In Proceedings of 3rd International Conference on Independent Component Analysis and blind Signal Separation, 2001.

SELECTION OF INDEPENDENT FACTOR MODEL IN FINANCE

Lai-Wan Chan and Siu-Ming Cha

Department of Computer Science and Engineering,
The Chinese University of Hong Kong,
Shatin, N.T., Hong Kong.
Email : lwchan@cse.cuhk.edu.hk
http://www.cse.cuhk.edu.hk/~lwchan/

ABSTRACT

In finance, factor model is a fundamental model to describe the return generation process. Traditionally, the factors are assumed to be uncorrelated with each other. We argue that independence is a better assumption to factor model from the viewpoint of portfolio mangement. Based on this assumption, we propose the independent factor model. As the factors are independent, construction of the model would be another application of Independent Component Analysis (ICA) in finance. In this paper, we illustrate how we select the factors in the independent factor models. Securities in the Hong Kong market were used in the experiment. Minimum description length (MDL) was used to select the number of factors. We examine four sorting criteria for factor selection. The resultant models were cross-examined by the runs test.

1. INTRODUCTION

Factor Model, also called Index Model, is one of the basic models in finance to analyze the risk/reward relationships of security returns [1]. It has been used extensively in finance. Applications of factor model include portfolio construction, sensitivity analysis. Besides, theories, such as Capital Asset Pricing Model (CAPM) and Arbitrage Pricing Theory (APT), are built upon factor models.

There are two approaches to factor models [2, 3, 4]. One is the fundamental approach which links the factors to some macro-economic measurements, such as unexpected changes in the rate of inflation, interest rate, rate of return on a treasury bill etc. The sensitivities, $\beta's$, are evaluated accordingly. However, it is very difficult to determine the appropriate model, include the number of factors and what the factors are. The other approach is the statistical factor model; for examples, factor analysis and PCA. Historical security returns are analyzed to generate uncorrelated factors. Under this approach, principle component analysis(PCA) is the most successful method [5, 6, 7]. It is used to find the factors and their sensitivities[8, 9]. However it has also been shown that the separated factors are not able to truly reflect the real case but only one meaningful factor, which corresponds to the market effect, is extracted. This is due to two limitations of PCA. First, the separated principal components must be orthogonal to each other. Second, PCA uses only up to second order statistics, *i.e.* the covariance and correlation matrix.

The motivation for us to apply ICA in factor model is more than a simple replacement of PCA by ICA. Traditionally, the factors in the factor model are assumed to be uncorrelated. It has recently been pointed out that uncorrelation is not an appropriate assumption for factor model [10]. Therefore, in this paper, we are proposing to restrict the factors to be independent. Under this assumption, Independent Component Analysis (ICA) is an ideal method for the extraction of the factors and hence the construction the factor models [11, 12, 13, 14]. The constructed factor models are named as independent factor models.

Previous studies have applied ICA to extract independent sources from stock data [15, 16, 17]. Their major focus is on the source signals. Factors are related to seasonal variations, and prediction on the source signals is also suggested. On one hand, it is useful to know what the exact underlying factors are. On the other hand, the financial market nowaday is extremely complex and dynamic, especially due to globalization and many newly introduced indices, such as IT index. It is not an easy task to decide which variables, among so many systematic factors and marco-economic variables, should be included in the model as factors. Our method serves as a data mining technique to automatically identify the hidden factors from historical data. Unlike the previous applications of ICA in finance, our focus is not on the source signals. The sensitivities are indeed the focus of attention. Our work relates ICA to the factor model, a basic theory in finance. This serves as a linkage to the current financial theories developed based on the factor model.

2. THE FACTOR MODELS

2.1. The uncorrelated Factor Models

Multifactor model is a general form of factor model [8, 18, 19], and is the most popular model for the return generating process. The return r_i on the ith security is represented as,

$$r_i = \alpha_i + \sum_{m=1}^{k} \beta_{im} F_m + u_i \tag{1}$$

where k is the number of factors and it is a positive integer larger than zero. $F_1, F_2, ..., F_k$ are the factors affecting the returns of ith security and β_{i1}, β_{i2}, ..., β_{ik} are the corresponding sensitivities. α_i is regarded as "zero" factor that is invariant with time; u_i is a zero mean random variable of ith security. It is generally assumed that the covariance between u_i and factors F_i are zero. The factors, F_i, are uncorrelated to each other. Also u_i and u_j for security i and j are independent if $i \neq j$. For simplicity, the multi-factor model with k factors is called k-factor models.

