

**A partitioning approach to the specification of large VAR models:
With an application to the DAX30 series**

Michael A. Hauser

Vienna University of Economics and Business Administration

hauser@wu-wien.ac.at

Presented at the EC² Conference on Causality and Exogeneity, Louvain-la-Neuve, Dec. 13-16, 2001. Preliminary version.

The paper proposes a new specification approach for large VAR models in standard form by partitioning the left hand side variables into a fixed number of groups, so that for each group a “best” set of explanatory (lagged endogenous) variables is given.

We start with a collection of different sets of lagged endogenous variables, which are valued with respect to the predictive power (Geweke(1982)) for all endogenous variables. The grouping of the left hand side variables and the selection of the “best” explanatory sets is performed simultaneously by the partitioning algorithm which solves the p -median problem of the uncapacitated facility location problem. In order to improve the first collection of explanatory variable sets a systematic iterative search is performed.

An application to 26 daily stock return series of the DAX30 is given.

Keywords: predictability; block causality; clustering. **JEL code:** C32

Acknowledgment: I am indebted to Manfred Deistler and Benedikt Pötscher for helpful comments.

Address of the author:

Department of Statistics, Vienna University of Economics and Business Administration, Augasse 2-6, A-1090 Vienna, Austria.

Tel.: +43 1 31336/4759, FAX: +43 1 31336/738, E-Mail: hauser@wu-wien.ac.at

1 Introduction

Large vector autoregressive regressions, VAR, are notoriously difficult to specify, especially if structural information is vague. For different solutions see Lütkepohl(1991, chp.5), Amisano and Giannini(1992), etc. Originally, Sims(1980, p.14f) has advocated for the inclusion of all variables as endogenous and construct a large system of dynamic equations in order to overcome haphazardly introduced restrictions in macroeconomic models. Unfortunately, this approach is limited in practice by the number of observations when the number of variables is large due to its “profligate parameterization”. Even in structural VAR and vector error correction models only a small number of variables is commonly used. For a recent reference see Jacobson et al.(2001).

A related but less demanding formulation of this problem is the construction of leading indicators for a set of (macro) economic time series. According to Burns and Mitchell(1946,p.3) a few indicators (e.g. dynamic linear combinations of some of the observed series) should capture common movements in most of the variables. Stock and Watson(1989), e.g., undertook this task and started with 280 preselected series and reduced them to 55 after investigation of the coherence, phase lead and Granger causation with respect to 15 main economic series. “Selecting a few “best” variables from this list is a daunting task: in theory over 200 million seven-variable indexes could be performed from these 55 series. We simplified this problem by adopting a modified stepwise regression procedure for constructing an ... indicator based on a relatively few series.”(p.365). (There are clearly several other approaches to this problem: Lahiri and Moore(1991), or recently, Forni et al. (2000).)

In this paper we will develop a procedure to detect a number of different driving forces behind a large number of variables. It works in an “automatic” and systematic way. We start with the notion that a series may be important to explain some but

usually not all of the variables included in the investigation, and some variables, on the other hand, may not explain any of the others. Further, we impose some structure on the relation between the endogenous and explanatory variables: Whole groups of the endogenous variables may be explained by one and the same set of explanatory variables. The groups of the endogenous ones are disjunctive, the explanatory ones need not be. So there may be common components in the explanatory sets among at least some of the groups, and clearly group specific effects. Individual characteristics of single endogenous variables may be specified a priori.

Explanatory power will be measured by Granger causality, predictability as proposed by Geweke(1982), respectively. At the beginning of our analysis we design a large collection of potential sets of right hand side variables (in our case lagged endogenous variables, but in general we are not limited to them) and choose a number of k groups in which we want to partition the left hand side variables. So, all variables in the same group are modeled by the same explanatory variables. A partitioning algorithm (it solves the p -median problem of the uncapacitated facility location problem) simultaneously chooses k sets of explanatory variables and allocates the endogenous variables into k groups by maximizing overall predictability. Further, a search for larger explanatory sets is performed by a forward and backward selection procedure based on already chosen sets by adding and discarding variables.

