Business Cycle Prediction Using Support Vector Methods

Kai Vogtländer
Fachbereich Statistik
Universität Dortmund

Claus Weihs *
Fachbereich Statistik
Universität Dortmund

March 2000

Abstract

This paper illustrates the Support Vector Method for the classification problem with two and more classes. In particular, the multi-class classification Support Vector Method of Weston and Watkins (1998) is correctly formulated as a quadratic optimization problem.
Then, the method is applied to the problem of predicting business phases of the German economy. The generated support vectors are interpreted, in particular with respect to whether they are able to characterize business phase switches. Finally, the classification power of the Support Vector Method and of Linear Discriminant Analysis are compared.
The results are two-fold. On the one hand, after the analysis of the results of this study it appears questionable that the Support Vector Method delivers an interpretable (dimension independent) data reduction by identifying the support vectors. Indeed, the support vectors did not appear to be sufficient to characterize the switches between the business phases.
On the other hand, the classification power of the Support Vector Method was distinctly better than with Linear Discriminant Analysis. Note however that the Support Vector Method needs very much more computation time than Linear Discriminant Analysis.

KEYWORDS: support vector method, multi-class classification linear discriminant analysis, business cycle analysis

* e-mail: weihs@statistik.uni-dortmund.de

1. Introduction

On the one hand, lately Support Vector Methods got more and more popular, especially in **computer science**, as an implementation of Vapnik's (1979, 1995, 1998) **learning theory** for binary classification. On the other hand, in **statistics** other classification techniques stay the most popular, namely discrimination methods and decision tree methods. In a way, computer science took the lead in a field occupied in history by statistics, because statistics did not prove to be flexible enough to realize the power of Support Vector Methods. In particular, Support Vector Methods deliver so-called **support vectors** which characterize the border between the classes to be separated. In this respect, the Support Vector Method promises to deliver (dimension independent) data reduction.

This paper illustrates the **Support Vector Method for the classification problem with 2 and more classes**. In particular, the underlying **optimization problem** and its practical solution are discussed.

Then, the method is applied to a **business cycle data set**. The generated support vectors are interpreted, in particular with respect to whether they are able to characterize business phase switches. Finally, the classification power of the Support Vector Method and of Linear Discriminant Analysis are compared.

2. Binary classification

The Support Vector Method is well developed for the solution of **binary classification problems** (cp. Vapnik (1979, 1995, 1998); Cortes, Vapnik (1995)). In this case the **data set** has the form

$$(\mathbf{x}_i, y_i) \in \mathrm{IR}^n \times \{-1, 1\}$$

where $\mathbf{x}_i$ is a vector of length n and $y_i \in \{-1,1\}$ represents the class of the observation $\mathbf{x}_i$, $i = 1, \ldots, N$.

The main **idea** of the Support Vector Method is to construct a **hyperplane** $\mathbf{w}'\mathbf{x} + b$ **to separate the two classes** so that the distance between the hyperplane and the nearest observation (the margin) is maximized. Note that $\mathbf{w}$ is the normal vector of the hyperplane. If the classes are not linearly separable, one simultaneously has to try to minimize the classification error.

This leads to the following (mixed) **constrained optimization problem**:

$$\min \left(\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i\right) \tag{1}$$

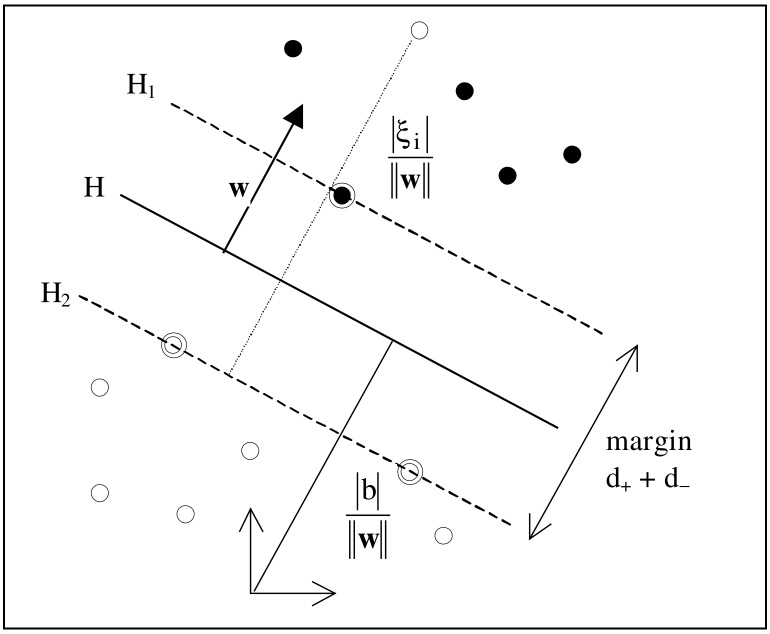
with respect to $\mathbf{w}$ and ξ_i constrained by

$$y_i(\mathbf{w}'\mathbf{x}_i + b) \geq 1 - \xi_i,\ i = 1, ..., N, \text{ and } \quad \xi_i \geq 0,\ i = 1, ..., N \qquad (2)$$

where ξ_i are so-called slack variables and C is a given parameter that controls the influence of possibly misclassified observations in the training set (cp. Cortes, Vapnik (1995)).

Indeed, $\xi_i > 0$, if and only if observation i lies at the 'wrong side' of the hyperplane parallel to the hyperplane $\mathbf{w}'\mathbf{x} + b$ which goes through the closest observations of the class of observation i in that half space of the hyperplane $\mathbf{w}'\mathbf{x} + b$ containing the most observations of this class (cp. Figure 1). All these 'closest' observations on the 'right side' of the hyperplane plus those observations with $\xi_i > 0$ together are called **support vectors**.

Figure 1: 2-class separation



This optimization problem is usually solved by using the method of Lagrange multipliers and the Kuhn-Tucker theorem. One can show that the corresponding **dual quadratic problem** is of the form:

$$\max \left(\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \cdot x_i' x_j \right) \text{ with respect to } \alpha = (\alpha_1, ..., \alpha_N)' \text{ restricted by}$$

$$\sum_{i=1}^{N} y_i \alpha_i = 0 \quad \text{and} \quad 0 \leq \alpha_i \leq C,\ i = 1, ..., N. \qquad (3)$$

Then, one can show that the **optimal Lagrange multipliers** α_i^* of the N first inequalities in (2) determine the solution $\mathbf{w}^*$ of (1), (2) as follows:

$$\mathbf{w}^* = \sum_{i=1}^{N} \alpha_i^* y_i \mathbf{x}_i \quad \text{(cp. Vapnik (1979, 1995, 1998)).}$$

For any vector $\mathbf{x}$ the **decision function** of the classification problem is

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i{}^* y_i \cdot \mathbf{x}_i{}'\mathbf{x} + b^*\right) \quad \text{where} \quad b^* = -\frac{1}{2}\mathbf{w}^{*\prime}(\mathbf{x}_+ + \mathbf{x}_-), \text{ and}$$

$\mathbf{x}_+$ and $\mathbf{x}_-$ are any **support vectors** of the classes +1 and −1, respectively, with $0 < \alpha_i^* < C$.

