An alternative approach to infomax and independent component analysis

Aapo Hyvärinen

Neural Networks Research Centre

Helsinki University of Technology

P.O. Box 5400, FIN-02015 HUT, Finland

http://www.cis.hut.fi/aapo/

18th August 2001

Abstract

Infomax means maximization of information flow in a neural system. A nonlinear version of infomax has been shown to be connected to independent component analysis and the receptive fields of neurons in the visual cortex. Here we show a problem of nonrobustness of nonlinear infomax: it is very sensitive to the choice the nonlinear neuronal transfer function. We consider an alternative approach in which the system is linear, but the noise level depends on the mean of the signal, as in a Poisson neuron model. This gives similar predictions as the nonlinear infomax, but seem to be more robust.

1 Introduction

Information-theoretic or statistical criteria can be used to model aspects of sensory information processing in the brain. This is based on investigating the optimal way of processing the natural data that is input to the sensory systems, assuming that some aspects of neural information processing are close to optimal. An important group of principles is based on redundancy reduction, sparseness and statistical independence [2, 8, 17, 9]. A related criterion that is more closely connected to the limited capacity of neural systems is the infomax principle [14, 15, 16, 3]. According to the infomax principle, sensory systems (or parts thereof) are adapted so that they transmit the maximum amount of information. This is typically measured as the mutual information between the input and the output (which is usually not the final output but some intermediate result).

Recently, a lot of attention has been paid to a particular kind of system, consisting of *nonlinear*, sigmoidal neurons. This system can be analyzed in the case of vanishing, isotropic noise [16, 3]. Maximization of information flow in

such a system has been shown, for example, to produce simple cell receptive fields when the input is similar to the natural input of the visual system [4]. Maximization of information flow is here closely related to a form of redundancy reduction [2, 8] in which the components of output of the system are made maximally independent, a method also called independent component analysis (ICA) [13, 12, 10], and closely related to sparse coding [17].

In this paper we argue, however, that this nonlinear infomax model has a serious problem. It is very sensitive to small changes in the nonlinear transfer function; even small, almost imperceptible changes can completely change the predictions of the model. Since the nonlinearity is only a crude approximation of what happens in real neurons, this kind of nonrobustness may mean that this model may not be very relevant in modelling biological systems.

We present an alternative formulation of infomax that leads to very similar predictions, but with quite different assumptions. Instead of concentrating on the nonlinearity, we concentrate on the *noise model*. Widely-used models of neuronal noise are based on a Poisson process, in which the noise level is dependent on the output level. Thus, we introduce an anisotropic noise whose variance is not constant. This leads to similar objective functions as the nonlinear infomax, but may be more robust from the modelling viewpoint. One can also consider these two aspects (nonlinear transfer function or Poisson noise) as two non-exclusive assumptions, both of which can be incorporated in a single model.

2 Nonlinear infomax

2.1 Definition

Let us assume that we have N neurons. Each receives the same set of inputs $x_1, ..., x_N$, denoted together as a random vector $\mathbf{x}$. The weight vectors of the neurons are denoted by $\mathbf{w}_i$, and the transfer functions by f_i. A noise term n_i is included in each output y_i as well. This can be expressed as

$$y_i = f_i(\mathbf{w}_i^T \mathbf{x}) + n_i \tag{1}$$

The goal is thus to maximize the mutual information

$$I(\mathbf{x}, \mathbf{y}) = H(\mathbf{x}) + H(\mathbf{y}) - H(\mathbf{x}, \mathbf{y}) \tag{2}$$

In the conventional framework [16, 3], the noise variances are assumed to be all equal. In the limit of zero noise, we then obtain

$$I(\mathbf{x}, \mathbf{y}) = \log |\det \mathbf{W}| + \sum_i E\{\log f_i'(\mathbf{w}_i^T \mathbf{x})\} + \text{const.} \tag{3}$$

where the last term depends only on the noise level and not on the weights. The weights are here all collected in a single matrix $\mathbf{W} = (\mathbf{w}_1, ..., \mathbf{w}_N)^T$. It can be shown that in this case, the maximization of information flow is equivalent

ot estimating the ICA model [5, 10], if the nonlinear functions f_i are suitable approximations of the cumulative distribution functions of the independent components.