Our approach may be seen as a step towards solving the specification problem of VARs. It yields the k most important driving forces, linear combinations of - hopefully few - observed variables. The driving forces – especially if they are common to some groups – may also turn out to be promising candidates for leading indicators. It may also be seen as an alternative approach to the clustering of time series, cp. Shumway and Stoffer(2000, chp.5) or Galeano and Pena(2000), and is not limited to VARs but may also be generalized to multidimensional models if single equation estimators are applicable. The procedure will be illustrated with returns of 26 of the

DAX30 series for the period April, 26 1999 to May, 21 2001.

Section 2 discusses the simple predictability measure based on the concept of causality and argues that it may be used to compare models of different endogenous variables. Section 3 states the partitioning problem. Section 4 describes the iterative search procedure for sets of explanatory variables. Section 5 illustrates our approach for the returns of the DAX30 series. The appendix lists the partitioning algorithm.

2 The predictability measure

We consider two nondegenerate, weakly stationary processes x_t and y_t . y_t is predictable by x_t if there exists a projection of y_t on $\{y_{t-s}, s \geq 1\}$ and $\{x_{t-s}, s \geq 1\}$

$$y_t = \sum_{i=1}^{\infty} \alpha_i y_{t-i} + \sum_{i=1}^{\infty} \beta_i x_{t-i} + u_{2t}$$

with some $\beta_i \neq 0$ and u_{2t} a possibly degenerate, weakly stationary process. The α_i may be zero. The term $\beta_0 x_t$ is not included, since we will focus on prediction. The possible constant is omitted to keep the presentation simple. The reference model with no predictability by the past of x_t is

$$y_t = \sum_{i=1}^{\infty} \alpha_i y_{t-i} + u_{1t}$$

The strength of predictability of y by x is then measured by

$$p(x; y) = \log\left(\frac{\sigma_1^2}{\sigma_2^2}\right)$$

where $\sigma_1^2 = V(u_{1t})$ and $\sigma_2^2 = V(u_{2t})$. This is the percentage reduction in the error variance, if lags of x_t are additionally included in the model. Geweke(1982) denoted $p(x; y)$ as measure of linear feedback which is a monotonic transformation of the strength of causality from x to y , $1 - \frac{\sigma_2^2}{\sigma_1^2}$, as defined by Granger(1963). Some advantageous properties of $p(x; y)$ are, cp. Geweke(1982): It is invariant with respect

to scaling of x and y . It remains unchanged if x and y are premultiplied by different invertible lag operators. It also allows the comparison in predictability of different y 's, y_1, y_2 , by different x 's, x_1, x_2 , $p(x_1; y_1)$ and $p(x_2; y_2)$. And, under Gaussianity $p(x; y)$ is asymptotically χ^2 -distributed under the null hypothesis of $\beta_i = 0$ for all i .

The finite sample approximations of our models are

$$y_t = \sum_{i=1}^{m_\alpha} \alpha_i y_{t-i} + u_{1t}$$

$$y_t = \sum_{i=1}^{m_\alpha} \alpha_i y_{t-i} + \sum_{i=1}^{m_\beta} \beta_i x_{t-i} + u_{2t}$$

In order to simplify notation we do not indicate in the following that the parameters and the error term in each model depend on the corresponding dependent variable as well as on the x -variables in the model.

Different numbers of parameters are taken into account by applying Akaike's information criterion, AIC, which has been recently shown to have good model selection properties in large dimensional systems, Gonzalo and Pitarakis(2000). The AIC will be approximated by

$$AIC(m) = -\log(\hat{\sigma}^2) - 2\frac{m}{n}$$

with $\hat{\sigma}^2$ the minimized residual variance, m the number of parameters in the model, and n the length of the series. Thereby estimation will be performed by ordinary least squares sacrificing the asymptotic efficiency in the VAR case below. Consistency and asymptotic normality is still guaranteed. Predictability is then measured by

$$p_{AIC}(x; y) = AIC(m_2) - AIC(m_1) = \log\left(\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}\right) - 2\left(\frac{m_2}{n_2} - \frac{m_1}{n_1}\right)$$

where $m_1 = m_\alpha + 1$ and $m_2 = m_\alpha + m_\beta + 1$, and n_1, n_2 are possibly different effective sample sizes. $p_{AIC}(x; y)$ is the increase of the AIC value if the reference model is augmented by m_β parameters.

