Blending Independent Components and Principal Components Analysis

[Nematrian website page: IndependentComponentsAnalysis, © Nematrian 2015]

Abstract

In these pages we describe the similarities and differences between independent components analysis and principal components analysis, particularly as applied to portfolio risk model construction. We describe how the best elements of each approach can be blended together to enhance the ability of portfolio risk models to handle fat-tailed behaviour.

Contents

1. Introduction and Conclusions
2. Independent Components Analysis (ICA)
3. Principal components analysis (PCA)
4. Blending together PCA and ICA in the construction of risk models

Tools
References

1. Introduction and Summary

[IndependentComponentsAnalysis1]

1.1 A widely held view across the financial services industry and within relevant academic circles is that financial markets exhibit 'fat-tails', i.e. extreme events occur more frequently than we might expect were market behaviour to be modelled in line with the normal (i.e. Gaussian) distributions often assumed in more straightforward academic texts. Despite this, commercial risk systems often still include assumptions of normality within their underlying formulations. This is partly because such assumptions often make the mathematical formulation of the underlying risk model framework more analytically tractable. It is also partly because the risk system providers may argue that the impact of deviation from normality is insufficient (or the evidence for such deviation is insufficiently compelling) to justify the necessary refinements to that part of the underlying risk model formulation. In short, there is a trade-off between, on the one hand, complexity, practicality and presumed model 'correctness' and, on the other hand, simplicity, analytical tractability and practical implementability.

1.2 In these pages, we explore two methodologies that are or ought to be relevant to the design of portfolio risk systems. One of these, *principal components analysis*, see Section 3, is a well-known and common tool used in the creation and validation of current portfolio risk system designs. It allows the risk system designer to identify potential 'factors' that best seem to explain individual stock variability. Underlying its applicability to portfolio risk model design is an implicit assumption of normality (or to be more precise an investor indifference to fat-tailed behaviour, i.e. a lack of need to include in the model a characterisation of the extent to which behaviour is non-normal).

1.3 The other is *independent components analysis*. It, and several variants motivated by the same underlying rationale, are described in Section 2. Given that it is a less well-known technique within the financial community we introduce it first, and when doing so we explore it in greater depth than the better-known principal components technique. Independent

components analysis has perhaps more commonly been applied to other signal extraction problems, e.g. image or voice recognition or differentiating between mobile phone signals. In its normal formulation it seeks to extract 'meaningful' signals however weak or strong these signals might be (with the remaining 'noise' discarded). 'Meaningful' here might be equated with extent of non-normality of behaviour, as is explicitly done in certain formulations of independent components analysis. In contrast, in the portfolio risk management context 'meaningfulness' also needs to be coupled with 'significance' (i.e. 'magnitude') for the source in question to be worth incorporating in the risk model formulation. Moreover, 'noise' is not to be discarded merely because it does not appear to be 'meaningful', because any such variability adds to portfolio risk.

1.4 Despite these apparent differences, we show in Section 4 that both of these techniques can be thought of as examples of a more generic approach in which we grade possible sources of observed behaviour according to some specified importance criterion. We show how we can use this insight to blend together the two techniques to enhance the portfolio risk model design. In particular, we can choose a blend that ought to cater better for fat-tails, i.e. extreme events, than more traditional risk model designs. However, we also highlight some of the challenges that arise when trying to cater in such a prescription for the time-varying nature of the world.

2. Independent components analysis (ICA)

[IndependentComponentsAnalysis2]

Section contents

2.1 What does ICA aim to achieve?
2.2 Linear combination versus distribution mixtures
2.3 The underlying rationale for ICA
2.4 Signal mixtures
2.5 Projection pursuit
2.6 Infomax and maximum likelihood Independent component analysis
2.7 Complexity pursuit
2.8 Gradient Ascent

2.1 What does ICA aim to achieve?

[IndependentComponentsAnalysis2a]

Independent component analysis (ICA) is a tool for extracting useful information from a large amount of data. It seeks to identify the driving forces that underlie a set of observed phenomena. The phenomena to which ICA could be applied are very wide ranging, including mobile phone signals, stock price returns, brain imaging or voice recognition. It belongs to a class of *blind source separation* (BSS) methods for separating data into underlying informational elements, where 'blind' here means that such methods endeavour to separate data into source signals even if very little is known about the nature of the source signals.

Generally speaking, ICA is applied in a situation where there are several distinct 'output' signals able to be measured, and it is reasonable to postulate that these outputs depend on combinations of distinct underlying 'input' signals or factors. The aim is, as far as possible, to estimate or characterise these underlying signals. For example, with voice recognition, the aim might be to differentiate

between different foreground contributors to the overall sound pattern and, additionally, to filter out, as far as possible, anything that appears to be background noise.

2.2 Linear combination versus distribution mixtures

[IndependentComponentsAnalysis2b]

Typically ICA assumes that the observed signals are a specific type of *mixture* of the underlying factors, namely a *linear combination* of these factors. The *mixing coefficients* applicable to this sort of mixture are the multipliers applied to each factor in the computation of the relevant output signal.