The characterization of a **support vector** is $\alpha_i^* > 0$. Note that vectors x_i with $\alpha_i^* = 0$ lie on the 'save' side of the separating hyperplane but not closest to the hyperplane. Vectors x_i with $C > \alpha_i^* > 0$ correspond to the closest observations on the 'save' side, and vectors x_i with $\alpha_i^* = C$ have the property $\xi_i > 0$, i.e. lie on the 'wrong' side of the hyperplane. Thus, only the support vectors determine the decision function.

3. Multi-class classification

To solve multi-class classification problems typically methods based on combination of many binary classification functions are used (i.e. the one-against-all method, cp. Schölkopf, Burges, Vapnik (1995)).

Weston and Watkins (1998) propose an extension to the SVM method to solve M-class problems in one step. In this case the classes of the sample are represented by $y_i \in \{1, ..., M\}$. This approach is to construct a decision function that considers all classes at once.

The generalization of the **minimization problem** (1) is

$$\min \frac{1}{2}\sum_{m=1}^{M}\|\mathbf{w}_m\|^2 + C\sum_{i=1}^{N}\sum_{\substack{m=1\\ m\neq y_i}}^{M}\xi_{i,m} \qquad (4)$$

with respect to **w** and ξ_i and with constraints

$\mathbf{w}_{y_i}{}'\mathbf{x}_i + b_{y_i} \geq 2 - \xi_{i,m} + \mathbf{w}_m{}'\mathbf{x}_i + b_m$, $\xi_{i,m} \geq 0$, $i = 1,...,N$, $m \in \{1,...,M\}\backslash y_i$ (cp. Figure 2).

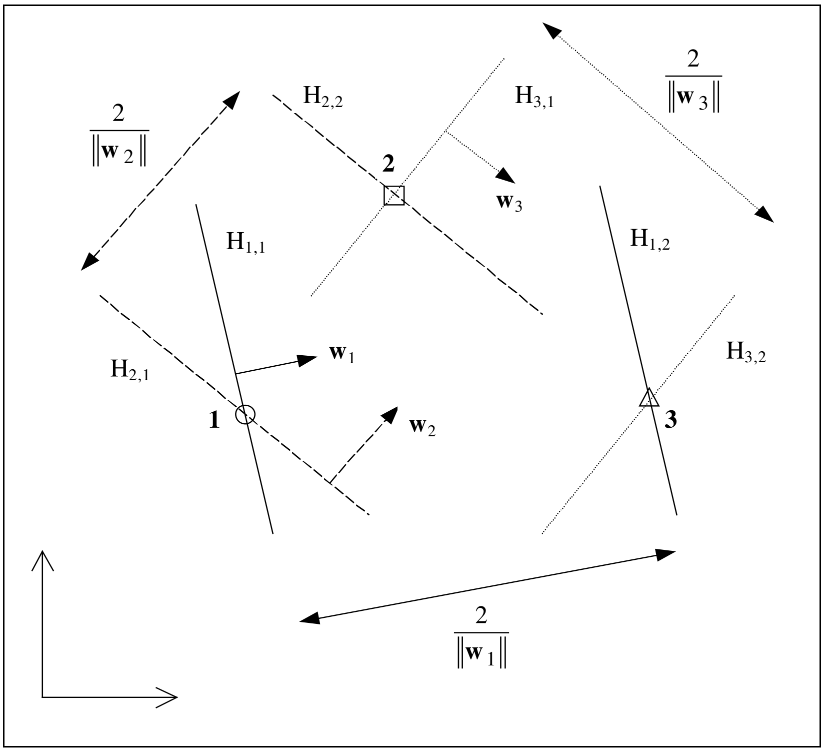
The corresponding **Lagrange function** is

$$L(\mathbf{w},b,\xi,\alpha,\beta) = \frac{1}{2}\sum_{m=1}^{M}\|\mathbf{w}_m\|^2 + C\sum_{i=1}^{N}\sum_{\substack{m=1\\ m\neq y_i}}^{M}\xi_{i,m} - \sum_{i=1}^{N}\sum_{m=1}^{M}\alpha_{i,m}\left(\mathbf{w}_{y_i}{}'\mathbf{x}_i - \mathbf{w}_m{}'\mathbf{x}_i + b_{y_i} - b_m - 2 + \xi_{i,m}\right)$$

$$-\sum_{i=1}^{N}\sum_{m=1}^{M}\beta_{i,m}\xi_{i,m}\,. \qquad (5)$$

Here, $\alpha_{i,y_i} = \beta_{i,y_i} = 0$, $\xi_{i,y_i} = 2$ are **pseudo variables** and the **constraints** $\alpha_{i,m} \geq 0$, $\beta_{i,m} \geq 0$, $\xi_{i,m} \geq 0$, $i = 1, ..., N$, $m \in \{1, ..., M\}\backslash y_i$ have to hold.

Figure 2: Multi-class separation



Note that the planes $H_{m,1}$ and $H_{m,2}$ correspond to the normal vector w_m, $m \in \{1,2,3\}$

Considering the derivatives of $L(\mathbf{w},b,\xi,\alpha,\beta)$ w.r.t. $\mathbf{w}_n$, b_n, and $\xi_{i,n}$, $n \in \{1, ..., M\}$, and using the equations

$$\sum_{m=1}^{M}\sum_{i=1}^{N}\alpha_{i,m}\cdot b_{y_i} = \sum_{m=1}^{M} b_m \sum_{i=1}^{N} c_{i,m}A_i = \sum_{m=1}^{M} b_m \sum_{i=1}^{N}\alpha_{i,m} = \sum_{m=1}^{M}\sum_{i=1}^{N}\alpha_{i,m}\cdot b_m\,,$$

$$\sum_{m=1}^{M} c_{i,m}c_{j,m} = c_{i,y_j} = c_{j,y_i}\,, \quad \sum_{m=1}^{M} c_{i,m}\alpha_{j,m} = \alpha_{j,y_i}\,, \quad \sum_{m=1}^{M} c_{j,m}\alpha_{i,m} = \alpha_{i,y_j}\,,$$

the Lagrange function (5) leads to the **dual quadratic problem**

$$\max\Big(2\sum_{m=1}^{M}\sum_{i=1}^{N}\alpha_{i,m} + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\Big(-c_{j,y_i}A_iA_j + \alpha_{j,y_i}A_i + \alpha_{i,y_j}A_j - \sum_{m=1}^{M}\alpha_{i,m}\alpha_{j,m}\Big)\mathbf{x}_i{}'\mathbf{x}_j\Big) \tag{6}$$

with respect to α and with constraints

$$\sum_{i=1}^{N}\alpha_{i,n} = \sum_{i=1}^{N} c_{i,n}A_i\,,\ n = 1,...,M,\ 0 \le \alpha_{i,m} \le C,\ \alpha_{i,y_i} = 0,\ i = 1,...,N,\ m \in \{1,...,M\}\backslash y_i. \tag{7}$$