We generalize in order to allow a comparison with more complex models of the type

$$y_t = \sum_{i=1}^{m_\alpha} \alpha_i y_{t-i} + \sum_{i=1}^{m_\beta^{(1)}} \beta_{1i} x_{1,t-i} + \dots + \sum_{i=1}^{m_\beta^{(l)}} \beta_{li} x_{l,t-i} + u_{3t}$$

m_α and $m_\beta^{(j)}$ are fixed for all j at the beginning of our investigation. p_{AIC} is - as above - the increase in the AIC with respect to the reference model

$$p_{AIC}(x_1, \dots, x_l; y) = AIC(m_3) - AIC(m_1) = \log\left(\frac{\hat{\sigma}_1^2}{\hat{\sigma}_3^2}\right) - 2\left(\frac{m_3}{n_3} - \frac{m_1}{n_1}\right)$$

with $m_3 = \sum_j m_\beta^{(j)} + 1$.

Given a set of endogenous variables $\{y_i, i = 1, \dots, N\}$, a set of explanatory (in our case lagged endogenous) variables $\{x_i, i = 1, \dots, P\}$ and a collection of $(M + 1)$ sets of combinations of explanatory variables $\{\mathcal{M}_i, i = 0, \dots, M\}$. Then we arrange the predictability measures for all pairs (\mathcal{M}_i, y_j) , $i = 0, \dots, M$, $j = 1, \dots, N$ in a $(M + 1) \times N$ predictability matrix. This matrix is in general not symmetric, even if the sets of explanatory variables consist only of single lagged endogenous variables.

Example: Given a selection of 5 daily return series of the German DAX30 stocks for the period April, 27 1999 to May, 21 2001. The length of the series is 525. We choose the same series as explanatory variables lagged once. The models considered are the simplest possible ones. $N = P = M = 5$.

$$y_t = c + u_{1t} \quad \text{and} \quad y_t = c + \beta_1 x_{i,t-1} + u_{2t} \quad y_t = y_{j,t} \quad x_{i,t} = y_{i,t} \quad i, j = 1, \dots, 5.$$

The associated predictability matrix is given in Table 1.

Table 1: Predictability measures ($\times 100$) for 5 stock return series (Labels: Reuter's Instrument Code (RIC) omitting the extension ".F")

$p_{AIC}(\mathcal{M}; y)$		y_1	y_2	y_3	y_4	y_5
		DRSDn	EONG	SAPG_p	TKAG	VOWG
\mathcal{M}_0		0.000	0.000	0.000	0.000	0.000
\mathcal{M}_1	DRSDn	0.028	-.009	-.009	-.004	0.030
\mathcal{M}_2	EONG	0.005	-.006	-.010	-.005	0.020
\mathcal{M}_3	SAPG_p	0.004	0.012	0.009	-.006	0.011
\mathcal{M}_4	TKAG	0.004	-.008	-.005	-.006	0.015
\mathcal{M}_5	VOWG	0.006	-.002	-.010	-.006	0.027

\mathcal{M}_0 denotes the reference models of the single y variables. Comparing the values of the first column, we obtain the best predictor variables for y_1 immediately, set \mathcal{M}_1 . For y_2 this is set \mathcal{M}_3 , for y_3 also set \mathcal{M}_3 , for y_4 set \mathcal{M}_0 , and for y_5 set \mathcal{M}_1 . So only three of the 6 sets are enough, if models with single explanatory variables are sought for each y : \mathcal{M}_1 may be used for forecasting y_1 and y_5 , \mathcal{M}_3 may be used for forecasting y_2 and y_3 . \mathcal{M}_0 is best for y_4 .

