Lesson 7

Towards the Structure

During the previous discussions, the role of the latent structure has become more and more emphasized. And, indeed, now we are taking yet another leap in that direction: It will be assumed that there really exists some underlying structure behind the observations (see Fig. 7.1)[1]. The observations x are used to determine the internal phenomena taking place within the system; the output variables are calculated only after that. Truly knowing what happens within the system no doubt helps to pinpoint the essential behavioral patterns, thus promising to enhance the accuracy of the regression model. In the earlier chapters the latent structure was just a conceptual tool for compressing the existing data, now it takes a central role in explaining the data.

As has been noticed, the methods presented this far do not offer us intuitively appealing ways to find the real structure: If simple scaling can essentially change the PCA model, for example (see (5.35), it cannot be the physical structure that is being revealed. On the other hand, somehow the idea of continuity between the methods (as utilized in CR) does not promise that a uniquely correct structure would be found. The mathematically motivated structure is not necessarily physically meaningful.

It is an undeniable truth that the underlying primary structure cannot be determined when only observations of the behavior are available. We can only make optimistic guesses — if we trust the benevolence of Nature these guesses are perhaps not all incorrect. However, remember Thomas Aquinas and his theories of "First Cause":

> "... And so we must reach a First Mover which is not moved by anything; and this all men think of as God."

[1] Note that the causal structure is now assumedly different as it was before: If both X and Y are only reflections of some internal system structure, so that no causal dependence is assumed between them, the applications of the final models should also recognize this fact. This means that control applications are somewhat questionable: If x values are altered in order to affect the y values according to the correlations as revealed by the model, it may be that the intended effects are not reached. On the other hand, different kinds of *soft sensor* applications are quite all right: The observed correlations justify us to make assumptions about y variables when only x has been observed (assuming invariant process conditions)

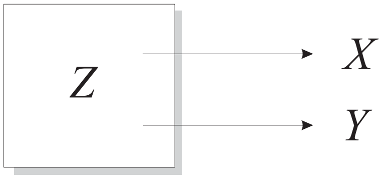

Figure 7.1: Yet another view of signal dependency structure

7.1 Factor analysis

An age-old method for feature extraction, or finding the underlying explanations beyond the observations is *Factor Analysis.* It has been applied widely in social sciences, etc. The basic model is familiar:

$$x(\kappa) = \theta z(\kappa), \tag{7.1}$$

or

$$X = Z\theta^T. \tag{7.2}$$

The goal is to find the basis θ and the scores Z (factors) so that the residual errors $E = X - Z\theta^T$ would be minimized. Nothing strange here — actually the goal sounds identical with the PCA problem formulation. However, now we have an additional *uncorrelatedness* constraint for the residual: The residual errors E_i should be uncorrelated[2]:

$$\mathrm{E}\{e(\kappa)e^T(\kappa)\} = \frac{1}{k} \cdot E^T E = \begin{pmatrix} \mathrm{var}\{e_1(\kappa)\} & & 0 \\ & \ddots & \\ 0 & & \mathrm{var}\{e_n(\kappa)\} \end{pmatrix}. \tag{7.3}$$

All dependencies between data should be explained by the factors alone. Assuming that the residual errors and factors are uncorrelated, the data covariance matrix can be written as

$$\begin{array}{rcl} \frac{1}{k} \cdot X^T X & = & \frac{1}{k} \cdot \left(\theta Z^T Z \theta^T + \theta Z^T E + E^T Z \theta^T + E^T E\right) \\ & & \frac{1}{k} \cdot \theta Z^T Z \theta^T + \frac{1}{k} \cdot E^T E. \end{array} \tag{7.4}$$

From this it follows that, if one defines

$$\begin{array}{rcl} \theta' & = & \theta M \\ Z'^T Z' & = & M^{-1} Z^T Z (M^T)^{-1}, \end{array} \tag{7.5}$$

the same residual errors are received for different factor structure; the new model is also equally valid factor model as the original one was for any invertible matrix M. This means that the results are not unique. Factor analysis is more like art than science; there are more or less heuristic basis *rotations* that can be applied to enhance the model. These algorithms will not be studied here.

[2] Note that this uncorrelatedness property is *not* fulfilled by the PCA basis

Note that the uniqueness of the PCA model (at least if the eigenvalues are distinct) is caused by the assumed ordering of the basis vectors according to their relevance in terms of explained variance; in the factor analysis model, this kind of ordering is not assumed and uniqueness is not reached in the same manner. As long as the rotations just operate in the same subspace, the selection of the factors does not affect the accuracy if regression model is to be implemented.

7.2 Independent components

Above, factor analysis tried to find the *original sources* by emphasizing uncorrelatedness — but the results were not quite satisfactory, uniqueness of the results remaining lost. Could we define more restrictive objectives that would fix the problems of traditional factor analysis? The key question here, again, is that of *ontological assumptions:* Just as in the case of information vs. noise (chapter 5), now one has to determine how the structure is manifested in the data.

