Data Decomposition: From Independent Component Analysis to Sparse Representations

E. Roussos
MLRG, University of Oxford, Parks Road, Oxford, OX1 3PJ, U.K.
eroussos@learning-machines.org

Abstract

This paper provides a unifying review of some recent approaches to decomposing data into sets of components. We start from the classical algebraic method of singular value decomposition and then introduce principal and independent component analysis. The text continues with the main subject of this paper, sparse representation and decomposition, emphasizing its biological plausibility. In this paper emphasis will be given to the geometric perspective, with the mathematics kept to an essential minimum.

1 Data modelling: A probabilistic approach

In an exploratory approach to data analysis, it is often useful to consider the observations as generated from a set of latent generators or 'sources' via a generally unknown mapping. Our goal is to recover the generators from the observations, an *inverse* problem. This can be often stated as a *data decomposition* problem: the data matrix is decomposed into factors, each one of them representing some salient characteristics of the data. In fact, many well known algorithms, such as singular value decomposition (SVD) and principal component analysis (PCA), indepedent component analysis (ICA), as well as k–means and many others can be stated under this formulation, providing a unifying framework for unsupervised learning. Another view is that of the *representation* of data sets in a new coordinate system such that certain properties hold. For example, in PCA we seek a new coordinate system in which the data become linearly uncorrelated. For the noisy overcomplete case, where we have more sources than observations, the problem of reconstructing the sources becomes extremely ill-posed. Solutions to such inverse problems can, in many cases, be achieved by incorporating prior knowledge about the problem, captured in the form of constraints.

When modelling complex systems we are unavoidably faced with *imperfect* or *missing* information, especially in the measurement and information sciences. This may have several causes, but it is mainly due to

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27456v1 | CC BY 4.0 Open Access | rec: 30 Dec 2018, publ: 30 Dec 2018

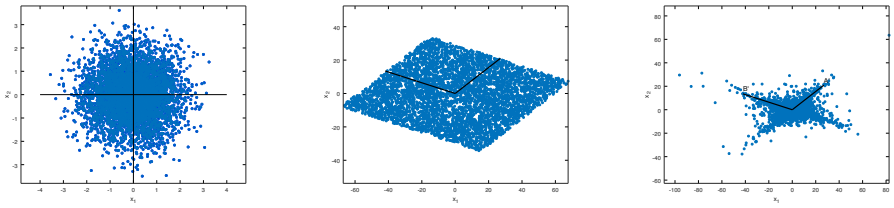


Figure 1: *Seeking structure in data.* Component analysis can be viewed as a family of data representation methods. The challenging task is to find *informative directions* in data space. These correspond to the column vectors of the observation (transformation, or 'mixing') matrix and form a new coordinate system. Their directions are non-orthogonal in general. (Left) Rotational invariance of the distribution of independent Gaussian random variables with equal variance. A scatterplot (point cloud) drawn from two such Gaussian sources illustrates the fact that there is not enough structure in the data in order to find characteristic directions in data space. Algebraically, we can only estimate the linear map up to an orthogonal transformation. (Center) Point cloud generated from a non-Gaussian distribution. (Right) The data cloud contains more structure in this case, which we want to exploit. In particular, the geometric shape of the point cloud of this figure is an example of a dataset that is *sparse* with respect to the coordinate axes shown by the two arrows.

- Lack of, or incompleteness in, our understanding or knowledge of the phenomena involved.
- The cost of obtaining and processing the vast amounts of information often needed for a more complete measurement of the phenomena.
- Inherent system complexity and stochasticity.

Probability theory is a conceptual and computational framework for reasoning under uncertainty. Probabilities model *uncertainty* regarding the occurence of *random* events. Assigning probability measures on uncertain quantities reflects precisely our lack of information about the quantities at hand. According to Cox's theorem [17], probability is the only consistent, universal logic framework for quantitatively reasoning under uncertainty. Moreover, probability theory offers a consistent framework for modelling and inference. Jaynes [38] viewed probability theory as a unifying tool for plausible reasoning in the presence of uncertainty. From a modeler's point of view, the greatest practical advantage of probability theory is perhaps that it offers modularity and extensibility: probability theory acts as "glue" for linking different models together.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27456v1 | CC BY 4.0 Open Access | rec: 30 Dec 2018, publ: 30 Dec 2018

2 Second order decompositions: Singular value decomposition and Principal Component Analysis

Singular value decomposition is an important method, originating in the Linear Algebra and Numerical Analysis communities, with a vast repertoire of applications in the Applied Sciences and Data Analysis. It is often used as a subroutine in more complicated models, and there exist versions of it that are very computationally efficient. We only present the basic ideas here; see [28] for a reference.

Let $\mathbf{X}$ be a $M \times N$ rectangular data matrix, where each row is a data point and each column is a "feature",

$$\underset{M\times N}{\mathbf{X}} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,N} \\ \vdots & \ddots & \vdots \\ x_{M,1} & \cdots & x_{M,N} \end{bmatrix} ,$$

and assume without loss of generality that $M \geq N$. The singular value decomposition (SVD) is a *factorization* of matrix $\mathbf{X}$ such that

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathsf{T}} \ , \tag{1}$$

where the $M \times M$ orthogonal matrix $\mathbf{U} = \left[\mathbf{u}_i\right]$ is called the left eigenvector matrix of $\mathbf{X}$, and the $N \times N$ orthogonal matrix $\mathbf{V}^{\mathsf{T}} = \left[\mathbf{v}_i^{\mathsf{T}}\right]$ is its right eigenvector matrix. The square roots of the N eigenvalues of the covariance matrices[1] $\mathbf{X}\mathbf{X}^{\mathsf{T}}$ and $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ are the singular values of $\mathbf{X}$, $\sigma_i = \sqrt{\lambda_i}$, forming the diagonal matrix $\mathbf{S} = \mathrm{diag}\left(\sigma_i\right)$. The singular values are nonnegative and sorted in decreasing order, such that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_N$, forming the *spectrum* of $\mathbf{X}$.

The singular value decomposition of X can be also written as

$$\mathbf{X} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathsf{T}} \ , \tag{2}$$

where $\mathbf{u}_i$ is the i–th eigenvector of $\mathbf{X}\mathbf{X}^{\mathsf{T}}$ and $\mathbf{v}_i$ is the i–th eigenvector of $\mathbf{X}^{\mathsf{T}}\mathbf{X}$, as above, and $r \leq N$ is the rank of $\mathbf{X}$. In other words, a matrix, $\mathbf{X}$, can be written as a linear superposition of its eigenimages, i.e. a sum of the outer products of its left and right eigenvectors, $\mathbf{u}_i\mathbf{v}_i^{\mathsf{T}}$, weighted by the square roots of the eigenvalues, σ_i. The important fact here is that often relatively few eigenvalues contain most of the 'energy' of matrix $\mathbf{X}$. Now if $r < N$, the energy of a data matrix, $\mathbf{X}$, can be captured with fewer variables than N, since the relevant information is contained in a lower-dimensional subspace of the measurement space. This is a form of *dimensionality reduction*. Note that due to the presence of noise in the data we may actually have $r = N$, though. In other words, in practice all eigenvalues may be non-vanishing. This, however, also hints at a *denoising* scheme in which one regards the smaller eigenvalues as corresponding

[1] Note that $\mathbf{X}^{\mathsf{T}}\mathbf{X} = \mathbf{V}\mathbf{S}^2\mathbf{V}^{\mathsf{T}}$ and $\mathbf{X}\mathbf{X}^{\mathsf{T}} = \mathbf{U}\mathbf{S}^2\mathbf{U}^{\mathsf{T}}$.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27456v1 | CC BY 4.0 Open Access | rec: 30 Dec 2018, publ: 30 Dec 2018

to the noise and then forms a *truncated* SVD, $\mathbf{X}_t = \mathbf{U}_t\mathbf{S}_t\mathbf{V}_t^{\mathsf{T}}$, where $t < r$. $\mathbf{X}_t$ is the unique minimizer of $\|\mathbf{X} - \mathbf{X}_t\|_{\mathrm{F}}$ among all rank-t matrices under the Frobenius norm and also a minimizer (perhaps not unique) under the 2–norm.

Remark 1 *For many important applications, such as fMRI and other biosignals, the signal of interest represents only a small part of what is measured (Lazar, [43]; Calhoun et al., [10]), in terms of signal power. Consequently, an optimization criterion that searches for components with maximum signal power, such as PCA, will fail to recover the signals we are looking for. Methods that exploit higher-order statistics in the data are therefore needed. Second-order methods can still be very useful as a preprocessing step, however, e.g. for dimensionality reduction, and are often used as such.*

3 Higher-order decompositions: Independent Component Analysis

In this section we review the independent component analysis (ICA) approach to source separation, with an emphasis on the aspect of *non-gaussianity*. Methodological and review literature includes [57], [31], [58], [25]. Additional resources are given below.

ICA is a family of data analysis methods that aims at decomposing datasets into maximally statistically independent components. In the noiseless setting, the observation model for linear ICA is

$$\mathbf{x} = \mathbf{A}\mathbf{s} \ , \tag{3}$$

where we have assumed that the observations have been de-meaned (i.e. we have translated the coordinate system to the data centroid). ICA employs the principle of *redundancy reduction* (Barlow, [5]) embodied in the requirement of *statistical independence* among the components (Nadal and Parga, [50]). In statistical language, this means that the joint density factorizes over latent sources:

$$P(\mathbf{s}) = \prod_{l=1}^{L} P_l(s_l) \ , \tag{4}$$

where $P(\mathbf{s})$ is the assumed distribution of the sources, $\mathbf{s} = (s_1, \ldots, s_L)$, regarded as stochastic variables, and $p_l(s_l)$ are appropriate *non-Gaussian* priors. Non-Gaussianity is the defining characteristic of the ICA family with respect to PCA. We seek non-Gaussian sources for two, complementary, reasons:

- Identifiability,
- "Interestingness".

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27456v1 | CC BY 4.0 Open Access | rec: 30 Dec 2018, publ: 30 Dec 2018

Gaussians are not interesting since the superposition of independent sources tends to be Gaussian. The concept of interestingness is directly exploited in the related method of *Projection Pursuit* (Friedman and Tukey, [24]), where the goal is to find the projection directions in a data set that show the least Gaussian distributions. An important result relating to the probability densities of the individual sources, due to Comon and based on the Darmois-Skitovitch theorem[2] [18], formalizes the above and states that for analysis in independent components at most one source may be Gaussian, in order for the model to be estimable [16]. Geometrically, this indeterminacy of Gaussian point clouds is due to the rotational invariance of the Gaussian distribution under orthogonal transformations (Hyvärinen, [34]). Gaussian point clouds are optimally described in terms of the PCA decomposition method (Figure 1 (left) [Lewicki and Sejnowski] (right)). This, geometric view of component analysis is a fundamental one in this paper.