If we want to allow only for one or two sets, the analysis becomes more difficult. As far as the number of variables is small, enumeration of all possible combinations is feasible. If k , the number of groups, which is equal to the number of different sets to be chosen, is one, then the six possibilities may be valued according to the sum of the predictability measures they yield. These are the sums of the lines of the predictability matrix resulting in a maximal line sum for \mathcal{M}_1 , 0.036.

In case of $k = 2$ there are $\binom{6}{2}$ possibilities to choose from the \mathcal{M} sets. And, for each choice there exist $(2^5 - 2)$ different mutually disjoint groups of the 5 y variables. The best choice according to the criterion of the maximization (over the different partitions) of the sum (over the groups) of the (within group) predictability measures turns out to be $\{\mathcal{M}_1, \mathcal{M}_3\}$ with a value of 0.075. $\{y_1, y_4, y_5\}$ and $\{y_2, y_3\}$ are the

corresponding groups.

The overall maximum of the objective function, 0.085, is obtained for a partition with three groups.

3 Partitioning algorithm

The problem to find the optimal partition for a general but fixed k is well known in the operations research literature as the p -median problem of the uncapacitated facility location problem. See e.g. Mirchandani and Francis(1990). The problem is usually formulated as follows: “Suppose that you plan to build a new chain of stores in a given city, and you have identified potential store sites in a different number of neighborhoods. Further assume that the demand in each neighborhood of the city is known. If you want to build exactly p stores, where should you locate them in order to minimize the traveling distances of your customers?”

In our case the potential store sites are the potential sets of explanatory variables, the different neighborhoods are the series to be predicted, the distance is a reciprocal function of our predictability measure, $-p_{AIC}(\mathcal{M}; y)$, and $p = k$. Formally the problem is an integer optimization problem. Formulated in $p_{AIC}(\mathcal{M}_i; y_j)$ it reads as maxisum problem

$$\max_I \sum_{j=1}^N \max_{\mathcal{M}_i \in I} p_{AIC}(\mathcal{M}_i; y_j) \quad \text{with} \quad |I| = k$$

Thereby $I \subset \{\mathcal{M}_0, \dots, \mathcal{M}_M\}$ and $|I|$ denotes the number of elements in I . This is equivalent to

$$\begin{aligned} & \max_{\gamma} \sum_{i,j} \gamma_{i,j} p_{AIC}(\mathcal{M}_i; y_j) \\ \text{s.t.:} \quad & \sum_i \gamma_{i,j} = 1, \quad \delta_i \geq \gamma_{ij} \quad \forall j \quad \text{and} \quad \sum_i \delta_i = k \quad \gamma_{i,j}, \delta_i \in \{0, 1\} \end{aligned}$$

The problem is known to be NP -hard, which means that any polynomial-time algorithm would also solve the associated NP -complete problems. However, no polynomial-

time algorithm is known.

There are large number of computational solutions (including commercial ones, e.g. CPLEX) for this problem, see e.g. Mirchandani and Francis(1990), or Guha and Khuller(1999), etc. An algorithm with a straightforward structure and good practical performance in empirical studies, Korupolu et al.(2000), is the local search (greedy) heuristic. We adapt one of this type, the PAM algorithm by Kaufman and Rousseeuw(1990), which has originally been developed for k -medoid clustering for our purpose. Our version copes with “similarity” matrices which have to be read only column wise and allows for “medoids” (the chosen sets of explanatory variables) which are no longer in the same “cluster” (group) as its “members” (endogenous variables). (A description of the algorithm is given in the appendix. The modified FORTRAN code is available from the author on request.)

4 Search for more complex models

The p -median solution supplies us with k optimal sets of explanatory variables chosen from a given collection together with k mutually disjoint groups of endogenous variables. In the following we perform a systematic search procedure for more complex sets of explanatory variables similar to Stock and Watson(1989), who are, however, not explicit in this respect. (Cp. also the forward selection and backward elimination technique of stepwise regression analysis. Draper and Smith(1981).) Our first collection consists of all sets with one x variable (appropriately lagged). The optimal solutions therefrom are each augmented by one x variable, yielding $k P$ different sets entering the partitioning procedure. (Second step). The resulting k sets are also augmented by one variable resulting in optimal sets with at most 3 predictor variables. Then a backward search technique is applied: The new collection consists out of all possible 2 variable sets of the 3 variable solution with the optimal 3 variable solution

included as well. (Third step). These forward search and backward elimination procedures are repeated until the chosen sets do not change any more.

