5 Dimension Reduction Methods

5.1 Principal Component Analysis

Principal components represent directions, in some vector space determined by some data. In Principal Component Analysis (PCA), we define the components to represent orthogonal - and hence uncorrelated - directions for which the data exhibits some level of variance that spans the data space. For instance, for some P-dimensional data; we can represent P directions [44] ordered by the amount of variance they capture. The first principal component corresponds to the direction of maximum variance, and each subsequent component captures orthogonal directions of decreasing variance.

Ultimately, principal component analysis is used as a dimension reduction technique by picking, among the PCs, the directions that exhibits most variance, and leaving out other directions.

Briefly, Principal Components estimation can be computed by following these three steps:

1. The initial step involves centering the design matrix. This is done for interpretational purposes and for simplicity, it is not necessary in practice [45].

2. Points are projected onto the first principal component, and the distance between the points and their projections is minimized to determine PC1.

3. Additional principal components are determined by selecting directions that are orthogonal to the initially found PC1. We iterate until the directions span the data space.

Mathematically, given some N-by-P design matrix, one can represent the principal components by some P-by-P dimensional unit vector $\vec{w}$ [46]; the points in the P-dimensional feature space are defined as $\overrightarrow{x_i}$; the projection of $\overrightarrow{x_i}$ on the PC is $(\vec{x}_i \cdot \vec{w})\,\vec{w}$ [47] and the distance between the point and its projection - also called the residual - is $\|\overrightarrow{x_i} - (\vec{w} \cdot \overrightarrow{x_i})\,\vec{w}\|^2$[48] . As mentioned, principal component analysis ultimately aims at minimizing the mean squared error of the residuals: $\text{MSE}(\vec{w}) = \frac{1}{n}\sum_{i=1}^{n} \|\vec{x}_i\|^2 - (\vec{w} \cdot \vec{x}_i)^2$;[49] Using matrix manipulations, one can prove that this mean squared residual

[44] That spans $\mathbb{R}^l$

[45] That is because centering X, allows us to interpret X^TX as the covariance matrix, which then permits the eigendecomposition of the X^TX matrix. I will discuss this later in the my analysis of PCA estimation

[46] $\vec{w}$ represent the directional vectors of the principal components. Its dimension is P-by-P as it represents P latent dimensions in some P dimensional feature space

[47] That is; we compute the dot product between $\overrightarrow{x_i}$ and $\overrightarrow{w}$, and since $\overrightarrow{w}$ is a unit vector, we get a scalar that represents the value on $\overrightarrow{w}$ which $\overrightarrow{x_i}$ gets projected to. To get the actual vector on which $\overrightarrow{x_i}$ gets projected, we multiply the inner product by the directional vector $\overrightarrow{w}$.

[48] The squared norm is used instead of the absolute value, because we want a differentiable residual

[49] That is because the residual can be reduced as such

$$\begin{aligned}
\|\overrightarrow{x_i} - (\vec{w} \cdot \overrightarrow{x_i})\,\vec{w}\|^2 &= (\overrightarrow{x_i} - (\vec{w} \cdot \overrightarrow{x_i})\,\vec{w}) \cdot (\overrightarrow{x_i} - (\vec{w} \cdot \overrightarrow{x_i})\,\vec{w}) \\
&= \overrightarrow{x_i} \cdot \overrightarrow{x_i} - \overrightarrow{x_i} \cdot (\vec{w} \cdot \overrightarrow{x_i})\,\vec{w} \\
&\quad - (\vec{w} \cdot \overrightarrow{x_i})\,\vec{w} \cdot \overrightarrow{x_i} + (\vec{w} \cdot \overrightarrow{x_i})\,\vec{w} \cdot (\vec{w} \cdot \overrightarrow{x_i})\,\vec{w} \\
&= \|\overrightarrow{x_i}\|^2 - 2\,(\vec{w} \cdot \overrightarrow{x_i})^2 + (\vec{w} \cdot \overrightarrow{x_i})^2\,\vec{w} \cdot \vec{w} \\
&= \overrightarrow{x_i} \cdot \overrightarrow{x_i} - (\vec{w} \cdot \overrightarrow{x_i})^2
\end{aligned}$$

