Machine learning for vision

Winter 2015

Roland Memisevic

Lecture 5, February 10, 2015

Correlation vs. dependence

Statistical independence implies uncorrelatedness.

Uncorrelatedness does *not* imply statistical independence.

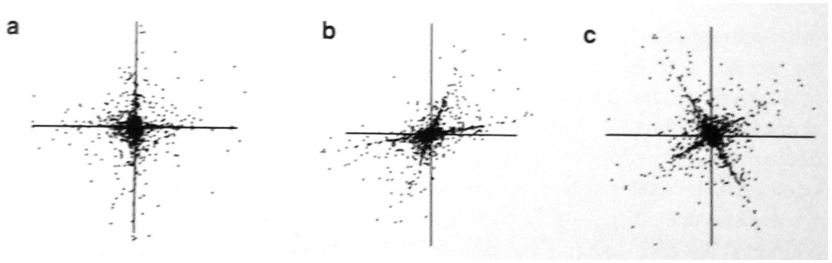
Gaussians and independence

- Spherical Gaussians have independent marginals:

$$p(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\|\boldsymbol{x}\|^2\right)$$
$$= \prod_i \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x_i^2\right)$$

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

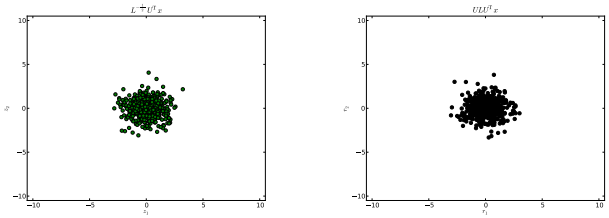
A counter example



- 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

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

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

with

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

- The corresponding analysis equation is thus

$$\boldsymbol{s} = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}$$

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

Independent components analysis

- Here, $\boldsymbol{A}$ and $\boldsymbol{W}$ 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 $\boldsymbol{A}$ to have some fixed length.
- Applying the model to already whitened components, $\boldsymbol{z}$, (e.g. from PCA or ZCA), will greatly simplify learning because it allows us to search for $\boldsymbol{A}$ among orthogonal matrices.

Maximum likelihood ICA

Densities under linear transformation

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

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

is:

Maximum likelihood ICA

Densities under linear transformation

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

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

is:

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

Maximum likelihood ICA

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

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

- So we can write

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

Maximum likelihood ICA

- For IID observations, we get the following likelihood:

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

- The log-likelihood is

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

- For whitened data, $\boldsymbol{W}$ must be orthonormal, so

$$\left|\det \boldsymbol{W}\right| = 1$$

- In that case it is sufficient to maximize

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

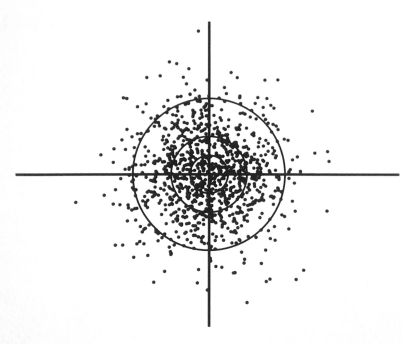
Maximum likelihood ICA

- This leads to the constrained optimization problem

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

  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 $\boldsymbol{W}$) will have no effect on the objective.
- Any *non*-Gaussian source distribution will.

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.

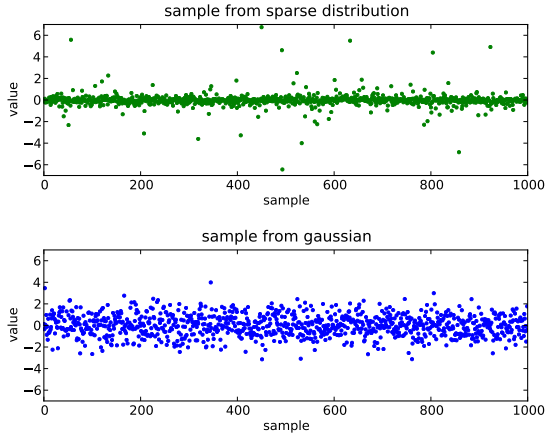
Gaussian scale mixtures

- 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:

$$s_i = g_i d_i$$

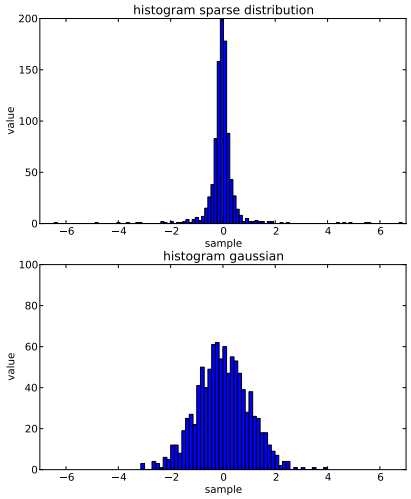
- This yields a super-Gaussian distribution, because $p(s_i)$ will be a superposition of Gaussians each with different variance.
- In fact, contrast normalization seems to reduce sparsity (a bit).

Sparseness



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

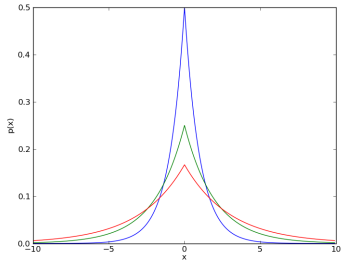
Histograms



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 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 coding

- With a Laplacian source density, the optimization problem becomes

$$\begin{aligned} \text{minimize} \quad & \sum_t \sum_i |\boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{z}_t| \\ \text{s.t.} \quad & \boldsymbol{W}^{\mathrm{T}} \boldsymbol{W} = \boldsymbol{I} \end{aligned}$$

- 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

$$\boldsymbol{W}^{-1} = \boldsymbol{W}^{\mathrm{T}}$$

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

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

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

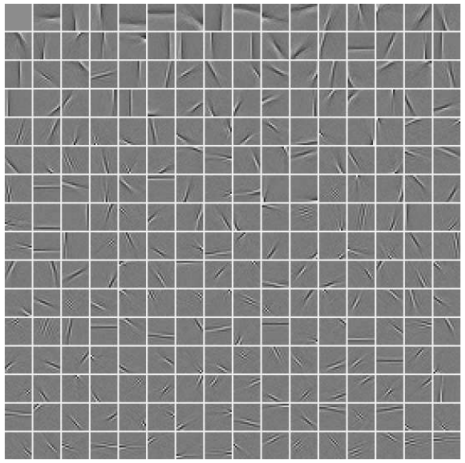
Search based inference

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

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

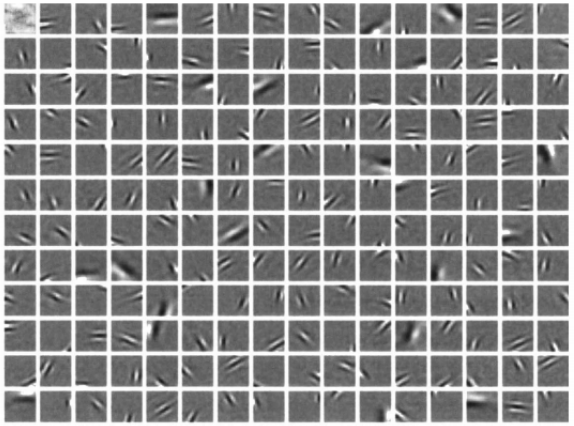
  where the optimization is over $\boldsymbol{A}$ and all the $\boldsymbol{s}_t$ during training and over $\boldsymbol{s}$ during inference.

ICA filters

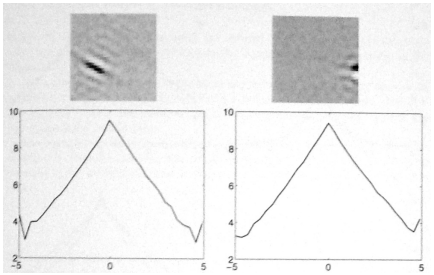


from: Natural Image Statistics (Hyvarinen, Hurri, Hoyer; 2009)

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!