Alternatively, we may implement the bottom-up or the top-down subset finding procedure of Penm and Terrell(1982), which gives for each equation an “optimal” model. The model for variable y_{it} is found by the bottom-up search by entering the first variable lagged once and testing whether it should be included in the model, resulting possibly in a new better model. Using this as reference model it tests for the second variable once lagged, etc., until the last variable, y_{Nt} , with the maximal lag considered. The procedure is repeated for all y_{it} , $i = 1, \dots, N$.

Using these N “optimal” sets of x variables also for the other endogenous variables we obtain a single predictability matrix and the partition algorithm has to be executed only once.

So far we obtain an “optimal” solution for fixed k and fixed lag structure, $m_\alpha, m_\beta(1), \dots, m_\beta(l)$. In applications neither k nor the maximal lag orders are known. The lag orders may be chosen simply according to the maximized objective function for fixed k , since the selection rule of the AIC is incorporated therein.

The choice between different k , however, is essentially the same problem as the determination of the number of clusters in cluster analysis. However, this is one of the insufficiently solved problems, see e.g. Jain and Dubes(1988). One possibility is to inspect the maximized objective functions for increasing values of k . Then the number of clusters is viewed as optimal where the last large increase in the objective function is observed. Then it may be conjectured that reasonably coherent groups are obtained. Any group more might split the already existing ones in an artificial way.

5 Dynamics in the DAX30 stock returns

In the following we apply our approach to 26 daily stock return series of the DAX30 for the period April, 27 1999 to May, 21 2001. The length of the series is 525. (Four of the DAX30 series have been omitted due to a smaller number of observations.) $N = 26, n = 525$. The series are labeled according to Reuter's Instrument Code (RIC), omitting the extension ".F". The main purpose is to illustrate the method developed above and not to give a detailed exploratory analysis of the German stock market during the last two years. The reference models and the augmented ones are

$$y_t = c + u_{1t} \quad \text{and} \quad y_t = c + \sum_{i=1}^{m_\beta^{(1)}} \beta_{1i} x_{1,t-i} + \dots + \sum_{i=1}^{m_\beta^{(l)}} \beta_{li} x_{l,t-i} + u_{3t}$$

since stock returns are hardly autocorrelated. The lag orders are assumed to be the same for all variables: $m_\beta^{(1)} = \dots = m_\beta^{(l)}$. As model search procedure forward selection and backward elimination is used. We will repeat the experiment for lag orders between 1 and 6 and $k = 2, \dots, 8$. Table 2 gives the maximized values of the objective function and the associated number of parameters.

Table 2: Maximal value of the objective function ($\times 100$) and number of variables for different choices of k and lag orders

k	Lag order					
	1	2	3	4	5	6
2	0.6004	0.6017	0.5225	0.4403	0.4424	0.3752
	104	118	104	66	86	98
3	0.9265	0.9166	0.7642	0.6044	0.6893	0.6046
	142	128	152	86	126	128
4	1.1013	1.0713	0.8642	0.7295	0.8341	0.7158
	144	146	107	134	166	164
5	1.2672	1.3198	1.1593	0.8905	0.9111	0.8623
	138	206	161	182	161	212
6	1.3694	1.5054	1.2638	1.0187	1.0155	0.9690
	122	192	227	178	176	224
7	1.4225	1.5940	1.4280	1.0830	1.2193	1.0368
	115	198	209	178	206	248
8	1.5190	1.6948	1.5418	1.1616	1.3480	1.0657
	138	206	215	190	226	254

(Computing time for the solution with 2 lags and 5 groups on a 500MHz PC using FORTRAN is 1 minute 30 seconds.)