A related concept is that of *linear structure* (Rao, [55]; Beckmann and Smith, [7]). A vector, $\mathbf{x}$, is said to have a linear structure if it can be decomposed as $\mathbf{x} = \boldsymbol{\mu} + \mathbf{A}\mathbf{s}$, where $\mathbf{s}$ is a vector of statistically independent random variables and the matrix $\mathbf{A}$ is of full column rank. Beckmann and Smith use results from Rao [55] in order to ensure uniqueness of their ICA decomposition. In particular, they use the fact that conditioned on knowing the number of sources and the assumption of non-Gaussianity, there is no non-equivalent decomposition into a pair $(\mathbf{A}, \mathbf{s})$, that is, there is no other decomposition with mixing matrix that is not a rescaling and permutation of $\mathbf{A}$.

Equation (4) is equivalent to minimizing the *mutual information* among the inferred sources[3] [8],

$$\begin{cases} \min I(s_1, \ldots, s_L), \quad \text{where} \\ I(s_1, \ldots, s_L) = \int p(\mathbf{s}) \log \frac{p(\mathbf{s})}{\prod_l p(s_l)} \mathrm{d}\mathbf{s} \end{cases} ,$$

or, equivalently, the "distance" between the distribution $p(\mathbf{s})$ and the fully factorized one, $\prod_l p(s_l)$, measured in terms of the Kullback-Leibler divergence, $KL\big[p(\mathbf{s})||\prod_l p(s_l)\big]$. This is defined as $KL\big[p(x)||q(x)\big] = \mathbb{E}_{p(x)}\left[\log \frac{p(x)}{q(x)}\right]$. This

[2]The Darmois-Skitovitch theorem reads:

Theorem 1 (Darmois-Skitovitch) *Let $\xi_1, \ldots, \xi_n$ be independent random variables and let α_i and β_i, $i = 1, \ldots, n$ be nonzero real numbers such that the random variables $\sum_{i=1}^n \alpha_i \xi_i$ and $\sum_{i=1}^n \beta_i \xi_i$ are independent. Then the ξ_i's are Gaussian.*

See, for example, V. Bogachev, *'Gaussian measures'* [9], p. 13.

[3]For two stochastic variables X and Y to be independent, it is necessary and sufficient that their mutual information equals zero:

$$\begin{aligned} I(X,Y) &= H(X) + H(Y) - H(X,Y) \\ &= \int \mathrm{d}X\mathrm{d}Y P_{X,Y}(X,Y) \log P_{X,Y}(X,Y) \\ &\quad - \int \mathrm{d}X P_X(X) \log P_X(X) - \int \mathrm{d}Y P_Y(Y) \log P_Y(Y) = 0 \ , \end{aligned}$$

where the quantity $H(Z)$ is the 'differential' entropy of the random variable Z.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27456v1 | CC BY 4.0 Open Access | rec: 30 Dec 2018, publ: 30 Dec 2018

enables ICA algorithms to separate statistically independent sources, up to possible permutations and scalings of the components [16]. The mutual information ("redundancy") can be equivalently computed as

$$I(s_1, \ldots, s_L) = \left(\sum_{l=1}^{L} H(s_l) \right) - H(s_1, \ldots, s_L) \ ,$$

where the first term at the RHS is the sum of the entropies of the individual sources and the second the joint entropy of $(s_1, \ldots, s_L)$. As shown by Bell & Sejnowski (1995), independence can lead to separation because the method exploits higher-order statistics in the data, something that cannot be done with methods such as PCA.

In practice, many ICA algorithms minimize a variety of 'proxy' functionals. Bell and Sejnowski's ICA approach uses the InfoMax principle (Linsker, [46]), maximizing *information transfer* in a network of nonlinear units (Bell & Sejnowski, [8]). Based on this, Bell and Sejnowski derive their very successful Infomax-ICA algorithm. The sources are estimated as

$$\hat{\mathbf{s}} = \mathbf{u} = \mathbf{W}\mathbf{x} \ , \tag{5}$$

where $\mathbf{W}$ is the separating (unmixing) matrix that is iteratively learned by the rule

$$\mathbf{W} \leftarrow \mathbf{W} + \eta \Big(\mathbf{I} - \mathbb{E}\big[\boldsymbol{\phi}(\mathbf{u})\big] \mathbf{u}^{\mathsf{T}} \Big) \mathbf{W} \ , \tag{6}$$

where the vector valued map $\boldsymbol{\phi}(\mathbf{u}) = (\phi_1(u_1), \ldots, \phi_L(u_L))$ is an appropriate nonlinear function of the output, $\mathbf{u}$, such as a sigmoidal 'squashing' function, applied component-wise. Popular choices are the logistic transfer function, $\phi(u) = \frac{1}{1+e^{-u}}$, and hyperbolic tangent, $\phi(u) = \tanh(u)$. The expectation operator, $\mathbb{E}[\,\cdot\,]$, is approximated by an average over samples in practice. Finally, the factor η is an appropriate learning rate. The above equation incorporates Amari et al.'s natural gradient descent approach [1]. Bell and Sejnowski show that optimal information transfer, that is maximum mutual information between inputs and outputs, or equivalently maximum entropy for the output, is obtained when highly-slopping parts of the transfer function are aligned with high-density parts of the probability density function of the inputs.

Hyvärinen chooses to focus explicitly on non-Gaussianity and derives a fixed-point algorithm, dubbed FastICA [33]. Non-Gaussianity can be quantified using the *negentropy*, J,

$$J(\mathbf{u}) = H(\mathbf{u}_{\text{Gauss}}) - H(\mathbf{u}) \ ,$$

where $\mathbf{u}_{\text{Gauss}}$ is a Gaussian random variable with the same covariance as $\mathbf{u}$. The FastICA algorihm maximizes an approximation of J using the estimate

$$J(u_l) \approx \Big\{ \mathbb{E}\big[G(u_l)\big] - \mathbb{E}\big[G(u_{\text{Gauss}})\big] \Big\}^2 \ ,$$

where $G(\cdot)$ is an appropriate nonlinearity, such as the non-quadratic function $G(z) = z^4$, and that is implicitly related to the source distributions (see below),

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27456v1 | CC BY 4.0 Open Access | rec: 30 Dec 2018, publ: 30 Dec 2018

u_{Gauss} is a standardized Gaussian r.v., and $u_1, \ldots, u_l, \ldots, u_L$ are also of mean zero and unit variance. The unknown sources, $\{u_l\}_{l=1}^{L}$, are again estimated using the projections $u_l = \mathbf{w}_l^{\mathsf{T}}\mathbf{x}$, where $\mathbf{w}_l$ is the l–th separating vector (column of $\mathbf{W}$), found by the iteration

$$\mathbf{w} \leftarrow \mathbb{E}\left[\mathbf{x}g(\mathbf{w}^{\mathsf{T}}\mathbf{x})\right] - \mathbb{E}\left[g'(\mathbf{w}^{\mathsf{T}}\mathbf{x})\right]\mathbf{w} \ ,$$

where $g(\cdot)$ is the derivative of $G(\cdot)$ and $g'(\cdot)$ is the derivative of $g(\cdot)$ and $\mathbf{w}$ is each time rescaled as $\mathbf{w} \leftarrow \frac{\mathbf{w}}{\|\mathbf{w}\|}$. For an application of the non-Gaussianity principle to fMRI see the Probabilistic ICA algorithm of Beckman and Smith [7].

ICA as Unfolding plus Rotation of a Dataset

An important result in the theory of ICA, with practical value, is that the ICA decomposition can be written as a factorization of an *"unfolding"* matrix times a *rotation* matrix. The former is usually implemented by pre-whitening (pre-sphering) the observations, such that $\mathbb{E}\left[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathsf{T}}\right] = \mathbf{I}_D$, where $\tilde{\mathbf{x}}$ now denotes the whitened observations:

$$\tilde{\mathbf{x}} = \mathbf{W}_{\text{sph}}\mathbf{x} \ .$$

$\mathbf{W}_{\text{sph}}$ can be computed from the eigendecomposition of the data covariance matrix, $\mathbf{C}_{\mathbf{xx}} = \mathbb{E}\left[\mathbf{x}\mathbf{x}^{\mathsf{T}}\right] \doteq \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathsf{T}}$, where the matrix $\mathbf{U}$ is a unitary matrix[4] containing the eigenvectors of $\mathbf{C}_{\mathbf{xx}}$ and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_D)$ is the diagonal matrix of eigenvalues. Then the decomposition problem can be written (taking the "square root" and inverting) as

$$\tilde{\mathbf{x}} = \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}^{\mathsf{T}}\mathbf{A}\mathbf{s} = \mathbf{W}_{\text{sph}}\mathbf{A}\mathbf{s} = \tilde{\mathbf{A}}\mathbf{s}, \quad \text{i.e.} \quad \mathbf{A} = \mathbf{W}_{\text{sph}}^{-1}\tilde{\mathbf{A}} \ .$$

That $\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}^{\mathsf{T}}$ spheres the data can be seen by simply performing the operations for $\mathbb{E}\left[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathsf{T}}\right]$, taking into account that $\mathbf{U}$ is an orthogonal matrix [29]. The above whitening operation transforms the original data vectors to the space of the eigenvalues and rescales the axes by the singular values. Alternatively, one may use $\mathbf{U}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}^{\mathsf{T}}$ for whitening, which maps the data back to the original space. This often makes further processing easier. In any case, since the whitening transformation removes any second-order statistics (correlations) in the data, learning the ICA matrix $\tilde{\mathbf{A}}$ is equivalent to learning a pure orthogonal rotation matrix:

$$\mathbb{E}\big[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^{\mathsf{T}}\big] = \tilde{\mathbf{A}}\mathbb{E}\big[\mathbf{s}\mathbf{s}^{\mathsf{T}}\big]\tilde{\mathbf{A}}^{\mathsf{T}} = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^{\mathsf{T}} = \mathbf{I} \ .$$

3.1 Probabilistic Inference for ICA

Note that until now, while we have used probabilistic concepts to define information-theoretic quantities such as the negentropy and the mutual information, we have taken the view that the solution of the blind source separation problem can be