minimization is equivalent to maximizing the variance of the projections $\widehat{\sigma^2}\left(\vec{w}\cdot\overrightarrow{x_i}\right)$ (i.e. the variance of the distance between the origin and the points' projection) ,which, in turn, is dependent on the covariance matrix $v = \frac{X^T X}{N}$ of the design matrix [50] [51] Principal Component Analysis can thus be mathematically summarized as performing:

$$\text{Argmax}_w w^T v w \quad \text{such that} \quad \mathbf{w}^T\mathbf{w} = 1$$

That is, we maximize the variance of the projections $w^T v w$ such that the direction vectors defining the principal components are unit vectors. This can be expressed as a Lagrangian problem : $L(\precapprox, \lambda) \equiv \precapprox^T \succapprox\precapprox - \lambda\left(\precapprox^T\precapprox - 1\right)$ Interestingly, we get that: $\mathbf{vw} = \lambda\mathbf{w}$: Principal components are thus the eigenvectors of the covariance matrix $v = \frac{X^T X}{N}$, with $\lambda \geq 0$ the corresponding eigenvalue matrix [52]. Since the covariance matrix is symmetric then its eigenvector matrix is orthogonal; thus w , the principal components' directions, are the eigenvectors of the covariance matrix that spans the whole p-dimensional space and λ corresponds to the magnitude of variance in the direction of each corresponding eigenvector. Hence, the (PC1) defined as the principal component along which the the data exhibits the highest variance level, is the w_j for which λ_j is the highest - the same logic follows for subsequent PCs.

Dimension reduction using Principal Component Analysis consists of choosing the most relevant direction among all the different directions. That is, for some P dimensional data space we would like to find a Q dimensional subspace - defined by a subset of orthogonal PCs - that summarizes best the data (More specifically , we would like to find the number of eigenvalues, Q) . There are different ways to achieve this: One can for example compute the $R^2 \equiv \frac{\sum_{i=1}^{q}\lambda_i}{\sum_{j=1}^{p}\lambda_j}$ metric that quantifies the ratio of variance explained by the subset of Principal components over the total variations in our model [53] ; graph a scree plot of R^2 with respect to λ and choose the lambda according to the shape of the plot [54]; however this method is not rigorous. Instead, one can also treat Q (or, equivalently, the number of

since $\vec{w}\cdot\vec{w} = \|\vec{w}\|^2 = 1$.

[50] $\text{MSE}(\vec{w}) = \frac{1}{n}\sum_{i=1}^{n}\|\vec{x}_i\|^2 - (\vec{w}\cdot\vec{x}_i)^2 = \frac{1}{n}\left(\sum_{i=1}^{n}\|\vec{x}_i\|^2 - \sum_{i=1}^{n}\left(\vec{w}\cdot\overrightarrow{x_i}\right)^2\right)$ the first term is independent of w , it is thus not relevant to our Minimization problem , we end up with $-\frac{1}{n}\sum_{i=1}^{n}\left(\vec{w}\cdot\overrightarrow{x_i}\right)^2$; Minimizing a concave function is equivalent to maximizing a convex one, hence, we will instead maximize $\frac{1}{n}\sum_{i=1}^{n}\left(\vec{w}\cdot\overrightarrow{x_i}\right)^2$. Since $\mathbf{Var}[X] = E[X^2] - E[X]^2$; we decompose the MSE Problem: as such $\left(\frac{1}{n}\sum_{i=1}^{n}\overrightarrow{x_i}\cdot\vec{w}\right)^2 + \widehat{\sigma^2}\left(\vec{w}\cdot\overrightarrow{x_i}\right)$; and since the X is centered the first term cancels out and we end up by maximizing the variance of the projection: $\widehat{\sigma^2}\left(\vec{w}\cdot\overrightarrow{x_i}\right)$. Minimizing the residuals is thus equivalent to maximizing the variance of the projections. In turn, the variance of the projection can be expressed with respect to X's covariance matrix: $\widehat{\sigma^2}\left(\vec{w}\cdot\overrightarrow{x_i}\right) = \frac{1}{n}\sum_i\left(\overrightarrow{x_i}\cdot\vec{w}\right)^2 = \frac{1}{n}(\mathbf{xw})^T(\mathbf{xw}) = \frac{1}{n}\mathbf{w}^T\mathbf{x}^T\mathbf{xw} = \mathbf{w}^T\frac{\mathbf{x}^T\mathbf{x}}{n}\mathbf{w} = \mathbf{w}^T\mathbf{vw}$

