A unified probabilistic model for independent and principal component analysis

Aapo Hyvärinen*
Dept of Computer Science and HIIT
University of Helsinki, Finland

A chapter in
Advances in Independent Component Analysis and Learning Machines
(Festschrift to Erkki Oja)
Academic Press, 2015

Abstract

Principal component analysis (PCA) and independent component analysis (ICA) are both based on a linear model of multivariate data. They are often seen as complementary tools, PCA providing dimension reduction and ICA separating underlying components or sources. In practice, a two-stage approach is often followed, where first PCA and then ICA is applied. Here, we show how PCA and ICA can be seen as special cases of the same probabilistic generative model. In contrast to conventional ICA theory, we model the variances of the components as further parameters. Such variance parameters can be integrated out in a Bayesian framework, or estimated in a more classic framework. In both cases, we find a simple objective function whose maximization enables estimation of PCA and ICA. Specifically, maximization of the objective under Gaussian assumption performs PCA, while its maximization for whitened data, under assumption of non-Gaussianity, performs ICA.

Keywords: Independent component analysis; Principal component analysis; Multivariate statistics; Blind source separation; Bayesian analysis.

1 Introduction

Principal component analysis (PCA) and independent component analysis (ICA) are two fundamental methods for unsupervised learning. In the machine learning and neural networks literature, they has a relatively long history. Neural

*Contact address: Dept of Computer Science, P.O.Box 68, FIN-00014 University of Helsinki, Finland. Fax: +358 2941 51234 email: Aapo.Hyvarinen@helsinki.fi.

learning for PCA started by Oja's Rule [3] and its extensions [5]. An early connection between PCA and ICA was given by nonlinear versions of PCA criteria [4]. The general theory ICA is explained in [1].

Let us consider a linear generative model

$$x_i = \sum_{j=1}^{n} a_{ij} s_j, i = 1, \ldots, n \tag{1}$$

where the x_i are observed random variables, the s_j are latent random variables (components) which are assumed mutually independent, and the a_{ij} are parameters. Depending on further assumptions, this framework can implement ICA or PCA. In particular, if we assume that the s_j are non-Gaussian, we obtain the basic version of ICA. The typical goal of ICA is to "separate sources" in the sense that we want to recover the original s_i.

On the other hand, if we assume that the components s_j are Gaussian and have different variances, we obtain a model which may be related to PCA, depending on what further assumptions, such as the orthogonality of the matrix $\mathbf{A}$ which collects the coefficients a_{ij}, are made. This approach to PCA is slightly unconventional, but we will see below that it is equivalent to the classic one. The goal in PCA is not so much to recover (all) the original s_i but to find the subspace spanned by a limited number of the s_i (and the corresponding columns of the matrix $\mathbf{A}$) which explains the largest amount of variance of the data.

In this paper, our purpose is to develop a probabilistic model based on (1) which unifies PCA and ICA in the sense maximization of the likelihood performs either PCA or ICA depending on the specific constraints and the data. The basic idea in our model is to modify the ICA assumptions so that we explicitly model the variances of the components, and then integrate them out in a Bayesian framework.

Using such a variant of the linear generative model, we show the following. First, if the components are assumed Gaussian in the model, and we constrain $\mathbf{A}$ to be orthogonal, maximization of the likelihood performs PCA. Second, if the components are assumed non-Gaussian in the model, maximization of the likelihood separates original non-Gaussian components, i.e. recovers the mixing matrix up to trivial indeterminacies like ICA, again assuming that $\mathbf{A}$ is constrained orthogonal, and further that the data is prewhitened.

2 Variance of components as separate parameter

2.1 Definition of new model

It is well-known that in the linear model (1), the variances of the components cannot be recovered. This is because we can always rescale a component s_j as $\gamma_j s_j$, redefine the mixing coefficients as a_{ij}/γ_j, and the model is equivalent in the sense that the observed data has the same distribution.

The conventional approach to ICA is to define that the variances of the components are equal to one. This approach simplifies the problem, but it

seems to have the drawback that the connection to PCA is lost, because PCA is dependent on the principal components having distinct variances.

We propose here to consider the variances of the components as separate parameters. Further, we propose to integrate those parameters out in a Bayesian approach.