It is worth noting that the term 'mixture' has a variety of meanings in statistics. Here it is being used to mean that the output signals, $y_{i,t}$, are derived from the input signals, $x_{j,t}$ using the following formula (where t indexes the observed values of a given signal, and so might correspond to points in time or space, and j indexes the signals themselves):

$$y_{i,t} = \sum_j w_{i,j} x_{j,t}$$

The term 'mixture' can also be used to describe a probability distribution for a random variable that is sequentially drawn with given probabilities from one of a number of different underlying distributions. The 'mixing coefficients' for such *distribution* mixtures relate to the probability of the given draw coming from the relevant underlying probability distribution. Such distribution mixtures can behave quite differently to the sorts of linear combination mixtures described above. We will explore the relevance of this latter type of mixture to portfolio risk management in Section 4.3.

2.3 The underlying rationale for ICA

[IndependentComponentsAnalysis2c]

ICA is based on the generic yet often physically realistic assumption that if different input signals are coming from different underlying physical processes then these input signals will be largely independent of each other. ICA aims to identify how to decompose output signals into (linear combination) mixtures of different input signals that are as independent as possible of each other. Several variants exist, which we also describe below, where 'independent' is replaced by an alternative statistical property that we might also expect might differentiate between input signals.

ICA is related to more traditional methods of analysing large data sets believed to involve linear combinations of underlying factors, such as *principal components analysis* (PCA) and *factor analysis* (FA). However, it arguably differs from them in important ways. ICA seeks to find a set of *independent* source signals. In contrast, PCA and FA seek to find a set of signals which are merely *uncorrelated* with each other. By uncorrelated we mean that the correlation coefficients between the different supposed input signals are zero. Lack of correlation is a potentially much weaker property than independence. Independence implies a lack of correlation, but lack of correlation does not imply independence. The correlation coefficient in effect 'averages' the correlation across the entire distributional form. For example, two signals might be strongly positively correlated in one tail, strongly negatively correlated in another tail, and show little correspondence in the middle of the distribution. The correlation between them, as measured by their correlation coefficient, might thus be zero, but it would be wrong then to conclude that the behaviour of the two signals were independent of each other (particularly, in this instance, in the tails of the distributional form).

How this works in practice can perhaps best be introduced, as in Stone (2004), by using the example of two people speaking into two different microphones, the aim of the exercise being to differentiate, as far as possible, between the two voices. The microphones give different weights to the different voices (e.g. there might be a muffler between one of the speakers and one of the microphones). To simplify matters the microphones are assumed to be equidistant from each source, so that phase differentials are not relevant to the problem at hand. ICA and related techniques rely on the following observations:

(a) The two input signals, i.e. the two individual voices, are likely to be largely independent of each other, when examined at fine time intervals. However, the two output signals, i.e. the signals coming from the microphones will not be as independent, since they involve mixtures (albeit differently weighted) of the same underlying input signals.

(b) If histograms of the amplitudes of each voice (when examined at these fine time intervals) are plotted then they will most probably differ from the traditional bell-shaped histogram corresponding to random noise. Conversely, the signal mixtures are likely to be more normal in nature.

(c) The temporal complexity of any mixture is typically greater than (or equal to) that of its simplest, i.e. least complex, constituent source signal.

These observations lead to the following algorithm for source signal extraction:

If source signals have some property X and signal mixtures do not (or have less of it) then given a set of signal mixtures we should attempt to extract signals with as much X as possible, since these extracted signals are then likely to correspond as closely as possible to the original source signals.

Different variants of ICA and its related techniques 'unmix' output signals, thus aiming to recover the original input signals, by substituting 'independence', 'non-normality' and 'lack of complexity' for *X* in the above prescription.

2.4 Signal Mixtures

[IndependentComponentsAnalysis2d]

As noted earlier, ICA typically assumes that outputs are (linear combination) *mixtures* of inputs, i.e. are derived by adding together input signals in fixed proportions. If there are n input (i.e. source) signals then there need to be at least n different mixtures for us to be able to differentiate between the sources. In practice, the number of signal mixtures is often larger than the number of source signals. For example, with electroencephalography (EEG), the number of signal mixtures is equal to the number of electrodes placed on the head (typically at least 10) but there are typically fewer sources than this. If the number of signals is known to be less than the number of signal mixes then the number of signals extracted by ICA can be reduced by dimension reduction, either by preprocessing the signal mixtures, e.g. by using PCA techniques, or by arranging for the ICA algorithm to return only a specified number of signals.

Such mixtures can be expressed succinctly in matrix form, W, where the coefficients of W are the $w_{i,j}$ referred to above. We can then write the formula deriving the output signals from the input signals as:

$$y_t = W x_t$$

We note that we have implicitly assumed a model of the world involving *time homogeneity* (or *spatial homogeneity*, if the signals involve a spatial dimension rather than a time dimension) i.e. that the $w_{i,j}$ and hence W are constant through time.

If the mixing coefficients, i.e. the elements of W, are known already then we can typically easily derive the input signals from the output matrix by inverting this matrix equation, i.e. $x = Ay$ where $A = W^{-1}$.

However, we are more usually interested in the situation where the mixing coefficients are *unknown* (although as we have stated, we do implicitly assume with ICA that the mixture can be expressed in the above form, and that the corresponding coefficients are constant through time). We therefore seek an algorithm that estimates the unmixing coefficients, i.e. the coefficients $a_{i,j}$ of A, directly from the data, allowing us then to recover the signals themselves (and the original mixing coefficients).