[4] If we restrict ourselves to the field of real numbers, $\mathbb{R}$, then the matrices $\mathbf{U}$ become orthogonal matrices.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27456v1 | CC BY 4.0 Open Access | rec: 30 Dec 2018, publ: 30 Dec 2018

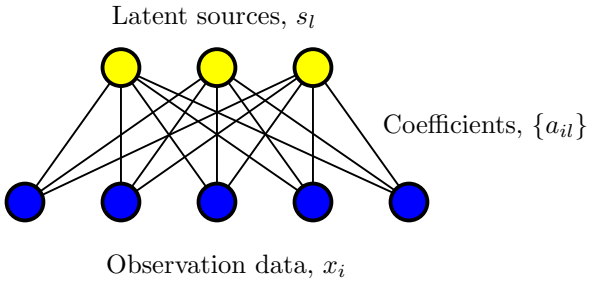


Figure 2: Graphical probabilistic model of the generative approach to component analysis. All models in this paper can be represented in this form.

achieved by transforming the observed signals through nonlinear functions in a bottom-up, filtering manner. Many classical component analysis algorithms, however, including ICA, can also be interpreted under the same probabilistic framework as top-down, *generative* models. This requires the construction of a density model. The model we consider here is the noisy transformation

$$\mathbf{s} \mapsto \mathbf{x} = \mathbf{A}\mathbf{s} + \underbrace{\boldsymbol{\varepsilon}}_{\text{noise}} , \tag{7}$$

where an L–dimensional vector of *latent variables*, $\mathbf{s}$, is linearly related to a D-dimensional vector of observations via the observation operator $\mathbf{A}$. Observation noise, $\boldsymbol{\varepsilon}$, may in general be added to the observations. In other words, the observed data is 'explained' by the unobserved latent variables, while the mismatch between the observations and the model predictions, $\mathbf{x} - \mathbf{A}\hat{\mathbf{s}}$, is explained by the additive noise. The fundamental equation of ICA, which we write again below,

$$P(\mathbf{s}) = \prod_{l=1}^{L} P_l(s_l) , \tag{8}$$

can be seen as a modelling assumption, i.e. a *working hypothesis*, as a factorization of a multi-dimensional distribution into a product of simpler one-dimensional distributions, in another interpretation. Classical ICA models such as Infomax ICA and FastICA assume noiseless and square mixing. This restriction is removed in more recent algorithms. A representation of the generative model for component analysis as a graphical probabilistic model is shown in Fig. 2.

Remark 2 *The generative model of Eqns (7), (8) defines a* constrained probability distribution *in data space. Referring back to Fig. 1, the "arms" of the point-cloud are oriented along the directions of the "regressors", which are encoded in the column vectors of the mixing matrix. Thus, when defining and learning a probabilistic ICA model, we are are in fact defining at least three things: the source distributions, the mixing matrix, and the noise model, given the constraints of Eqns 7 and 8.*

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27456v1 | CC BY 4.0 Open Access | rec: 30 Dec 2018, publ: 30 Dec 2018

This remark is important, as it gives an insight into why ICA algorithms are so successful in decomposing certain types of data such as fMRI [19].

In the general, noisy and non-square mixing case, one can formulate the penalized optimization problem (see e.g. [47], [11], [54], [61], and [59] for a nice concise review)

$$\hat{\mathbf{s}} = \underset{\mathbf{s}}{\operatorname{argmax}} \left\{ -\frac{1}{2\sigma^2}\|\mathbf{x} - \mathbf{A}\mathbf{s}\|^2 + \sum_{l=1}^{L} \log p_l(s_l) \right\} \ , \tag{9}$$

assuming spherical Gaussian noise, $\boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}_L\right)$, for example, in order to reconstruct the sources from the inputs at their most probable value.

As shown by MacKay [47] and Pearlmutter and Parra [54], Infomax-ICA can be interpreted as a maximum likelihood model. Assuming square mixing (i.e. as many latent dimensions as observations, $L = D$), and invertibility of the mixing matrix, the separating matrix is $\mathbf{W} = \mathbf{A}^{-1}$. We can then immediately write down the probability of the data, as

$$p(\mathbf{x}) = |\det(\mathbf{J})| \ p(\mathbf{s}) \ ,$$

where $\mathbf{J}$ is the Jacobian matrix of the transformation, with $J_{li} = \frac{\partial s_l}{\partial x_i}$. Under the linear model, and using the fundamental assumption of ICA, of mutual independence of the latent variables, $p(\mathbf{s}) = \prod_{l=1}^{L} p(s_l)$, we have

$$p(\mathbf{x}) = |\det(\mathbf{W})| \prod_{l} p(s_l) \ .$$

Then, the log–likelihood of an *i.i.d.* data set, $\mathcal{X} = \{\mathbf{x}_n\}_{n=1}^{N}$, under the model can then be written as

$$\mathcal{L}(\boldsymbol{\theta}) \overset{\text{def}}{=} \log p(\mathcal{X}|\boldsymbol{\theta}) = \log \left(\prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\theta}) \right) = N \log |\det(\mathbf{W})| + \sum_{n=1}^{N} \sum_{l=1}^{L} \log\big(p_l\big(\mathbf{w}_l^{\mathsf{T}} \mathbf{x}_n\big)\big) \ ,$$

where we have substituted $s_{l,n}$ with $u_{l,n} = \mathbf{w}_l^{\mathsf{T}} \mathbf{x}_n = \sum_{i=1}^{D} w_{l,i} x_{i,n}$. The parameter vector, $\boldsymbol{\theta}$, here contains the matrix, $\mathbf{A}$, or equivelently the unmixing one, $\mathbf{W} = \mathbf{A}^{-1}$, since these are uniquely related in this case.

We can now derive a maximum likelihood algorithm for ICA via gradient descent, in order to learn the separating matrix, $\mathbf{W}$. Taking the derivative of $\mathcal{L}(\boldsymbol{\theta})$ with respect to $\mathbf{W}$ and using well-known derivative rules we finally find the learning rule

$$\frac{\partial}{\partial W_{li}} \mathcal{L}(\boldsymbol{\theta}) = A_{li} + z_l x_i \ ,$$

where we have used the shorthand notation $z_l = \phi_l(u_l)$, where the ICA nonlinearity is the *score* function of the sources, $\phi_l(s_l) = -\frac{\partial}{\partial s_l} \log p_l(s_l)$, where $p_l(s_l)$ are the *assumed* source priors. Multiplying with $\mathbf{W}^{\mathsf{T}}\mathbf{W}$, to make the algorithm covariant [47], we get exactly the Infomax-ICA update rule, Eq. (6). Note that

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27456v1 | CC BY 4.0 Open Access | rec: 30 Dec 2018, publ: 30 Dec 2018

the above multiplication is equivalent to using the 'natural gradient' approach of Amari [1], a learning algorithm based on the concept of information geometry.

The FastICA algorithm can be also interpreted as an instance of the EM algorithm [20], an iterative method for finding maximum likelihood or maximum a-posteriori solutions of statistical estimation problems. (See the "The EM Algorithm" sidebar.) Lappalainen [41] derives it as an algorithm that filters Gaussian noise. This is an important interpretation, as it leads us to a conceptually new framework for ICA, that of source separation via *denoising*. Here, the term 'denoising' is interpreted as filtering out irrelevant information. It is worth going through the main steps of the derivation.

The EM Algorithm

The general idea of the EM algorithm is to estimate the latent variables, $\mathbf{Y}$, and model parameters, $\boldsymbol{\theta}$, of a probabilistic model (which in this case are the sources, $\mathbf{S}$, and mixing matrix, $\mathbf{A}$, of the BSS problem, respectively), in two alternating steps. The 'E' (expectation) step computes the expectation of the log–likelihood with respect to the posterior distribution $p\left(\mathbf{Y}\middle|\mathbf{X},\boldsymbol{\theta}^{(\tau)}\right)$, using the current ($\tau$th) estimate of the parameters, $\boldsymbol{\theta}^{(\tau)}$, giving the so-called '$Q$–function',

$$Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(\tau)}\right) = \mathbb{E}_{\mathbf{Y}|\mathbf{X},\boldsymbol{\theta}^{(\tau)}}\left[\log \mathcal{L}(\boldsymbol{\theta};\mathbf{X},\mathbf{Y})\right] \ ;$$

this is a function of $\boldsymbol{\theta}$ only. (Recall that $\mathbf{X}$ is observed and $\boldsymbol{\theta}^{(\tau)}$ is temporarily fixed to it current point estimate.) The 'M' (maximization) step then computes the model parameters that maximize the expected log–likelihood,

$$\boldsymbol{\theta}^{(\tau+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(\tau)}\right) \ .$$

This scheme is iterated until the algorithm converges. It can be shown that the EM algorithm is guaranteed to increase the observed data likelihood at each iteration [20].

Applying the above generic EM recipe, we can compute the maximum likelihood estimate of the mixing matrix of our ICA model as

$$\hat{\mathbf{A}} = \left(\mathbf{X}\,\mathbb{E}\left[\mathbf{S}\right]^{\mathsf{T}}\right)\left(\mathbb{E}\left[\mathbf{S}\mathbf{S}^{\mathsf{T}}\right]\right)^{-1} \ ,$$

where the expected *sufficient statistics*[5] of the sources, $\mathbb{E}\left[\mathbf{S}\right]$ and $\mathbb{E}\left[\mathbf{S}\mathbf{S}^{\mathsf{T}}\right]$, are computed with respect to their posterior[6]. In the low sensor noise ($\sigma^2 \to 0$) and square-mixing case of FastICA, Lappalainen approximates the posterior mean

[5] A sufficient statistic is the minimal statistic that provides sufficient information about a statistical model. Typically, the sufficient statistic is a simple function of the data, e.g. the sum of all the data points, sum of squares of the data points, etc.

[6] These are relationships that will become useful later in the paper as well.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27456v1 | CC BY 4.0 Open Access | rec: 30 Dec 2018, publ: 30 Dec 2018

of the sources as

$$\hat{\mathbf{s}} = \mathbb{E}\left[\mathbf{s}\,\middle|\,\mathbf{A}, \mathbf{x}, \sigma^2\mathbf{I}_D\right] \approx \mathbf{s}_0 + \sigma^2 \left(\mathbf{A}^\mathsf{T}\mathbf{A}\right)^{-1} \boldsymbol{\phi}(\mathbf{s}_0) \ ,$$

where $\mathbf{s}_0 \stackrel{\text{def}}{=} \mathbf{A}^{-1}\mathbf{x}$ and the function $\boldsymbol{\phi}(\cdot)$ is defined as before, as the vector of the logarithmic derivatives of $p_l(\mathbf{s}_l)$. For prewhitened data, this expression simplifies even more, since $\mathbf{A}$ is orthogonal, and therefore, $\left(\mathbf{A}^\mathsf{T}\mathbf{A}\right)^{-1} = \mathbf{I}_L$. Then $\hat{\mathbf{A}} \approx \mathbf{A} + \sigma^2\mathbf{X}\boldsymbol{\phi}(\mathbf{S}_0)/M$.