[51] This equivalence between residual minimizing and variance maximizing can be also demonstrated by the Pythagorean theorem: Consider for simplicity finding PC1 only: since x_is are fixed and since the projections $(\vec{x}_i\cdot\vec{w})\,\vec{w}$ form a triangular rectangle, we apply the Pythagorean theorem: $\underbrace{\left[(\vec{x}_i\cdot\vec{w})\,\vec{w}\right]}_{\text{A}}^2 + \underbrace{\left[\overrightarrow{x_i} - \left(\vec{w}\cdot\overrightarrow{x_i}\right)\vec{w}\right]}_{\text{B}}^2 = \underbrace{\left[x_i^2\right]}_{\text{C}}$; Since A is fixed any increase in B corresponds to a decrease in C and vice versa.

[52] Since the elements of covariance matrices are always positive. Then the eigenvalues will be positive

[53] This metric is $\in [0, 1]$; The bigger it is the more is the sub-model representative

[54] One can choose the λ , on the "elbow" of the scree plot. That is the point above which the screen plot does not

eigenvalues) as a hyper parameter of the model and use cross validation for optimal selection (Detailed illustration is available in the empirical analysis chapter).

Principal Component Regression is simply a linear regression method applied on the Principal Components of the predictors. This method is used mitigate the problems of ill-posedness and overfitting already discussed ; by using a subset of principal components. Instead of using a penalized L_p regularization method, regularization is explicit and is done prior to fitting. This method solves overfitting, because, as for lp regularization methods, we reduce the complexity and hence control overfitting by choosing a subset of predictors. And ill-posedness, because, by applying PCA and selecting a subset of features, we mitigate some issues related to ill-posedness; Notably, linear dependence in the Design matrix is automatically corrected as the Principal components are uncorrelated. In addition,choosing a subset of predictors can solve underdetermined ; $P >> N$; design matrices.

Principal Component Regression first requires finding Q Principal Components using PCA. The subset of principal component spans a "latent" feature space. PCR consists of performing a linear regression using the new latent factors as covariates. Concretly, after identifying the orthogonal principal components (PCs), we perform a linear regression in a space defined by $PC_1, ..., PC_Q$. In this transformed space, each point from the original feature space is now represented by coordinates determined by their projections onto the PCs. The resulting Q-dimensional estimated parameter can then be projected back to the original P- dimensional space by multiplying it with w_Q[55] .This multiplication is intuitive when considering w_Q as a Factor Loadings, transforming latent factors to observed ones (This interpretation of w_Q as Factor Loadings is typical of Factor Analysis and Probabilistic Principal Component Analysis, however I will not delve into these concepts as they are beyond the intended scope of this study). In order to test/tune this regression, the testing data (i.e. the testing design matrix) is regressed on the already retrieved PCs.