And, indeed, the answer to the question whether structure can be characterized in a reasonable way or not is *yes:* During the last decade, it has turned out that the *independence* of sources is a good starting point. This approach is called *Independent Component Analysis (ICA),* and it has lately been studied specially in the neural networks community. It has been successfully applied for blind source separation, image coding, etc. (see [16], [28]).

7.2.1 Why independence?

Intuituively, the original sources are those that are independent of other sources. Finding the underlying structure can be based on this idea: Search for data that is maximally independent. In mathematical terms, two variables x_1 and x_2 can be said to be independent if there holds[3]

$$\mathrm{E}\{f_1(x_1(\kappa))f_2(x_2(\kappa)\} = \mathrm{E}\{f_1(x_1(\kappa))\} \cdot \mathrm{E}\{f_2(x_2(\kappa)\}. \tag{7.6}$$

According to the above formulation, it can be said that maximizing independence between signals simultaneously minimizes the *mutual information* between them.

In a way, the idea of ICA is to *invert the central limit theorem:* When various independent variables are mixed, the net distribution more or less approximates normal distribution. So, when searching for the original, unmixed signals, one can search for *maximally non-normal* projections of the data distribution!

7.2.2 Measures for independence

Probability distributions can uniquely be determined in terms of *moments* or *cumulants.* Gaussian distribution is determined by the first order cumulant

[3] Note that independence is much more than simple uncorrelatedness, where the formula (7.6) holds only when both of the functions are identities, $f_1(x_1) = x_1$ and $f_2(x_2) = x_2$. Because independence is so much more restricting condition than what uncorrelatedness is, one is capable of finding more unique solutions than what is the case with traditional factor analysis

(mean value) and the second order cumulant (variance) alone; for this distribution, all higher order cumulants vanish. This means that the "non-normality" of a distribution can be measured (in some sense) by selecting *any* of the higher order cumulants; the farther this cumulant value is from zero (in positive or negative direction), the more the distribution differs from Gaussian. For example, non-normality in the sense of "non-symmetricity" can be measured using the third-order cumulant *skewness.* In ICA, the standard selection is the fourth-order cumulant called *kurtosis* that measures the "peakedness" of the distribution:

$$\mathrm{kurt}\{x_i(\kappa)\} = E\{x_i^4(\kappa)\} - 3 \cdot E^2\{x_i^2(\kappa)\}. \tag{7.7}$$

For normalized data this becomes

$$\mathrm{kurt}\{x_i(\kappa)\} = E\{x_i^4(\kappa)\} - 3. \tag{7.8}$$

If the data is appropriately normalized, the essence of kurtosis is captured in the fourth power properties of the data; this fact will be utilized later.

After the ICA basis has been determined somehow, regression based on the independent components can be implemented (this method could be called "ICR"). Note that the expressions are somewhat involved because the basis vectors are non-orthogonal.

7.2.3 ICA vs. PCA

Figs. 7.3 and 7.2 illustrate the difference between the principal components and the independent components in a two-dimensional case. The data is assumed to have uniform distribution within the diamond-shaped region, and in these figures, ICA and PCA bases for this data are shown, respectively. It really seems that independence means non-Gaussianity: Note that the trapetzoidal marginal distributions in the non-independent PCA case are much more Gaussian than the "flat", negatively kurtotic uniform distributions in the ICA case. The "mixing matrix" (using the ICA terminology) in the case of Fig. 7.3 is

$$\theta = \begin{pmatrix} 1/\sqrt{2} & 1 \\ 1/\sqrt{2} & 0 \end{pmatrix}, \tag{7.9}$$

meaning that $x = \theta z$. Note that, as compared to the Gaussian distribution, uniform distribution is rather "flat"; in this case the kurtosis is maximally negative in the directions of the original sources, other projections having smoother, more Gaussian distributions.

7.3 Eigenproblem-oriented ICA algorithms

Normally independent component analysis is carried out in an algorithmic, iterative framework [16]; there are good reasons for this, but in this context we would like to bring ICA into the same eigenproblem-oriented framework as all the other approaches before.