2.2 Problem of nonrobustness

The problem with this formulation is that if the transfer function is just a little bit different, the predictions may change completely. Usually, the logistic function [3]

$$f_i(u) = \frac{1}{1 + \exp(-u)} \tag{4}$$

is used. This can be shown to estimate the ICA model in the case of supergaussian (sparse) independent components. This is a correct way of estimating the ICA model for natural image data in which the components really are supergaussian. However, if the transfer function is changed to

$$f_i(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{\xi^2}{2\sigma^2}) d\xi \tag{5}$$

that is, the gaussian cumulative distribution function, the method does not estimate the ICA model anymore, since this would amount to assuming gaussian independent components, which makes the estimation impossible [6, 10]. An even worse situation arises if we change the function to, for example,

$$f_i(u) = \int_{-\infty}^{u} \frac{c_1}{c_2} \exp(-\frac{\xi^4}{c_2^4}) d\xi \tag{6}$$

where c_1 and c_2 are appropriate scaling constant to make this the cumulative distribution function of given variance. This amounts to estimating the ICA model assuming subgaussian independent components. If the components are actually supergaussian, as is the case with natural image data [17, 4], the estimation fails completely [11].

The three nonlinear functions in Eqs (4,5,6) look all very similar, however. This is illustrated in Fig. 1. (Note that we can freely scale the functions along the x-axis since this has no influence on the behaviour in ICA estimation. Here we have chosen the scaling parameters σ and c_2 in these three functions to emphasize the similarity.) All the three functions have the same kind of qualitative behavior. In fact, all cumulative distribution functions look very similar after appropriate scaling along the x-axis.

It is not very likely that the neural transfer functions (which are only crude approximations anyway) would consistently be of the type in Eq. (4), and not closer to the two other transfer functions. Thus, *the model can be considered to be nonrobust, that is, too sensitive to small fluctuations in its parameters.*[1]

---

[1]It could be argued that the nonlinear transfer function can be estimated from the data and it need not be carefully chose beforehand, but this only modifies this robustness problem because then the nonlinearity must be estimated very precisely.
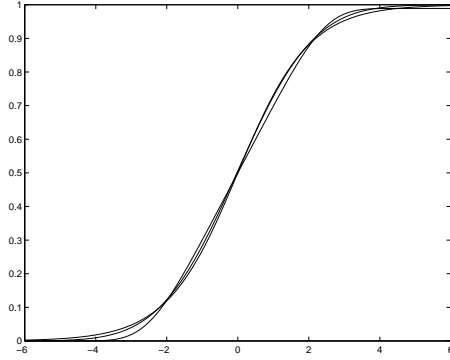
Figure 1: The three sigmoidal nonlinearities corresponding to logistic, gaussian, and subgaussian prior cumulative distributions in Eqs. (4,5,6) for the independent components. The nonlinearities are practically indistinguishable.

3   Linear infomax with noise of changing variance

3.1   Definition of noise model

One remedy for the nonrobustness of the nonlinear infomax model is to shift attention to the noise term in Eq. (1). Let us thus take the linear function as f_i, which is to be considered as a transfer function of minimum commitment, and change the definition of the noise term instead.

What would be a sensible model for the noise? If we consider a basic Poisson model of spike trains, the rate coding would have the corresponding Poisson noise. By Poisson noise we mean the error in the conventional estimation of the firing rate parameter when observing a Poisson spike train, that is, the difference between the observed firing rate and the true firing rate parameter in a given time interval. Such Poisson noise has a variance that is equal to its mean. Thus, we would have

$$\mathrm{var}(n_i|\mathbf{x}) \propto r + |\mathbf{w}_i^T \mathbf{x}| \tag{7}$$

where r is a constant that embodies the spontaneous firing rate that is not zero (and hence does not have zero noise). We take the absolute value of $\mathbf{w}_i^T \mathbf{x}$ because we consider the output of a signed neuron to be actually coded by two different neurons, one for the negative part and one for the positive part, as is conventional in models of the primary visual cortex. The neuron that is active always has a Poisson noise whose variance is proportional to $r + |\mathbf{w}_i^T \mathbf{x}|$.