Principal Component Analysis is widely used in high dimensional factor modeling; However, there exists a spectrum of efficient alternative dimension reduction techniques. The Probabilistic Principal Component Analysis (PPCA), for example, is a method derived from PCA that assumes a latent variable model with probabilistic assumptions on the latent variable. Factor analysis, for instance, is another ubiquitous method, that finds latent factors similar to PPCA, but assumes a non-isotropic Gaussian distribution for the covariance matrix. Fourier Analysis, for example, interprets the data as the sum of Fourier basis functions to compute latent factors. The Wavelet decomposition method too, is another notable feature selection method, where wavelets functions are exploited using topological methods to determine the latent components etc...

exhibit sharp variations

[55]Note, w_Q the principal components directional vector - or equivalently, the eigenvector matrix - is a matrix that defines Q dimensions in a P-dimensional space. It is thus a P-by-Q matrix, defined as a transformation from $R^Q \rightarrow R^P$. By multiplying β_Q by w_Q we are projecting the matrix to the original P dimensional space.

Principal Component Analysis, however, remains the most documented method in asset pricing feature selection applications, because it is interpretable, has applications in pure linear algebra, and importantly, because it provides an ordered list of Principal components: Unlike other dimension reduction techniques where the "relevance" of one latent dimension with respect to another is not explicitly indicated; PCA provides an ordered list of uncorrelated dimensions based on the level of variance in each PC direction, which facilitates dimension reduction.

One major drawback of Principal Component Analysis, is that while the resulting Principal Components (PCs) are uncorrelated, they are not guaranteed to be statistically independent.[56] Although the resulting PCs from PCA are linearly independent orthogonal directions, ensuring null correlation when feature points are projected onto them, this does not imply statistical independence. In fact, a null correlation is a second order degree of independence, while statistical independence can imply higher orders of dependencies. From an information theory point of view, for example, one can argue that, correlation "does not reflect the information distance" between two variables (Taleb, 2023) [27]. It's important to note that variables can be both statistically independent and uncorrelated when exhibiting only second-order dependence, but this is unlikely in the case of the features used in this dataset, given their intricate nature and their complexity. To address this limitation, I perform another dimension reduction method that specifically aims at finding higher orders of statistical independence between the latent factors.

5.2 Independent Component Analysis

Independent component analysis is a blind source separation method typically used in the field of signal processing in order to retrieve statistical independent sources of some set of signals (See appendix, for a graphical representation of ICA). I use Independent Component Analysis to retrieve statistically independent factors. Statistical independence is a strong measure of dependence[57]: It quantifies how much the occurrence (or non occurrence) of an event affects the occurrence (or non occurrence) of another. We define $X_1, ..., X_N$ random variables to be "statistically independent" when $P(x_1, \ldots, x_N) = \prod_{i=1}^{N} P(x_i)$.

The independent component analysis is built upon the following framework: Consider some observed multidimensional data represented by x (P-by-N) , to be some linear mixture of statistically independent sources s of dimension Q-by-N. Thus, $x = As$ with A a P-by-Q linear mixing matrix with both A and s unknowns. This method assumes that what we observe is an unknown linear transfor-

[56] Uncorrelated variables implies linear independence . While statistical independence, between variables X and Y for instance, implies that $f(\mathbf{X}, \mathbf{Y}) = f_X(\mathbf{X}) \cdot f_Y(\mathbf{Y})$: This means that none of the variables explains the other. Clearly, statistical independence is a much stronger assumption

[57] This is in contrast with the correlation metric. Which is considered a weak dependence measure from a statistical point of view

mation of unknown statistically independent latent factors; and the goal of ICA is to retrieve $\hat{s} = Wx$ with $W = A^{+}$, and hence obtain the statistically independent source factors defining the observed features. The issue with $x = As$ however, is that it is an $Ax = b$ problem with both X and A being unknown, and this is, a priori not solvable.