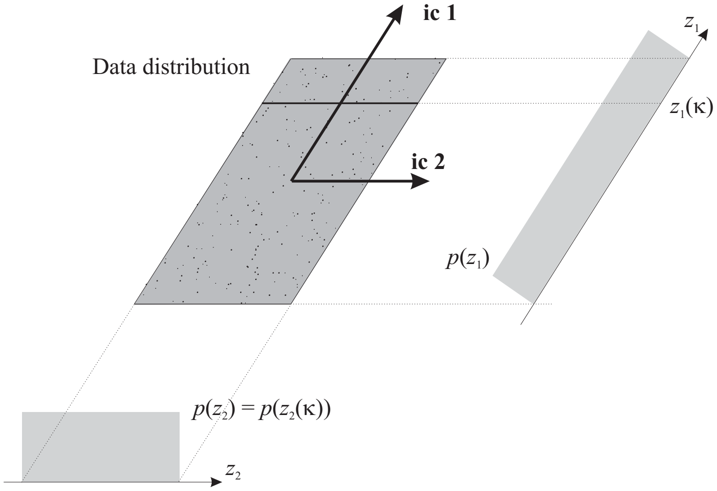

Figure 7.2: The ICA basis vectors, or "independent components". Knowing the value of $z_1(\kappa)$, say, nothing about the value of $z_2(\kappa)$ can be said. The distribution remains intact, or $p(z_2(\kappa)) = p(z_2(\kappa)|z_1(\kappa))$, and the two projected variables really are independent (compare to the PCA case below: information about $z_1(\kappa)$ affects the posteriori probabilities of $z_2(\kappa)$)

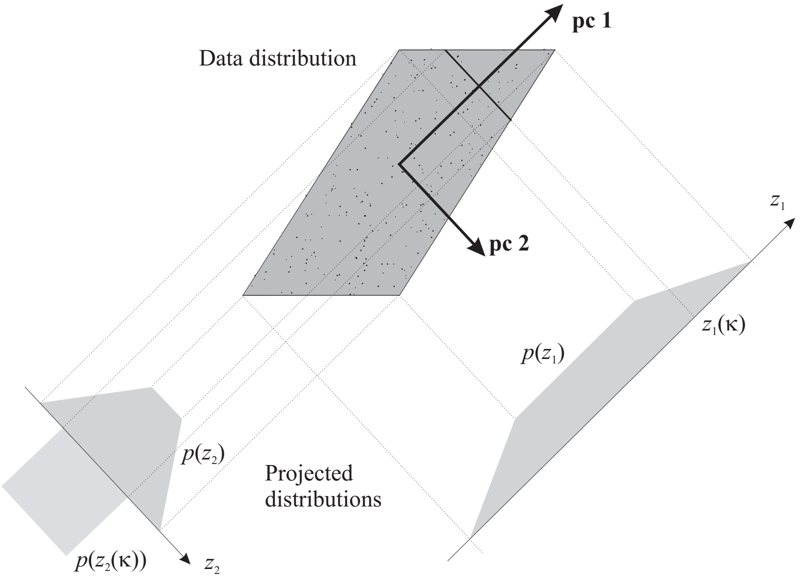

Figure 7.3: The PCA basis vectors, or "principal components": the first of them captures the maximum variance direction, and the other one is perpendicular to it. Variables are not independent

In what follows, kurtosis (or, equally, the fourth moment of data, as shown in (7.8)) as a measure of independence is concentrated on (even though other *contrast functions* can also be defined). The problem with the eigenproblem framework is that it naturally emerges only when the second-order data properties, covariances and correlations, are studied. It is now asked whether the higher-order statistical properties like kurtosis could somehow be captured in the same way. And, indeed, the *tensor methods* for ICA have been found[4]. In principle, the tensors are linear operators just as normal matrices are, and the eigenstructure can be defined also for the four-dimensional tensors; however, the procedures are computationally involved, tensors consisting of $n \cdot n \cdot n \cdot n$ elements, and also the mathematical theory is cumbersome (the "eigenvectors" now being $n \times n$ matrices!). Here the excessive growth of search space (and the sophisticated mathematics) is avoided and some alternative approaches are studied.

7.3.1 Data whitening

The key point is to modify the data distribution so that the structural features — as assumedly being revealed by the fourth-order properties — become visible. To reach this, the lower-order properties have to be compensated, because they typically outweight the higher-order properties:

- First-order properties are eliminated by only studying mean-centered data, that is, $\mathrm{E}\{x_i(\kappa)\} = 0$ for all i;
- Third-order properties (or "skewness") vanish if one *assumes* that the distributions are symmetric, so that $\mathrm{E}\{x_i(\kappa)x_j(\kappa)x_l(\kappa)\} = 0$ for all i, j, l; and
- Second-order properties are eliminated if the data is *whitened.*

The data whitening means that the data is preprocessed so that its covariance matrix becomes an identity matrix. This can be accomplished by

$$x(\kappa) = \left(\sqrt{\mathrm{E}\{\mathbf{x}(\kappa)\mathbf{x}^T(\kappa)\}}\right)^{-1} \cdot \mathbf{x}(\kappa), \tag{7.10}$$

where the square root of a matrix is here defined so that $M = \sqrt{M}^T\sqrt{M}$. After this modification there holds $\mathrm{E}\{x(\kappa)x^T(\kappa)\} = I$. No matter what kind of additional preprocessing is needed, the above elimination of lower-order statistics is assumed in what follows[5].