Now Lappalainen makes the crucial observation that while the EM algorithm has not yet converged to the optimal values, the sources, $\mathbf{s}_0$, can be written as a "mixture"

$$\mathbf{s}_0 = \alpha\mathbf{s}_{\text{opt}} + \beta\mathbf{s}_\text{G}, \qquad \text{with} \quad \alpha^2 + \beta^2 = 1 \ ,$$

where the "noise" $\mathbf{s}_\text{G}$ is mostly due to the other sources not having been perfectly unmixed. When far from the optimal solution, we have $\beta \approx 1$ and $\alpha \approx 0$. Using an argument based on the central limit theorem, as the number of the other sources becomes large he then approximates the mixing matrix corresponding to those other sources as

$$\hat{\mathbf{a}}_\text{G} \approx \mathbf{a} + \sigma^2\mathbf{X}_\text{G}\boldsymbol{\phi}(\mathbf{s}_{0\text{G}})^\mathsf{T}/L \ ,$$

where $\mathbf{X}_\text{G}$ are Gaussian-distributed "sources" with the same covariance as $\mathbf{X}$, as is done in the standard FastICA algorithm, and the sources $\mathbf{s}_{0\text{G}}$ are $\mathbf{s}_{0\text{G}} \stackrel{\text{def}}{=} \mathbf{a}^\mathsf{T}\mathbf{X}_\text{G}$. Then the update equation for the mixing matrix, normalized to unity, is estimated by

$$\hat{\mathbf{a}}_\text{new} = \frac{\hat{\mathbf{a}} - \hat{\mathbf{a}}_\text{G}}{\|\hat{\mathbf{a}} - \hat{\mathbf{a}}_\text{G}\|} \approx \frac{\sigma^2\left[\mathbf{X}\boldsymbol{\phi}(\mathbf{s}_0)^\mathsf{T} - \mathbf{X}_\text{G}\boldsymbol{\phi}(\mathbf{s}_{0\text{G}})^\mathsf{T}\right]/L}{\|\hat{\mathbf{a}} - \hat{\mathbf{a}}_\text{G}\|} \ .$$

Lappalainen interprets the above E-step as *filtering* Gaussian noise.

The final step that will bring us to the standard FastICA is to note that the term $\mathbf{X}_\text{G}\boldsymbol{\phi}(\mathbf{s}_{0\text{G}})^\mathsf{T}/L$ is equal to $\mathbf{a}\mathbf{s}_{0\text{G}}\boldsymbol{\phi}(\mathbf{s}_{0\text{G}})^\mathsf{T}/L$, where the factor $\mathbf{s}_{0\text{G}}\boldsymbol{\phi}(\mathbf{s}_{0\text{G}})^\mathsf{T}/L$ is constant, and therefore the numerator of the update equation becomes the standard FastICA update, $\hat{\mathbf{a}} - \hat{\mathbf{a}}_\text{G} = \mathbf{X}\boldsymbol{\phi}(\mathbf{s}_0)^\mathsf{T} - c\mathbf{a}$.

While Teh [59] computes the data likelihood in a maximum likelihood framework, Knuth [39] uses a maximum a-posteriori framework. The latter allows us to impose constraints on the model parameters as well. This was further explored in Hyvarinen and Karthikesh in [32] in order to impose sparsity on the *mixing* matrix.

Up to now we have either assumed equal number of sources and sensors or we have implicitly assumed that their number is somehow given. Roberts [56] derives a Bayesian algorithm for ICA under the evidence framework that estimates tha most probable number of sources as a model order estimation problem. The evidence framework, as applied in [56], makes a local Gaussian approximation to the likelihood conditioned on the mixing matrix using a nested Laplace approximation, but takes into account the local curvature by estimating the Hessian. Due to computational reasons, this is approximated by a diagonal

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27456v1 | CC BY 4.0 Open Access | rec: 30 Dec 2018, publ: 30 Dec 2018

matrix here, setting the off-diagonal elements to zero. The noise width, regarded as a *hyperparameter*, is computed at its maximum likelihood value.

Finally, Choudrey et al. [15] and Miskin and MacKay [49] propose a fully Bayesian approach to ICA using a variational, *ensemble learning* approach under a mean-field approximation. They use a flexible source model based on mixtures of Gaussians and perform model order estimation using a variety of techniques.

We can now select an appropriate functional form for the individual marginal distributions, $p_l(s_{l,n})$, based on our prior knowledge about the problem, as was done in the original formulation of InfoMax ICA of Bell & Sejnowski for the separation of speech signals, for example. The source model should model the real source distributions as accurately as possible. Many natural signals exhibit characteristic amplitude distributions, which can provide some guidance and indeed should be exploited when possible. This allows us to utilize fixed source models in our separation algorithms. Bell and Sejnowski, for example, use several nonlinearities (recall that these are uniquely related to the assumed PDFs of the sources), such as $1/(1+e^{-u_i})$, $\tanh(u_i)$, $e^{-u_i^2}$, etc., as well as propose general-purpose 'score functions' (see Figure 2 of Ref. [8]) in their Infomax-ICA algorithm. FastICA uses nonlinearities such as u_i^3, $\tanh(\alpha u)$, $u_i e^{-\alpha u_i^2/2}$, and u_i^2. However, this is not always possible. The problems that can arise from an incorrect latent signal model and possible solutions are discussed in section 4.

4 The Importance of using Appropriate Latent Signal Models

Many classical ICA algorithms, such as Infomax-ICA and FastICA, allow the plug-in setting of the respective nonlinearity function in the system, as mentioned above. For successful separation, the form of the nonlinearity must somehow *match*, as far as possible, the underlying (unknown) statistical properties of the sources, such as their super- or sub-gaussianity. This was first stated as "matching the neuron's input-output function to the expected distribution of the signals" in [8]. Since the estimating equations for the mixing matrix and sources are coupled, the functional form of the nonlinearity is critical for their correct estimation: an incorrect choice of nonlinearity will lead to an incorrect estimation of the (un-)mixing matrix, which will map the observations back to the source space incorrectly, etc. Cardoso [12] gives a compelling example of how estimation can go wrong. Another example of how classical ICA fails in separating sources in an image processing context is given in Fig. 4 (from Tonazzini et al., [60]).

Remark 3 *Tonazzini et al. use a Markov random field in order to impose an image prior. However, the images of Fig. 4 (left) are actually also prime examples of* sparse *sources. In [27] and [19], an extensive study of how justified and robust are ICA algorithms for functional MR imaging of the brain was conducted and various simulations of fMRI "brain" activations under well-controllable situations with shapes similar to that of ref. [60] were performed that highlighted*

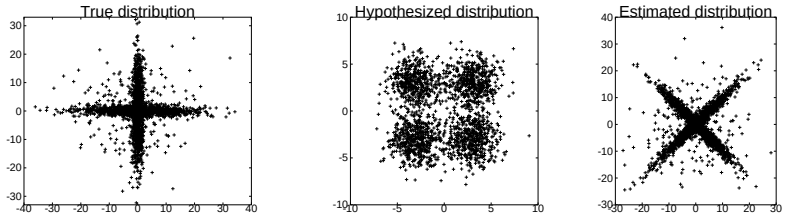


Figure 3: Effect of an incorrect source model specification [12]. Left: true distribution; Middle: Hypothesized distribution; Right: Estimated distribution.

the need for alternative decomposition algorithms that are effective for fMRI, based on sparsity.

It can be shown that the Infomax-ICA as well as the FastICA algorithms are instances of maximum likelihood estimation [47], [54], [30], [41]. Under this interpretation, one can see that the nonlinearity, $\phi(\cdot)$, is actually the logarithmic derivative of the (hypothesized) probability density of the sources (the 'score' function): for the l–th source, $\mathbf{s}_l$,

$$l: \quad \phi_l\left([\mathbf{W}\mathbf{x}]_l\right) = -\frac{\partial}{\partial \mathbf{s}_l}\log p_l(\mathbf{s}_l) = -\frac{p_l'(\mathbf{s}_l)}{p_l(\mathbf{s}_l)} \ ,$$

where the symbol $\mathbf{W}$ denotes the separating operator from observation space to source space and $\mathbf{x}$ is an observation. In other words, in a perfect match the nonlinearity is exactly the cumulative distribution function of the sources. Of course we do not know the *actual* source PDFs, since the sources themselves are unobserved, but we may try to estimate them from the data. For this purpose, we can employ a parameterized model source PDF, $p_l(\mathbf{s}_l;\theta_{\mathbf{s}_l})$, and *learn*, instead of fix, its parameters, $\theta_{\mathbf{s}_l}$, from the data. A flexible prior that is at the same time mathematically tractable is a mixture distribution. Lawrence and Bishop [42] uses a Mixture of Gaussians (MoG) prior for ICA, albeit in a fixed form. Attias [3] has used MoGs as source models for blind source separation under a maximum likelihood framework, leading to a flexible algorithm dubbed 'Independent Factor Analysis' (IFA). Choudrey et al. [15] and Lappalainen [40] use the same prior under a Bayesian ensemble learning approach, i.e. with a factorized posterior (the so-called 'naive' mean-field method).

5 Sparse Decompositions

As noted by Cardoso [13], non-Gaussianity is not the only possible route to independent component analysis, and indeed to blind source separation in general; other possibilities also exist—including exploiting non-stationarity and time-correlation in signals. Such a different paradigm, *sparsity*, in combination with doing away with the assumption of independence, will be explored next.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27456v1 | CC BY 4.0 Open Access | rec: 30 Dec 2018, publ: 30 Dec 2018

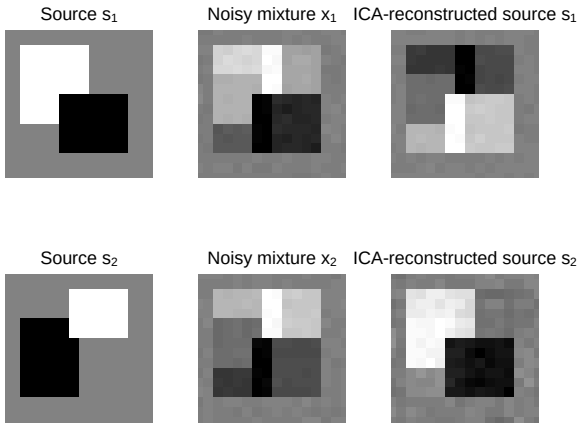


Figure 4: Effect of an incorrect source model specification for a blind image separation problem. Left: true sources; Middle: noisy mixtures; Right: Estimated sources from ICA. The model clearly fails to recover the sources. In particular, one of the sources is not recovered at all.