Independent Component Analysis, proposes a strategy to find W (and hence retrieve $\hat{s}$): First, Consider the Singular Value Decomposition of A: $A = U\Sigma V^{\top}$,which can be studied as a rotation-stretching-rotation transformation, and its "inverse" [58] expressed as $W = V\Sigma^{-1}U^{\top}$

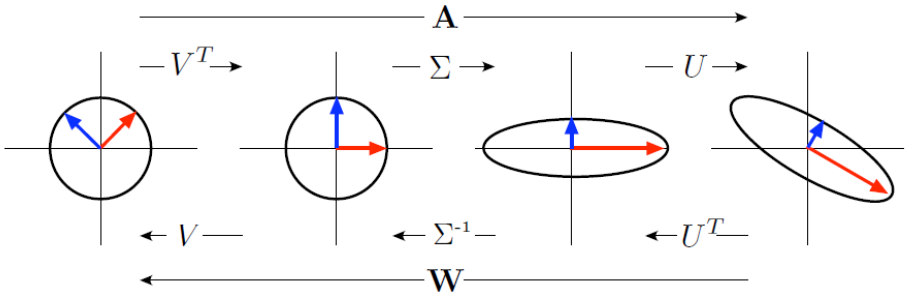


Figure 6: Graphical representation of the Singular Value Decomposition of the linear mixing matrix A, as well as its Pseudoinverse W. Hence, $x = As$ can be viewed as an SVD transformation from x to s (and vice versa)

Importantly this Singular Value Decomposition of the mixing matrix illustrates the idea that $x = As$ can be viewed as an SVD transformation from s to x (and vice versa). Independent Component Analysis's strategy unfolds in two stages: First, one needs to study the covariance of the observed data to find U and Σ. Independent Component Analysis assumes demeaned variables and introduces a crucial assumption: whitened covariance of sources, i.e. $E(s^T s) = I$.

By doing so, we have simplified our $Ax = b$ problem : In fact, we can now express the covariance of our observed variable $\mathbb{E}\left(xx^{\top}\right) = U\Sigma^2 U^{\top}$ independently of v and s [59] and since $\mathbb{E}(xx^T)$ is a symmetric matrix and it is always diagonalizable such that $E\left[xx^{\top}\right] = EDE$,with E and D corresponding to the eigenvectors and eigenvalues matrices. Thus, by imposing the whitening assumption we have now found both U and Σ such that $\hat{s} = Wx = V\Sigma^{-1}U^{\top}x$ gets reduced to $\hat{s} = VD^{-\frac{1}{2}}E^T x$ with now only V being unknown.

Independent Component Analysis's second stage is to exploit the statistical independence of sources in order to find V. In fact,V is a rotation matrix determined by some angle parameter θ (In a two dimensional space ,for instance, a rotation matrix V takes the form $\begin{pmatrix} \cos(\theta) & -\sin(\theta)) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}$ where θ is the only defining variable). In the $\hat{s} = VD^{-\frac{1}{2}}E^T x$ problem, V represents the last rotation from a

[58] "Pseudo-inverse" is a technically better suited term since W is a double sided inverse if and only if it has Full Rank . I use the term "inverse" for simplicity

[59] $\mathbb{E}\left(xx^{\top}\right) = \mathbb{E}\left(Ass^{\top}A^{\top}\right) = \mathbb{E}\left(U\Sigma V ss^t V\Sigma U^{\top}\right) = U\Sigma V^{\top}\mathbb{E}\left(ss^{\top}\right)V\Sigma U^{\top}$, since $V^T = V^{-1}$ (Since it is an orthogonal matrix) and $\mathbb{E}(ss^T) = I$ then $\mathbb{E}\left(xx^{\top}\right) = U\Sigma^2 U^{\top}$

rotated-then-stretched source s to x (refer to figure). Hence, computing V can be viewed as finding the angle parameter of the last rotation that gives us statistically independent sources. ICA exploits the independence assumption to find the angle: Finding V can be formalized as finding θ such that some metric of independence is minimized. Information theory provides us with this metric: In fact, mutual information is an adequate metric here as it computes the information distance between two distributions; i.e. quantifies the amount of information one variable provides about another; and is thus a good proxy for statistical independence. However, since sources are usually more than two, "multi-information", a generalization of the mutual information, is better suited. Defined as $I(y) = \int p(y) \log_2 \frac{P(y)}{\prod_{i=1}^{N} p(y_i)} dy$, this metric is an ideal proxy to statistical independence. Now, one can find rotation matrix V and solve ICA by minimizing $I(\hat{s})$ where $\hat{s} = VD^{-\frac{1}{2}}E^T x$. That is, finding the θ such $I(\hat{s})$ is minimized. This minimization problem is not trivial, a reduced form of $I(\hat{s})$ is used instead [60]. Thus $V = \text{Argmin}_v \sum_i H\left[\left(VD^{-\frac{1}{2}}E^\top X\right)_i\right]$. Having found V, one can find W and s the statistically independent factors.