We are again searching for a basis θ so that $x(\kappa) = \theta z(\kappa)$, signals z_i now hopefully being independent; and, again, we assume that in the whitened data space

[4] Note that the first-order statistical properties of a distribution are captured by the one-dimensional mean value vector, and the second-order properties are captured by the two-dimensional covariance matrix — similarly, the fourth-order properties can be captured by the four-dimensional *tensor*

[5] If one is capable of finding some structure in the data after this prewhitening, this structure cannot be dependent of the measurement scaling, thus reflecting the *real* structure in a more plausible way — this dependency of the scaling was one of the arguments against the PCA model

the basis is orthogonal (of course, when expressed in the original coordinates, the orthogonality does not hold — see Fig. 7.2).

7.3.2 Deformation of the distribution

One way to reduce the fourth-order properties to second-order properties is to explicitly change the distribution. In *Fourth-Order Blind Identification (FOBI)* the data is preprocessed (after first being whitened) so that the samples are either stretched or contracted about the origin. This can be accomplished as

$$x'(\kappa) = f(\|x(\kappa)\|) \cdot x(\kappa), \tag{7.11}$$

where f is some function. For example, selecting $f(\|x\|) = \|x\|$ means that analyzing the variance properties of x' the fourth order properties of the original x are modeled. This can be seen when the new covariance matrix is studied:

$$\begin{aligned} \mathrm{E}\{x'(\kappa)x'^T(\kappa)\} &= \mathrm{E}\{x(\kappa)x^T(\kappa) \cdot \|x(\kappa)\|^2\} \\ &= \mathrm{E}\{\Theta z(\kappa)z^T(\kappa)\Theta^T \cdot z^T(\kappa)\Theta^T\Theta z(\kappa)\} \\ &= \Theta \cdot \mathrm{E}\{z(\kappa)z^T(\kappa) \cdot z^T(\kappa)z(\kappa)\} \cdot \Theta^T. \end{aligned} \tag{7.12}$$

This formulation is justified because one assumes that there exists an orthogonal basis Θ and independent signals z_i. Let us study the matrix $\mathrm{E}\{z(\kappa)z^T(\kappa) \cdot z^T(\kappa)z(\kappa)\}$ closer. The element i, j has the form

$$\begin{aligned} &\mathrm{E}\{z_i(\kappa)z_j(\kappa) \cdot z^T(\kappa)z(\kappa)\} \\ &\quad = \mathrm{E}\{z_i(\kappa)z_j(\kappa) \cdot (z_1^2(\kappa) + \cdots + z_n^2(\kappa))\} \\ &\quad = \mathrm{E}\{z_i(\kappa)z_j(\kappa) \cdot (z_1^2(\kappa)\} + \cdots + \mathrm{E}\{z_i(\kappa)z_j(\kappa)(z_n^2(\kappa)\} \\ &\quad = \begin{cases} \mathrm{E}\{z_i^4(\kappa)\} + \mathrm{E}\{z_i^2(\kappa)\} \cdot \sum_{l \neq i} \mathrm{E}\{z_l^2(\kappa)\} = \mathrm{E}\{z_i^4(\kappa)\} + n - 1, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } i = j \text{, and} \\ \mathrm{E}\{z_i^3(\kappa)z_j(\kappa)\} + \mathrm{E}\{z_i(\kappa)z_j^3(\kappa)\} + \\ \qquad \mathrm{E}\{z_i(\kappa)z_j(\kappa)\} \cdot \sum_{l \neq i, l \neq j} \mathrm{E}\{z_l^2(\kappa)\} = 0, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{otherwise.} \end{cases} \end{aligned}$$

The above simplifications are justified because of the assumed independence of the signals z_i — for example, $\mathrm{E}\{z_i^\xi(\kappa)z_j^\zeta(\kappa)\} = \mathrm{E}\{z_i^\xi(\kappa)\} \cdot \mathrm{E}\{z_j^\zeta(\kappa)\}$ for $i \neq j$. Also, because of centering, $\mathrm{E}\{z_i(\kappa)\} = 0$, and because of whitening, $\mathrm{E}\{z_i^2(\kappa)\} = 1$. Additionally, taking into account the assumed orthogonality of Θ (in the whitened data space), there holds $\Theta^T = \Theta^{-1}$, and

$$\begin{aligned} &\mathrm{E}\{x'(\kappa)x'^T(\kappa)\} \\ &\quad = \Theta \cdot \begin{pmatrix} \mathrm{E}\{z_1^4(\kappa)\} + n - 1 & & 0 \\ & \ddots & \\ 0 & & \mathrm{E}\{z_n^4(\kappa)\} + n - 1 \end{pmatrix} \cdot \Theta^T \\ &\quad = \Theta \cdot \Lambda \cdot \Theta^{-1}. \end{aligned} \tag{7.13}$$