A_i and $c_{i,\lambda}$ are defined as $\quad A_i = \sum_{m=1}^{M}\alpha_{i,m} \quad$ and $\quad c_{i,\lambda} = \begin{cases} 1, & y_i = \lambda \\ 0, & y_i \ne \lambda \end{cases}$.

Note that Weston and Watkins (1998) mistakenly did not arrive at the dual quadratic problem (6).

Solving the quadratic maximization problem (6) with respect to $\alpha_{i,m}$ for any vector $\mathbf{x}$ the **decision function** is

$$f(\mathbf{x}) = \arg\max_{m}\left(\sum_{i=1}^{N}(c_{i,m}A_i^* - \alpha_{i,m}^*)\mathbf{x}_i'\mathbf{x} + b_m^*\right). \tag{8}$$

If $\alpha_{i,m}^* \in (0; C]$, the vector $\mathbf{x}_i$ is called a **support vector** with regard to class m.

The matrix form of (6) is

$$\mathrm{LIN} - \frac{1}{2}\sum_{m=1}^{M}(\mathbf{a}_m - \mathbf{e}_m)'\mathbf{X}_S(\mathbf{a}_m - \mathbf{e}_m) \tag{9}$$

with $(\mathbf{a}_m)_i = \begin{cases} \sum_{m'=1}^{M} \alpha_{i,m}^*, & y_i = m \\ 0 & , otherwise \end{cases}$ and $(\mathbf{e}_m)_i = \alpha_{i,m}^*$.

The matrix $\mathbf{X}_S = (\mathbf{x}_i'\mathbf{x}_j)_{ij}$, i, j = 1, ..., N, contains the scalar products of the observation vectors. The term LIN denotes the linear term of (6).

4. Quadratic optimization

Expression (9) is not quite in the **standard matrix form of a quadratic optimization problem**

$$g(\alpha) = \mathbf{p}'\alpha + \alpha'\mathbf{X}\alpha = \max! \text{ w.r.t.}$$

$$\alpha = (\alpha_{1,1}, ..., \alpha_{N,1}, ..., \alpha_{1,M}, ..., \alpha_{N,M})' \text{ with } M\cdot N \text{ entries,} \tag{10}$$

where $\mathbf{p}$ is a coefficient vector and $\mathbf{X}$ is a coefficient matrix.
In the literature many solution methods for these problems are suggested (e.g. cp. Fletcher (1981)).

One can show that one can **fill X according to the following rules**:

1) The coefficients of the parameters α_{i,y_i} can be set to 0, i = 1, .., N. This means that the corresponding rows and columns of $\mathbf{X}$ are 0.
2) The coefficients of parameter products $\alpha_{i,m}\cdot\alpha_{i,m}$ ($m \neq y_i$) on the main diagonal of $\mathbf{X}$ are $-\mathbf{x}_i'\mathbf{x}_i$, i = 1, .., N.
3) The coefficients of the mixed terms $\alpha_{i,m}\cdot\alpha_{i,q}$ ($m \neq q$) are $-0.5\mathbf{x}_i'\mathbf{x}_i$ for i = 1, .., N and m, q $\in$ {1, .., M}\{y_i}.

4) The coefficients of the products $\alpha_{i,y_j} \cdot \alpha_{j,y_i}$ $(i \neq j)$ are $\mathbf{x}_i'\mathbf{x}_j$ for i,j = 1, .., N.
5) The coefficients of the products $\alpha_{i,m} \cdot \alpha_{j,y_i}$ $(i \neq j)$ are $0.5\mathbf{x}_i'\mathbf{x}_j$ for i,j = 1, .., N and $m \in \{1, .., M\} \backslash \{y_i, y_i\}$.
6) The coefficients of the products $\alpha_{i,m} \cdot \alpha_{j,m}$ $(i \neq j)$ are $-0.5\mathbf{x}_i'\mathbf{x}_j$ for i,j = 1, .., N and $m \in \{1, .., M\} \backslash \{y_i, y_j\}$.
7) The coefficients of the products $\alpha_{i,m} \cdot \alpha_{j,q}$ $(i \neq j,\ m \neq q)$ are 0 for i,j = 1, .., N and $m,q \in \{1, .., M\} \backslash \{y_i, y_j\}$.

Example: 4-class problem with N = 4 observations. For simplicity let $y_1 = 1$, $y_2 = 2$, $y_3 = 3$, and $y_4 = 4$. Thus $c_{1,1} = c_{2,2} = c_{3,3} = c_{4,4} = 1$, otherwise $c_{i,j} = 0$ (i,j = 1, .., 4), and $\alpha_{1,1} = \alpha_{2,2} = \alpha_{3,3} = \alpha_{4,4} = 0$. Then, the quadratic term of equation (6) is given by the following expression:

$$\frac{1}{2} \cdot [\ (-A_1A_1 - \alpha_{1,2}^2 - \alpha_{1,3}^2 - \alpha_{1,4}^2) \cdot \mathbf{x}_1'\mathbf{x}_1 \qquad \text{(i)}$$

$$+ (\alpha_{2,1}A_1 + \alpha_{1,2}A_2 - \alpha_{1,3}\alpha_{2,3} - \alpha_{1,4}\alpha_{2,4}) \cdot \mathbf{x}_1'\mathbf{x}_2 \qquad \text{(ii)}$$

$$+ (\alpha_{3,1}A_1 + \alpha_{1,3}A_3 - \alpha_{1,2}\alpha_{3,2} - \alpha_{1,4}\alpha_{3,4}) \cdot \mathbf{x}_1'\mathbf{x}_3 \qquad \text{(ii)}$$