We note that there is no way in such a framework of distinguishing between two input signals that are constant multiples of each other. Thus ICA and its variants will only generally identify signals up to a scalar multiplier (although we might in practice impose some standardised scaling criteria when presenting the answers or deciding which signals to retain and which to discard as 'insignificant' or unlikely to correspond to a true input signal).

If the number of input and output signals is the same then a matrix such as W can be viewed as corresponding to a vector transformation in an n-dimensional vector space (n being the number of input signals) spanned by vectors corresponding to the input signals. In this representation, each input signal would be characterised by a vector of unit length in the direction of a particular axis in this n-dimensional space, with each different pair of input signals being *orthogonal* to each other (in geometric terms, 'perpendicular' to each other). Any possible (linear combination) mixture of these signals then corresponds to some vector in the same vector space, and a set of n of them corresponds to a set of n vectors in such a space. An $n \times n$ matrix thus defines how simultaneously to map one set of n vectors to another in a way that respects underlying linear combinations. The types of transformations that are catered for by such matrices include rotation, shearing, and expansion away from or contraction towards the origin. Inverting such a matrix (if this is possible) corresponds to identifying the corresponding inverse transformation that returns a set of n transformed signal vectors to their original positions.

ICA uses this insight by identifying which (orthogonal) mixtures of the output series seem to exhibit the largest amount of 'independence', 'non-normality' or 'lack of randomness', since these mixtures can then be expected to correspond to the original input series (or scalar multiples of them). We note that not only will ICA not be able to differentiate between two signals that are scalar multiples of each other, but also that it won't be able to differentiate between, say, two different pairs of signals in which the ordering of the signals is reversed. This is because it doesn't directly include any prescription that ensures that any particular input signal will be mapped back to its own original axis. Instead, it is merely expected to arrange for each input signal to be mapped back to any one of the original (orthogonal) axes (but for no two different input signals to be mapped back onto the same axis). However, there may be a natural ordering that can be applied to the extracted signals, e.g. if input signals are expected to be strongly non-normal then we might order the extracted signals so that the first one is the most non-normal according to the relevant measure of non-normality that we have chosen, etc.

2.5 Projection pursuit

[IndependentComponentsAnalysis2e]

Suppose we focus further on the property of non-normality, which we might measure via the (excess) kurtosis of a distribution. Kurtosis has two properties relevant to ICA:

(a) All linear combinations of independent distributions have a smaller kurtosis than the largest kurtosis of any of the individual distributions (a result that can be derived using the Cauchy-Schwarz inequality).

(b) Kurtosis is invariant to scalar multiplication, i.e. if the kurtosis of distribution s is κ then the kurtosis of the distribution defined by $q = ks$ where k is constant is also κ.

Suppose we also want to identify the input signals (up to a scalar multiple) one at a time, starting with the one with the highest kurtosis. This can be done via *projection pursuit*.

Given (a) and (b) we can expect the kurtosis of $z_t = \sum_i p_i y_{i,t} = \sum_i \sum_j p_i w_{i,j} x_t$ to be maximised when this results in $z_t = k x_{q,t}$ for the q corresponding to the signal $x_{i,t}$ which has the largest kurtosis, where k is arbitrary. Without loss of generality, we can reorder the input signals so that this one is deemed the first one, and thus we expect the kurtosis of z_t to be maximised with respect to p when $pW = (k, 0, \ldots, 0)^T$ and $z_t = k x_{1,t}$. Although we do not at this stage know the full form of W we have still managed to extract out one signal (namely the one with the largest kurtosis) and found out something about W.

In principle, the appropriate value of p can be found using brute force exhaustive search, but in practice more efficient gradient based approaches would be used instead, see Section 2.8.

We can then remove the recovered source signal from the set of signal mixtures and repeat the above procedure to recover the next source signal from the 'reduced' set of signal mixtures. Repeating this iteratively, we should extract all available source signals (assuming that they are all lepto-kurtotic, i.e. all have kurtosis larger than any residual noise, which we might assume is merely normally distributed). The removal of each recovered source signal involves a projection of an m-dimensional space onto one with $m - 1$ dimensions and can be carried out using *Gram-Schmidt orthogonalisation*.

With any blind source separation method, a fundamental issue that has no simple answer is when to truncate such a search. If the mixing processes were noise free then the truncation should stop having extracted exactly the right number of signals (as long as there are at least as many distinct output signals as there are input signals). However, outputs are rarely noise free. In practice, therefore, we might truncate the signal search once the signals we seem to be extracting via it appear to be largely artefacts of noise in the signals or mixing process, rather than suggestive of additional true underlying source signals.

We see similarities with random matrix theory, an approach used to truncate the output of a principal components analysis to merely those principal components that are probably not just artefacts of noise in the input data.

2.6 Infomax and maximum likelihood independent components analysis

[IndependentComponentsAnalysis2f]

ICA as normally understood can be thought of as a multivariate, parallel version of projection pursuit, i.e. an algorithm that returns 'all at once' all of the unmixing weights applicable to all of the input signals. Indeed, if ICA uses the same measure of 'signal-likeness' (i.e. 'independence', 'non-normality', 'lack of complexity') and assumes the same number of signals exist as is used in the corresponding projection pursuit methodology then the two should extract the same signals.