In this paper, observations x are the features (Those are P, N dimensional feature vectors) and sources s represent the reduced latent features (hence , s is a Q by N matrix). The linear transformation A, essentially transforms the sources which are points in the R^Q reduced latent feature space to signals (Those are the points we observe in the original feature space) in a new separate R^p space spanned by the known original features. A's columns thus represent the latent direction of the observations in the original feature space.

While principal components are orthogonal, the independent components resulting from an ICA are not (Unless sources dependence is limited to second order). This can be clearly illustrated in a 2 Dimensional ICA; where the signals are two dimensional features (Above 3 Dimensions this cannot be graphically illustrated). Below is a graphical representation of the difference between PCA and ICA, using my dataset, on two factors.

[60] The reduced form is obtained by noting that the multi-information metric can be expressed as $I(\mathbf{y}) = \sum_i H[y_i] - H[\mathbf{y}]$, thus $I(\hat{s}) = \sum_i H\left(\left(VD^{-\frac{1}{2}}E^T X\right)_i\right) - H\left(VD^{-\frac{1}{2}}E^T X\right)$. Given the following entropy property: for any continuous random variable X and transformation A the differential entropy $H(AX + b) = H(X) + log|A|$. Then $I(\hat{s}) = \sum_i H\left(VD^{-\frac{1}{2}}E^\top x\right)_i\Big) - \left(H\left(D^{-\frac{1}{2}}E^\top x\right) + \log_2|V|\right)$, since $\det(V) = 1$ (Property of a rotation matrix) then $\log_2|V| = 0$ and since $D^{-\frac{1}{2}}E^\top X$ is constant and independent of V we can neglect it. We end up with a simplified version of $I(\hat{s})$: $I(\hat{s})|_{\text{simple}} = \sum_i H\left(VD^{-1/2}E^\top X\right)_i\Big)$