2.2. The portfolio construction

One application of the factor model is on portfolio analysis. As we have pointed out in Section 1, uncorrelation is not a good assumption on factor model. In this section, we further illustrate our point using portfolio analysis.

Suppose we have two securities, A and B. Let r_A and r_B be their returns respectively. Without lose of generosity, we use a simple two-factor model to determine their returns as below

$$\begin{aligned} r_A &= \alpha_A + \beta_{A1} F_1 + \beta_{A2} F_2 + \mu_A \\ r_B &= \alpha_B + \beta_{B1} F_1 + \beta_{B2} F_2 + \mu_B \end{aligned}$$

The main objective of portfolio management is to construct a diversified portfolio, p, composing of a number of securities. With these two securities, we construct a portfolio, p, and its return, r_p, is defined as

$$r_p = w_A r_A + w_B r_B$$

where w_A and w_B are the weightings of the securities A and B respectively. If the portfolio is constructed a way that we hedge out the effect due to F_1, the weightings should be assigned as

$$w_A = \frac{\beta_{B1}}{\beta_{B1} - \beta_{A1}}$$

and $w_B = 1 - w_A$. In this case the return of the portfolio becomes

$$r_p = \alpha_p + \frac{\beta_{B1}\beta_{A2} - \beta_{A1}\beta_{B2}}{\beta_{B1} - \beta_{A1}} F_2 + \mu_p$$

where $\alpha_p = w_A \alpha_A + w_B \alpha_B$ and $\mu_p = w_A \mu_A + w_B \mu_B$ In this way, the portfolio return does not directly relate to F_1 any more. However, in traditional factor models, we require the factors are uncorrelated to each other. It is possible that F_1 and F_2 are uncorrelated but not independent. If F_2 depends on F_1, it is obvious that the portfolio return is still under the influence of F_1. Therefore, the typical assumption on uncorrelated factors in the factor models cannot guarantee the return of the portfolio be free from the influence of F_1. On the contrary, if the factors F_1 and F_2 are independent to each other. It is possible to construct a portfolio which is free from the influence of neither factors.

2.3. The Independent Factor Model

With the assumption of independent factors, we name the factor model as "independent factor models". Independent factor models can still be applicable in the current financial theories, which are derived based on the uncorrelation properties of the factor models. As all independent signals are also uncorrelated (the converse is not true), the factors in the independent factor models are still uncorrelated.

With independent factors, ICA is an ideal candidate for the extraction of factors. Though there are certain concerns in ICA, such as the independence of the signal extracted, our main focus lies on the application of ICA in factor models and the linkage between ICA and factor models. Any deficit in the independence of the extract signals has to be relied either on better learning algorithms proposed or better factor selection methods.

Applying ICA to independent factor models is straight forward. We have illustrated the details of the factor model construction in [20]. The security prices are first transformed into return series. Then we zero-mean the series and apply an ICA algorithm to extract the independent source signals. To construct the factor model, we select the appropriate source signals as factors and the remaining signals are regarded as residues. The expected return is also included back into the model at this stage.

3. SELECTION CRITERIA FOR FACTOR MODELS

As illustrated in [20], we have demonstrated the construction of independent factor models using N stock series. The number of factors used in the independent factor models is left undecided. The fundamentalists tend to decide k manually. Using the ICA approach, it is possible to construct k-factor model, where k is any integer value between 0 and $N-1$. Now, we have two questions. One is the choice of k. The second one is the selection of k factors from N sources. To select the value of k, we apply the minimum description length method [21, 22]. After the value of k has been determined, we sort the source signals according to certain criteria, and the first k signals are picked as factors. In the following sections, we will discuss these two steps in details.

3.1. Minimum Description Length

Under the framework of ICA, Ikeda used the minimum description length principle to select m factors in factor analysis [21, 22]. The MDL derived is shown as below

$$MDL = -L(A, \Sigma) + \frac{logN}{N}(n(m+1) - \frac{m(m-1)}{2}) \quad (2)$$

where A is the mixing matrix to the factors, Σ is the unique variance matrix of data, *i.e.* it is a diagonal matrix, N and n are the number and dimension of the observations respectively. And $L(A, \Sigma)$ is defined as,

$$\begin{aligned} L(A, \Sigma) &= -\frac{1}{2}\{tr(C(\Sigma + AA^T)^{-1}) \\ &\quad +log(det(\Sigma + AA^T)) + nlog2\pi\} \quad (3) \end{aligned}$$

where C is the covariance matrix of the observations x, *i.e.* $C = \sum xx^T/N$. There is a necessary condition for A to be estimable and this provides a bound of the number of factors, m.