5.1 Parsimonious representation of data

ICA works well for a variety of blind source separation problems. However, in order for the decomposition to make sense the true sources must themselves indeed be (nearly) independent. This may make sense in the separation of voice signals that are independently generated by people with no interaction among them, for example. For other problems, however, searching for components that are maximally independent may not be so meaningful. Recently, another paradigm for BSS, and inverse problems in general, *sparsity*, has emerged as an alternative. Sparsity refers to the property of a representation to form compact encodings of signals, data, or functions, using a small number of basis functions. Those basis functions are used as "building blocks" to build more complex signals.

There has been a variety of algorithms for sparse representation, or sparse coding, originating from the computational neuroscience and neural networks communities as well as several others from a signal processing perspective. Sparse decomposition, and ways to impose sparsity constraints, has recently also been a topic of much research in the statistics and machine learning literature.

Sparse coding. In the study of the visual system, Field [23] proposed sparsity as an organization principle of the visual receptive field. He conjectured that

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27456v1 | CC BY 4.0 Open Access | rec: 30 Dec 2018, publ: 30 Dec 2018

populations of neurons optimize the representation of their visual environment by forming *sparse representations* of natural scenes, a hypothesis that has high biological plausibility since it is based on the general idea of a system using its available resources efficiently. According to his theory, the visual system performs efficient coding of natural scenes in terms of natural scene statistics by finding the sparse structure available in the input. Field's theory directly reflects the principle of *redundancy reduction* of Barlow [5], [6].

Dictionary learning. Olshausen and Field [51] further test the above theory, seeking experimental evidence for sparsity in the primary visual cortex (V1) by building a predictive (mathematical) model of sparse coding. In their model, images are formed as a linear combination of local basis functions with corresponding activations that are as sparse as possible. These bases model the V1 receptive fields and form overcomplete sets *adapted* to the statistics of natural images. Olshausen and Field's model is an early example of *dictionary learning*. Formally, the model of Olshausen and Field is described by:

$$\mathbf{x}_p \simeq \sum_i a_{p,i} \boldsymbol{\phi}_i \ ,$$

where $\mathbf{x}_p$ is an image "patch" (i.e. a small image window) and $\{\boldsymbol{\phi}_i\}$ are the underlying basis elements. A network representation of their model, `Sparsenet`, is shown in Fig. 5. They proposed the following objective:

$$\mathcal{I}(\boldsymbol{\Phi}) = \min_{a_{p,i}} \left\{ \sum_p \left\| \mathbf{x}_p - \sum_i a_{p,i} \boldsymbol{\phi}_i \right\| + \lambda \sum_i \log\left(1 + {a_{p,i}}^2\right) \right\} \ ,$$

to be minimized over bases, $\boldsymbol{\Phi}$, learned by searching for bases that optimized the sparsity of the coefficients, $\{a_{p,i}\}$, (subject to appropriate scale normalization of $\{\boldsymbol{\phi}_i\}$). In general, the basis set can be overcomplete. That is, the number of bases, $|\boldsymbol{\Phi}|$, can be greater than the dimensionality of the 'input' data space, D (see for example [52]). The reason for this is that the 'code' can be more sparse if one allows an overcomplete basis set, as the algorithm can select the bases that better match the structures contained in the signal (the "active" elements). See also Asari, [2]. As shown in Fig. 6 this objective results in highly sparse distributions for the coefficients. Astonishingly, the learned receptive fields (filters) have properties that resemble the properties of natural simple-cell receptive fields, that is they are spatially localized, oriented and bandpass, i.e. selective to structure at different spatial scales (Fig. 7).

In the signal processing community, Mallat and Zhang [48] proposed a greedy algorithm analogous to the projection pursuit in statistics, called 'matching pursuit', that iteratively finds the best matching projections of signals onto a fixed overcomplete dictionary of time-frequency 'atoms'. Linear combinations of those atoms form compact representations of the given signal.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27456v1 | CC BY 4.0 Open Access | rec: 30 Dec 2018, publ: 30 Dec 2018

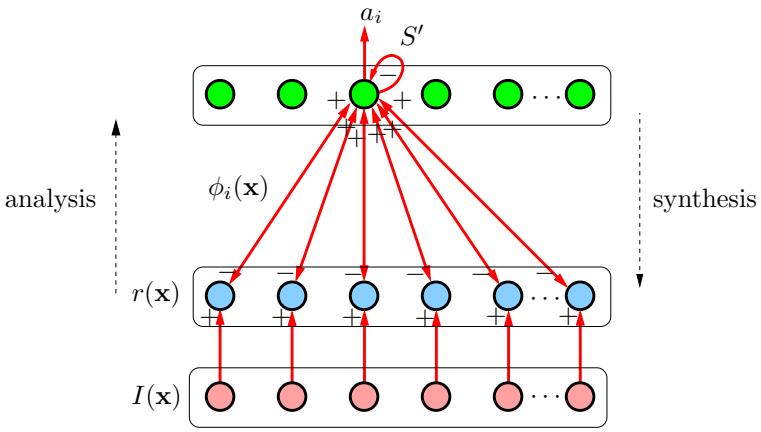


Figure 5: The Olshausen and Field model [51] as a neural network, `Sparsenet`. The inputs to the network are images, $I(\mathbf{x})$, where $\mathbf{x}$ denotes picture elements (pixels) over an image domain, Ω, and the outputs are the coefficients of the representation, a_i. The symbol $r(\mathbf{x})$ is the residual image, $r(\mathbf{x}) = I(\mathbf{x}) - \sum_i a_i \boldsymbol{\phi}_i(\mathbf{x})$. Each output neuron evolves according to the differential equation $\dot{a}_i = \sum_{\mathbf{x}\in\Omega} \boldsymbol{\phi}_i(\mathbf{x}) r(\mathbf{x}) - \lambda S'(a_i)$, where the derivative of the sparsity activation function $S(\cdot)$ induces non-linear self-inhibition, and the multiplier $\lambda \geq 0$ is a regularization parameter. This enfoces *sparsity*, as it drives activities towards zero. The regularization parameter balances the first, data fidelity term, which ensures accurate reconstruction. During the 'analysis' ("filtering") phase, a given image, $I(\mathbf{x})$, is decomposed in a dictionary, $\boldsymbol{\Phi}$, and its corresponding coefficients, a_i, are computed. During the 'synthesis' phase a learned dictionary predicts an estimate of an image, $\hat{I}(\mathbf{x})$, with residuals $\mathbf{r}(\mathbf{x})$. The optimal value of each a_i is determined from the corresponding equilibrium solution.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27456v1 | CC BY 4.0 Open Access | rec: 30 Dec 2018, publ: 30 Dec 2018

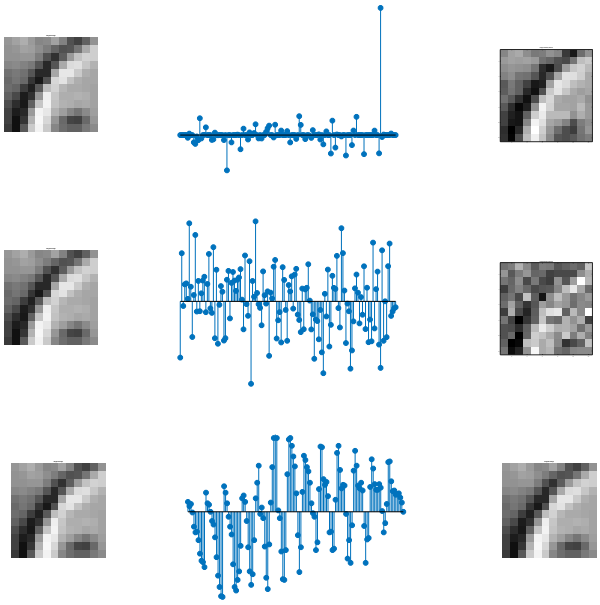


Figure 6: Activities, a_i, resulting from the model of Olshausen and Field [51]. The input image on the left is reconstructed from learned bases using their algorithm. Note how the coefficients a_i resulting from the model (first row) are highly sparse, compared to reconstructing the image patch using random bases (second row) or pixel (canonical) bases (third row). The canonical basis offers no compression at all, as it is merely a copy of the original image.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27456v1 | CC BY 4.0 Open Access | rec: 30 Dec 2018, publ: 30 Dec 2018

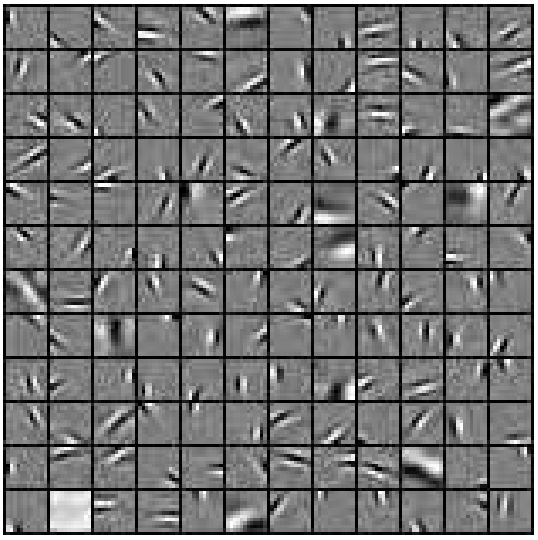


Figure 7: Learned receptive fields (filters) from the sparse coding algorithm of Olshausen and Field `Sparsenet` [51]. These filters exhibit properties of simple-cell receptive fields such as locality, orientation and spatial selectivity.

Geometric interpretation of sparse representation. A geometric interpretation of sparse representation is depicted in Fig. 8. Each data vector can be viewed as a point in a D–dimensional vector space, the whole dataset forming a cloud of points. We now seek a linear transformation of the dataset such that the inferred "projections" on to the new coordinate system defined by the column vectors of the learned transformation matrix, $\mathbf{A} = \left[\mathbf{a}_l\right]_{l=1}^{L}$, are as sparse as possible.