Thus, define σ_j^2 to be the variance of the j-th independent component. Denote the vector collecting the σ_j as $\boldsymbol{\sigma}$, and denote $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n)^T = \mathbf{A}^{-1}$. Then, we have, using the well-known derivation of the ICA likelihood [1]:

$$p(\mathbf{x}|\mathbf{A}, \boldsymbol{\sigma}) = |\det \mathbf{W}| \prod_j \frac{1}{\sigma_j} p_j(\frac{\mathbf{w}_j^T \mathbf{x}}{\sigma_j}) \tag{2}$$

where p_i denotes the pdf of the s_i when it is standardized to unit variance.

In order to be able to integrate out the σ_j in closed form, we have to restrict ourselves to a special form of the p_i. We consider the generalized Laplacian distribution, also called the generalized Gaussian distribution. The pdf is given by

$$p_j(s|\alpha_j) = \frac{1}{Z(\alpha_j)} \exp(-|s|^\alpha C(\alpha_j)) \tag{3}$$

where α is the parameter controlling the shape of the density. For $\alpha = 2$, we obtain the Gaussian density, for $\alpha < 2$, we obtain (highly) peaked, super-gaussian densities, and for $\alpha > 2$, flat, sub-gaussian densities. The constants Z and C are needed to normalize the pdf and to make its variance equal to one, and they are well-known (although different conventions of parameterization exist), but irrelevant for our purposes.

2.2 Integrating out the variance parameter

To handle the variances in a Bayesian framework, we first need to define a prior for the σ_j. We choose to use the non-informative (Jeffreys') prior:

$$p(\boldsymbol{\sigma}) = \prod_j \frac{1}{\sigma_j} \tag{4}$$

where obviously σ_j are constrained positive.

Consider an i.i.d. sample of $\mathbf{x}$, denoted individually as $\mathbf{x}(t), t = 1, \ldots, T$ and as a whole in matrix form as $\mathbf{X} = (\mathbf{x}(1), \ldots, \mathbf{x}(T))$. We have

$$\begin{aligned} p(\mathbf{X}, \boldsymbol{\sigma}|\mathbf{A}, \boldsymbol{\alpha}) &= p(\mathbf{X}|\mathbf{A}, \boldsymbol{\sigma}, \boldsymbol{\alpha}) p(\boldsymbol{\sigma}) = p(\boldsymbol{\sigma}) \prod_t p(\mathbf{x}(t)|\mathbf{A}, \boldsymbol{\sigma}, \boldsymbol{\alpha}) \\ &= |\det \mathbf{W}|^T \prod_j \frac{1}{\sigma_j^{T+1} Z(\alpha_j)^T} \exp(-\sum_t |\frac{\mathbf{w}_j^T \mathbf{x}(t)}{\sigma_j}|^{\alpha_j} C(\alpha_j)) \end{aligned} \tag{5}$$

To integrate out $\boldsymbol{\sigma}$, make the following change of variables:

$$u_j = \sum_t |\frac{\mathbf{w}_j^T \mathbf{x}(t)}{\sigma_j}|^{\alpha_j} C(\alpha_j) \tag{6}$$

$$\Leftrightarrow \sigma_j = [\sum_t |\mathbf{w}_j^T \mathbf{x}(t)|^{\alpha_j} C(\alpha_j)]^{1/\alpha_j} u_j^{-1/\alpha_j} \tag{7}$$

$$\Rightarrow \frac{d\sigma_j}{du_j} = -\frac{1}{\alpha_j}[\sum_t |\mathbf{w}_j^T \mathbf{x}(t)|^{\alpha_j} C(\alpha_j)]^{1/\alpha_j} u_j^{-1/\alpha_j - 1} \tag{8}$$

which enables the integration out as

$$\begin{aligned}
p(\mathbf{X}|\mathbf{A}, \boldsymbol{\alpha}) &= \int p(\mathbf{X}, \boldsymbol{\sigma}|\mathbf{A}, \boldsymbol{\alpha}) d\boldsymbol{\sigma} \\
&= \int |\det \mathbf{W}|^T \prod_j \frac{1}{Z(\alpha_j)^T} [\sum_t |\mathbf{w}_j^T \mathbf{x}(t)|^{\alpha_j} C(\alpha_j)]^{-(T+1)/\alpha_j} u_j^{(T+1)/\alpha_j} \exp(-u_j) \\
&\qquad \times \frac{1}{\alpha_j} [\sum_t |\mathbf{w}_j^T \mathbf{x}(t)|^{\alpha_j} C(\alpha_j)]^{1/\alpha_j} u_j^{-1/\alpha_j - 1} d\mathbf{u} \\
&= |\det \mathbf{W}|^T \prod_j \frac{C(\alpha_j)^{-T/\alpha_j}}{Z(\alpha_j)^T} [\sum_t |\mathbf{w}_j^T \mathbf{x}(t)|^{\alpha_j}]^{-T/\alpha_j} \int u_j^{T/\alpha_j - 1} \exp(-u_j) du_j \\
&= |\det \mathbf{W}|^T \prod_j \frac{C(\alpha_j)^{-T/\alpha_j}}{Z(\alpha_j)^T} [\sum_t |\mathbf{w}_j^T \mathbf{x}(t)|^{\alpha_j}]^{-T/\alpha_j} \Gamma(T/\alpha_j)
\end{aligned} \tag{9}$$