$$m \leq \frac{1}{2}\{2n + 1 - \sqrt{8n+1}\} \quad (4)$$

3.2. Factor Selection

Once we have determined the value of k, we have to select the factors from the source. Up to date, a number of criteria have been used to measure the properties of the source signals. Euclidean norm is used in JADE. It measures the energetic significance of the component so that the most energetically significant component appear first[23, 24, 25]. L_∞ norm is another criterion which has been used. It focuses on the maximum value of the factors, F_i. L_∞ norm measures those ICs causing the maximum price change in the stock[17]. Kurtosis, the fourth-order cumulant, on the other hand, has also been widely used in the ICA community to measure the nongaussianity of a signal [26, 27]. A nongaussian signal is unlikely the resultant of a mixture of signals [14]. So kurtosis is also introduced to select those nongaussian signals. A gaussian random variable has zero kurtosis. Subgaussian and supergaussian variable would have positive and negative kurtosis respectively. In this paper, the absolute value of kurtosis is used to sort the factors because we want to measure nongaussianity, and we do not care if the signal is supergaussian nor subgaussian.

4. RANDOM RESIDUES

Suppose we have successfully extracted independent factors from the security prices. One remaining requirement for a factor model which we have not yet addressed is that the residue has to be random. For those models with nonrandom residues are invalid and should be rejected. Therefore, we have to check if this requirement is satisfied and the randomness of residue is estimated by the "runs test".

4.1. Runs Test

The Runs Test, also known as Wald-Wolfowitz Test, is used to test the randomness of a sequence at $100(1-\alpha)\%$ confidence level. A run is a succession of an identical class[28]. For a time series with continuous values, each data point is compared with the mean to see if it is above or below the mean We denote a point as "ABOVE" if its value is above or equals to the mean value of the whole series; otherwise it is denoted as "BELOW". If the hypothesis, H_0, that a series is random, is true, the number of runs should following a particular probability distribution. The following summarizes the testing procedure.

1. Decide the level of significance, α. In this paper, we put $\alpha = 0.05$.

2. Calculate the number of runs, u, in the series.

3. Calculate n_1 and n_2, the numbers of ABOVEs and BELOWs respectively. When n_1 and n_2 are both sufficiently large, it is reasonable to assume that the number of runs follows a normal curve with mean, μ, and standard deviation, σ; where μ and σ are defined as follows [29],

$$\mu = \frac{2n_1n_2}{n_1+n_2} + 1 \quad (5)$$

$$and$$

$$\sigma = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}} \quad (6)$$

4. Put $z = \frac{u-\mu}{\sigma}$. Using 5% level of significance, if $z \leq -1.96$ or $z \geq 1.96$, we reject H_0. Otherwise, we accept H_0.

4.2. Interpretation of z-value

Under the hypothesis test, if there are too few runs relative to the gaussian mean and standard deviation, the z-value is small and it implies that the series is having a trend. If there are too many runs, the z-value is large and the series contains many ups and downs. Therefore, the absolute z-value of the series gives us some information on randomness of the series. It is natural to suggest the use of the z values as a sorting criterion. In this respect, we include the "non-random" source signals as factors and the remaining "random" source signals would be left as residues. It is necessary to clarify that we sort the source signals according to their individual z values; whereas the runs test is applied to test the randomness of the residues, the combinations of the unused signals.

5. EXPERIMENTS AND RESULTS

In the experiment, we used 22 stocks, selected from the Hang Seng Index constitutes in Hong Kong. Daily closing prices started form 2/1/1992 to 16/5/2000 were used. Figure 1 shows the stocks' price series.