Note that it is the sparseness of the components (and the selection of a suitable model prior) that drives learning of the new representation (unmixing) directions. This sparseness is reflected in the *shape* of the point-cloud: referring to the above figure (where $D = L \doteq 2$), sparse data mapped in to the latent space produce a highly-peaked and heavy-tailed distribution for both axes (Fig. 8 (lower right)). This is indeed a result of the sparseness property of the dataset: the two 'arms' of the sparse data cloud are *tightly packed* around the directions of the unmixing vectors, $\mathbf{a}_l$. Algebraically, this means that for a particular point, n, either the coefficient $s_{1,n}$ $(l = 1)$ or the coefficient $s_{2,n}$ $(l = 2)$ is almost zero, as the particular datum is well described by the $\mathbf{a}_2$ or the $\mathbf{a}_1$ "regressor", respectively. On the contrary, non-sparse data will typically produce a projection that corresponds to a "fat" empirical histogram, as shown in Fig. 8 (upper-right).

[7] Field studied the statistics of natural scenes and their relation to computer vision and perception in [23]. The 'state-space' in this context is a state-space of neural activation amplitudes.

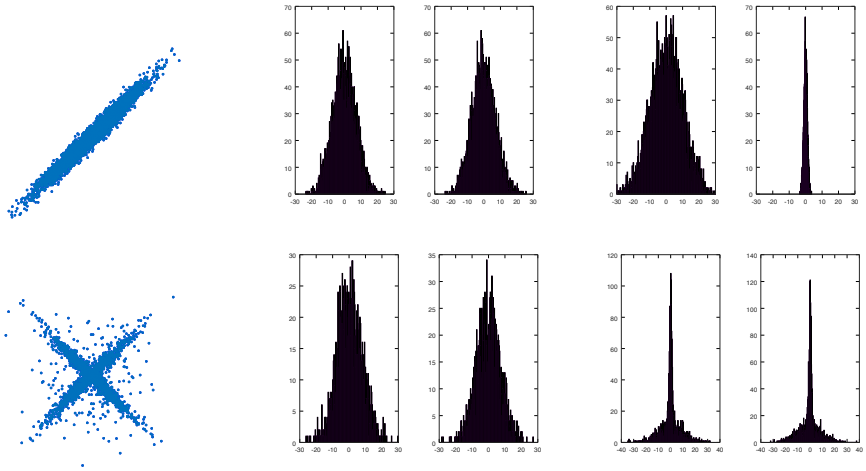


Figure 8: *Geometric interpretation of sparse representation.* State-spaces (in the terminology of Field[7][23]) and projections of two datasets, one sparse (lower row) and the other non-sparse (upper row), are shown. Each dataset, plotted in the measurement coordinate system, xy, produces a point cloud (left part of the figure) — for visualization purposes, both observation and latent dimensionalities are equal to $D = L = 2$ in this figure. By projecting the point clouds on to each coordinate we can produce the corresponding empirical histograms of 'state' amplitudes (middle part of the figure). We now seek a linear transformation to a latent space, uv, such that it optimizes some suitable criterion (this is shown in the right part of the figure). Sparse data mapped in the latent space produce heavy-tailed distributions for both latent dimensions (lower right), while for non-sparse data this is not the case (upper right).

With respect to the soft clustering view of component analysis (Miskin, [36]), discussed in the Introduction of the paper, if the data vectors are sufficiently sparse, their images on the unit hypersphere $\mathbb{S}^{D-1}$, i.e. the radial sections of their position vectors with the unit hypersphere, mapped as

$$\mathbf{x}_n \in \mathbb{E}^D \mapsto \hat{\mathbf{x}}_n \in \mathbb{S}^{D-1} \ ,$$

where the projection operator $P : \mathbf{u} \mapsto \hat{\mathbf{u}} = \frac{\mathbf{u}}{\|\mathbf{u}\|}$ maps vectors along their radii, *concentrate* around the unit vectors $\{\hat{\mathbf{a}}_l\}_{l=1}^{L}$; see Fig. 9 and Ref. [62]. While Miskin did not use this property per se for sparse decomposition, one can design separation algorithms that exploit it [45].

5.2 Sparse Decomposition of Data Matrices

Inspired by the model of Olshausen and Field, Donoho [21] first points out the connection and differences between the two lines of research, independent component analysis and sparse decompositions, and he promotes the idea of

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27456v1 | CC BY 4.0 Open Access | rec: 30 Dec 2018, publ: 30 Dec 2018

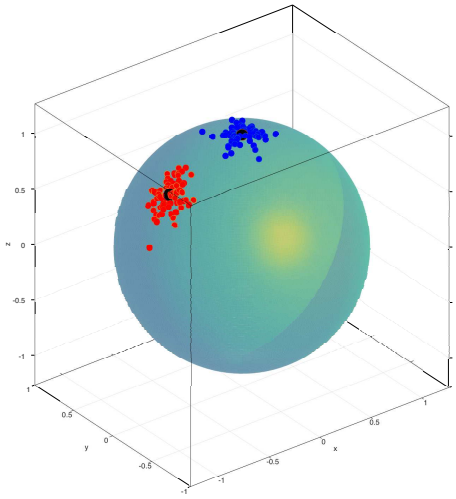


Figure 9: *Clustering of a sparse set of points on the unit hypersphere,* $\mathbb{S}^{D-1}$, embedded in a D–dimensional space. The points cluster around the direction vectors corresponding to the columns of the mixing matrix.

sparsity, overcompleteness, and optimal atomic decompositions as a better goal than independence. He provides a rationale of why sparsity is a more plausible principle, being "intrinsically important and fundamental", due to both biological and modelling reasons. Regarding the former, he too cites the extremely efficient sparse representation achieved by the human visual system, and its higher compression performance compared to the best engineered systems. With respect to the latter, he notes that independence is inherently a probabilistic assumption and of unknown interpretability (with respect to vision) because natural images are composed by occlusion. Occlusion inevitably creates *dependent* components. He finally suggests that one of the future challenges of 'sparse components analysis' would be to search over spaces of objects of much larger scale than the image patches of Olshausen and Field.

It turns out (see Olshausen, [52]) that the Infomax-ICA algorithm becomes, in fact, a special case of the sparse linear algorithm of Olshausen and Field when there is an equal number of basis functions/latent dimensions and inputs, the $\boldsymbol{\phi}_i$s are linearly independent, and there is no observation noise. In this case, there is a unique set of coefficients $\{a_i\}$ that is the root of $\|\mathbf{X} - \boldsymbol{\Phi}\mathbf{a}\|$, and we can write $\mathbf{a}$ as $\mathbf{a} = \mathbf{W}\mathbf{X}$, where $\mathbf{W} = \boldsymbol{\Phi}^{-1}$ (note that based upon the above assumptions, $\boldsymbol{\Phi}$ becomes invertible). If, in addition, the ICA nonlinearity is chosen to be the cumulative density function of the sparse components, then

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27456v1 | CC BY 4.0 Open Access | rec: 30 Dec 2018, publ: 30 Dec 2018

the sparse algorithm gives exactly the algorithm of Bell and Sejonwski. The point here is actually to show that sparsity constraints can lead to separation. Many researchers have indeed shown that this can be indeed the case. Indeed, as pointed out by Li, Cichocki and Amari [45],

Remark 4 *Sparse decompositions of data matrices* can *be used for the blind source separation problem.*

They provide various examples from simulations and EEG data analysis that demonstrate the performance of sparse decompositions in signal separation. Li, Cichocki and Amari performed a sophisticated mathematical analysis for the case of sparse representation of data matrices under the ℓ_1 prior, for given basis matrices. They tackle the two-step decomposition problem of learning the base matrix first, via clustering, and then estimating the coefficients of the decomposition. If $\mathbf{X}$ is a data matrix and $\mathbf{A} = \{\mathbf{a}_l\}$ is a given basis, Li et al. start from the mathematical model shown below:

$$\min \Big\{ \underbrace{\sum_{l=1}^{L} \sum_{n=1}^{N} |s_{ln}|}_{\mathrm{S}(\mathbf{S})} \;\Big|\; \text{subject to } \mathbf{AS} = \mathbf{X} \Big\} \;, \tag{10}$$

with $\mathrm{S}(\cdot)$ the *sparsity function* on the sources. This particular case of optimization problem can then be solved using linear programming. While the ℓ_0–norm solution is the sparsest one in general, its optimization is a non-trivial combinatorial problem. Li et al. show that, for sufficiently sparse signals, the solutions to the problem of sparse representation of data matrices that are obtained using the ℓ_0 and ℓ_1 norms are equivalent. This fact was previously shown by Donoho and Elad [22] but Li et al. [45] give a less strict sparseness ratio (i.e. the ratio of zero versus non-zero elements).

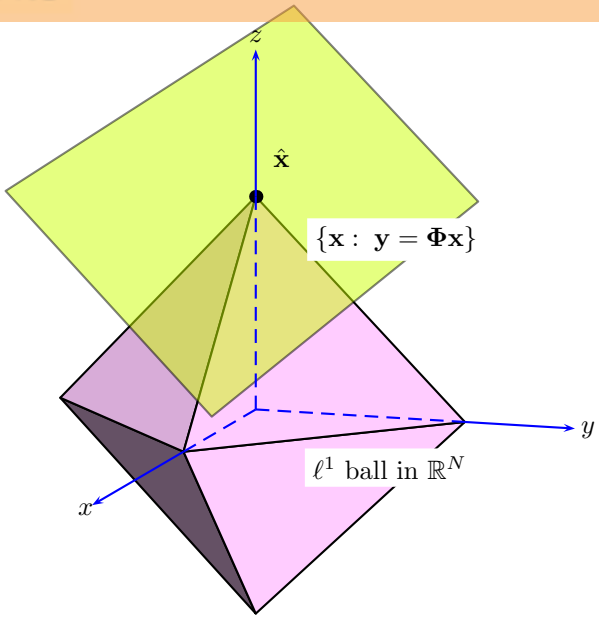
Uniqueness. Importantly, Li et al. [45] also show that the above problem has a *unique* solution. While in general there would be an infinite number of solutions for the underdetermined system of equations

$$\mathbf{As} = \mathbf{x} \;,$$

where the $D \times L$ matrix $\mathbf{A}$ (observation operator) with $L > D$ maps the unknown signal $\mathbf{s}$ in to the observed signal $\mathbf{x}$, the sparsity constraint makes the particular linear inverse problem well-posed. A geometric interpretation of why ℓ_1–type sparsity regularization works well for signal recovery under sparsity constraints is shown in Fig. 10. We want to find the optimal $\mathbf{x}$ as the minimum-norm vector that satisfies the constraint $\mathbf{x} = \mathbf{As}$, i.e. such that the hyperplane does not intersect the ℓ_1 ball. More generally, the problem can be stated (in the deterministic framework) as:

$$\min_{\mathbf{s}} \Big\{ \|\mathbf{s}\|_1 : \;\; \|\mathbf{As} - \mathbf{x}\| < c \Big\}$$