$$+ (\alpha_{4,1}A_1 + \alpha_{1,4}A_4 - \alpha_{1,2}\alpha_{4,2} - \alpha_{1,3}\alpha_{4,3}) \cdot \mathbf{x}_1'\mathbf{x}_4 \qquad \text{(ii)}$$

$$+ (\alpha_{1,2}A_2 + \alpha_{2,1}A_1 - \alpha_{2,3}\alpha_{1,3} - \alpha_{2,4}\alpha_{1,4}) \cdot \mathbf{x}_2'\mathbf{x}_1 \qquad \text{(ii)}$$

$$+ (-A_2A_2 - \alpha_{2,1}^2 - \alpha_{2,3}^2 - \alpha_{2,4}^2) \cdot \mathbf{x}_2'\mathbf{x}_2 \qquad \text{(i)}$$

$$+ (\alpha_{3,2}A_2 + \alpha_{2,3}A_3 - \alpha_{2,1}\alpha_{3,1} - \alpha_{2,4}\alpha_{3,4}) \cdot \mathbf{x}_2'\mathbf{x}_3 \qquad \text{(ii)}$$

$$+ (\alpha_{4,2}A_2 + \alpha_{2,4}A_4 - \alpha_{2,1}\alpha_{4,1} - \alpha_{2,3}\alpha_{4,3}) \cdot \mathbf{x}_2'\mathbf{x}_4 \qquad \text{(ii)}$$

$$+ \ldots\ldots\ ]$$

where $A_1 = \alpha_{1,2} + \alpha_{1,3} + \alpha_{1,4}$, $A_2 = \alpha_{2,1} + \alpha_{2,3} + \alpha_{2,4}$, $A_3 = \alpha_{3,1} + \alpha_{3,2} + \alpha_{3,4}$, $A_4 = \alpha_{4,1} + \alpha_{4,2} + \alpha_{4,3}$.

Rule 1 follows from constrains (7). The lines marked with (i) are related to the rules 2 and 3. They contain the quadratic and the accompanying mixed coefficients. The rules 4 to 7 are associated to the lines denoted by (ii).

The result of the seven rules for the given 4-class problem is the following **symmetric coefficient matrix X** corresponding to the coefficients vector

$$\alpha = (\alpha_{11}\ \ \alpha_{21}\ \ \alpha_{31}\ \ \alpha_{41}\ \ \alpha_{12}\ \ \alpha_{22}\ \ \alpha_{32}\ \ \alpha_{42}\ \ \alpha_{13}\ \ \alpha_{23}\ \ \alpha_{33}\ \ \alpha_{43}\ \ \alpha_{14}\ \ \alpha_{24}\ \ \alpha_{34}\ \ \alpha_{44})'$$

Note that the abbreviation $x_{ij} := x_i'x_j$ is used.

$$
\mathbf{X} = \frac{1}{2}\begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -2x_{22} & -x_{23} & -x_{24} & 2x_{12} & 0 & x_{23} & x_{24} & x_{12} & -x_{22} & 0 & 0 & x_{12} & -x_{22} & 0 & 0 \\
0 & -x_{23} & -2x_{33} & -x_{34} & x_{13} & 0 & -x_{33} & 0 & 2x_{13} & x_{23} & 0 & x_{34} & x_{13} & 0 & -x_{33} & 0 \\
0 & -x_{24} & -x_{34} & -2x_{44} & x_{14} & 0 & 0 & -x_{44} & x_{14} & 0 & 0 & -x_{44} & 2x_{14} & x_{24} & x_{34} & 0 \\
0 & 2x_{12} & x_{13} & x_{14} & -2x_{11} & 0 & -x_{13} & -x_{14} & -x_{11} & x_{12} & 0 & 0 & -x_{11} & x_{14} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & x_{23} & -x_{33} & 0 & -x_{31} & 0 & -2x_{33} & -x_{34} & x_{13} & 2x_{23} & 0 & x_{34} & 0 & x_{23} & -x_{33} & 0 \\
0 & x_{24} & 0 & -x_{44} & -x_{41} & 0 & -x_{43} & -2x_{44} & 0 & x_{24} & 0 & -x_{44} & x_{14} & 2x_{24} & x_{34} & 0 \\
0 & x_{12} & 2x_{13} & x_{14} & -x_{11} & 0 & x_{13} & 0 & -2x_{11} & -x_{12} & 0 & -x_{14} & -x_{11} & 0 & x_{13} & 0 \\
0 & -x_{22} & x_{23} & 0 & x_{12} & 0 & 2x_{23} & x_{24} & -x_{11} & -2x_{22} & 0 & -x_{24} & 0 & -x_{22} & x_{23} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & x_{34} & -x_{44} & 0 & 0 & x_{34} & -x_{44} & -x_{14} & -x_{24} & 0 & -2x_{44} & x_{14} & x_{24} & 2x_{34} & 0 \\
0 & x_{12} & x_{13} & 2x_{14} & -x_{11} & 0 & 0 & x_{14} & -x_{11} & 0 & 0 & x_{14} & -2x_{11} & -x_{12} & -x_{13} & 0 \\
0 & -x_{22} & 0 & x_{24} & x_{12} & 0 & x_{23} & 2x_{24} & 0 & -x_{22} & 0 & x_{24} & -x_{12} & -2x_{22} & -x_{23} & 0 \\
0 & 0 & -x_{33} & x_{34} & 0 & 0 & -x_{33} & x_{34} & x_{13} & x_{23} & 0 & 2x_{34} & -x_{13} & -x_{23} & -2x_{33} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix} \tag{11}
$$

5. Example Data and Models

The data set consists of 13 **"stylized facts"** (cp. Lucas (1983)) **for the German business cycle** and 157 quarterly observations from 1955/4 to 1994/4 (price index base is 1991). The stylized facts are real GNP (gr), real private consumption (gr), government deficit, wage and salary earners (gr), net exports, money supply M1 (gr), real investment in equipment (gr), real investment in construction (gr), unit labor cost (gr), GNP price deflator (gr), consumer price index (gr), nominal short term interest rate and real long term interest rate. The abbreviation 'gr' stands for growth rates corresponding the last years corresponding quarter.

For the investigation of the data with respect to business cycle phases we use the same **4-phase scheme** as Heilemann and Münch (1996) where phases are called "upswing", "upper turning points", "downswing", and "lower turning points" (**model 1**). Table 1 shows the number of observations of each phase.