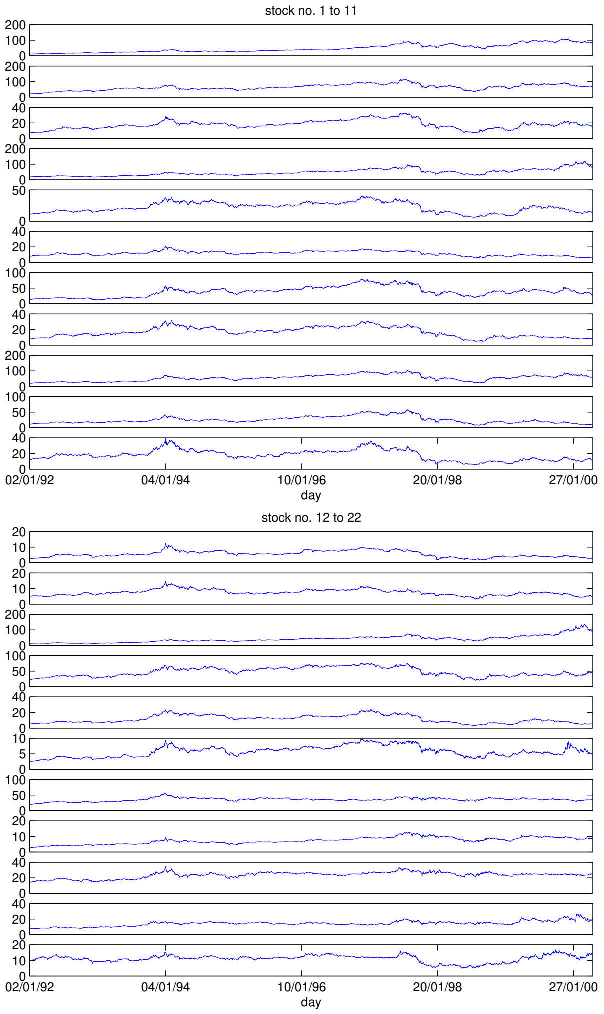


Fig. 1. The price series of the Stocks used.

5.1. Determination of k

In our experiment, we transformed our daily security prices into sequences of return series. We then applied ICA to construct our independent factor models [20]. Both JADE and FastICA had been used and they gave similar results. The next step is to perform the factor model selection. As we have illustrated in Section 3, there are two steps in this process. The first is to select the appropriate value of k for the k factor model. We computed the MDL of the factor models with different number of factors. According to equation 4, m (or k in our notation) must be less than or equal to 15. Figure 2 shows the results of factor models with different number of factors. It is observed that 8-factor model has the smallest minimum description length, 0.5796.

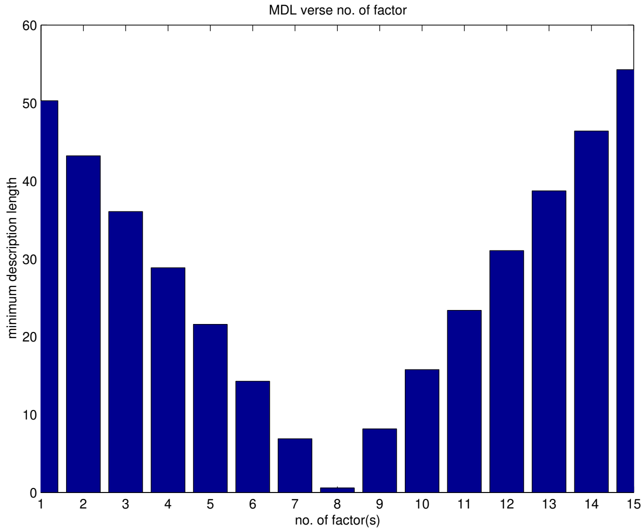


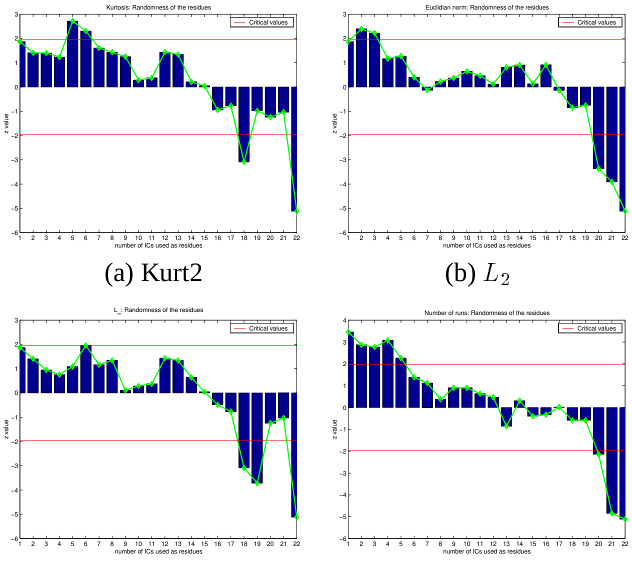
Fig. 2. Minimum description lengths of factor models with different number of factors.

