ICA: algorithms & applications by Hyvärinen & Oja

2. ICA

2.1 def

1. Observe n linear mixtures $x_1, \dots, x_n$ of ind. components:

$$x_j = a_{j1}s_1 + a_{j2}s_2 + \dots + a_{jn}s_n \quad \forall j$$

2. Assume the x & s are zero-mean r.v.

3. Write as $x = As$
   - . x : observed data
   - . s : latent ind. components

4. estimate A, s given X

5. Assume the s_i are statistically independent

6. let $W = A^{-1}$, then $s = Wx$.

2.2 Ambiguities of ICA

1. Can not determine the variances of s_j.
   - L assume unit variance $E[s_j^2] = 1$.

2. Can not determine the order of s_j

3.3 Why Gaussian variables are forbidden

1. The fundamental restriction in ICA is that the independent components must be non-gaussian.

2. We can prove that the distribution of any orthogonal transformation of the gaussian (x_1, x_2) has exactly the same distribution as (x_1, x_2).

3. We can only estimate the ICA model up to an orthogonal transformation.

4. Orthogonal transformation is a linear transformation:

$$T: V \to V$$

which preserves a symmetric inner product.

$$\langle v, w \rangle = \langle Tv, Tw \rangle$$

preserves the length of vectors & angles b/t vectors.

4. Principles of ICA estimation:

4.1 Intuition

1. From CLT, the sum of 2 independent r.v. usually has a distribution that is closer to a gaussian than any of the 2 ~~ote~~ orig. r.v.

2. Let y be the estimate of one of the ind. comp. :

$$y = w^T x \quad , w \text{ is a vector to be determined.}$$

3. Define $z = A^T w$, then $y = w^T x = w^T A s = z^T s$.

$$\Downarrow$$

- $z^T s$ is more gaussian than any of the s_i.
- become least gaussian when $z^T s$ is one of the s_i.

$$\Downarrow$$

- find w to maximize the non-gaussianity of $w^T x$.
- then $w^T x = z^T s$ becomes one of the independent component.

4.2 Measure of nongaussianity

- Assume Y is centered, has unit variance.

4.2.1 Kurtosis

1. $\text{kurt}(y) = E[y^4] - 3(E[y^2])^2$

$$= E[y^4] - 3 \quad (\text{since } y \text{ has unit var})$$

2. kurtosis is zero for gaussian r.v.
3. sub-gaussian : r.v. with neg. kurtosis
   super-gaussian : r.v. with pos. kurtosis
4. the kurtosis can be estimated by using the 4th moment of the sample data.
5. But the estimate is sentitive to outliers, not a robust measures of non-gaussianity

4.2.2 Negentropy

1. Under assumptions, entropy is the coding length of a r.v.

2. The differential entropy:

$$H(y) = -\int f(y) \log f(y) \, dy$$

3. A gaussian var. has the largest entropy among r.v. of equal variances.

4. Negentropy: $J(y) = H(y_{gauss}) - H(y)$
   - y_{gauss} is Gaussian with the same cov. matrix as y.
   - 0 for Gaussian r.v. and always nonneg ~~otherwise~~ otherwise

5. J is invariant for invertible linear tranformation.

6. Negentropy is in some sense the optimal estimator of nongaussianity, as far as statistical properties are concerned.

7. But H is computationally very diff.

4.2.3 Approximations of negentropy

1. The classical method: $J(y) \approx \frac{1}{12} E[y^3]^2 + \frac{1}{48} \text{kurt}(y)^2$

2. Based on MaxEnt principle, we obtain the approximations:

$$J(y) \approx \sum_{i=1}^{p} k_i \left[E[G_i(y)] - E[G_i(v)] \right]^2$$

   - k_i : pos. constant
   - $v \sim N(0,1)$
   - G_i are q non-quadratic functions

3. Always non-negative, and 0 if y is Gaussian.

4. In the case of using only 1 G,

$$J(y) \propto \left[E[G(y)] - E[G(v)] \right]^2$$

a generalization of kurtosis (take $G(y) = y^4$).

5. Choosing G that does not grow too fast, we can obtain more robust estimators.

4.3 Minimization of mutual information

1. This approach is inspired by information theory for ICA estimation.
2. Provide a rigorous justification for the heuristics principle in prev. section.
3. This will lead to the same principles of finding the most nongaussian 1D-subspace.

4.3.1 Mutual information

1. $I(y_1, y_2, \cdots, y_m) = \sum_{i=1}^{m} H(y_i) - H(y)$, the y_i are scalar r.v.
2. equivalent to the KL b/t the joint $f(y)$ & the prod. of the marginal $f(y_i)$.
3. Always non-negative & zero iff the variables are statistically independent.
4. For invertible linear trans. $y = Wx$:

$$I(y_1, \cdots, y_m) = \sum_i H(y_i) - H(x) - \log|\det W|$$

5. Constraint y_i to be uncorrelated & unit var.

$$\Rightarrow E[yy^T] = W E[xx^T] W^T = I$$

$$\Rightarrow \det I = 1 = (\det W)(\det E[xx^T])(\det W^T)$$

$$\Rightarrow \det W \text{ is constant}$$

6. For y_i of unit var. & uncorrelated: is this necessary? the paper is ambiguous

$$I(y_1, \ldots, y_n) = C - \sum_i J(y_i).$$

mutual info differs from the negent. by only the sign & a constant.

4.3.2 Defining ICA by MI

1. Define ICA of a random vector x as an invertible transformation W

$$s = Wx$$

where the W is determined so that $I(s_1, \dots, s_n)$ is minimized.

2. From (30): $I(y_1, \dots, y_n) = C - \sum_i J(y_i)$

then finding W that minimizes MI is roughly equi. to finding direction where the negentropy is maximized.