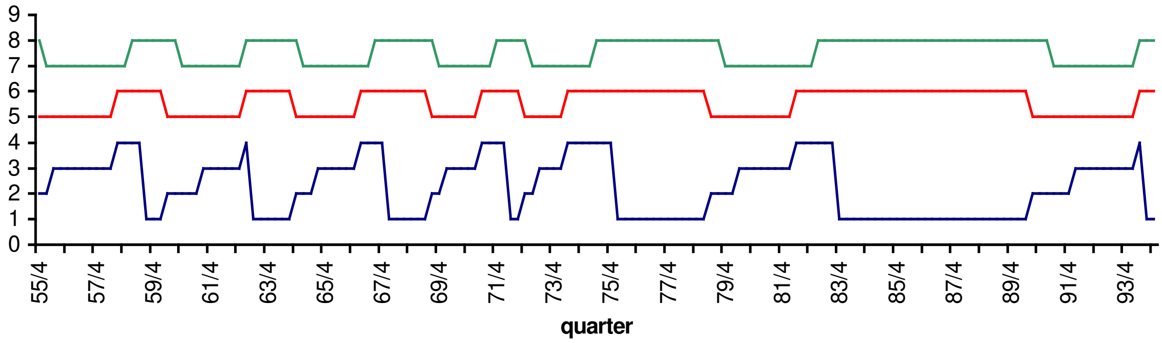
This 4-phase-model can be considered as an extension of a **2-phase-model** containing only the phases upswing and downswing. The turning points will be handled in two different ways:

- For **model 2** the phases "lower turning points" and "upswing" are joined as well as "upper turning points" and "downswing" since each turning point phase can be understood as the beginning of an upswing or a downswing, respectively.
- For **model 3** the separation of phases takes place in the middle of the upper and lower turning phases. This leads to two classes called "long upswing" and "long downswing". The term "long" is added to indicate that these phases are longer than the same classes in the 4-phase-model 1. Figure 3 illustrates the phases.

Table 1: Number of observations in phases (including phase code)

<table>
<tr><th></th><th>model 1
4-phase</th><th>model 2
2-phase
(joined phases)</th><th>model 3
2-phase
(separated turning phases)</th></tr>
<tr><td>lower turning points</td><td>27 (4)</td><td rowspan="2">86 (+1)</td><td rowspan="2">84 (+1)</td></tr>
<tr><td>upswing</td><td>59 (1)</td></tr>
<tr><td>upper turning points</td><td>24 (2)</td><td rowspan="2">71 (−1)</td><td rowspan="2">73 (−1)</td></tr>
<tr><td>downswing</td><td>47 (3)</td></tr>
</table>

Figure 3: Phases



model 3 (top): phase codes: long downswing (7), long upswing (8)
model 2 (middle): upper turning point + downswing (5), lower turning point + upswing (6)
model 1 (below): upswing (1), upper turning point (2), downswing (3), lower turning point (4)

The idea is that the classification of phases depends on the stylized facts. Unclear is the **influence of time** on the classification. Adding a variable TIME does not promise a gain of information because time increases monotonously. Therefore time is modeled by using the **lag 1 phase**. Thus for each model we consider two submodels without (a) and with (b) the lag 1 phase as an additional explanatory variable.

6. Results for the 2-phase-models

The SVM includes a parameter C to be optimized. **The goal is to minimize the error rate.**

Table 2 shows the error rates for both kinds of the models 2 and 3. The columns "tr.set" contain the error rates for the training set, the columns "cv" contain the rates computed with crossvalidation (leave-one-out). The **selection criterion for C** is the **crossvalidated error rate** because it is an unbiased estimator for the real misclassification rate (cp. Weiss and Kulikowski (1991)). In Table 2 the values for each model printed in bold have the **lowest crossvalidated error rate**: $C_{2a} = 5$, $C_{2b} = 100$, $C_{3a} = 10$, and $C_{3b} = 5$. For some values of C the SVM computes the same support vectors. In these cases, shaded in grey, the parameter C has no influence on the classification.

Table 2: Error rates for models 2 and 3

	model 2a		model 2b		model 3a		model 3b	
			phase lag 1				phase lag 1	
C	tr.set	*cv*	tr.set	*cv*	tr.set	*cv*	tr.set	*cv*
1	0.128	*0.172*	0.064	*0.134*	0.108	*0.185*	0.057	*0.089*
5	**0.108**	***0.159***	0.051	*0.102*	0.115	*0.178*	**0.038**	***0.089***
10	0.102	*0.166*	0.032	*0.102*	**0.115**	***0.172***	0.038	*0.096*
50	0.108	*0.172*	0.032	*0.102*	0.115	*0.185*	0.045	*0.089*
100	0.102	*0.178*	**0.026**	***0.064***	0.115	*0.185*	0.045	*0.089*
500	0.102	*0.178*	0.032	*0.083*	0.115	*0.185*	0.045	*0.089*
1000	0.102	*0.178*	0.032	*0.089*	0.115	*0.185*	0.045	*0.089*

Furthermore the error rates of the **models with the lag 1 phase** as an additional variable are lower than those of the models without this information. It is remarkable as well that the error rates of the **models 2** are lower than those of the **models 3**.

The number of **support vectors** for the computed models is different. The models 2a and 3a contain more support vectors than the models 2b and 3b. Depending of the choice of C model 2a contains between 48 and 61 support vectors, and 3a 51 up to 54. Model 2b has between 21 and 30, and in one case 42 support vectors (C = 1). The number of support vectors for model 3b is between 28 and 38. It is remarkable that for each model the choice C = 1 is coupled with the highest number of support vectors.

The **optimal models**, i.e. the models with the optimal choice of parameter C, i.e. with $C_{2a} = 5$, $C_{2b} = 100$, $C_{3a} = 10$, and $C_{3b} = 5$ will now be analyzed.
In particular, the position of support vectors and of misclassified vectors will be discussed.

The Support Vector Method estimates a hyperplane which marks the boundary between the two classes dependent on the variables. The normal vector of the hyperplane has one component for each economic variable. But the number of variables is too big to discuss each component. Therefore we only analyze the **support vectors in relation to the variable GNP** being the most important economic indicator.

Figures 4 and 5 show the variable **GNP together with the course of phases** for the models 2 and 3. The squares mark the support vectors which are not crossvalidated errors, and the crosses mark the crossvalidated errors. Note that **all crossvalidated errors have to be support vectors**.