5.2. Randomness of Residues using Various Sorting Criteria

Apart from the determination of the value of k, another issue we have to consider is the selection of factors into our factor model. In the rest of the paper, we demonstrate our results using only one stock, namely, New World Development Co. Ltd. The other stocks produced similar results and hence we do not display their graphs. New World Development Co. Ltd. is chosen as an example as it gives the most negative z values of -5.1134.

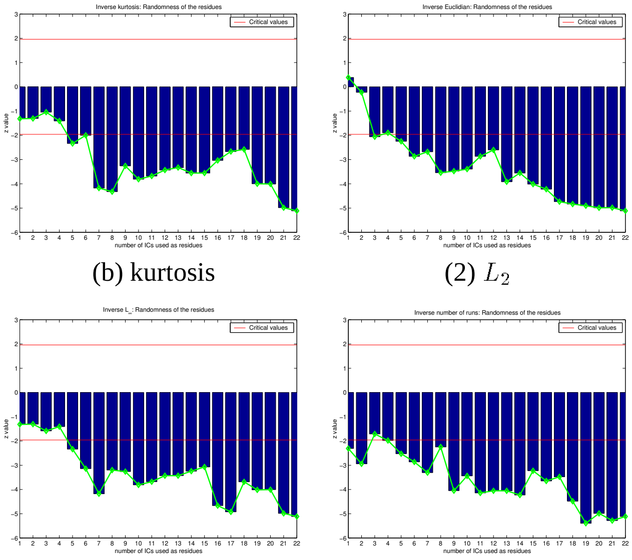
We constructed the independent factor models using the procedures in the previous section, and the factors were sorted by four different sorting criteria, kurtosis, euclidean norm, L_∞ norm and number of runs. We then examined whether the independent factor models show the property of having random residues. Although the MDL method suggested 8-factor model is the most appropriate one, we examined the residues produced by all factor models. We applied the runs test on their residues so as to investigate their randomness. Figure 3 shows the result of the runs test, *i.e.* z values of the residues of the independent factor models under different sorting criteria. Note that the x-axes of the graphs are “the number of ICs in residue”. In other words, if j is the number of independent components (IC) in residue, the corresponding k factor model is the one with $k = 22-j$. For the cases that 22 ICs are used as residue, they equivalent to applying the runs test to the original stock return. Those factor models with z values falling within the two red horizontal lines are regarded as valid factor models. From the figure, we can see the sorting criteria give similar results and that all factor models with $k = 5$ to 15 satisfy the random residue requirement, including the 8-factor model selected by MDL. By examining the results for all stocks, it is found that the results using kurtosis and L_∞ closely follow each other; whereas, the graphs corresponding to L_2 and z values show more monotonicity than the other two.

As a control experiment, we reversed the sorting orders for the four methods. The result is show in Figure 4. Here we clearly see that none of the models is valid. This gives us a positive indication that the sorting criteria play an important part in the factor selection.



(a) Kurt2 (b) L_2 (c) L_∞ (d) z-value

Fig. 3. z values from the runs tests applied to the residues of the factor models constructed under four sorting criteria.



(b) kurtosis (2) L_2 (c) L_∞ (d) z values

Fig. 4. z values from the runs tests applyed to the residues of the factor models constructed under four sorting criteria in reverse order.

6. CONCLUSIONS AND DISCUSSIONS

In financial analysis, it is more appropriate to assume the factors in factor models are independent rather than uncorrelated. Construction of this type of models, the independent factor models, is an applicational area of ICA in finance. We have applied MDL to extract 8-factor model from 22 stocks in the Hong Kong market. Among the four sorting criteria we have compared in this paper, L_∞ and L_2 have been used in previous applications. Kurtosis, is also another candidate in some general applications. The sorting using z values is particularly designed in our application. Although the four sorting methods appear to perform equally well to select the factors and it is not easy to specify which sorting method is superior, we have found that factors need to be carefully selected in order to turn them into valid factor models. This paper serves as a preliminary study of applications of ICA in factor models. In future, specially designed ICA algorithms can be proposed to replace the general ICA tools we use here. For example, we can incorporate the temporal knowledge or the random residue requirement while extracting the components.

7. ACKNOWLEDGEMENT

The work described in this paper was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administration Region, China. We would also thank Professors Oja and Cardoso for providing free downloads of FastICA and JADE respectively.

8. REFERENCES

[1] W. F. Sharpe, *Investments*, Prentice-Hall, 1981.

[2] B. Manly, *Multivariate statistical methods: A primer*, Chapman and Hall, 1994.