To the extent that the two differ, the core measure of 'signal-likeness' underlying most implementations of ICA is that of statistical independence. As we have noted earlier, this is a stronger concept than mere lack of correlation. To make use of this idea, we need a measure that tells us how close to independent are any given set of unmixed signals.

Perhaps the most common measure used for this purpose is *entropy*. This is often thought of as a measure of the uniformity of the distribution of a bounded set of values. However, more generally, it can also be thought of as the amount of 'surprise' associated with a given outcome. This requires some a priori view of what probability distribution of outcomes is to be 'expected'. Surprise can then equated with *relative entropy* (i.e. Kullback-Leibler divergence) which measures the similarity between two different probability density functions.

The ICA approach thus requires an assumed probability density function for the input signals and identifies the unmixing matrix that maximises the joint entropy of the resulting unmixed signals. This is called the infomax ICA approach. A common assumed probability density function ('pdf') used for this purpose is a very high-kurtosis one such as some suitably scaled version of $p_s = \left(1 - \tanh(\mathrm{s})\right)^2$.

ICA can also be thought of as a maximum likelihood method for estimating the optimal unmixing matrix. With maximum likelihood we again need to specify an a priori probability distribution, in this case the assumed joint pdf p_s of the unknown source signals, and we seek the unmixing matrix, A, that yields extracted signals $x = Ay$ with a joint pdf as similar as possible to p_s. In such contexts, 'as similar as possible' is usually defined via the log likelihood function, which results in the same answer as the equivalent infomax approach, since both involve logarithmic functions of the underlying assumed probability distribution.

Both methods appear to rely on the frankly unrealistic assumption that the model pdf is an exact match for the pdf of the source signals. Of course, in general, the pdf of the source signals is not known exactly. Despite this, ICA seems to work reasonably well. This is because we do not really care about the form of the pdf. Indeed, it could correspond to a quite extreme distribution. Instead all we really need for the approach to work is for the model pdf to have the property that the closer any given distribution is to it (in relative entropy or log likelihood terms), the more likely that distribution is to correspond to a true source input signal. A hyperbolic tangent ('tanh')-style pdf may be an unrealistic 'model' for a true signal source, but use within the algorithm typically means that distributional forms with high kurtosis will be preferentially selected versus ones with lower kurtosis (even though neither may have a kurtosis anywhere near as large as that exhibited by the hyperbolic tangent pdf itself). It is the relative ordering of distributional forms introduced by choice of model pdf that is important rather than the structure of the model pdf per se. As a tanh-style model pdf preferentially extracts signals exhibiting high kurtosis it will extract similar signals to those extracted by kurtosis-based projection pursuit methods. Indeed, it ought to be possible to select model pdfs (or at least definitions of how to order likenesses of distributional forms to the model pdfs) that exactly match whatever metric is used in a corresponding projection pursuit methodology (even if this isn't how the ICA methodology was originally developed).

To estimate the unmixing matrix A $(= W^{-1})$ that maximises the relative entropy or log likelihood and hence corresponds to the supposed input signals, we could again use brute force. However, again it is

more efficient to use some sort of gradient ascent method as per Section 2.8, iteratively adjusting the estimated W^{-1} in order to maximise the chosen metric.

2.7 Complexity pursuit

[IndependentComponentsAnalysis2g]

Most signals measured within a physical system can be expected to be a mixture of statistically independent source signals. The most parsimonious explanation for the complexity of an observed signal is thus that it consists of a mixture of simpler signals, each from a different source. Underpinning this is the assumption that a mixture of independent source signals is typically more complex than the simplest (i.e. least complex) of its constituent source signal. This *complexity conjecture* underpins the idea of *complexity pursuit*.

One simple measure of complexity is predictability. If each value of a signal is relatively easy to predict from previous signal values then we might characterise the signal as having low complexity. Conversely, if successive values of a signal are independent of each other then prediction is in principle impossible and we might characterise such a signal as having a high complexity.

Stone (2004), when discussing this technique, focuses on minimising *Kolmogorov complexity* and defines a measure, F, of temporal predictability for a given set of signal mixtures $\mathbf{y}$ and weights, a_i, applied to these signal mixtures as follows:

$$F(a_i, \mathbf{y}) = \log V_i - \log U_i$$

Here $x_{i,t} = a_{i,j} y_{j,t}$, $\tilde{x}_{i,t}$ is a suitable exponentially weighted moving average of $x_{i,t}$, i.e. $\tilde{x}_{i,t} = \eta \tilde{x}_{i,t-1} + (1-\eta) x_{i,t-1}$ for some suitable (perhaps predefined) value of η, $V_i = \frac{1}{N}\sum_{t=1}^{N}\left(x_{i,t} - \bar{x_i}\right)^2$ corresponds to the overall variance of the given linear combination and $U_i = \frac{1}{N}\sum_{t=1}^{N}(\tilde{x}_t - x_t)^2$ corresponds to the extent to which it is well predicted by its previous values (assuming here a first order autoregressive dependency).