This means that the right hand side can be interpreted as the eigenvalue decomposition of the covariance matrix of the modified data. The diagonal elements in the eigenvalue matrix are directly related to the fourth-order properties of the (assumed) independent components. The cumulant maximization/minimization task (for whitened data) is also transformed into the variance

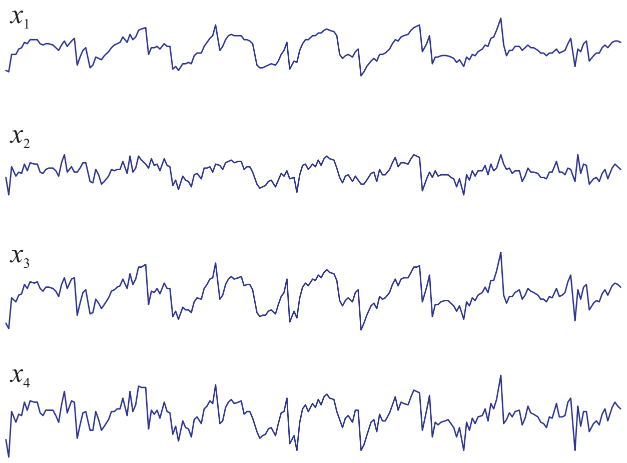

Figure 7.4: The original mixture of signals (200 of the original 1000 samples are shown as time series). Can you see any structure here?

maximization/minimization task for the modified variable[6]. This means that the standard PCA approach can be applied; simultaneously as the covariance structure of x' is analyzed, so is the kurtosis structure of the original variables x. However, contrary to the standard PCA, now the principal components carrying the least of the variance may be equally interesting as the first ones are — depending on whether one is searching for the latent basis of maximal or minimal kurtosis. The eigenvalues reveal the kurtoses of the signals z_i so that $\mathrm{kurt}\{z_i(\kappa)\} = \mathrm{E}\{z_i^4(\kappa)\} - 3 = \lambda_i - n - 2$.

As an example, study the four-dimensional data samples as shown in Fig. 7.4. Here, the data sequence is interpreted as constituting a continuous signal; however, note that this signal interpretation is only for visualization purposes. Using the above scheme, the underlying signals can be extracted — with no additional information, just based on the statistical properties of the samples (see Fig. 7.5)!

The exclusively input-oriented approach for determining the latent structure can again be extended: Note that the regression structure $y = F^T x$ remains formally intact if both sides are multiplied by the same factor, so that there holds $yf(x,y) = F^T x f(x,y)$. This means that extensions towards the directions of PLS and CCR, for example, can be proposed where both input and output data are preprocessed prior to determination of the latent structure; it seems that such possibilities have never been explored.

It is important to note that the curve continuity and periodicity, properties that are intuitively used as criteria for "interesting" signals, are not at all utilized by the ICA algorithms — indeed, the samples could be freely rearranged, the continuity and periodicity vanishing, but the analysis results would still remain the same. In fact, the traditional methods like some kind of harmonic analysis could reveal the underlying periodic signal structure, but ICA is specially

[6] Note that if the signals z_i are independent, kurtosis can be maximized/minimized using this algorithm even if the distributions are skewed, that is, $\mathrm{E}\{z_i^3(\kappa)\} \neq 0$

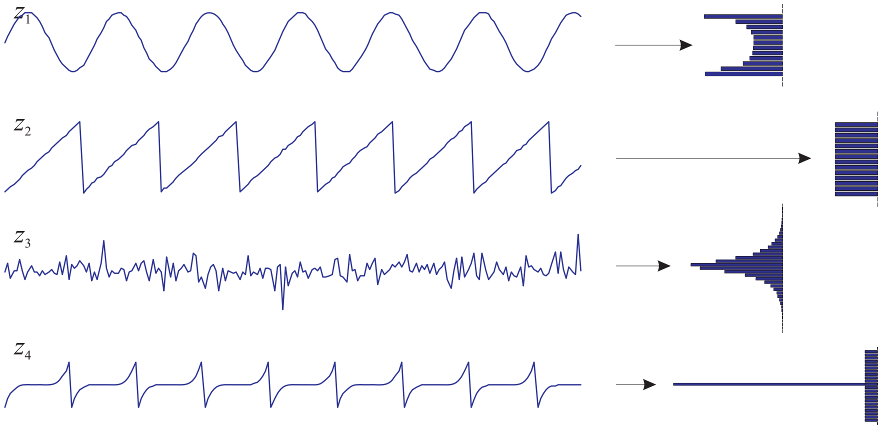

Figure 7.5: The extracted sources and their distributions

powerful when such periodicity or continuity properties cannot be assumed.

However, even though the above approach seems promising, it has to be recognized that in many cases the distribution properties are only visible on the local scale, they cannot be attacked applying global methods like PCA. For example, see Fig. 7.6: there it is shown how the peculiar data distribution is deformed in the data processing. The key observation here is that even after the data deformation (last image), the covariance properties remain identical in orthogonal directions, meaning that none of the directions can be selected by the PCA-based approaches. Typically, algorithmic approaches to ICA are superior, because locally there still exist gradients in the kurtosis-oriented design criterion.