Figure 4: Support vectors and crossvalidated errors (GNP, models 2a and 2b)

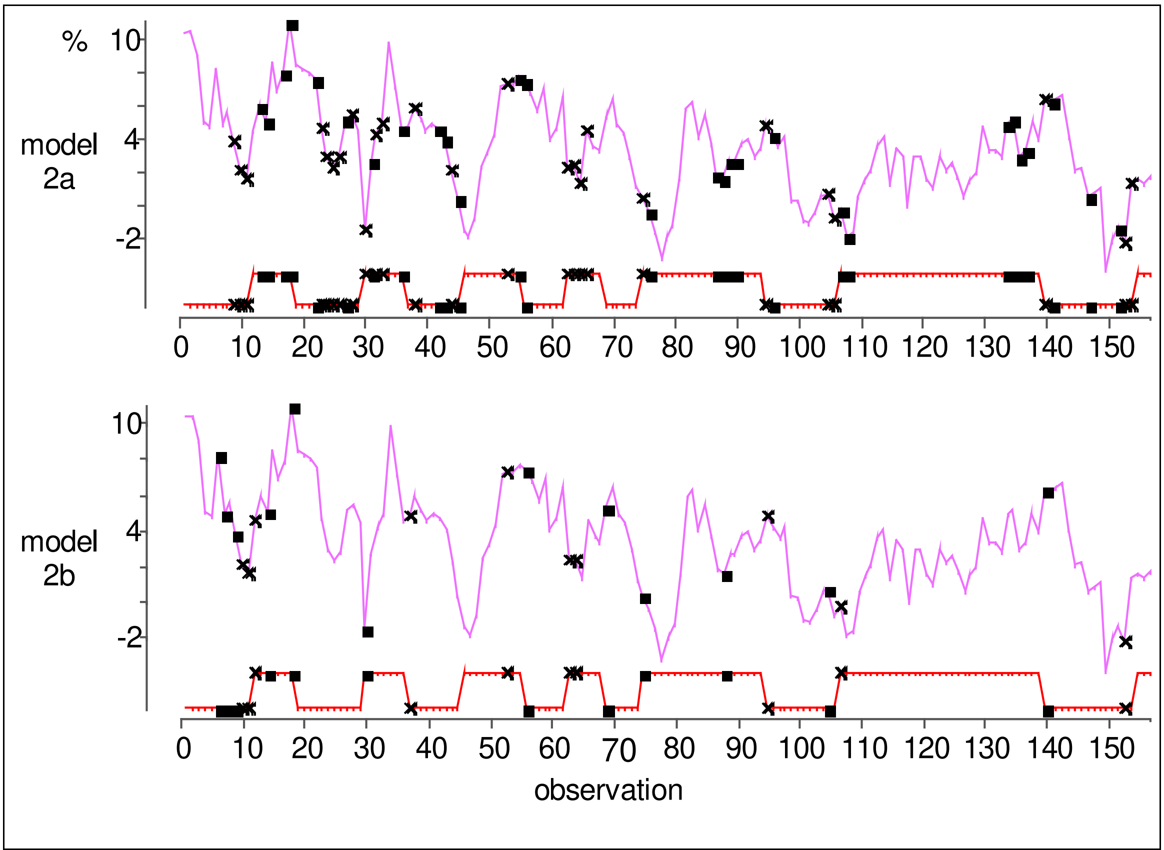
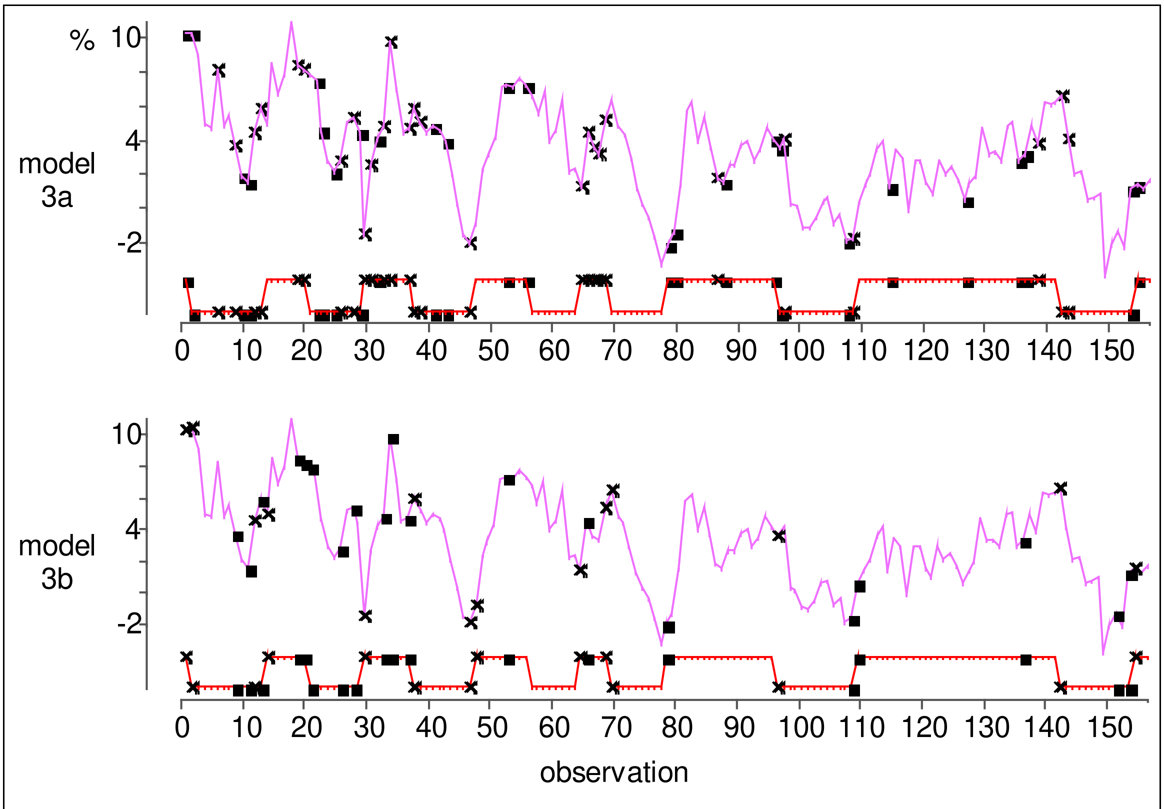


Figure 5: Support vectors and crossvalidated errors (GNP, models 3a and 3b)



The **number of support vectors** is 53 in the optimal model 2a, and 22 in the optimal model 2b. The **number of crossvalidated errors** is 25 in model 2a, and 10 in model 2b. In the optimal model 3a the number of support vectors is 52, and 33 in model 3b. The number of crossvalidated errors is 27 in model 3a, and 14 in model 3b.

The **support vectors** mainly appear in the first half of the observed time period, and the majority of the vectors are located near to a phase switch. This means that the boundary of classes in IR^N is marked by observations which are close to phase switches in the data set. One reason why this might have been expected is that observations near to phase switches will have related values independent of their phases.