Complexity pursuit has certain advantages and disadvantages over ICA and projection pursuit. Unlike ICA it does not appear explicitly to include an a priori model for the signal pdfs, but seems only to depend on the complexity of the signal. It ought therefore to be able to extract signals with different pdf types. It also does not ignore signal structure, e.g. its temporal nature if it is a time series. Conversely, 'complexity' is a less obviously well-defined concept than independence or non-normality. For example, the prescription introduced by Stone (2004) and described above seems to be very heavily dependent on 'lack of complexity' being validly equated with signals exhibiting strong one-period autodependency.

2.8 Gradient ascent

[IndependentComponentsAnalysis2h]

All of the above approaches require us to maximise (or minimise) some function (the kurtosis, the log likelihood, the Kolmogorov complexity etc.) with respect to different unmixing vectors (or unmixing matrices, i.e. simultaneously for several unmixing vectors all at once). Whilst brute force could be applied for simple problems this rapidly becomes impractical as the number of signals increases. Instead, we typically use *gradient ascent*, in which we head up the (possibly hyper-dimensional) surface formed by plotting the value of the function for different unmixing vectors in the direction of

steepest ascent. The direction of steepest ascent can be found from the first partial derivative of the function with respect to the different components of the unmixing vector/matrix. Second order methods can be used to estimate how far to go along that gradient before next evaluating the function and its derivatives, see e.g. Press *et al.* (2007).

3. Principal components analysis (PCA)

[IndependentComponentsAnalysis3]

Section contents

3.1 Characteristics of PCA
3.2 Weighting schemas

3.1 Characteristics of PCA

[IndependentComponentsAnalysis3a]

Earlier, we noted some similarities between ICA and PCA but also noted some differences. In particular, ICA focuses on 'independence', 'non-Normality' and 'lack of complexity' and assumes that source signals exhibit these features, whereas PCA focuses merely on lack of correlatedness. Indeed, PCA analysis will even 'unmix' pure Gaussian (i.e. normally distributed) signals. Or rather, it will decompose multiple Gaussian output signals into some presumed orthogonal Gaussian input signals, which can be ordered with ones higher up the ordering explaining more of the variability in the output signal ensemble than ones lower down the ordering.

PCA involves calculating the eigenvectors and eigenvalues of the covariance matrix, i.e. the matrix of correlation coefficients between the different signals. Usually, it would be assumed that the covariance matrix had been calculated in a manner that gives equal weight to each data point. However, this is not essential; we could equally use a computation approach in which different weights were given to different data points, e.g. an exponentially decaying weighting that gives greater weight to more recent observations.

For n different output signals $y_{i,t}$, with covariance matrix, V, between the output signals, PCA searches for the n (some possibly degenerate) eigenvectors, q, satisfying the following matrix equation for some scalar λ.

$$Vq = \lambda q$$

For a non-negative definite symmetric matrix (as V should be if it actually corresponds to a covariance matrix), the n values of λ are all non-negative. We can therefore order the eigenvalues in descending order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$. The corresponding eigenvectors q_i are orthogonal i.e. have $q_i^T q_j = 0$ if $i \neq j$ and are also typically normalised so that $|q_i| = q_i^T q_i = 1$ (i.e. so that they have 'unit length'). For any λ's that are equal, we need to choose a corresponding number of orthonormal eigenvectors that span the relevant subspace.

By q_i we mean the vector $\left(Q_{i,1}, \ldots, Q_{i,n}\right)^T$ such that the signal corresponding to eigenvalue λ_i is expressible as:

$$q_{i,t} = \sum_k Q_{i,k}\left(y_{k,t} - \overline{y_k}\right)$$

If Q is the matrix with coefficients $Q_{i,j}$ then the orthonomalisation convention adopted above means that $Q^T Q = I$ where I is the identity matrix. This means that $Q^{-1} = Q^T$ and hence we may also write the output signals as a linear combination of the eigenvector signals as follows (in each case up to a constant value, since the covariances do not depend on means of series):

$$\left(y_{i,t} - \bar{y_j}\right) = \sum_k Q_{k,i} q_{k,t}$$

Additionally, we have $q_i^T V q_j = \lambda_j q_i^T q_j = 0$ if $i \neq j$ and $= \lambda_i$ if $i = j$. We also note that if $V_{i,j}$ are the coefficients of V then each individual $V_{i,j}$ is (here assuming that we have been using 'sample' rather than 'population' values for variances:

$$V_{i,j} = \frac{1}{N-1} \sum_t \left(\sum_k Q_{k,i} q_{i,t} \right) \left(\sum_k Q_{k,j} q_{j,t} \right) = \sum_k Q_{k,i} Q_{k,j} \lambda_k$$

Hence the sum of the variances of each output signal, i.e. the trace of the covariance matrix, satisfies:

$$tr(V) \equiv \sum_i V_{i,i} = \sum_i \sum_k Q_{k,i} Q_{k,i} \lambda_k = \sum_k \sum_i Q_{k,i} Q_{k,i} \lambda_k = \sum_k \lambda_k$$

We can interpret this as indicating that the aggregate variability of the output signals (i.e. the sum of their individual variabilities) is equal to the sum of the eigenvalues. Hence the larger the eigenvalue the more the corresponding eigenvector signal 'contributes' to the aggregate variability across the ensemble of possible output signals.

3.2 Weighting schemas

[IndependentComponentsAnalysis3b]

There are certain types of problem where it is particularly desirable to ascribe different levels of 'importance' to different input signals according to their contribution to aggregate variability in the output signal ensemble.