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27456v1 | CC BY 4.0 Open Access | rec: 30 Dec 2018, publ: 30 Dec 2018



1

Figure 10: *Why ℓ_1 works: A geometric intuition into sparse priors.* We seek the sparsest vector $\mathbf{x} \in \mathbb{R}^N$ under the ℓ_1 norm, in this case, that satisfies the linear constraint $\mathbf{y} = \mathbf{\Phi x}$, where $\mathbf{\Phi}$ is a dictionary. The ℓ_1 penalty corresponds geometrically to a cross-polytope (the 'ℓ_1 ball' in $\mathbb{R}^N$) and the linear constraint to a hyperplane. The shape of the polytope dictates the form of the solution. The optimal vector, $\hat{\mathbf{x}}$, is the one that touches the hyperplane without the latter intersecting the cross-polytope. Mathematically, this is the solution to the problem $\hat{\mathbf{x}} = \text{argmin}_{\mathbf{y}=\mathbf{\Phi x}} \|\mathbf{x}\|_1$. As can be seen from the figure, the inclusion of ℓ_1 norm necessarily drives all components of $\mathbf{x}$ but one towards zero, leading to sparse solutions.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27456v1 | CC BY 4.0 Open Access | rec: 30 Dec 2018, publ: 30 Dec 2018

(Chen and Haykin, [14]), where $\mathbf{x}$ can be a "corrupted" (noisy, blurred, etc) version of the original signal and c is a positive scalar constant that plays a role similar to the noise variance in the probablistic framework (Li et al., [45]). In this case, the hyperplane becomes an orthotope (hyperrectangle), defining a "zone" in which the vertex of the ℓ_1 ball must fall. In addition, Li et al. [45] use k–means clustering to get an estimate of the basis, which is then used in a linear programming algorithm in order to estimate the coefficients of the representation.

5.2.1 Probabilistic Solutions

Lewicki and Sejnowski [44], introduce a probabilistic method for sparse overcomplete representations. A Laplacian prior on the coefficients of the basis was used, $p(s_l) \propto \mathrm{e}^{-\theta|s_l|}$, enforcing parsimonious representations. They then propose a gradient optimization scheme for maximum a-posteriori (MAP) learning. For the linear model $\mathbf{x} = \mathbf{A}\mathbf{s} + \boldsymbol{\varepsilon}$, with Gaussian observation noise with variance σ^2, we seek the most probable decomposition coefficients, $\hat{\mathbf{s}}$, such that

$$\hat{\mathbf{s}} = \underset{\mathbf{s}}{\operatorname{argmax}} \left\{ p(\mathbf{x}|\mathbf{A}, \mathbf{s}) p(\mathbf{s}) \right\} \ . \tag{11}$$

The probability of a single data point is obtained by integrating out the unknown signals, $\mathbf{s}$:

$$p(\mathbf{x}|\mathbf{A}) = \int p(\mathbf{x}|\mathbf{A}, \mathbf{s}) p(\mathbf{s}) \mathrm{d}\mathbf{s} \ .$$

In order to derive a tractable algorithm, they make a Laplace approximation to the data likelihood, by assuming that the posterior is Gaussian around the posterior mode. This involves computing the Hessian $\mathbf{H} = \nabla_{\mathbf{s}} \nabla_{\mathbf{s}} \left\{ -\log \left[p(\mathbf{s}) p(\mathbf{x}|\mathbf{A}, \mathbf{s}) \right] \right\} = \frac{1}{\sigma^2} \mathbf{A}^{\mathsf{T}} \mathbf{A} - \nabla_{\mathbf{s}} \nabla_{\mathbf{s}} \log p(\mathbf{s})$. To make a smooth approximation of the derivative of the log–prior, and a diagonal approximation to the Hessian, they then take $p(s_l) \approx \cosh^{-\theta/\beta}(\beta s_l)$, which asymptotically approximates the Laplacian prior for $\beta \to \infty$. Moreover, a low noise level is assumed. The above approximations finally lead to the gradient learning rule

$$\Delta \mathbf{A} = \mathbf{A}^{\mathsf{T}} \mathbf{A} \ \nabla_{\mathbf{A}} \log p(\mathbf{x}|\mathbf{A}) \approx -\mathbf{A} \left(\mathbf{I} + \mathbf{z} \hat{\mathbf{s}}^{\mathsf{T}} \right) \ ,$$

where, again, $z_l = \partial \log p(s_l) / \partial s_l$. Note that this has the same functional form as the Infomax-ICA learning rule, however the basis matrix is generally non-square in this case. In contrast to the standard ICA learning rule, and where the sources are estimated simply by $\mathbf{s} = \mathbf{W}\mathbf{x}$, where the unmixing matrix is $\mathbf{W} = \mathbf{A}^{-1}$, here we must use a nonlinear optimization algorithm in order to estimate the coefficients, using Eq. (11). Due to the low-noise assumption, the level of the observation noise is not estimated from the data and has to be set manually. Lewicki and Sejnowski's algorihm, however, is faster in obtaining good approximate solutions than the linear programming method and is more easily generalizable to other priors.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27456v1 | CC BY 4.0 Open Access | rec: 30 Dec 2018, publ: 30 Dec 2018

Girolami [26] proposes a variational method for learning sparse represenations. In particular, his method offers a solution to the problem of analytically integrating the data likelihood, for a range of heavy-tailed distributions. Starting from the heavy-tailed distribution $p(s) \propto \cosh^{-\frac{1}{\beta}}(\beta s)$, he derives a variational approximation to the Laplacian prior by introducing a variational parameter, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_L)$, such that the prior $p(\mathbf{s}) = \prod_{l=1}^{L} \exp(-|s_l|)$ becomes $p(\mathbf{s}; \boldsymbol{\xi})$, with $\mathbf{s}|\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{s}; \mathbf{0}, \boldsymbol{\Lambda})$ and $\boldsymbol{\Lambda} = \mathrm{diag}\,(|\xi_l|)$. Then $p(\mathbf{s})$ is the supremum

$$p(\mathbf{s}) = \sup_{\boldsymbol{\xi}} \left\{ \left[\prod_{l=1}^{L} \varphi(\xi_l) \right] \mathcal{N}(\mathbf{s}; \mathbf{0}, \boldsymbol{\Lambda}) \right\} \quad ,$$

with $\varphi(\xi) \to \exp(-\frac{1}{2}|\xi|)\sqrt{2\pi|\xi|}$ as $\beta \to \infty$. The above is derived using a variational argument and using convex duality [37], [53]. In essence, what this approximation means is that, at each point of its domain, the intractable prior is lower-bounded tightly by a best-matching Gaussian with width parameter ξ, with this variational parameter being estimated by the algorithm along with the model parameters. Using the above, the posterior takes a Gaussian form. This enables him to derive an EM algorithm in order to infer the sparse coefficients and learn the overcomplete basis vectors of the representation. Girolami applies his sparse representation algorithm to the problem of overcomplete source separation and achieves superior results compared to the algorithm of Lewicki and Sejnowski.

The problem of sparsely representating a data matrix described above is a special case of the more general problem of recovering latent signals that themselves have a sparse representation in a signal *dictionary* (Zibulevsky et al., [62]). Many real-world signals have sparse representations in a proper signal dictionary but not in the physical domain. The discussion in Zibulevsky et al. is motivated by starting from the case of representing sparse signals in the physical domain, depicted in Fig. 8, and then noting that the intuition there carries over to the situation of sparsely recovering signals in a transform domain.

6 Conclusion

This paper provided a high-level overview of the philosophy and basic principles of the data decomposition approach to data analysis. Starting from the classical Singular Value Decomposition method of Linar Algebra and progressing towards newer and more powerful methods, such as Independent Component Analysis, we showed how the interplay of a geometric depiction of the data space and the use of prior constraints on the unknowns can lead to stable solutions to the inverse problem of reconstructing the sources. Moreover, we gradually lifted the biologically implausible priors imposed by earlier methods and focused on the principle of parsimony and on sparsity. These have already given exciting results in the field of Computational Neuroscience and promise to give analogous results in other fields of Science and Engineering as well.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27456v1 | CC BY 4.0 Open Access | rec: 30 Dec 2018, publ: 30 Dec 2018

Acknowledgements

The author is indebted to Prof. Steven Roberts for his invaluable discussions and stimulating ideas.

A A Primer on Probability Theory

A.1 Probability Space

The axiomatic formulation of probability starts by defining a probability space, which is a tuple, (Ω, P), that describes our idea about uncertainty with respect to a random experiment. It defines:

- A *sample space*, Ω, of possible outcomes, $\{\omega_i\}$, of a random experiment and
- A probability *measure*, P, which describes how likely an outcome is.

Now, let $\mathcal{A}$ be a collection of subsets of Ω, called random events. Then for $A \in \mathcal{A}$ the two following conditions must hold:

- Probabilities must be non-negative, $P(A) \geq 0$, and $P(\Omega) = 1$,
- Probabilities must be additive: for two disjoint events, A, B,

$$P(A \cap B) = P(A) + P(B) \ .$$

We also define the *conditional probability*, which can be thought of as "a probability within a probability",

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) \neq 0 \ .$$

Then *random variables* (r.v.'s) are defined as functions from Ω to a range, $\mathcal{R}$, e.g. a subset of $\mathbb{R}$ or $\mathbb{N}$, etc. These can, inversely, define events as:

$$\mathcal{R} \to \Omega : \quad A(x) = \Big\{ \omega \in \Omega \ : \ \big[x(\omega)\big] \Big\} \ ,$$

where $\big[\,\cdot\,\big]$ denotes a "predicate"[8] (e.g. the event '$x > 2$'), and therefore act as "filters" of certain experimental outcomes.

Probability *densities* are defined as densities of probability measures:

$$p(x) = \frac{\mathrm{d}}{\mathrm{d}x} P(A(x))|_x, \quad \text{with} \quad A(x) = \Big\{ x' \in [x, x + \mathrm{d}x] \Big\}, \ \ x \in \mathcal{R} \ .$$

Finally *joint densities* (e.g. for the case of two random variables X, Y) are defined as

$$p_{XY}(x, y) = p\left(\Big\{ \omega : X(\omega) = x \wedge Y(\omega) = y \Big\} \right) \ .$$

Joint densities of more than two r.v.'s are defined analogously.