7.3.3 Further explorations*

One of the disadvantages of the above algorithm is that it cannot distinguish between independent components that have equal kurtosis[7]. Let us try to find another approach offering more flexibility.

First, study the basic properties of the fourth power of the data point norm:

$$\begin{aligned} \|x\|^4 &= \left(\sqrt{x_1^2+\cdots+x_n^2}\right)^4 \\ &= \left(x_1^2+\cdots+x_n^2\right)^2 \\ &= x_1^4+\cdots+x_n^4+2x_1^2x_2^2+2x_1^2x_3^2+\cdots+2x_{n-1}^2x_n^2. \end{aligned} \qquad (7.14)$$

[7] Note that the non-uniqueness problem is the same with PCA if there are equal eigenvalues; however, in this case when we are searching for the real explanations beneath the observations, not only some compression of information, this problem is more acute

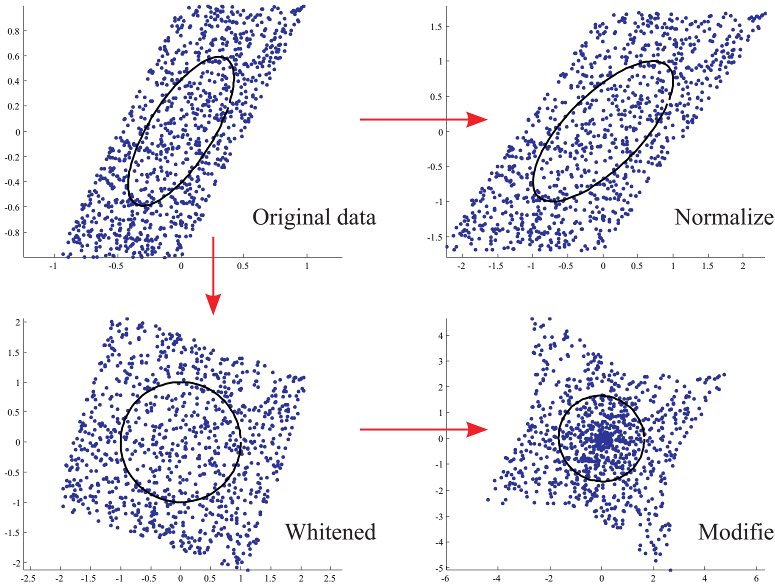

Figure 7.6: Modifications of the data distribution. The covariance structures are shown as ellipses (circles) in the figures

Let us define a modified $(n^2 + n)/2$ dimensional data vector as follows:

$$x' = \begin{pmatrix} x_1^2 \\ \vdots \\ x_n^2 \\ \sqrt{2} \cdot x_1 x_2 \\ \vdots \\ \sqrt{2} \cdot x_{n-1} x_n \end{pmatrix}, \tag{7.15}$$

containing all possible second order cross-products between the x elements. Using this kind of modified data vector one can express the fourth moment of the original data vector simply as

$$\|x\|^4 = \|x'\|^2 = x'^T \cdot x'. \tag{7.16}$$

Now in the x' space one can project the point onto an axis l as $x'^T l$, and, further, it is possible to express the fourth moment of this projected data as

$$l^T \cdot x' x'^T \cdot l. \tag{7.17}$$

One should find such an axis l that the average of this quantity over all the modified samples $x'(\kappa)$, where $1 \leq \kappa \leq k$, would be maximized (or minimized). First, construct the expression for the average of projected fourth moment values:

$$\frac{1}{k} \cdot l^T \cdot \sum_{\kappa=1}^{k} x'(\kappa) x^T(\kappa) \cdot l = \frac{1}{k} \cdot l^T \cdot X'^T X' \cdot l, \tag{7.18}$$

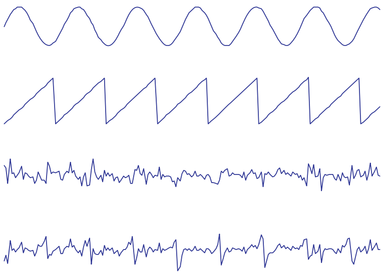

Figure 7.7: The basis vectors corresponding to l of lowest kurtosis. Note that only first two are “correct” signals, these sources being sub-Gaussian