where Γ denotes the conventional gamma function. Note that the integrals here are in the positive quadrant since u_j as well as σ_j are by definition positive.

Thus, we have the following log-likelihood:

$$\frac{1}{T} \log p(\mathbf{X}|\mathbf{A}, \boldsymbol{\alpha}) = \log |\det \mathbf{W}| - \sum_j \frac{1}{\alpha_j} \log \left[\sum_t |\mathbf{w}_j^T \mathbf{x}(t)|^{\alpha_j} \right] + f(\alpha_j, T) \tag{10}$$

where f denotes a function depending on α_j and T alone:

$$f(\alpha_j, T) = \frac{1}{T} \log \Gamma(T/\alpha_j) - \frac{1}{\alpha_j} \log C(\alpha_j) - \log Z(\alpha_j) \tag{11}$$

2.3 Alternative approach maximizing joint likelihood

An alternative, non-Bayesian approach is possible by considering the joint likelihood of $\mathbf{A}$ and $\boldsymbol{\sigma}$, directly given in (2) when evaluated for the whole sample. Again, define the p_i as in (3). Thus, we have the joint log-likelihood

$$\frac{1}{T} \log(\mathbf{X}|\mathbf{A}, \boldsymbol{\sigma}) = \log |\det \mathbf{W}| + \frac{1}{T} \sum_j \sum_t -|\frac{\mathbf{w}_j^T \mathbf{x}(t)}{\sigma_j}|^{\alpha_j} C(\alpha_j) - \log \sigma_j - \log Z(\alpha_j) \tag{12}$$

Now, for a fixed $\mathbf{W}$, we can find the maxima of this likelihood with respect to $\boldsymbol{\sigma}$ in closed form as

$$\hat{\sigma}_j(\mathbf{w}_j) = \left[\frac{\alpha_j C(\alpha_j)}{T} \sum_t |\mathbf{w}_j^T \mathbf{x}(t)|^\alpha \right]^{1/\alpha} \tag{13}$$

and we can plug this in the joint likelihood to obtain after some manipulations

$$\frac{1}{T} \log(\mathbf{X}|\mathbf{A}, \hat{\boldsymbol{\sigma}}(\mathbf{W})) = \log|\det \mathbf{W}| - \sum_j \frac{1}{\alpha_j} \log \left[\sum_t |\mathbf{w}_j^T \mathbf{x}(t)|^{\alpha_j} \right] + \tilde{f}(\alpha_j, T) \tag{14}$$

with

$$\tilde{f}(\alpha_j, T) = \frac{1}{\alpha_j}[\log \frac{T}{\alpha_j} - 1] - \frac{1}{\alpha_j} \log C(\alpha_j) - \log Z(\alpha_j) \tag{15}$$

We see that (14) is equal to (10) expect for the additive functions f and $\tilde{f}$, which do not depend on the sample or $\mathbf{W}$, although they do depend on the parameters α_j. Simple numerical simulations show that in fact f and $\tilde{f}$ are practically equal for any reasonable α and $T \geq 100$.

2.4 Comparison with conventional likelihood

The likelihood in (10) is formally rather similar to the conventional log-likelihood of ICA, which in the case of the generalized gaussian density can be written as

$$\frac{1}{T} \log \tilde{p}(\mathbf{X}|\mathbf{A}, \boldsymbol{\alpha}) = \log|\det \mathbf{W}| - \frac{1}{T} \sum_j \sum_t |\mathbf{w}_j^T \mathbf{x}(t)|^{\alpha_j} C(\alpha_j) + \bar{f}(\alpha_j, T) \tag{16}$$

where the function $\bar{f}$ is defined as

$$\bar{f}(\alpha_j, T) = -\log Z(\alpha_j) \tag{17}$$

Thus, we see the interesting phenomenon that our new likelihood in (10) contain the logarithmic function between the two summations. If the α_j are fixed, this logarithm is the main difference between the two likelihoods, in addition to the different "weighting" factors $C(\alpha_j)$ and $1/\alpha_j$.