A good example is portfolio risk measurement. We might characterise the return series coming from each individual stock as an output series and we might be seeking a parsimonious way of explaining the variability across the stock universe by assuming that there are a relatively modest number of underlying factors driving the behaviour of multiple stocks, together with some residual idiosyncratic risk factors applicable to each stock in isolation. The ultimate aim is to estimate some measure of the likely spread of returns that might arise from one particular portfolio (the actual portfolio chosen by the fund manager) relative to those on a benchmark portfolio (also drawn from the same universe but with the stocks differently weighted). A common proxy for spread here might be the standard deviation (or variance) of the relative return. However, this may not be a good proxy for fat-tailed distributions.

Commercial statistical factor risk models typically derive estimates of these underlying factor signals using principal components analysis. Suitably averaged across possible portfolios that might be chosen, the factors exhibiting the highest eigenvalues really are the 'most important' ones, because they explain the most variability across the universe as a whole, see Section 3.1. At least they do if variability and standard deviation/variance are equated as would the case for normally distributed

random variables, but not necessarily for fat-tailed distributions. For these types of distributions, some refinement may be desirable, see Section 4.

Implicit in PCA is thus a weighting schema being applied to the different output signals. Suppose we multiply each individual output signal $y_{i,t}$ by a different weighting factor, w_i, i.e. we now recast the problem as if the output signals were $z_{i,t} = w_i y_{i,t}$. This does not, in some sense, alter the available information we have to identify input signals. But what it does do is alter how much variability each given output series contributes to the total. It will therefore alter the coefficients defining the eigenvectors and which ones are deemed most important. Hence the results of PCA are not *scale invariant* in relation to individual stocks, since one of the implicit assumptions we are adopting is that a given quantum of output from any given signal has the same intrinsic 'importance' (in variability terms) as the same quantum of output from any other signal.

How does this compare with ICA? The projection pursuit method introduced earlier (and corresponding infomax and maximum likelihood ICA approaches) grade signal importance by reference to kurtosis, rather than by reference to contribution to overall variability. As we noted earlier, kurtosis is scale invariant. Thus ICA should identify 'meaningful' signals that influence the ensemble of output signals (if we are correct to ascribe 'meaning' to signals that appear to exhibit 'independence', 'non-Normality' or 'lack of complexity'), but it will not necessarily preferentially select ones whose behaviours contribute *significantly* to the behaviour of the output signal ensemble.

4. Blending together PCA and ICA in the construction of risk models

[IndependentComponentsAnalysis4]

Section contents

4.1 Similarities between PCA and ICA
4.2 Different blended importance criteria
4.3 Time-varying volatility
4.4 Caveats

4.1 Similarities between PCA and ICA

[IndependentComponentsAnalysis4a]

Despite the different scale properties of PCA and ICA there are actually many similarities. In particular, we note that the value of $a^T V a$ for any arbitrary mixture of output signals $a = (a_1, \ldots, a_n)^T$ of unit length, i.e. where $|a| = a^T a = 1$ is somewhere between the largest eigenvector λ_1 and the smallest one λ_n (if the eigenvalues are ordered appropriately) and takes its largest value when $a = q_1$. Moreover, if we remove the signal corresponding to the largest eigenvalue using Gram-Schmidt orthogonalisation then the remaining vector space is spanned by the remaining eigenvectors (all of which are orthogonal to the eigenvector being removed).

Thus PCA can be re-expressed as an example of a projection pursuit methodology but using as the importance criterion the contribution of the input signal to aggregate output signal ensemble variance rather than the magnitude of the input signal kurtosis. This explains the close analogy between methods for deciding when to stop a projection pursuit algorithm and when to truncate a PCA, i.e. random matrix theory. The two involve the same underlying mathematics, but applied to different importance criteria.

This suggests that we can blend PCA and ICA together to better capture the strengths of each, by adopting a projection pursuit type methodology applied to an importance criterion that blends together variance as well as some suitable measure(s) of independence, non-Normality and/or lack of complexity.

4.2 Different blended importance criteria

[IndependentComponentsAnalysis4b]

One possible approach would be to use an importance criterion involving a function $P(a)$ that includes both variance and factors like kurtosis that characterise the extent to which the data seems to be coming from a non-normal distribution. For example, we might use the following:

$$P(a) = v(a)\big(1 + c.\kappa(a)\big)^2$$

(N.B. We could also use any function that varied monotonically in line with this function, e.g. its square root , if $P(a)$ is positive, and we would derive exactly the same input signals)

Here $v(a)$ is the variance of the time series corresponding to the mixture of output signals characterised by a, $\kappa(a)$ is its kurtosis, and c is a constant that indicates the extent to which we want to focus on kurtosis rather than variance in the derivation of which signals might be 'important'. Again we would constrain a to be of 'unit length', i.e. to have $|a| = a^T a = 1$.

The larger (i.e. more positive) c is, the more we might expect such an approach to tend to highlight signals that exhibit positive kurtosis. Thus the closer the computed unmixed input signals should be to those that would be derived by applying ICA to the mixed signals (if the ICA was formulated using model pdfs with high kurtosis). We here need to assume that $v(a)$ does not vary 'too much' with respect to a, so that in the limit as $c \to \infty$ any signal exhibiting suitably positive kurtosis will be selected at some stage in the iterative process, although we might expect variation in $v(a)$ to 'blur' together some signals that ICA might otherwise distinguish. The smaller (i.e. closer to zero) c is, the closer the result should be to a PCA analysis.