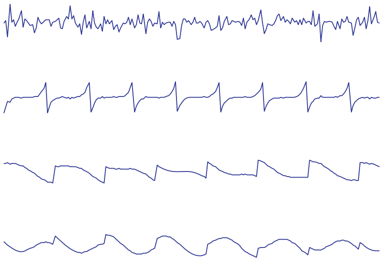

Figure 7.8: The basis vectors corresponding to l of highest kurtosis. Only the first two are correct, these sources being super-Gaussian

where the modified data vectors are written in the matrix form. Letting $\|l\| = 1$, Lagrangian constrained optimization problem results:

$$J(l) = \frac{1}{k} \cdot l^T \cdot X'^T X' \cdot l + \lambda \cdot \left(1 - l^T l\right), \tag{7.19}$$

so that

$$\frac{d\,J(l)}{d\,l} = \frac{2}{k} \cdot X'^T X' \cdot l - 2\lambda \cdot l = 0, \tag{7.20}$$

again resulting in an eigenproblem:

$$\frac{1}{k} \cdot X'^T X' \cdot l = \lambda \cdot l. \tag{7.21}$$

Substituting (7.21) in (7.19) one can see that the eigenvalue equals the cost criterion value, that is, λ is the average of the projected fourth moments of the samples. Note that here the least significant eigenvector can be more important than the most significant one, depending whether one is searching for sub-Gaussian or super-Gaussian distribution. This principal component is now presented in the high-dimensional x' space, and to make it useful as a basis vector, one needs to approximate it in the lower-dimensional space of x vectors. For this purpose, remember what is the interpretation of each of the elements in l:

$$\left\{ \begin{array}{l} l_1 \sim x_1^2 \\ \quad\vdots \\ l_n \sim x_n^2 \\ l_{n+1} \sim \sqrt{2} x_1 x_2 \\ \quad\vdots \\ l_{(n^2+n)/2} \sim \sqrt{2} x_{n-1} x_n, \end{array} \right. \qquad \rightarrow \qquad \left\{ \begin{array}{l} x_1^2 \sim l_1 \\ \quad\vdots \\ x_n^2 \sim l_n \\ x_1 x_2 = x_2 x_1 \sim \frac{1}{\sqrt{2}} \cdot l_{n+1} \\ \quad\vdots \\ x_{n-1} x_n = x_n x_{n-1} \sim \frac{1}{\sqrt{2}} \cdot l_{(n^2+n)/2}, \end{array} \right.$$

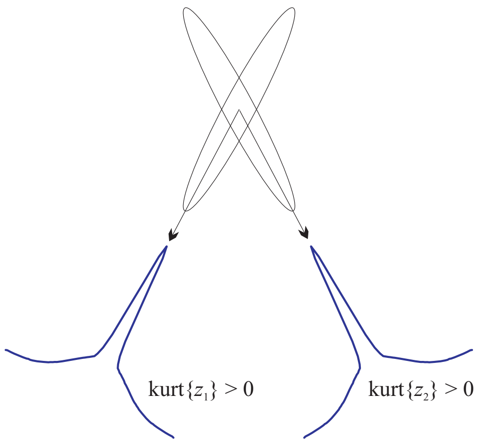

Figure 7.9: Another (more difficult) example of underlying structure: Positive kurtosis in the basis that is suggested by the structure of the distribution (on the left), and zero kurtosis using PCA (two equal projections of normal distributions summed together). In this case, for example, the PCA analysis hides the underlying structure altogether — all samples belonging to different distribution regions are mixed up. However, for this distribution the independence assumption also collapses (see text)

These are not expectation values, but they still tell something about the connections between the variables for some hypothetical data; from the elements of l one can construct an association matrix

$$R = \begin{pmatrix} l_1 & \frac{1}{\sqrt{2}} \cdot l_{n+1} & \cdots & \frac{1}{\sqrt{2}} \cdot l_{2n-1} \\ \frac{1}{\sqrt{2}} \cdot l_{n+1} & l_2 & & \\ \vdots & & \ddots & \\ \frac{1}{\sqrt{2}} \cdot l_{2n-1} & & & l_n \end{pmatrix}. \tag{7.22}$$

Using this matrix, one can determine the n dimensional basis vectors θ_i that best can span the higher-dimensional space; the answers must be given by the principal components of R. Note that the eigenvalues may now be negative, as well as the diagonal elements; this could be explained assuming that data is *complex-valued.* However, because the matrix is symmetric (in this case, actually, Hermitian) the eigenvalues and vectors are real-valued.

7.4 Beyond independence

Study the distributions in Fig. 7.9: The intuitively correct basis vectors fulfill the non-Gaussianity goal, the marginal distributions being peaked, or positively kurtotic. However, note that the variables z_1 and z_2 are in this case *not independent:* knowing, for example, that z_1 has high value, one immediately knows that z_2 must be near zero, whereas low values of z_1 leave much more freedom for z_2; in a way, these variables are rather *mutually exclusive* than indepen-

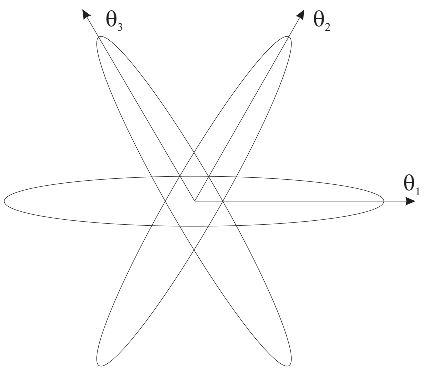

Figure 7.10: Three basis vectors in a two-dimensional space: The basis set is *overcomplete*

dent, either of them telling very much about the other one. The independence objective does not seem to always work when searching for good basis vectors. What kind of alternatives do exist?