[3] N.F. Chen, R. Roll, and S. Ross, "Economic forces and the stock market," *Journal of Business*, vol. 59, no. 3, pp. 383–403, July 1986.

[4] A. Gordon, W. Sharp, and B. Jeffery, *Fundamentals of investments*, Prentice Hall, 1993.

[5] G. Feeney and D. Hester, "Stock market indices: A principal component analysis," *Cowles Foundation*, vol. Monograph 19(39), pp. 110–138, 1967.

[6] H. Schneeweiss and H. Mathes, "Factor analysis and principal components," *Journal of multivariate analysis*, vol. 55, pp. 105–124, 1995.

[7] J. Utans, W.T. Holt, and A.N. Refenes, "Principal components analysis for modeling multi-currency portfolios," in *Proceedings of the Fourth International Conference on Neurals Networks in the Capital Markets, NNCM-96*, 1997.

[8] S. Brown, "The number of factors in security returns," *The Journal of Finance*, vol. 44, no. 5, pp. 1247–1262, December 1989.

[9] G. Connor and R. Korajczyk, "Performance measurement with the arbitrage pricing theory a new framework for analysis," *Journal of financial economics*, vol. 15, pp. 373–394, 1986.

[10] P. Embrechts, A.J. McNeil, and D. Straumann, "Correlation and dependence in risk management: Properties and pitfalls," in *to appear in RISK Management: Value at Risk and Beyond*, M. Dempster, Ed. 2001, Cambridge University Press.

[11] P. Comon, "Independent component analysis, a new concept ?," *Signal Processing*, vol. 36, no.3, pp. 287–314, April 1994.

[12] T.-W. Lee, *Independent Component Analysis: Theory and Applications*, Kluwer Academic Publishers, Boston, first edition, 1998.

[13] A. Hyvärinen, "Survey on independent component analysis," in *Neural Computing Surveys 2*, 1999, pp. 94–128.

[14] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, Issue 4, pp. 411–430, 2000.

[15] K. Kiviluoto and E. Oja, "Independent component analysis for parallel financial time series," in *International Conference on Neural Information Processing, ICONIP'98*, October 1998, pp. 895–898.

[16] S. Mălăroiu, K. Kiviluoto, and E. Oja, "ICA preprocessing for time series prediction," in *Proceedings of ICA2000*, May 2000.

[17] A. Back and A. Weigend, "A first application of independent component analysis to extracting structure from stock returns," *International Journal of Neural Systems*, vol. 8, pp. 473–484, 1997.

[18] G. Connor and R. Korajczyk, "A test for the number of factors in an approximate factor model," *The Journal of Finance*, vol. 48, no. 4, pp. 1263–1291, 1993.

[19] H. Markowitz, *Portfolio selection, efficient diversification of investment*, Blackwell Publishers Ltd, 1991.

[20] S.M. Cha and L.W. Chan, "Applying independent component analysis to factor model," in *Intelligent Data Engineering and Automated Learning - IDEAL 2000, Data Mining, Financial Engineering and Intelligent Agents*, L.W. Chan K.S. Leung and H. Meng, Eds. 2000, pp. 538–544, Springer.

[21] S. Ikeda, "ICA on noisy data: A factor analysis approach," in *Advances in independent component analysis*, M. Girolami, Ed., chapter 11, pp. 201–215. Springer, 2000.

[22] S. Ikeda, "Factor analysis preprocessing for ICA," in *Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation*, 2000, pp. 1249–1252.

[23] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non-gaussian signals," *IEE Proceedings-F*, vol. 140, no. 6, pp. 362–370, December 1993.

[24] J.-F. Cardoso, "High-order contrasts for independent component analysis," *Neural Computation*, vol. 11(1), pp. 157–192, 1999.

[25] J.-F. Cardoso, "Blind signal separation: statistical principles," in *Proceedings of the IEEE, special issue on blind identification and estimation*, R.-W Liu and L Tong, Eds., October 1998, vol. 86, Issue 10, pp. 2009–2025.

[26] P. Huber, "Project pursuit," *The Annals of Statistics*, vol. 13(2), pp. 435–475, 1985.

[27] M. Jones and R. Sibson, "What is project pursuit," *Journal of the Royal Statistical Society, ser. A*, vol. 150, pp. 1–36, 1987.

[28] J. Freund, *Mathematical statistics*, Prentice Hall, Upper Saddle River, New Jersey, sixth edition, 1999.

[29] J. Gibbons, *Nonparametric methods for Quantitative Analysis*, American Sciences Press, 1985.