However, there are several possible weaknesses with such an approach:

(a) There is no immediately obvious reason to choose any particular value of c. This is because we have not introduced into the problem specification any particular relative importance to ascribe to variance versus kurtosis. One possible solution to this problem is to focus application of such a methodology onto a problem that does potentially provide some guidance in this area. The most obvious such application would be portfolio risk measurement in a situation where we wanted to measure risk not by reference to variance of relative return (or a monotonically equivalent measure such as standard deviation) but by reference to some other metric such as Value-at-Risk or Expected Shortfall that places greater weight on tail behaviour. We could for example 'extrapolate' into the tail based on observed variance and kurtosis (and also skew) using the 4th order Cornish Fisher asymptotic expansion. According to this expansion, we can estimate the quantile of a distribution relative to that which would apply were the distribution to have no skew or variance using the following formula, see e.g. Kemp (2009):

$$y_{CF4} = m + \sigma\left(x + \frac{\gamma_1(x^2-1)}{6} + \frac{3\gamma_2(x^3-3x) - 2\gamma_1(2x^3-5x)}{72}\right)$$

Here, $x = N^{-1}(\alpha)$, γ_1 is the skew of the distribution and $\gamma_2 = \kappa$ is the kurtosis of the distribution, where α is the probability to which x applies and $N^{-1}(z)$ is the inverse normal distribution function.

For example, we might adopt a 1 in 200 quantile cut-off, in which case $x = N^{-1}(0.005) = -2.576$. For a distribution with zero skew, we might thus apply an importance criterion that sought to maximise:

$$P(a) = \frac{v(a)}{x^2}\left(x + \frac{3\gamma_2(x^3 - 3x)}{72}\right)^2 = v(a)(1 + 0.39\kappa)^2$$

The physical interpretation of this is that, if these assumptions apply, then the 1 in 200 quantile is a factor of $(1 + 0.39\kappa)$ further into the tail than we might otherwise expect purely from the standard deviation of the distribution.

(b) Unfortunately, the 4th order Cornish-Fisher expansion is not in general very good at estimating the shape of the distributional form in regions in which we might be most interested, see e.g. Kemp (2009). In effect, the computation of skew and kurtosis gives 'too much' weight to the extent of non-normality in the centre of the distributional form whereas typically for risk management purposes we are most interested in the extent of non-normality in the tail of the distribution. He proposes an alternative approach, more directly akin to fitting a curve through the observed (ordered) distributional form, to 'extrapolate into the tail'.

Such an approach is more computationally intensive than the Cornish-Fisher approach, particularly if the data series in question involve a large number of terms. The approach requires the return series to be sorted, in order to work out which observations to give most weight to in the curve fitting algorithm. Sorting large data sets is intrinsically much slower than merely calculating their moments since it typically involves a number of computations that scales in line with approximately $O(N \log N)$ rather than merely $O(N)$. It may be that such a refinement would not in practice lead to a much enhanced risk model, as non-zero kurtosis is still typically a good indicator of the presence of fat-tailed behaviour, even if it is not a particularly good indicator of exactly how fat-tailed it is in the particular part of the distributional form in which we might be most interested.

(c) More problematic, perhaps, is another topic that Kemp explores in Kemp (2009) and Kemp (2010). He notes, as implicitly have earlier authors, that much of the fat-tailed behaviour observed in practice in return series (both when viewed singly and when viewed jointly) seems to derive from time-varying volatility, see Section 4.3.

4.3 Time-varying volatility

[IndependentComponentsAnalysis4c]

4.3.1 Introduction

Time-varying volatility is an example of a phenomenon that cannot easily be modelled via an approach involving *linear combination* mixtures. Instead, it can be thought of as an example a *distributional* mixture as referred to in Section 2.2. This is because we can characterise a world which exhibits time-varying volatility as one in which returns are coming from different distributions (the distributions differing by reference to their variance) depending on when the return occurs. If it occurs at a time

when ‘volatility’ is high then, all other things being equal we would expect the return to be more spread out than when ‘volatility’ is low. Another name for time-varying volatility is heteroscedasticity.

It is widely accepted within the financial services industry and within relevant academic circles that markets exhibit time-varying volatility. Markets can, for possibly extended periods of time, appear to be quite ‘quiet’, e.g. without many large daily movements, but then to move to a different regime in which, say, daily movements are more significant. However, volatility does not necessarily always appear to move in tandem across markets or even across parts of the same market.

We set out below some ideas for how the blending of PCA and ICA might be refined to cater for time-varying volatility and also some of the challenges that might arise in practice.

There are several possible ways of catering for time-varying volatility. One approach would be to assume that the market (or sub-components of it) might be in two or more relatively stable *discrete* ‘regimes’, the regimes being differentiated by some postulated underlying state variable that reveals itself by reference to the level of volatility that a market is exhibiting. Probabilities of movement between such the different regimes might then be built up, most probably incorporating some sort of autoregressive characteristics as in *threshold autoregressive time series models*.