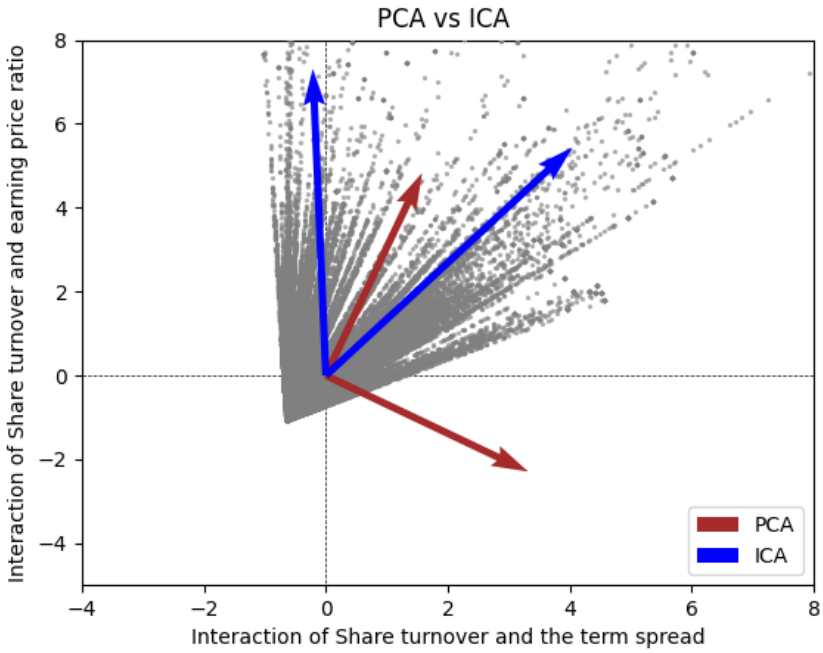
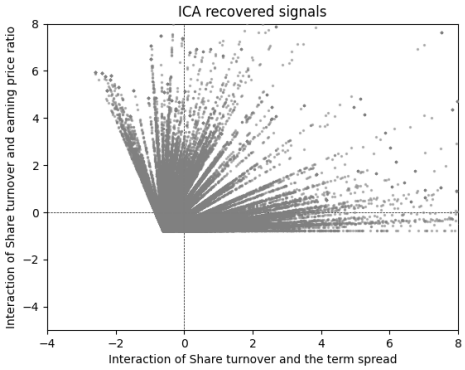
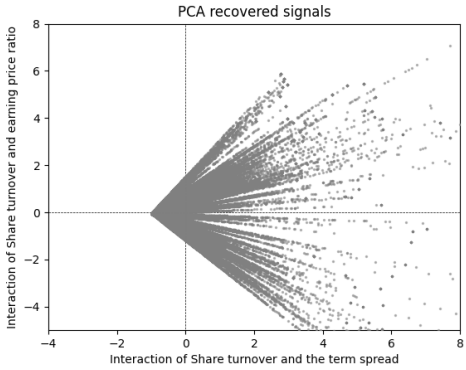


Figure 7: Graphical Representation of PCA vs ICA using a subset of Factors; notably Share turnover, Earnings to Price Ratio and the Term Spread on returns between 2001 and 2020. The Latent directions are different. In this example, Independent Components seem more relevant as they represent better the data.



((a)) Independent Component Analysis Recovered Signals



((b)) Principal Component Analysis Recovered Signals

Figure 8: Principal Components vs Independent Component Tested on a subset of the Dataset

This application of ICA on a subset of the data is indicative of the power and potential relevance of this method relative to the PCA. In fact, one can see that the PCA might not be particularly effective for orders of correlation above 2 as the above data seems to exhibit.

As for the Principal Component Regression, I use the Independent Components found in ICA as covariates of a linear regression. And as for PCR, the IC-based regression model is tuned on the validation set by changing the Number of Independent Components. However, the main drawback of ICA, is

that the independent components are not explicitly sorted with respect to their relevance , unlike PCA (This is, in fact, is the reason for which PCA is ubiquituous in Dimension reduction methods applications). ICA provides no indication of the relative relevance among subsets of independent components, and trying for different combinations is computationally infeasible [61]. I thus use the power data method proposed by Hendrikse et al. (2007)[10] . The rationale is as follow: The variance of the signals (i.e. the observed features) can be expressed with respect to the different independent component contributions. In fact, for some 1-Dimensional signal (P-by-1) we can express the variance of standardized signals as $Var(X) = E(X^2) = \sum_{i=1}^{P} \mathrm{E}\left[x_i^2\right] = \sum_{i=1}^{P} \mathrm{E}\left[(\mathrm{a}_i \cdot \mathrm{s})^2\right] = \sum_{j=1}^{Q} \left\{\mathrm{E}\left[s_j^2\right] \cdot \sum_{i=1}^{P} a_{i,j}^2\right\}$ for P the number of signals and Q the number of latent sources [62]. Hence, I compute for each $j \in \{1, Q\}$, $(E\left[s_j^2\right] \cdot \sum_{i=1}^{P} a_{i,j}^2)$ and choose the component j for which the contribution is the highest.

[61] For my dataset comprising 912 factors, testing all different of features combinations requires $\sum_{k=1}^{912} \begin{pmatrix} 912 \\ k \end{pmatrix}$ iterations. This is infeasible

[62] Note: In this paper, the signals are multidimensional. I use 1-Dimensional signals for simplicity