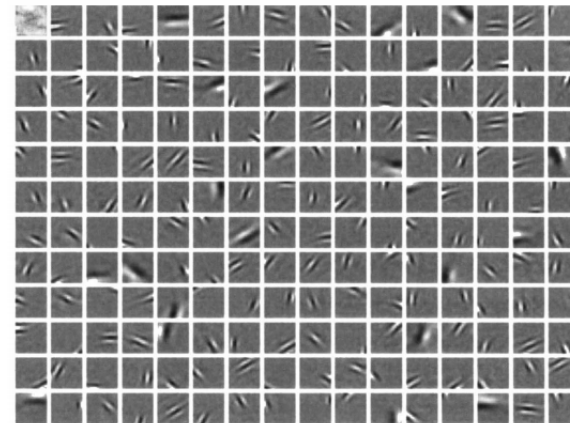
- ▶ All of the above versions of ICA make use of an encoder (\mathbf{W}) and decoder (\mathbf{A}).
- ▶ An alternative formulation one can find in the literature is to ignore the decoder and to *search* for the right \mathbf{s} during learning (and inference as well):

$$\text{minimize} \quad \sum_t \|\mathbf{A}^T \mathbf{s}_t - \mathbf{z}_t\|^2 + \sum_t \sum_i |s_{ti}|$$

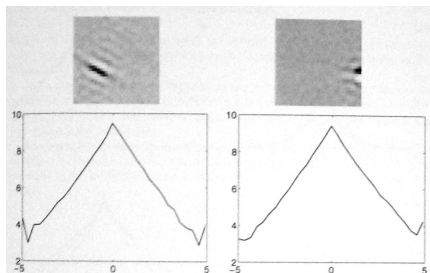
where the optimization is over \mathbf{A} and all the \mathbf{s}_t during training and over \mathbf{s} during inference.



Sparse coding components (Olshausen/Field)

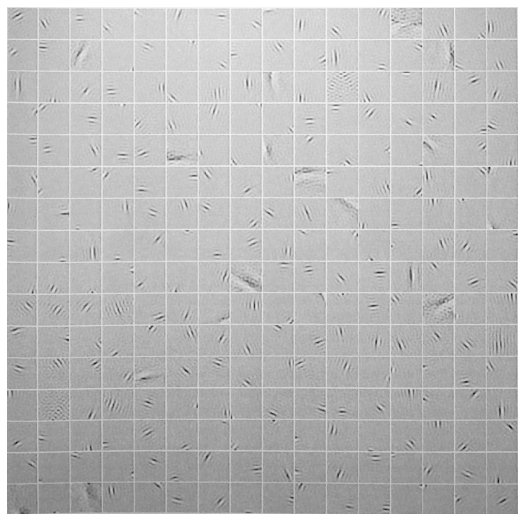


Estimating the source densities



- ▶ One can estimate the source densities from data as well.
- ▶ It turns out that not all components are sparse.
- ▶ The DC component, for example, tends to be sub-Gaussian.
- ▶ So to keep it or not can make a difference in practice!

Example analysis filters



from: Natural Image Statistics (Hyvarinen, Hurri, Hoyer; 2009)

Relation between analysis and synthesis weights

- ▶ Since the s_i are independent and have unit variance, the covariance matrix, \mathbf{C} , over input images can be written

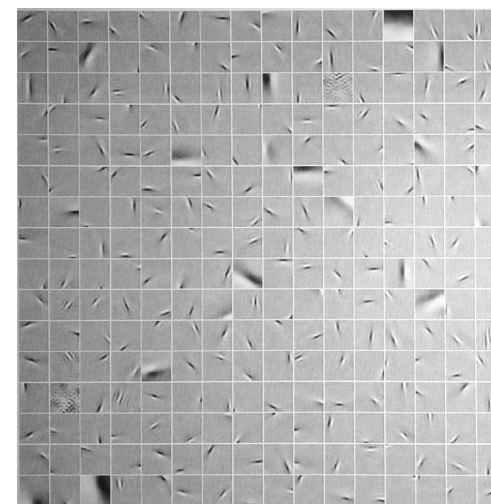
$$\mathbf{C} = \mathbf{A}\mathbf{A}^T$$

- ▶ From this, and $\mathbf{W} = \mathbf{A}^{-1}$, it follows that:

$$\begin{aligned}\mathbf{A}^T &= \mathbf{I} \mathbf{A}^T \\ &= (\mathbf{A}^{-1} \mathbf{A}) \mathbf{A}^T \\ &= \mathbf{W}^T \mathbf{A} \mathbf{A}^T \\ &= \mathbf{W}^T \mathbf{C}\end{aligned}$$

- ▶ In other words, synthesis weights, $\mathbf{A} = \mathbf{C}\mathbf{W}$, are equal to input covariance times analysis weights!

Example synthesis filters



from: Natural Image Statistics (Hyvarinen, Hurri, Hoyer; 2009)

Frequency channels

- ▶ The emergence of bandpass filters from whitening shows that 2d Fourier features (frequencies/orientations) are uncorrelated components.
- ▶ The emergence of Gabor features from ICA shows that (frequencies/orientations/positions) are independent components.
- ▶ What independence/sparsity adds is *locality*.
- ▶ This suggests that natural images are mainly invariant to *local translations*.