The new likelihood in (10) has the interesting property that it is homogeneous with respect to the rows norms of $\mathbf{W}$. That is, if we multiply the rows of $\mathbf{W}$ by any scalar factors, the likelihood is constant. This seems to be an interesting reflection of the fact that the scales of the rows of $\mathbf{W}$ cannot be determined in the generative model.

To recapitulate, we have derived an alternative likelihood, given in (10), for the linear generative model in (1). The likelihood was obtained in closed form for the case of the generalized gaussian density with parameters α_j controlling the shape of the densities.

3 Analysis of maximum likelihood estimation

Next we show how maximization of the new likelihood in (10) can perform PCA or ICA in special circumstances.

3.1 Constraint on separating matrix

Since the objective function is constant with respect to the norms of the rows $\mathbf{w}_i$, we can constrain them, purely for reasons of numerical stability, to be equal to unity. In fact, we decide to constrain the matrix $\mathbf{W}$ to be orthogonal in what follows:

$$\mathbf{W}\mathbf{W}^T = \mathbf{I}. \tag{18}$$

This is justified to the extent that such a constraint is often used in PCA and in ICA, assuming whitened data in ICA. Furthermore, the constraint allows for stronger theoretical results below. The constraint is equivalent to constraining $\mathbf{A}$ to be orthogonal.

3.2 Estimation for data modelled as Gaussian

First we show that if the data is modelled as (or assumed to be) Gaussian, maximization of the new likelihood in (10) performs PCA. This is given in the following theorem:

Theorem 1 *Fix $\alpha_j = 2$ for all j, which means modelling the latent components as Gaussian. Assume that the eigenvalues of the data (sample) covariance matrix are distinct. When the likelihood in (10) is maximized under the constraint of orthogonality of $\mathbf{A}$, the global maximum is attained when $\mathbf{A}$ contains the eigenvectors of the covariance matrix of the data $\mathbf{X}$ as its columns.*

Proof: Denote by $\mathbf{C}$ the covariance matrix of the sample. Ignoring the additive constant f for notational simplicity, The log-likelihood in (10) then becomes

$$L = -\frac{1}{2}\sum_j \log[\mathbf{w}_j^T \mathbf{C} \mathbf{w}_j] \tag{19}$$

since the log-determinant of $\mathbf{W}$ is zero. Denote by $\mathbf{C} = \mathbf{U}\mathrm{diag}(\lambda_i)\mathbf{U}^T$ the eigenvalue decomposition of the covariance matrix, and make the change of variables $\mathbf{Q} = \mathbf{W}\mathbf{U}$. Then, we have

$$L = -\frac{1}{2}\sum_j \log[\mathbf{q}_j^T \mathrm{diag}(\lambda_i)\mathbf{q}_j] = -\frac{1}{2}\sum_j \log[\sum_i q_{ij}^2 \lambda_i] = -\frac{1}{2}\sum_j \log[\sum_i b_{ij}\lambda_i] \tag{20}$$

where we denote $b_{ij} = q_{ij}^2$. Due to orthogonality of $\mathbf{W}$ and $\mathbf{U}$, the matrix $\mathbf{B}$ is doubly stochastic, which means its rows and columns have sum equal to one. Let us write $f(u) = -\frac{1}{2}\log(u)$, so we have

$$L(\mathbf{B}) = \sum_{j=1}^{n} f(\sum_i b_{ij}\lambda_i) \tag{21}$$

The function f is strictly convex. For any strictly convex function f, for any j, and for any set of distinct λ_i, we have

$$f(\sum_i b_{ij}\lambda_i) \leq \sum_i b_{ij} f(\lambda_i) \tag{22}$$

with equality if and only if exactly one of the b_{ij} is non-zero. Thus, we have

$$L(\mathbf{B}) \leq \sum_j \sum_i b_{ij} f(\lambda_i) = \sum_i f(\lambda_i) \tag{23}$$

with equality in the $\leq$ only if the b_{ij} has exactly one non-zero element for each j, which implies that $\mathbf{B}$ is a permutation matrix. Thus, we see that L is maximized when $\mathbf{B}$ is a permutation matrix. This corresponds to $\mathbf{Q}$ being a signed permutation matrix, and $\mathbf{A} = \mathbf{W}^T = \mathbf{UQ}$ thus contains the eigenvectors in $\mathbf{U}$ as its columns, and the Theorem is proven.