The distribution of Poisson noise is, of course, nongaussian in the single neuron case. However, in the following we approximate it as gaussian noise. The fundamental property of this new type of noise is considered to be the variance behavior given in Eq. (7), and not its nongaussianity. This is, in fact, not necessarily an approximation, since one could consider a neuronal ensemble

whose mean firing rate is given by $r + |\mathbf{w}_i^T \mathbf{x}|$. The Poisson noises of different neurons are them averaged, which gives a gaussian noise distribution with the same kind of variance behavior. Therefore, we call noise with this kind of variance behavior "noise with *Poisson-like variance*" instead of Poisson noise.

Of course, a more general form of the model can be obtained by defining the variance to be a nonlinear function of the quantity in Eq. (7). To investigate the robustness of our model, we do in the following all the computations in the more general case where

$$\text{var}(n_i|\mathbf{x}) = h(\mathbf{w}_i^T \mathbf{x}) \tag{8}$$

where h is some arbitrary function with nonnegative values, for example

$$h(u) = r + |u|, \tag{9}$$

which gives Eq. (7).

A similar noise model for infomax was introduced by Vreeswijk [18], using a rather different motivation.

3.2 Calculation of information transfer

Let us now compute the mutual information between inputs and outputs in the case of noise of Poisson-like variance. This section can be skipped if you're not interested in the exact derivation.

To do this, let us express the noise with Poisson-like variance as a product of a hypothetical constant-variance noise variable n_i^o and $\mathbf{w}_i^T \mathbf{x}$:

$$n_i = n_i^o \sqrt{h(\mathbf{w}_i^T \mathbf{x})} \tag{10}$$

where $\text{var}(n_i^o)$ equals some constant σ^2. We can express $\mathbf{x}$ and $\mathbf{y}$ as a transformation $\mathbf{T}$ of $\mathbf{x}$ and $\mathbf{n}^o$:

$$\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \mathbf{T} \begin{pmatrix} \mathbf{x} \\ \mathbf{n}^o \end{pmatrix} \tag{11}$$

According to the entropy transformation formula [7], the entropy $H(\mathbf{x}, \mathbf{y})$ is equal to the entropy of $H(\mathbf{x}, \mathbf{n}^o)$ plus $E\{\log|\det \mathbf{J}|\}$, where $\mathbf{J}$ is the matrix of partial derivatives (Jacobian) of this function $\mathbf{T}$. Considering that we have defined

$$y_i = \mathbf{w}_i^T \mathbf{x} + n_i^o \sqrt{h(\mathbf{w}_i^T \mathbf{x})}, \tag{12}$$

the partial derivatives of $\mathbf{x}$ and $\mathbf{y}$ with respect to $\mathbf{x}$ and $\mathbf{n}_o$ can be computed as

$$\frac{\partial \mathbf{x}}{\partial \mathbf{x}} = \mathbf{I} \tag{13}$$

$$\frac{\partial \mathbf{x}}{\partial \mathbf{n}^o} = \mathbf{0} \tag{14}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{n}^o} = \text{diag}(\sqrt{h(\mathbf{w}_i^T \mathbf{x})}) \tag{15}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \mathbf{M} \tag{16}$$

$$\tag{17}$$

5

where $\mathbf{M}$ is a matrix that does not need to be computed. The determinant of the Jacobian can then be obtained as

$$\det \mathbf{J} = \det \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{M} & \mathrm{diag}(\sqrt{h(\mathbf{w}_i^T \mathbf{x})}) \end{pmatrix} = 1 \times \prod_i \sqrt{h(\mathbf{w}_i^T \mathbf{x})} - 0 \times \det \mathbf{M} \quad (18)$$

Thus, we have

$$H(\mathbf{x}, \mathbf{y}) = H(\mathbf{x}, \mathbf{n}^o) + E\{\sum_i \log \sqrt{h(\mathbf{w}_i^T \mathbf{x})}\} \quad (19)$$

where, further, we have $H(\mathbf{x}, \mathbf{n}^o) = H(\mathbf{x}) + H(\mathbf{n}^o)$ by independence.