[8] This is called an 'Iverson bracket' in Iverson notation [35].

A.2 Three Simple Rules

Probability theory is a mathematically elegant theory. The whole construction can be based on the following three simple rules:

1. The Product rule, which gives the probability of the logical conjunction of two events A and B,

$$P(A \cap B) = P(A|B)P(B) \ .$$

This can be generalized for N events, giving the chain rule

$$P\left(\bigcap_{i=1}^{N} A_i\right) = \prod_{i'=1}^{i-1} P\left(A_i \middle| \bigcap_{i'=1}^{i} A_{i'}\right), \quad i' < i \ .$$

This will be valuable for reasoning in Bayesian networks later.

2. Bayes' rule, which is a recipe that tells us how to update our knowledge in the presence of new information, and can directly be derived from the definition of conditional probability and the product rule,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \quad P(B) \neq 0 \ .$$

3. Marginalization: given a joint density, $p_{XY}(x, y)$, get the marginal density of X or Y by integration (i.e. 'integrate out' the uncertainty in one variable):

$$p_X(x) = \int_{\{Y \in \mathcal{Y}\}} p_{XY}(x, y) \mathrm{d}y \ .$$

In principle, this is everything we need to know in order to perform probabilistic modelling and inference.

References

[1] S.-I. Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10:251–276, 1998.

[2] H. Asari, B. Pearlmutter, and A. Zador. Sparse representations for the cocktail party problem. *Journal of Neuroscience*, 26(28):7477–7490, 2006.

[3] H. Attias. Independent Factor Analysis. *Neural Computation*, 11(4):803–851, May 1999.

[4] R. Baraniuk. Compressive Signal Processing, 2010.

[5] H. B. Barlow. Unsupervised Learning. *Neural Computation*, 1(3):295–311, 1989.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27456v1 | CC BY 4.0 Open Access | rec: 30 Dec 2018, publ: 30 Dec 2018

[6] H.B. Barlow, T P Kaushal, and G.J. Mitchison. Finding minimum entropy codes. *Neural Computation*, 1:412–423, 1989.

[7] C. Beckmann and S. Smith. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Trans. Med. Imag.*, 23:137–152, 2004.

[8] A. Bell and T. J. Sejnowski. An information-maximisation approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[9] V.I. Bogachev. *Gaussian measures*. Amer. Math. Soc., 1998.

[10] V. D. Calhoun, T. Adali, L. K. Hansen, J. Larsen, and J. J. Pekar. ICA of functional MRI data: An overview, Apr. 2003. Invited Paper.

[11] J.-F. Cardoso. Infomax and maximum likelihood for blind separation. *IEEE Signal Process. Lett.*, 4(4):112–114, 1997.

[12] J.-F. Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 90(8):2009–2026, Oct. 1998.

[13] J.-F. Cardoso. The three easy routes to independent component analysis; contrasts and geometry. In *Proc. of the ICA 2001 workshop*, San Diego, Dec. 2001.

[14] Z. Chen and S. Haykin. On different facets of regularization theory. *Neural Computation*, 14:2791–2846, December 2002.

[15] R. Choudrey, W.D. Penny, and S.J. Roberts. An ensemble learning approach to independent component analysis. In *Proceedings of the 2000 IEEE Signal Processing Society Workshop*, pages 435–444. IEEE Neural Networks for Signal Processing X, 2000.

[16] Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

[17] R. T. Cox. Probability, Frequency, and Reasonable Expectation. *Am. Jour. Phys.*, 14:1–13, 1946.

[18] G. Darmois. Analyse Générale des Liaisons Stochastiques. *Rev. Inst. Internat. Stat.*, 21:2–8, 1953.

[19] I. Daubechies, E. Roussos, S. Takerkart, M. Benharrosh, C. Golden, K. D'Ardenne, W. Richter, J. Cohen, and J. Haxby. Independent component analysis for brain fMRI does not select for independence. *PNAS*, 106(26):10415–10412, 2009.

[20] A.P. Dempster, N.M. Laird, and D.B Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27456v1 | CC BY 4.0 Open Access | rec: 30 Dec 2018, publ: 30 Dec 2018

[21] D. L. Donoho. Sparse components of images and optimal atomic decompositions. *Constructive Approximation*, 17:353–382, 2001.

[22] D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *PNAS*, 100(5):2197–2202, March 2003.

[23] D. J. Field. Wavelets, vision and the statistics of natural scenes. *Phil. Trans. R. Soc. Lond. A*, 357:2527–2542, 1999.

[24] J. H. Friedman and J. W. Tukey. A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE Transactions on Computers*, 23(9):881–890, Sep. 1974.

[25] M. Girolami, editor. *Advances in Independent Component Analysis*. Perspective on Neural Computing. Springer-Verlag, New York, Aug. 2000.

[26] M. Girolami. A Variational Method for Learning Sparse and Overcomplete Representations. *Neural Computation*, 13(11):2517–2532, Nov. 2001.

[27] C. Golden. *Spatio-Temporal Methods in the Analysis of fMRI Data in Neuroscience*. PhD thesis, Princeton University, 2005.

[28] Charles F. Golub, Gene H.; Van Loan. *Matrix Computations*. Johns Hopkins, 1996.

[29] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, second edition, February 2009.

[30] A. Hyvarinen. The Fixed-Point Algorithm and Maximum Likelihood Estimation for Independent Component Analysis. *Neural Processing Letters*, 10(1):1–5, 1999.

[31] A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, June 2001.

[32] A. Hyvarinen and R. Karthikesh. Imposing sparsity on the mixing matrix in independent component analysis. *Neurocomputing*, 49:151–162, 2002. Special Issue on ICA and BSS.

[33] A. Hyvarinen and E. Oja. A Fast Fixed-Point Algorithm for Independent Component Analysis. *Neural Computation*, 9(7):1483–1492, October 1997.

[34] A. Hyvarinen and E. Oja. Independent Component Analysis: Algorithms and Applications. *Neural Networks*, 13(4–5):411–430, 2000.

[35] Kenneth Iverson. *A Programming Language*. Wiley, 1962.

[36] D. J. C. Mackay J. W. Miskin. Application of Ensemble Learning ICA to Infrared Imaging. In *Conference: Independent Component Analysis - ICA*, 2000.

[37] T. Jaakkola. *Variational Methods for Inference and Estimation in Graphical Models*. PhD thesis, Mass. Inst. of Techn., 1997.

[38] E.T. Jaynes and G.L. Bretthorst (editor). *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[39] K.H. Knuth. A Bayesian approach to source separation. In C. Jutten J.-F. Cardoso and P. Loubaton, editors, *Proceedings of the First International Workshop on Independent Component Analysis and Signal Separation: ICA'99*, pages 283–288, Aussios, France, Jan 1999.

[40] H. Lappalainen. Ensemble learning for independent component analysis. In *Proceedings of the First International Workshop on Independent Component Analyis*, pages 7–12, 1999.

[41] H. Lappalainen. Fast Fixed-Point Algorithms for Bayesian Blind Source Separation, 1999.

[42] N. D. Lawrence and C. M. Bishop. Variational Bayesian Independent Component Analysis. Technical report, , 1999.

[43] N. Lazar. *The Statistical Analysis of Functional MRI Data*. Statistics for Biology and Health. Springer, 2008.

[44] M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12(2):337–365, February 2000.

[45] Y. Li, A. Cichocki, and S.-I. Amari. Analysis of sparse representation and blind source separation. *Neural Computation*, 16(6):1193–1234, June 1994.

[46] R. Linsker. A local learning rule that enables information maximization for arbitrary input distributions. *Neural Computation*, 9:1661–1665, November 1997.

[47] D. J. C. MacKay. Maximum likelihood and covariant algorithms for independent component analysis. Technical report, University of Cambridge, 1996.

[48] S. G. Mallat and Z. Zhang. Matching Pursuits with Time-Frequency Dictionaries. *IEEE Transactions on Signal Processing*, pages 3397–3415, December 1993.

[49] J. W. Miskin and D. J. C. MacKay. Ensemble Learning for Blind Source Separation. In S. J. Roberts and R. M. Everson, editors, *Independent Components Analysis: Principles and Practice*. Cambridge University Press, 2001.

[50] J.-P. Nadal and N. Parga. Nonlinear neurons in the low-noise limit: a factorial code maximizes information transfer. *Network: Computation in Neural Systems*, 5(4):565–581, 1994.

[51] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, June 1996.

[52] B. A. Olshausen and D. J. Field. Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1? *Vision Research*, 37(23):3311–3325, 1997.

[53] J. A. Palmer, D. P. Wipf, K. Kreutz-delgado, and B. D. Rao. Variational EM algorithms for non-Gaussian latent variable models. In *Advances in Neural Information Processing Systems 18*, pages 1059–1066. MIT Press, 2006.

[54] B. Pearlmutter and L. Parra. Maximum Likelihood Blind Source Separation: A context-sensitive generalization of ICA. In *Advances in Neural Information Processing Systems 9*, pages 613–619, 1997.

[55] C. R. Rao. A decomposition theorem for vector variables with a linear structure. *Ann. Math. Statist.*, 40(5):1845–1849, 1969.

[56] S. Roberts. Independent component analysis: source assessment and separation, a Bayesian approach. *Vision, Image and Signal Processing, IEE Proceedings*, 145(3):149–154, Jun. 1998.

[57] S. Roberts and R. Everson, editors. *Independent Component Analysis: Principles and Practice*. Cambridge University Press, 2001.

[58] J. V. Stone. *Independent Component Analysis: A Tutorial Introduction*. MIT Press, September 2004.

[59] Y. W. Teh, M. Welling, S. Osindero, G. E. Hinton, T.-W. Lee, J.-F. Cardoso, E. Oja, and S.-I. Amari. Energy-based models for sparse overcomplete representations. *Journal of Machine Learning Research*, 4:2003, 2003.

[60] A. Tonazzini, L. Bedini, and E. Salerno. A Markov Model for Blind Image Separation by a Mean-Field EM algorithm. *IEEE Trans. Image Proc.*, 15(2), 2006.

[61] M. Welling and M. Weber. A constrained EM algorithm for Independent Component Analysis. *Neural Computation*, 13:677–689, 2001.

[62] M. Zibulevsky and B. A. Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural Computation*, 13:863–882, April 2001.