Somewhat more surprising is that the **support vectors** appear in the whole region of the (growth rates of) GNP. In some cases the support vectors are located at striking positions of the time plot of GNP (e.g. the observations 18 and 108 with model 2a). But apparently **no rule** exists concerning the relationship of the value of GNP and the location of a support vector. In particular, support vectors cannot be found near all phase switches. Thus, the idea of **data reduction to support vectors** appears **questionable** if one is interested **to characterize phase switches**.

Also most of the **crossvalidated errors** appear in the first half of the time period. Maybe one reason for this that economic growth rates changed more erratically during the period of the so-called 'economic miracle' ("Wirtschaftswunder"). This might lead to the observed misclassification errors.

Model 2a has 25 **crossvalidated errors**, model 2b only 10. Most of these errors are located near to switch phases. Also many errors lie in the periods from observation 23 (1962/2) to 33 (1963/4) and from number 63 (1971/2) up to 66 (1972/1). The last periods coincide with the first oil crisis and are often misclassified also by clustering techiques (cp. Theis and Weihs (1999)). Model 3a has 27 **misclassified observations** and model 3b only 14. The main periods with errors are located from 26 (1962/1) to 39 (1965/2) and from 65 (1971/4) to observation 69 (1972/4), similar as in models 2a, 2b.

Thus, it is remarkable that for all 4 models **misclassified observations** nearly lie in the same area, although the corresponding decision functions are very different. Obviously it is difficult to classify these time periods.

7. Results for the 4-phase-model

For model 1 the model optimization with respect to the constant C is repeated (cp. Table 3). First, it appears remarkable that the **optimal error rates** are higher than with the 2-phase models. This might indicate that there is not enough evidence in the data to separate 4 phases.

Note in particular that the turning point phases are supported by only a small number of observations.

Table 3: Error rates for model 1

C	**model 1a** tr.set	*cv*	**model 1b** phase before tr.set	*cv*
1	0.172	*0.267*	0.070	*0.204*
5	**0.166**	**0.229**	**0.076**	**0.178**
10	0.166	0.261	0.064	0.204
50	0.159	*0.255*	0.038	0.217
100	0.140	0.274	0.025	0.222
500	0.127	0.280	0	0.236
1000	0.134	0.274	0	0.242

The **optimal error rate** 0.229 found by the Support Vector Method for model 1a might be compared with the error rate 0.285 found by Weihs et al. (1999) by means of **Linear Discriminant Analysis** (LDA) in the whole 13 dimensional space using Bayes decision rules based on estimated normal densities with identical covariance matrices for all 4 classes to construct separating hyperplanes for all pairs of classes. Thus, the **Support Vector Method** has a distinctly **better** error rate than LDA. This result might have been expected since the Support Vector Method was constructed to find optimal separating hyperplanes.

Figures 6 and 7, analogous to figures 4 and 5 in the analysis of models 1 and 2, show the GNP curve and the course of the business phases together with the **support vectors and classification errors** for models 1a, 1b.

In the optimal model 1a (C = 5), overall 76 of the 157 observations are **support vectors**. Thereof, 36 observations are **misclassified**. In the optimal model 1b (C = 5, again), 57 observations are support vectors, and 28 of them are not correctly classified. Note that the classification errors mainly lie near to phase switches, whereas the other support vectors more often appear inside of phases.

The majority of the **crossvalidated errors** again appear in the first halve of the observed time period. Moreover, those observations wrongly allocated by models 2 and 3 are again falsely classified by model 1. The errors particularly appear in the time periods 8-15, 28-33, and 68-75.

The number of support vectors, number of crossvalidated errors, and the corresponding crossvalidated error rate of the optimal models 1a (C = 5), 2a (C = 5), 3a (C = 10), 1b (C = 5), 2b (C = 100), and 3b (C = 5) are contrasted in Table 4. Again, the best entries are marked for both ‘static’ and ‘dynamic’ model versions. Note the superiority of models 2.

Figure 6: Support vectors and crossvalidated errors (GNP, model 1a)

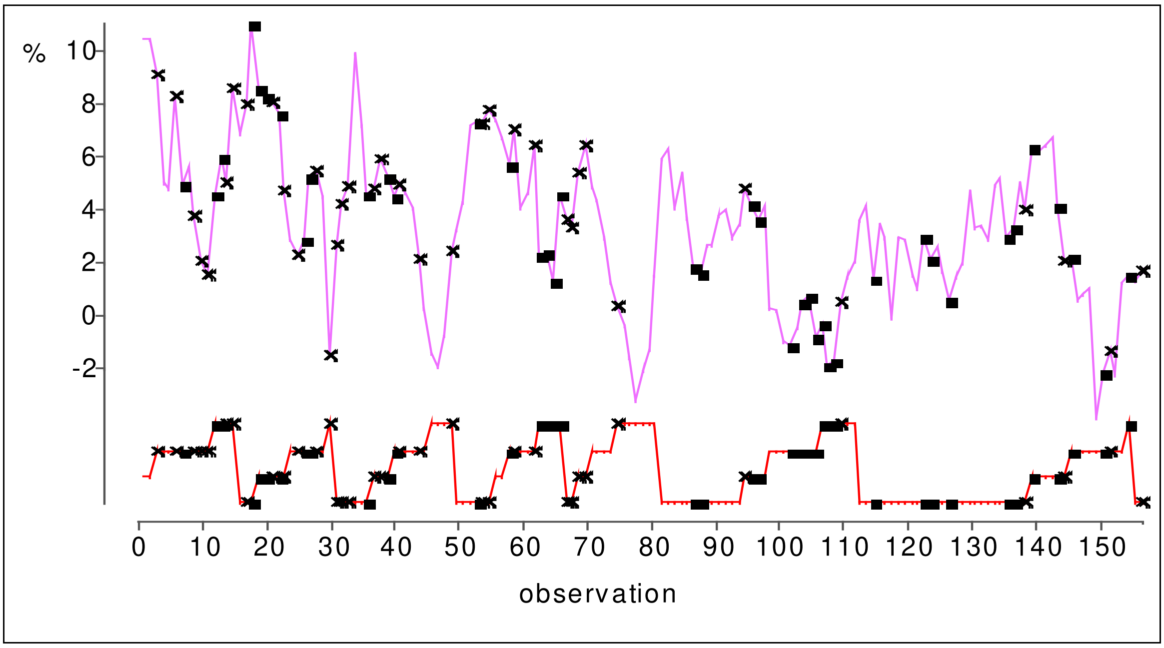


Figure 7: Support vectors and crossvalidated errors (GNP, model 1b)