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍

Information theoretic interpretation

- ▶ To measure the independence of the sources we can use *mutual information* as follows:

$$\begin{aligned} \text{MI}(\mathbf{s}_1, \dots, \mathbf{s}_K) &= \int_{\mathbf{s}_1, \dots, \mathbf{s}_K} p(\mathbf{s}_1, \dots, \mathbf{s}_K) \log \frac{p(\mathbf{s}_1, \dots, \mathbf{s}_K)}{p(\mathbf{s}_1) \cdots p(\mathbf{s}_K)} d\mathbf{s}_1 \cdots d\mathbf{s}_K \\ &= \sum_{i=1}^K H(\mathbf{s}_i) - H(\mathbf{s}) \\ &= \sum_{i=1}^K H(\mathbf{w}_i^T \mathbf{z}) - H(\mathbf{W}\mathbf{z}) \end{aligned}$$

- ▶ For orthogonal \mathbf{W} , the joint entropy is constant.
- ▶ So to minimize MI, minimize the entropies of the individual components.

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍

Information theoretic interpretation

- ▶ To measure the independence of the sources we can use *mutual information* as follows:

$$\begin{aligned} \text{MI}(\mathbf{s}_1, \dots, \mathbf{s}_K) &= \int_{\mathbf{s}_1, \dots, \mathbf{s}_K} p(\mathbf{s}_1, \dots, \mathbf{s}_K) \log \frac{p(\mathbf{s}_1, \dots, \mathbf{s}_K)}{p(\mathbf{s}_1) \cdots p(\mathbf{s}_K)} d\mathbf{s}_1 \cdots d\mathbf{s}_K \\ &= \sum_{i=1}^K H(\mathbf{s}_i) - H(\mathbf{s}) \\ &= \sum_{i=1}^K H(\mathbf{w}_i^T \mathbf{z}) - H(\mathbf{W}\mathbf{z}) \end{aligned}$$

- ▶ For orthogonal \mathbf{W} , the joint entropy is constant.
- ▶ So to minimize MI, minimize the entropies of the individual components. → Make them less Gaussian!

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍

Infomax ICA

- ▶ Yet another information theoretic approach to ICA can be derived as follows:
- ▶ Maximum likelihood ($\hat{=}$ minimum entropy) under some density $f(x)$ can also be viewed as maximizing the *derivative of its cumulative distribution*
 $F(x) = \int f(x) dx$
- ▶ But this is equivalent to making the values under the cumulative distribution more uniform. (Also known as “maximum spacing” estimation in the literature.)
- ▶ To maximize uniformity of $F(x)$ we may *maximize* the entropy of the random variable $F(x)$ ($\hat{=}$ minimize its likelihood).
- ▶ (We can make the same argument for multivariate data \mathbf{x})

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ ⏸ ⏹ ⏺ ⏻ ⏼ ⏽ ⏾ ⏿ 🔍

Infomax ICA

- ▶ But $F(\mathbf{x})$ is a deterministic function of \mathbf{x} , so we can write its density as a function of \mathbf{x} .
- ▶ In particular, the transformation of densities under non-linear transformations is

$$p_F(F(\mathbf{x})) = \frac{1}{|\det J(\mathbf{x})|} p_x(\mathbf{x})$$

where $J(\mathbf{x}_t)$ is the Jacobian of $F(\mathbf{x})$ at \mathbf{x}

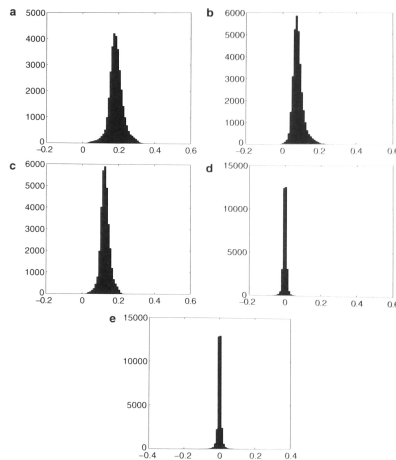
- ▶ By eliminating terms that do not depend on the parameters of F , we get the optimization problem:

$$\text{minimize} \quad - \sum_t \log |\det J(\mathbf{x}_t)|$$

- ▶ (Bell, Sejnowski 1997)



Non-linear correlations, limitations of ICA



- ▶ a: $f(s) = |s|$
- ▶ b: $f(s) = s^2$
- ▶ c: $f(s) = |s| > 1$
- ▶ d: $f(s) = \text{sign}(s)$
- ▶ e: $f(s) = s^3$

from: Natural Image Statistics
(Hyvarinen, Hurri, Hoyer; 2009)



Non-linear correlations, limitations of ICA

- ▶ For independent random variables, s_1, s_2 , the following must be true:

$$\text{cov}(f(s_1), f(s_2)) = 0$$

where f is any non-linear function.

- ▶ There are infinitely many f -functions we could choose from.
- ▶ This hints at the fact that a linear ICA transformation with its n^2 parameters may not be able to yield perfectly independent components.



Blind source separation

- ▶ A classic application of ICA is separating (unmixing) audio sources:
- ▶ Eg. [http://cni.salk.edu/~tewon/Blind/blind audio.html](http://cni.salk.edu/~tewon/Blind/blind%20audio.html)