Relation between analysis and synthesis weights

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

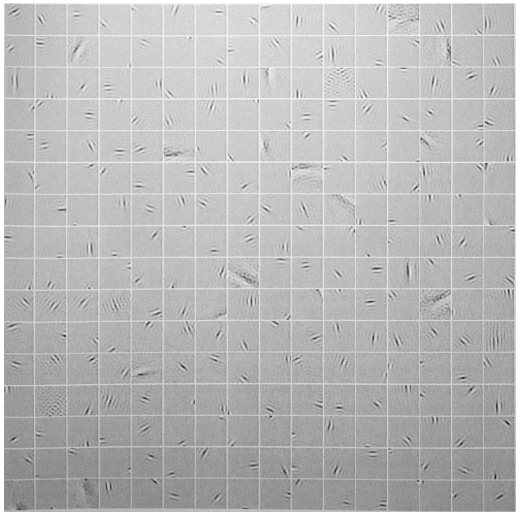
$$\boldsymbol{C} = \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$$

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

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

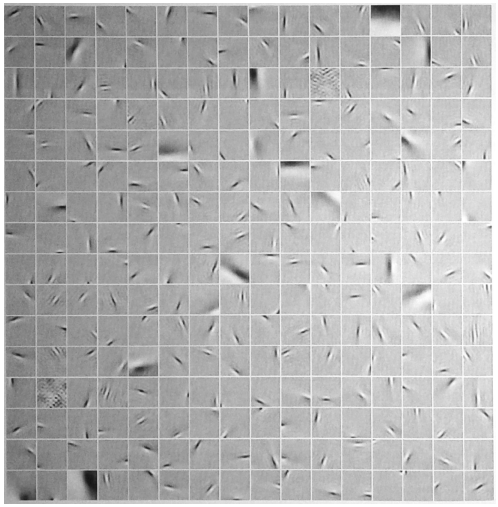
- In other words, synthesis weights, $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{W}$, are equal to input covariance times analysis weights!

Example analysis filters



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

Example synthesis filters



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

Frequency channels

- The emergence of bandpath 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:

$$\begin{aligned} & \mathrm{MI}(s_1, \ldots, s_K) \\ = & \int_{s_1,\ldots,s_K} p(s_1, \ldots, s_K) \log \frac{p(s_1, \ldots, s_K)}{p(s_1) \cdots p(s_K)} \, \mathrm{d}s_1 \cdots \mathrm{d}s_K \\ = & \sum_{i=1}^{K} H(s_i) - H(\boldsymbol{s}) \\ = & \sum_{i=1}^{K} H(\boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{z}) - H(\boldsymbol{W}\boldsymbol{z}) \end{aligned}$$

- For orthogonal $\boldsymbol{W}$, 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:

$$\begin{aligned} & \mathrm{MI}(s_1, \ldots, s_K) \\ = & \int_{s_1,\ldots,s_K} p(s_1, \ldots, s_K) \log \frac{p(s_1, \ldots, s_K)}{p(s_1) \cdots p(s_K)} \, \mathrm{d}s_1 \cdots \mathrm{d}s_K \\ = & \sum_{i=1}^{K} H(s_i) - H(\boldsymbol{s}) \\ = & \sum_{i=1}^{K} H(\boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{z}) - H(\boldsymbol{W}\boldsymbol{z}) \end{aligned}$$

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

Infomax ICA

- Yet another information theoretic approach to ICA can be derived as follows:
- Maximum likelihood ($\widehat{=}$ 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 $F(x)$ ($\widehat{=}$ minimize its likelihood).
- (We can make the same argument for multivariate data $\boldsymbol{x}$)

Infomax ICA

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

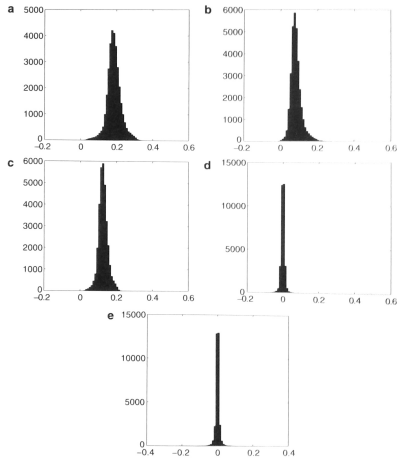
$$p_F(F(\boldsymbol{x})) = \frac{1}{|\det J(\boldsymbol{x})|} p_x(\boldsymbol{x})$$

  where $J(\boldsymbol{x}_t)$ is the Jacobian of $F(\boldsymbol{x})$ at $\boldsymbol{x}$
- By eliminating terms that do not depend on the parameters of F, we get the optimization problem:

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

- (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:

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

  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://cnl.salk.edu/˜ tewon/Blind/blind audio.html