3. Thus, forumlating ICA as min. of MI provides a rigorous justification of the heuristics of finding maximally non-gaussian direction.

4.4 MLE

4.4.1 the likelihood

1. Essentially equi. to min. of MI.
2. Let $W = (w_1, \dots, w_n)^T = A^{-1}$, the log-likelihood is:

$$L = \sum_t \sum_i \log f_i \left(w_i^T x(t) \right) + T \log \left| \det W \right|$$

- f_i is the density of s_i (assumed known).
- $x(t)$ are the realizations of x.

What is this the log-likelihood of ?

3. For any r.v x with density p, and for matrix W, the density of $y = Wx$ is given by $p_x(Wx) \left| \det W \right|$.

4.4.2 The infomax principle

1. Another contrast function is based on maximizing the output entropy of a NN with non-linear outputs.

2. The function is:
$$L_2 = H(\phi_1(w_1^T x), \dots, \phi_n(w_n^T x))$$
(maximize the entropy of the output)

- x : input
- $\phi_i(w_i^T x)$: output of the NN
- ϕ_i : non-linear scalar functions
- w_i : weight vectors

3. If $\phi_i'(\cdot) = f_i(\cdot)$, that is the non-linearities ϕ_i are chosen to be the cdf of the corresponding densities f_i, then the principles of network entropy maximization, or "infomax", is equivalent to MLE.

4.4.3 Connections to MI

1. Consider the expectation of the log-likelihood:
$$\frac{1}{T} E[L] = \sum_i E\left[\log f_i(w_i^T x)\right] + \log|\det W|.$$

2. If f_i is the density of $w_i^T x$, then the first term equals $-\sum_i H(w_i^T x)$, then the likelihood equals the neg. of MI, up to an additive constant, as given in (28).

3. They argue that in practice, we do not know the dist. of the ind. components f_i
   - ⇒ estimate ~~them~~ density of $w_i^T x$ as part of the ML estimation method
   - ⇒ use as approx. to density of s_i
   - ⇒ likelihood & MI are equivalent.

Scanned with CamScanner

5. Preprocessing for ICA

5.1 Centering:

1. Center x by subtracting its mean $m = E[x]$ to make x a zero-mean r.v.

2. Then the latent s is also zero-mean.

3. This step is solely to simplify the ICA algorithm.

5.2 Whitening:

1. White r.v.: components are uncorrelated & has unity variances.
2. After centering, we transform the observed vector $\tilde{x}$ linearly to obtain a new vector $\tilde{x}$ which is white, that is

$$E[\tilde{x}\tilde{x}^T] = I \quad \text{(covariance matrix of } \tilde{x} \text{ equal } I\text{)}$$

3. One method for whitening is to use eigen-value decomposition:
   - estimate $E[xx^T]$ from the sample $x(1), \dots, x(T)$
   - perform EVD $E[xx^T] = EDE^T$
   - whitening $\tilde{x} = ED^{-\frac{1}{2}}E^T x$ (then easy to check $E[\tilde{x}\tilde{x}^T] = I$)
4. Whitening transforms the mixing matrix A into $\tilde{A}$:

$$\tilde{x} = ED^{-\frac{1}{2}}E^T x = ED^{-\frac{1}{2}}E^T As = \tilde{A}s$$

5. $\tilde{A}$ is orthogonal:

$$E[\tilde{x}\tilde{x}^T] = \tilde{A}\, E[ss^T]\, \tilde{A}^T = \tilde{A}\tilde{A}^T = I$$

6. An orthogonal matrix has $\frac{n(n-1)}{2}$ dof compared to n^2 of an arbitrary $n \times n$ matrix

what does this mean & why?

7. Then whitening reduces the no. of parameters to be estimated.

8. Assume data is centered & whiten for the remaining & remove tilde.

6. The FastICA algorithm

6.1 FastICA for one unit

1. The var of $w^T x$ must be constrained to unity.
2. For whitened x, this is equivalent to constraining the norm of w to be 1.
3. Goal: find w s.t. $w^T x$ maximizes nongaussianity.
4. Recall (25), an approx. of negent.:

$$J(y) \propto \left[E[G(y)] - E[G(v)] \right]^2 \quad , v \sim \mathcal{N}(0,1)$$

5. Denote g as the derivative of G
6. The basic form of the FastICA algo is:
   1. Choose initial w
   2. let $w^+ = E\left[x g(w^T x) \right] - E\left[g'(w^T x) \right] w$
   3. let $w = w^+ / \| w^+ \|$
   4. If not converge, go back to 2.

      this means that the old & new values of w point in the same direction, (dot prod. is almost 1).

7. Derivation of ~~FastICA~~:

. maxima of $J(w^T x)$ obtained at optima of $E[G(w^T x)]$

$$\Downarrow$$

. KKT & $\|w\|^2 = E\left[(w^T x)^2 \right] = 1 \Rightarrow \overbrace{E\left[x g(w^T x) \right] - \beta w}^{F} = 0$

$$\Downarrow \text{ solve with Newton's method}$$

Jacobian

. $JF(w) = E\left[x x^T g'(w^T x) \right] - \beta I$ x is whitened.

$$\Downarrow$$

. Approximate $E\left[x x^T g'(w^T x) \right] \approx E\left[x x^T \right] E\left[g'(w^T x) \right] = E\left[g'(w^T x) \right] I$

$$\Downarrow$$

. $JF(w)$ approx. is diagonal & can easily be inverted (the Newton method is fast)

$$\Downarrow \text{ obtain approx. Newton iteration}$$

. $w^+ = w - \left[E[x g(w^T x)] - \beta w \right] / \left[E[g'(w^T x)] - \beta \right]$

. . . . further simplification.