A perhaps simpler approach is to assume that there is some underlying (and relatively slowly changing) *continuous* variable characterising the variability of the market or of a segment of it, which we might estimate at any particular point in time by applying moving average techniques to recent past observations. This sort of approach is in effect the one used in Kemp (2009) and Kemp (2010). It is also the one implicit in GARCH models and their variants. We also immediately recognise a link with the complexity pursuit variant of ICA described in Section 2.7, which also focused on moving averages as a sign of ‘predictability’ of a time-ordered series. It would be possible to use a moving average that applied equal weights to observations within a fixed length window. However, as in Section 2.7 it might be preferable to focus on an exponentially damped moving average, potentially allowing flexibility in the decay factor involved.

We should then bear in mind that there are several possible ways in which we might define time-varying ‘variability’ (even in the context of blind source separation when there is no obvious differentiator between individual output signals, here return series). We can think of any particular return series as possessing its own *individual* volatility which is somehow evolving through time. The average of these individual time-evolving volatilities, i.e. *average* volatility, might itself also be somehow evolving through time, in a possibly more reliably predictable manner, given the greater number of data points contributing to its calculation. However, we can also characterise the ensemble of return series as exhibiting a potentially time-varying *cross-sectional* volatility. Own/market average volatility and cross-sectional volatility in effect characterise different parts of the covariance matrix between different stocks. The former corresponds to the elements along the leading diagonal or their average (the ‘variance’ terms), whilst the latter corresponds to an average of the off-diagonal elements (the ‘covariance’ terms). When we talk about ‘average’ stock correlation some of the same types of topics also arise, see e.g. Measuring Average Stock Correlation.

4.3.2 Higher moments

If a partial parameterisation of the evolution of ‘variability’ through time can include elements bearing the hallmarks of the structure of a covariance matrix then a more complete parameterisation might introduce further elements akin to the higher moment structure of a multidimensional probability distribution. This would probably involve a rather sophisticated to model and that would lack

parsimony and it may be better to limit ourselves merely to models that incorporate one of three types of time-varying volatility adjustments, namely ones involving:

(a) An exponentially weighted moving average estimate of volatility for a given individual stock (measured, say, by the standard deviation of past returns for that stock in isolation, with greater weight given to more recent observations);

(b) An exponentially weighted moving average estimate of volatility for the average of all stocks (calculated, say, in a manner similar to (a) but applied to the average return for the market as a whole); and

(c) An exponentially weighted moving average estimate of cross-sectional volatility between stocks (measured, say, by calculating for each time period the cross-sectional standard deviation of returns across the stock universe and then determining a suitable exponentially weighted moving average through time of these standard deviations).

4.3.2 Contemporaneous estimates of future volatility

It is also worth bearing in mind that there may be available to us contemporaneous estimates of future volatility that may be more reliable than exponentially weighted moving averages of past data. For example, we might be able to source market implied volatilities (and correlations) from options markets. The relevance of this sort of data to risk model design is discussed further in Kemp (2009) as is the more fundamental topic of whether it is better when trying to measure risk to use 'market implied' probabilities of occurrence rather than or in addition to estimated 'real world' probabilities of occurrence. An introduction to how it is possible to calibrate probability distributions used for risk measurement purposes to market implied data is given here.

4.4 Caveats

[IndependentComponentsAnalysis4d]

When carrying out blended PCA/ICA analyses of financial series date it may be worth bearing in mind that:

(a) It is common, at least for equity-based risk models, for there to be far more series than there are datapoints in each series. This alters eigenvector dynamics and therefore we might presume the dynamics of the equivalent importance rated factors extracted using a blended PCA/ICA approach. For example, however many series there are, the number of eigenvectors it is possible to distinguish is limited by the number of points in the series (and is no larger than $N-1$ if there are N datapoints in each series. Moreover, practical risk model design requires additional elements able to cater for incomplete data series (e.g. for stocks that have only recently listed or have recently demerged).

(b) Any added robustness that we might appear to identify for a blended approach may merely be an artefact of look-back bias. See e.g. Kemp (2009) for an explanation of look-back bias and why it may apply here, even when any backtesting approach we might have used is 'out of sample'.

(c) In any case, past market dynamics may not turn out to be a good predictor of future market dynamics, even if weight of academic opinion seems to believe that volatility is 'more predictable' than, say, future return.

Relevant Nematrian web service tools

[IndependentComponentsAnalysisTools]

The main tools currently made available by the Nematrian website for carrying out independent components and principal components analyses are:

- MnPrincipalComponents, MnPrincipalComponentsSizes, MnPrincipalComponentsWeights. These carry out various activities linked to derivation of principal components
- MnBlendedPCAICA. This is a tool that allows users to carry out the blended PCA/ICA computations described in Section 4b.

References

[IndependentComponentsAnalysisRefs]

Kemp, M.H.D. (2009). *Market consistency: Model calibration in imperfect markets*. John Wiley & Sons [for further information on this book please see Market Consistency]

Kemp, M.H.D. (2010). *Extreme Events: Robust portfolio construction in the presence of fat tails*. John Wiley & Sons [for further information on this book please see Extreme Events]

Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (2007). *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press

Stone, J.V. (2004). *Independent Component Analysis: A Tutorial Introduction*. MIT Press