Note that the Theorem does not apply to the conventional ICA likelihood in (16), because in that case we would have the log-likelihood without the additional logarithm as

$$\tilde{L}(\mathbf{W}) = -\sum_j \frac{1}{2}(\mathbf{w}_j^T \mathbf{C} \mathbf{w}_j) = -\frac{1}{2}\mathrm{tr}(\mathbf{WCW}^T) = -\frac{1}{2}\mathrm{tr}(\mathbf{C}) \tag{24}$$

Thus, the conventional likelihood is constant under the assumption of Gaussianity and the constraint of orthogonality, as is well-known.

Further note that for the Theorem to hold, we do not need to assume that the data actually is Gaussian, we only need to assume that we model the data as Gaussian in the sense of setting $\alpha_j = 2$ in the estimation procedure. Of course, if the α_j are estimated from the data instead of being fixed a priori, the assumption could presumably be replaced by assuming that the data actually is Gaussian.

3.3 Estimation for data assumed to be non-Gaussian

Next, we analyse the behaviour of the new likelihood when the data follows the conventional ICA model, with non-Gaussian components; this is equivalent to our model with non-Gaussian components. Furthermore, we assume that in the estimation, we use a non-Gaussian version of the likelihood, i.e. $\alpha_j \neq 2$.

Importantly, we assume the data is whitened in contrast to the preceding section. Like in the preceding section, we constrain $\mathbf{W}$ to be orthogonal. For simplicity, we consider the case where the α_j (non-Gaussianity models) are fixed a priori, although the result is unlikely to change essentially if we estimate the α_j.

A complication in the analysis is that our log-likelihood is not smooth, while most related analysis [4, 2] assumes smooth functions. We restrict ourselves here to an approximative analysis, where we apply the existing smoothness-based analysis to our method, essentially assuming that we use a smooth approximation of our new likelihood.

Using such a smoothness approximation, the consistency of our new likelihood can be easily shown. Consider each summand in the log-likelihood, of the form $\log\left[\sum_t |\mathbf{w}_j^T\mathbf{x}(t)|^{\alpha_j}\right]$. This is a logarithm of an objective of the form $\sum_t G(\mathbf{w}_j^T\mathbf{x}(t))$. Such an objective (without logarithm) was analysed, for example, in [2], where it was shown that it reaches the maximum at the independent components, assuming the data is white, the norm of $\mathbf{w}_j$ is constrained to unity, and crucially, that a certain "non-polynomial cumulant" is positive. The non-polynomial cumulant is positive if the model of non-Gaussianity is reasonable for the data; in our case, the generalized Laplacian distribution with the given α must be a reasonable approximation of the distribution of the component. Typically, this is not very restrictive: If the data is sparse (super-Gaussian), taking $\alpha < 2$ usually makes this condition hold.

Assuming that the condition on the reasonable non-Gaussianity model holds, we can easily see that our likelihood enables estimation of the model. The likelihood is simply a sum of logarithms of functions which are each maximized at the independent components. The situation only differs from the ordinary case of the ordinary ICA likelihood in the existence of the logarithm. Since logarithm is a monotonic function, the maxima are the same, and we have thus shown that maximization of the new likelihood estimates the ICA model (under the reservations and approximations given above).

4 Conclusion

We proposed to consider the variances of components in a linear mixing model as independent parameters. This enabled a unification of PCA and ICA in the form of the likelihood in (10). In fact, it is intuitively clear that the conventional assumption of unit variance of the components in ICA makes it impossible to analyse the variances of the components. As we have shown here, the conventional assumption can be removed by considering the variances as additional parameters to be estimated, and eventually integrated out.

The unified model is primarily proposed here as an interesting theoretical framework. Future research is needed to see if it is useful in practice. The two-stage approach of first doing PCA and then ICA has been quite successful in practice, so it remains to be seen if a unified approach could be better any practical applications.

References

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

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

[3] E. Oja. A simplified neuron model as a principal component analyzer. *J. of Mathematical Biology*, 15:267–273, 1982.

[4] E. Oja. The nonlinear PCA learning rule in independent component analysis. *Neurocomputing*, 17(1):25–46, 1997.

[5] E. Oja and J. Karhunen. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of Math. Analysis and Applications*, 106:69–84, 1985.