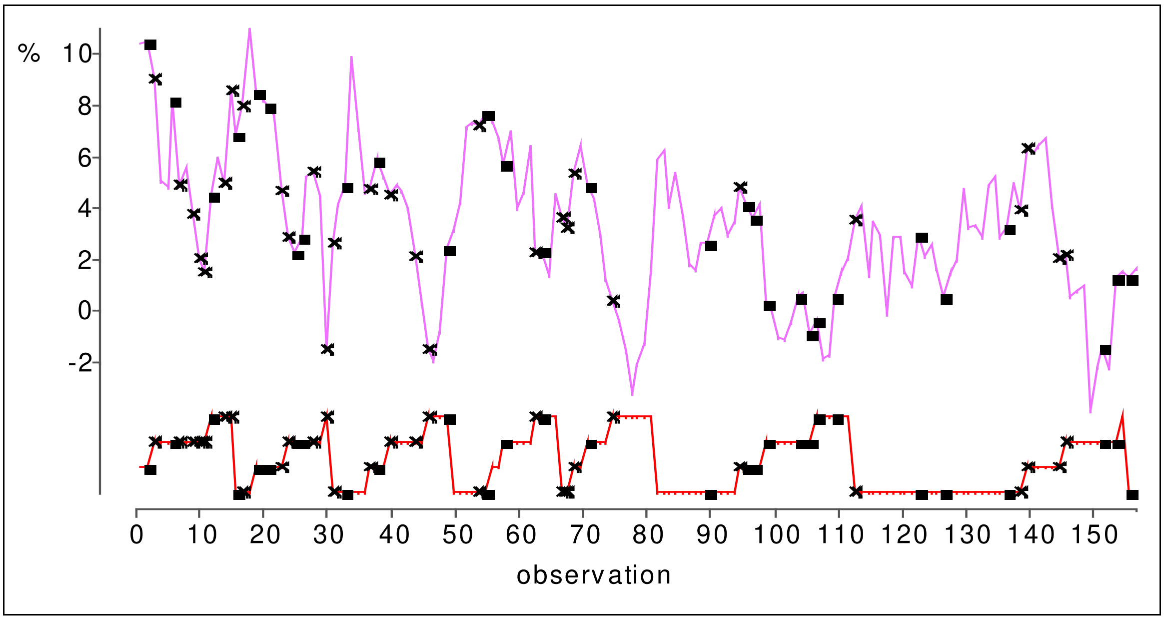


Table 4: Number of support vectors and crossvalidated errors, as well as error rates for optimal models

model	1a	**2a**	3a	1b	**2b**	3b
no. of support vectors	76	53	**52**	57	**22**	33
no. of errors	36	**25**	27	28	**10**	14
error rate	0.229	**0.159**	0.172	0.178	**0.064**	0.089

8. Computational Aspects

The computation of the support vectors takes very much **computer time** in the case of more than $M > 2$ classes, especially since an optimization problem in $M \cdot N$ dimensions has to be solved, N = number of observations.

We utilized an active sets algorithm (cp. Fletcher, 1981) in SAS/IML to solve the quadratic optimization problem. The program needs around 2.5 minutes on a 300 MHz PC for one optimization. Cross validation with 157 observations thus needed around 7 hours which is by any means unacceptable. One should check alternatives, at least concerning the programming language and the resampling algorithm.

9. Conclusion

In this paper the **multi-class classification** Support Vector Method of Weston and Watkins (1998) is correctly formulated as a quadratic optimization problem. The standard binary classification Support Vector Method and this multi-class classification method were applied to the problem of **predicting business phases** of the German economy.

The results are two-fold. On the one hand, after the analysis of the results of this study it appears questionable that the Support Vector Method delivers a meaningful (dimension independent) **data reduction** by means of identifying the support vectors only. Indeed, the support vectors did not appear to be sufficient to characterize the switches between the business phases. Note however that there might be arguments not to expect that all phase switches are ‘covered’ by support vectors since in such a case the reasons for a phase switch would never be similar!

On the other hand, the **classification power** of the Support Vector Method was somewhat better than with Linear Discriminant Analysis. Note however that the Support Vector Method needs very much more computation time than Linear Discriminant Analysis.

Overall, the **properties of the Support Vector Method** have to be analyzed in greater detail in order to decide in which situations the bigger effort to construct a classification rule can be justified. Especially the notion of a support vector might have to revised. For this the interpretation of support vectors should be analyzed more thoroughly, e.g. by means of simulation studies.

Acknowledgment

This work has been supported by the Collaborative Research Centre "Reduction of Complexity in Multivariate Data Structures" (SFB 475) of the German Research Foundation (DFG). We thank Dipl.Inf. T. Joachims for his constructive comments.

References

Cortes, C., Vapnik, V.N. (1995). Support Vector Networks, *Machine Learning,* **20**, 273-297

Fletcher, R. (1981). *Practical Methods of Optimization*, Wiley & Sons, New York

Heilemann, U., Münch, H.J. (1996). West German Buisness Cycles 1963-1994: A Multivariate Discriminant Analysis, CIRET-Studien 50, Singapur

Lucas, R.E. (1983). Understanding Buisness-Cycles, In: *Studies in Business-Cycle Theory*, Blackwell, Oxford, 215-239

Schölkopf, B., Burges, C.J.C., Vapnik, V.N. (1995). Extracting Support Data for a Given Task, In: Fayyad, U.M., Uthurusamy, R. (eds.). *Proceedings of the First International Conference on Knowledge Discovery and Data Mining.* AAAI Press, Menlo Park, CA, 252-257

Theis, W., Weihs, C. (1999). Clustering techniques for the detection of business cycles, In: *Studies in Classification, Data Analysis, and Knowledge Organization, Proceedings of the 23rd Annual Conference of the GfKl* (submitted), Technical Report **40**/99, SFB 475, Universität Dortmund

Vapnik, V.N. (1979). *Estimation of dependences based on empirical data,* Nauka, Moskau (in Russian), (English translation: 1982, Springer, New York)

Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory,* Springer, New York

Vapnik, V.N. (1998). *Statistical Learning Theory,* Wiley & Sons, New York

Weihs, C., Röhl, M.C., Theis, W.: (1999). *Multivariate Classification of Business Phases*, Technical Report 26/99, SFB 475, Universität Dortmund (submitted to Applied Econometrics)

Weiss, S.M., Kulikowski, C. (1991). *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems,* Kaufmann, San Francisco

Weston, J., Watkins, C. (1998). Multi-class Support Vector Machines, Technical Report CSD-TR-98-04 Royal Holloway University of London