Comparing the maximal values for fixed k two lags seem to be appropriate (except for $k = 3, 4$). In absence of a better simple criterion we compare the increases in the objective functions for the best choice of the lag order for increasing k . They are: 0.325, 0.175, 0.218, 0.186, 0.089 and 0.101. The largest increase (apart from the first value) is observed for the step from 4 to 5 groups, followed by the step from 5 to 6. (We have also checked larger k values and found that the increases are getting smaller, with no more potential interesting change in the objective function.) So we tend to propose either partition with $k = 5$ or 6. The detailed results for partition 5 are given in Tables 3.

Table 3: The chosen models of the partition with 5 groups: Explanatory variables and group members

Group	Explanatory variables (lagged) (x_1, \dots, x_l)	Group members (y_j)
1	BASF, SCHG	ADSG, BAYG, DTEGn, PRSG, TKAG
2	EONG, TKAG	FMEG, HNKG_p, LHAG, MANG, SCHG
3	ALVG, DCXGn, DRSDn	BASF, DCXGn, DBKG, VOWG
4	ADSG, CBKG, DTEGn, DRSDn	ALVG, CBKG, DRSDn, HVMG, LING, MUVGn, RWEG
5	BAYG, BMWG, DGXG, SAPG_p, TKAG, VOWG	BMWG, DGXG, EONG, SAPG_p, SIEGn

None of the chosen models yields a negative predictability value. So, predictability is improved for all endogenous variables with respect to the reference models. (This need not be so in general as seen in Table 1.) Though the groups are mutually disjoint per definition, the sets of the explanatory variables need not be: DRSDn e.g. appears in set 3 and 4. Despite the fact that single return series are hardly autocorrelated, there seem to be relevant dynamic interactions between the stock returns during the observation period. For members of group 5 even 6 variable models are chosen, and for members of group 4 models with 4 explanatory variables. On the other hand, 11 of the 26 variables are not used for predictive purposes. The total number of parameters of the so specified VAR is only 206 compared to 1404 of the complete model. A detailed interpretation of the results is difficult due to the heterogeneity of the stocks. They may be considered, e.g., to belong to at least 13 different sectors. The algorithm, however, is designed to map each endogenous variable to a group, even if it is an atypical one.

6 Discussion

Our partitioning approach assumes a kind of block causality: Variables of each group are caused by the same set of variables. However, the sets of explanatory variables need not be disjunctive. (This might be an interesting qualification of vector causality

as discussed by Dufour and Renault(1998).)

Our objective function measures the percentage decrease in the error variance after adding variables to the reference model. Alternatively, one could maximize the sum of the levels of the AIC instead of their increases. In this respect Geweke's measure may be seen as percentage criterion of error variance reduction, while the levels of the AIC correspond to the level of error variance reduction.

Some modifications of our procedure could be installed without effort: (a) Different reference models for different endogenous variables may be assumed. E.g. some variables should be explained by their own past, others not. (b) The partition algorithm could restrict the group membership to endogenous variables which can be explained to a sufficient extent, i.e. $p_{AIC}(.,.) > p_0$ should hold within each group. This would help to obtain more clear cut groups, and to identify atypical variables.

However, a caveat should be kept in mind: The number of models estimated and compared in the search procedure should be in a reasonable relation to the number of observations available in order to avoid overfitting. If observations are scarce the top-down or bottom-up search procedure of Penm and Terrell(1982) seem to be feasible. They require $N^2(m_{max} + 1)$ model comparisons, where m_{max} ist the maximal lag considered.

In extension of this paper we are going to compare also the projection modulus approach by Zhang and Terrell(1997) which promises to be very efficient for subset VARs.

Potential empirical applications of our approach seem to be numerous: Dynamics between industry sectors, between the monetary and real sector, between stock return volatilities, exchange rates, etc. Our example, e.g., may be viewed as direct generalization of the multi-index models for forecasting stock markets, cp. Eun and Resnick(1992) and Lam et al. (1994).