7.4.1 Sparse coding

It turns out that in those types of distributions that seem to be characteristic to measurement data and that we are specially interested in, meaning mixture models as studied in Sec. 2.4, this mutual exclusiveness is more like a rule rather than exception: If a sample belongs to some specific subdistribution, the other subdistributions do have no role in explaining it. And there are more surprises: It may be so that the correct number of latent structures is higher than what is the dimension of the data space (see Fig. 7.10). Perhaps it is this *exclusiveness* that could be taken as starting point? And, indeed, this approach results in a framework that could be called *Sparse Component Analysis (SCA).*

In sparse coding it is assumed that a sample x is represented in latent basis so that most of the scores are zeros. Sparse coding is a rather general framework: For example, the various submodels constituting a mixture model can be presented within the same sparse structure. But sparse models are *more general* than the mixture models are: Whereas the constructs in mixture models strictly belong to one submodel only, in the sparse framework the *components may be shared,* so that the submodels can have common substructures. This exchange of substructures is the key to the expressional power of sparse models. Unfortunately, this power also suggests that there exist no explicit algortihms for constructing sparse models[8]. Also the *varimax, quartimax,* and *infomax* rotation algorithms resemble sparse coding; these approaches are commonly used within the factor analysis community for maximizing the *score variance,* thus distributing the activity in more specialized factors).

As compared to the modeling methods discussed before, ICA is typically not seen as a compression technique; rather, it carries out data reorganization, so that the z vectors often do have the same dimension as x. In the sparse coding

[8] However, various iterative approaches exist; for example, see next chapter

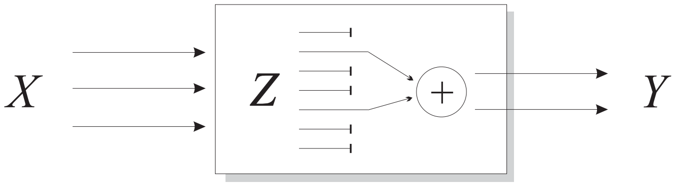

Figure 7.11: The structure of a sparse model

case, the goal is still different: The model may even be *inflated,* so that the number of constructs is higher than the data dimension, hoping that the structural phenomena become better visible when the data is less compactly packed.

One of the implicit general objectives in modeling is *simplicity,* in the spirit of *Occam's razor.* Now this simplicity objective has to be interpreted in another way: Usually, "model of minimum size" means minimum number of substructures; in sparse coding it means minimum number of simultaneously *active* units (see Fig. 7.11).

Regression based on a sparse model is nonlinear; however, the nonlinearity is concentrated on the selection of appropriate latent vectors among the candidates — after they are selected, the model *is* linear. The latent variables can be selected so that together they can explain the data sample as well as possible. The abrupt switching between latent structures means that the model behavior is discontinuous if no additional measures are applied.

When the most specialized constructs are only used, it seems that sparse representations often seem to be "easily interpreted", being sometimes connected to intuitive mental (subsymbolic) constructs. There is some evidence that the human brain organizes at least sensory information in this way: In visual cortex, there are areas, groups of neurons that have specialized in very narrow tasks, like detecting tilted lines in the visual image. The observed image is mentally reconstructed using the low-level visual features — and what is more, it seems that similar principles may be governing the higher level perception, too. There is perhaps room for fruitful cooperation between cognitive science and multivariate statistics.

Computer exercises

1. Study the robustness of the eigenproblem-formulated ICA by iteratively running the two commands below with different selections of parameter `alpha` (here α denotes the power used in data preprocessing: $x' = \|x\|^{\alpha} \cdot x$. Note that the default value $\alpha = 1$ resulting in the nominal strictly kurtosis-oriented algorithm is *not* necessarily the best choice — for example, try $\alpha = -1$ for this data):

   ```
   X1 = dataIndep;
   regrICA(X1,alpha);
   ```

   Define data as

   ```
   X2 = dataIndep(1000,...
         'randn(1000,1)',...
         'sign(randn(1000,1)).*(rand(1000,1)<1/3)');
   regrICA(X2);
   ```

   and analyze the independent components. Change the threshold value (the peak probability; value "1/3" above) between 0 and 1, and explain the results.

2. Download the `FastICA Toolbox` for `Matlab` through the Internet address `http://www.cis.hut.fi/projects/ica/fastica/`, and install it. Apply the FastICA algorithm to the above data sets.