To obtain the mutual information, we still need to compute $H(\mathbf{y})$ (there is no point in computing $H(\mathbf{x})$ which does not depend on $\mathbf{W}$). In the limit of infinitesimal noise, we have by continuity (assuming all the relevant functions to be sufficiently smooth):

$$H(\mathbf{y}) = H(\mathbf{Wx}) = \log |\det \mathbf{W}| + H(\mathbf{x}) \quad (20)$$

We now obtain the mutual information in the limit of zero noise as

$$I(\mathbf{x}, \mathbf{y}) = H(\mathbf{x}) + H(\mathbf{y}) - H(\mathbf{x}, \mathbf{y})$$
$$= H(\mathbf{x}) + \log |\det \mathbf{W}| + H(\mathbf{x}) - [H(\mathbf{x}) + H(\mathbf{n}^o) + E\{\sum_i \log \sqrt{h(\mathbf{w}_i^T \mathbf{x})}\}] \quad (21)$$

3.3 Information transfer and ICA

Thus, we finally have the following result:

$$I(\mathbf{x}, \mathbf{y}) = \log |\det \mathbf{W}| - \sum_i E\{\log \sqrt{h(\mathbf{w}_i^T \mathbf{x})}\} + \text{const.} \quad (22)$$

where terms that do not depend on $\mathbf{W}$ are grouped in the constant. A comparison of Eq. (22) with Eq. (3) reveals that in fact, mutual information is of the same algebraic form in the two cases. By taking $h(u) = 1/f'(u)^2$, we obtain an expression of the same form. Thus, we see that *considering noise with non-constant noise, we are able to reproduce the same results as with a nonlinear transfer function.*

If we consider the basic case of Poisson-like variance, which means defining the function h as in Eq. (7), this is equivalent to the nonlinear infomax with

$$f'(u) = \frac{1}{\sqrt{r + |u|}} \quad (23)$$

In the nonlinear infomax, f' corresponds to the probability density function assumed for the independent components. The function in (23) is an improper

probability density function, since it is not integrable. However, its qualitative behaviour is typically supergaussian: very heavy tails and a peak at zero.

Thus, in the basic case of Poisson-like variance, the infomax principle is equivalent to estimation of the ICA model with this improper prior density for the components. Since the choice of nonlinearity is usually critical only along the subgaussian vs. supergaussian axis [11], this improper prior distribution should still properly estimate the ICA model for most supergaussian components. We have performed simulations in which the corresponding learning rule (obtained as the natural gradient rule [1]) did properly estimate supergaussian independent components.[2]

To investigate the robustness of our model, we can consider what the noise variance structure should be like to make the estimation of supergaussian components fail. As with the nonlinear infomax, we can find a noise structure that corresponds to the estimation of gaussian independent components. From the relation $h(u) = 1/f'(u)^2$, we see that the gaussian case corresponds to

$$h(u) \propto \exp(u^2) \tag{24}$$

This is an exploding function clearly very different from the Poisson-like variance structure given by the essentially linear function in Eq. (9). In the space of possible functions h that define the noise structure in our model, the function in Eq. (24) can be considered as a borderline between those variance structures that enable the estimation of supergaussian independent components, and those that do not. These two different choices for h, together with the one corresponding to subgaussian independent components as in Eq. (6) are plotted in Fig. 2. The Poisson-like variance is clearly very different from the other two cases.