References

- Amisano, Gianni, and Giannini, Carlo, 1992, Topics in Structural VAR Econometrics, Springer, New York.
- Brillinger, D.R., 1981, Time Series: Data Analysis and Theory, 2nd ed., San Francisco: Holden-Day.
- Diebold, Francis X. and Kilian, Lutz, 2001, Measuring Predictability: Theory and Macroeconomic Applications, Journal of Applied Econometrics, forthcoming.
- Draper, Norman and Smith, Harry, 1981, Applied Regression Analysis, New York: Wiley.
- Dufour, Jean-Marie and Renault, Eric, 1998, Short run and long run causality in time series: Theory, Econometrica, 66, 1099-1125.
- Eun, Cheol S., and Resnick, Bruce G., 1992, Forecasting the correlation structure of share prices: A test of new models, Journal of Banking and Finance 16, 643-656.
- Forni, Mario, Hallin, Marc, Lippi, Marco and Reichlin, Lucrezia, 2000, Coincident and leading indicators for the EURO area, Economic Journal, forthcoming.
- Galeano, Pedro, and Pena, Daniel, 2000, Multivariate Analysis in Vector Time Series, Working paper, Universidad Carlos III de Madrid.
- Geweke, John, 1982, Measurement of linear dependence and feedback between multiple time series, J. Am. Stat. Assoc., 77, 304-313.
- Gonzales, Jesus and Pitarakis, Jean-Yves, 2000, Lag Length Estimation in Large Dimensional Systems, Working Paper, Department of Economics, University of Reading.
- Granger, Clive W.J., 1963, Economic processes involving feedback, Information and Control, 6, 28-48.
- Guha, Sudipto and Khuller, Samir, 1999, Greedy strikes back: Improved Facility Location Algorithms, Journal of Algorithms, 31, 228-248.
- Jain, Anil K., and Dubes, Richard C., 1988, Algorithms for Clustering Data, Prentice-Hall, Englewood Cliffs.
- Kaufman, Leonard and Rousseeuw, Peter J., 1990, Finding Groups in Data, Wiley,

New York.

Kakizawa, Yoshihide, Shumway, Robert H. and Taniguchi, Masanobu, 1998, Discrimination and clustering for multivariate time series. *J. Am. Stat. Assoc.* 93, No.441, 328-340.

Korupolu, Madhukar R., Plaxton, C. Greg, and Rajaraman, Rajmohan, 2000, Analysis of a Local Search Heuristic for Facility Location Problems, *Journal of Algorithms*, 37, 146-188.

Lahiri, Kajal, and Moore, Geoffrey H., (eds.) 1991, *Leading economic indicators, New approaches and forecasting records*, Cambridge University Press, Cambridge.

Lam, Kin, Mok, Henry M.K., Cheung, Iris, and Yam, H.C., 1994, Family groupings on performance of portfolio in the Hong Kong stock market, *Journal of Banking and Finance* 18, 725-742.

Lütkepohl, H., 1991, *Introduction to Multiple Time Series Analysis*, Springer, New York.

Mirchandani Pitu B. and Francis, Richard L., 1990, *Discrete Location Theory*, Wiley, New York.

Penm, J.H.W., and Terrell, R.D., 1982, On the recursive fitting of subset autoregressions, *Journal of Time Series Analysis*, 3, 43-59.

Penm, J.H.W., and Terrell, R.D., 1984, Multivariate subset autoregressive modelling with zero constraints for detecting "overall causality", *Journal of Econometrics*, 24, 311-330.

Reichlin, Lucrezia, 2000, Extracting business cycle indexes from large data sets: aggregation estimation, identification, Paper presented at the World Congress of the Econometric Society, Seattle, Aug. 2000.

Rissanen, Jorma, 1989, *Stochastic Complexity in Statistical Inquiry*, World Scientific Co. Pte. Ltd., Singapore.

Stock, J.H. and Watson, M.W., 1989, New Indexes of Coincident and Leading Economic Indicators, *NBER Macroeconomics Annual* 1989, 351-94.