Thus, we may conclude that our model with Poisson-like variance is quite robust against changes of parameters in the model, since the main parameter is the function h, and this can change qualitatively quite a lot before the behaviour of the model with respect to ICA estimation changes. This is in contrast to the nonlinear infomax principle where the nonlinearity has to be very carefully chosen according to the distribution of the data.

4   Conclusion

We claimed that the nonlinear infomax framework suffers from a problem of non-robustness and offered an alternative approach. In our approach, the transfer functions are linear but the variance of the noise depends on the signal amplitude, as in Poisson processes. Our approach is mathematically equivalent to the nonlinear infomax but it may be more plausible as a model of the physical phenomena involved. In must be noted, however, that the two versions of infomax

---

[2]There is, however, the problem of scaling the components. Since the improper density has infinite variance, the estimates of the components (and the weight vectors) grow infinitely large. Such behavior can be prevented by adding a penalty term of the form $\alpha \sum_i \|\mathbf{w}_i\|^2$ in the objective function. An alternative approach would be to use a saturating nonlinearity, thus combining our model with the nonlinear infomax model.
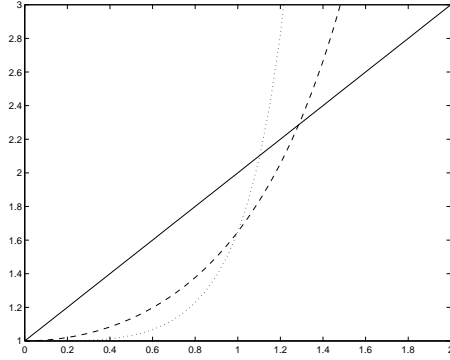
Figure 2: The three functions h that give the dependence of noise variance, for the same three distributions of the independent components as in Fig. 2. Solid line: Poisson like variance as in Eq. (9). Dashed line: the case of gaussian distribution. Dotted line: the subgaussian distribution corresponding to the cumulative distribution function in Eq. (6). Here, the functions are very different, which indicates better robustness for the model.

are not incompatible: one could use a model in which both nonlinearities and Poisson-like noise work together. In fact, our model needs an additional penalty mechanism for preventing explosive behavior, and this could be provided by a saturating nonlinearity.

References

[1] S.-I. Amari, A. Cichocki, and H.H. Yang. A new learning algorithm for blind source separation. In *Advances in Neural Information Processing Systems 8*, pages 757–763. MIT Press, 1996.

[2] H. B. Barlow. Single units and sensation: A neuron doctrine for perceptual psychology? *Perception*, 1:371–394, 1972.

[3] A.J. Bell and T.J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[4] A.J. Bell and T.J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37:3327–3338, 1997.

[5] J.-F. Cardoso and B. Hvam Laheld. Equivariant adaptive source separation. *IEEE Trans. on Signal Processing*, 44(12):3017–3030, 1996.

[6] P. Comon. Independent component analysis—a new concept? *Signal Processing*, 36:287–314, 1994.

[7] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, 1991.

[8] D.J. Field. What is the goal of sensory coding? *Neural Computation*, 6:559–601, 1994.

[9] A. Hyvärinen and P. O. Hoyer. A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Research*, 41(18):2413–2423, 2001.

[10] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* Wiley Interscience, 2001.

[11] A. Hyvärinen and E. Oja. Independent component analysis by general nonlinear Hebbian-like learning rules. *Signal Processing*, 64(3):301–313, 1998.

[12] A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.

[13] C. Jutten and J. Hérault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

[14] S. Laughlin. A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung*, 36 C:910–912, 1981.

[15] R. Linsker. Local synaptic learning rules suffice to maximize mutual information in a linear network. *Neural Computation*, 4:691–702, 1992.

[16] J.-P. Nadal and N. Parga. Non-linear neurons in the low noise limit: a factorial code maximizes information transfer. *Network*, 5:565–581, 1994.

[17] 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, 1996.

[18] Carl van Vreeswijk. Whence sparseness? In *Advances in Neural Information Processing Systems*, volume 13, pages 180–186. MIT Press, 2001.