Shumway, Robert H. and Stoffer, David, S., 2000, *Time Series Analysis and Its Ap-*

plications, Springer, New York.

Zhang, Xichuan, and Terrell, R. Deane, 1997, Projection Modulus: A new Direction for Selecting Subset Autoregressive Models, *Journal of Time Series Analysis*, 18, 195-212.

Appendix

The partitioning algorithm

We reformulate the description of the partitioning PAM algorithm (as given in Kaufman and Rousseeuw(1990, p.102-104) for unidirectional similarity matrices. We distinguish between reference objects, which are candidates for medoids, and objects, which are actually clustered. Both sets need not be the same. The dissimilarities are denoted $d(m, j)$, $m = 1, \dots, mm$, $j = 1, \dots, nn$ with m in reference object set, j in object set. ($d(j, m)$ is without a meaning.) The objective function is (as before) the sum of the dissimilarities within the clusters, which has to be minimized.

The algorithm consists of two phases the BUILD and SWAP phase. It is based on dissimilarity measures. In the BUILD step an initial set of K (fixed) reference medoids is chosen in order to perform a first clustering. The SWAP step improves this choice.

BUILD:

(1) Choose the central reference object. This minimizes the dissimilarities to all objects among the reference objects. We denote it by m_0 , $M_0 = \{m_0\}$.

(2)

(2.1) Consider a reference object m_1 which has not yet been selected. $m_1 \notin M_0$.

(2.2) Consider a non selected object j . Calculate its dissimilarity with the most similar previously selected object, $\min_m d(m, j)$. Find its dissimilarity with object m_1 , $d(m_1, j)$, and take the difference

$$D_j = \min_{m \in M_0} d(m, j) - d(m_1, j)$$

(2.3) Is this difference positive, object j will contribute to the decision to select object m_1 . Therefore we define

$$C_{m_1, j} = \max[\min_{m \in M_0} d(m, j) - d(m_1, j), 0]$$

(2.4) Calculate the total gain obtained by selecting reference object m_1 , $\sum_j C_{m_1, j}$.

And, choose the not yet selected reference object m which minimizes this sum

$$\min_{m \notin M_0} \sum_j C_{m,j}$$

Augment M_0 by the new element m .

(3) Repeat step (2) until k reference objects are found.

SWAP:

All reference objects $m_1 \notin M_0$ are investigated whether it would be better to have one of them in M_0 instead.

(1) Consider an $m_1 \notin M_0$.

(2) Consider an $m_0 \in M_0$.

(3) Consider an object j and calculate its contribution $C_{m_0, m_1, j}$ to the swap:

(3.1) If $d(m_0, j)$ and $d(m_1, j)$ are larger than any other $d(m, j)$, $m \in M_0$, $C_{m_0, m_1, j} = 0$.

(3.2) If $d(m_0, j) \leq \min_{m \in M_0} d(m, j) - d(m_1, j) = D_j$ two situations arise:

(3.2.a) Let E_j be defined as the second smallest dissimilarity from a m in M_0 to object j . If m_1 is closer to j than to the second closest reference object m , with $m \in M_0$: $d(m_1, j) < E_j$ then the contribution of j to the swap between objects m_0 and m_1 is

$$C_{m_0, m_1, j} = d(m_1, j) - d(m_0, j)$$

(3.2.b) If m_1 is at least as distant from j as the second best reference object, i.e. $d(m_1, j) \geq E_j$ then the contribution of object j to the swap is

$$C_{m_0, m_1, j} = E_j - D_j$$

(3.2.c) If m_0 is more distant from j than at least one of the other reference objects in M_0 , but with m_1 closer to j than any other reference objects in M_0 . then the

contribution of j to the swap is

$$C_{m_0, m_1, j} = d(m_1, j) - D_j$$

(3.3) Select the pair (m_0, m_1) which

$$\min_{m_0, m_1} \sum_j C_{m_0, m_1, j}$$

(3.4) If the minimum in (3.3) is negative the swap is carried out and the algorithm returns to step (1). If the value is positive or 0, the procedure stops.